JAZYKOVEDNÝ ČASOPIS

JAZYKOVEDNÝ ÚSTAV ĽUDOVÍTA ŠTÚRA

SLOVENSKEJ AKADÉMIE VIED

2

ROČNÍK 68, 2017

DE GRUYTER OPEN

SAP
SLOVAK ACADEMIC PRESS

JAZYKOVEDNÝ ČASOPIS
VEDECKÝ ČASOPIS PRE OTÁZKY TEÓRIE JAZYKA

JOURNAL OF LINGUISTICS
SCIENTIFIC JOURNAL FOR THE THEORY OF LANGUAGE

Vydáva/Published by: Jazykovedný ústav Ľudovíta Štúra Slovenskej akadémie vied
- v tlačenej podobe vo vydavateľstve SAP – Slovak Academic Press, s.r.o.
- elektronicky vo vydavateľstve Versita – De Gruyter (Open Access)
  www.degruyter.com/view/j/jazcas

Adresa redakcie/Editorial address: Jazykovedný ústav Ľ. Štúra SAV, Panská 26, 811 01 Bratislava
kontakt: gabam@juls.savba.sk
Elektronická verzia časopisu je dostupná na internetovej adrese/The electronic version of the journal is available at: http://www.juls.savba.sk/ediela/jc/

Vychádza trikrát ročne/Published triannually

Dátum vydania aktuálneho čísla (2017/68/2) – október 2017

SCImago Journal Ranking (SJR) 2015: 0,101
Source Normalized Impact per Paper (SNIP) 2015: 0,875
Impact per Publication (IPP) 2015: 0,111

JAZYKOVEDNÝ ČASOPIS je evidovaný v databázach/JOURNAL OF LINGUISTICS is indexed by the following services: CEJSH (The Central European Journal of Social Sciences and Humanities); Celdes; CNKI Scholar (China National Knowledge Infrastructure); CNPIEC; De Gruyter - IBR (International Bibliography of Reviews of Scholarly Literature in the Humanities and Social Sciences); De Gruyter - IBZ (International Bibliography of Periodical Literature in the Humanities and Social Sciences); DOAJ; EBSCO (relevant databases); EBSCO Discovery Service; Elsevier – SCOPUS; ERIH PLUS (European Reference Index for the Humanities and Social Sciences); Google Scholar; International Medieval Bibliography; J-Gate; JournalTOCs; Linguistic Bibliography Online; Linguistics Abstracts Online; MLA International Bibliography; Naviga (Softweco); Primo Central (ExLibris); ProQuest - International Bibliography of the Social Sciences (IBSS); ProQuest - Linguistics & Language Behavior Abstracts (LLBA); ProQuest - Research Library; ReadCube; SCImago (SJR); Summon (Serials Solutions/ProQuest); TDOne (TDNet); Ulrich's Periodicals Directory/ulrichsweb; WorldCat (OCLC).

ISSN 0021-5597 (tlačená verzia/print) ISSN 1338-4287 (verzia online) MIČ 49263

JAZYKOVEDNÝ ČASOPIS

JAZYKOVEDNÝ ÚSTAV ĽUDOVÍTA ŠTÚRA
SLOVENSKEJ AKADÉMIE VIED

2

ROČNÍK 68, 2017

Natural Language Processing, Corpus Linguistics, Terminology, e-Terminology

SLOVKO 2017

Tematické číslo Jazykovedného časopisu venované počítačovému spracovaniu prirodzeného jazyka, korpusovej lingvistike, terminológii a e-terminológii.

Prizvané editorky: Mgr. Katarína Gajdošová, Ph.D.
Mgr. Adriána Žáková

DE GRUYTER OPEN

SAP SLOVAK ACADEMIC PRESS

JAZYKOVEDNÝ ČASOPIS

ROČNÍK 68 2017 ČÍSLO 2

CONTENT

SLOVO NA ÚVOD

9. ročník bienálnej medzinárodnej konferencie **Slovko** sa koná v Bratislave 25. – 27. októbra 2017. Okrem tradičného zamerania na počítačové spracovanie prirodzeného jazyka (v písanej i hovorenej podobe) a na korpusovú lingvistiku sa osobitne venuje terminológii a e-terminológii. Tento tretí tematický okruh je v každom ročníku konferencie iný (`http://korpus.sk/slovko.html`) a tentoraz súvisí s riešením projektu *Analýza terminologickej práce Jána Horeckého ako inšpirácia pre terminologický manažment 21. storočia na Slovensku* (projekt VEGA 2/0114/15, zodpovedná riešiteľka Jana Levická).

Organizátori zo Slovenského národného korpusu Jazykovedného ústavu Ľ. Štúra SAV dostali na prvú výzvu vyše 90 prihlášok záujemcov o prezentáciu svojich najnovších výsledkov v uvedených oblastiach. Do stanoveného termínu prišlo 44 príspevkov, ktoré starostlivo posúdili členovia vedeckého výboru a na publikovanie vybrali 31 z nich. Recenzentom aj na tomto mieste ďakujeme za vyjadrenia, ktorými prispeli k zlepšeniu kvality publikovaných príspevkov a celého priebehu konferencie.

Osobitné poďakovanie patrí redakčnej rade a hlavnej redaktorke Jazykovedného časopisu za poskytnutie publikačného priestoru vedecko-výskumným témam, ktoré v slovenskej lingvistike nemajú dlhú tradíciu. Postupy a výsledky korpusovo a počítačovolingvisticky orientovaných výskumov v oblasti gramatiky, lexikológie, terminológie a analýzy hovorenej reči môžu byť obohatením existujúcich metodologických postupov pri poznávaní a opise jazyka a rozšírením doterajších poznatkov.

Počas rokovania odznejú 2 plenárne prednášky a 32 prezentácií od 69 autorov – 18 príspevkov je v individuálnom autorstve, 16 v spoluautorstve dvoch, troch i viacerých autorov. Spomedzi 11 zúčastnených krajín je najviac autorov z ČR (32), SR (14) a Ruska (8), ďalšie krajiny sú zastúpené menším počtom (Gruzínsko, Chorvátsko, Fínsko, Rakúsko, Ukrajina, Nemecko, Poľsko, Švédsko). Všetkým účastníkom konferencie Slovko 2017 želáme úspešný priebeh rokovania, vzájomne užitočné a obohacujúce diskusie, ako aj nadviazanie nových kontaktov na prípadné budúce spolupráce.

Nasledujúce Slovko v roku 2019 bude jubilejné 10. a plánujeme ho v samostatnom tematickom okruhu špecifickejšie zamerať na témy, ktoré sa objavujú už v tomto ročníku: dynamika jazyka a gramatické zmeny v súčasnom jazyku na báze korpusových dát.

Mária Šimková

FOREWORD

The 9th edition of the biannual conference **Slovko** is held in Bratislava from 25–27 October 2017. In addition to the traditional focus on natural language processing (in its written and spoken form) and on corpus linguistics it is specifically devoted to terminology and e-terminology. The third thematic area being different in each edition (`http://korpus.sk/slovko.html`) was chosen in accordance with the organizers' project *Analysis of terminology work of Ján Horecký as an inspiration for the terminology management of the 21st century in Slovakia* (VEGA project nº 2/0114/15, coordinator Jana Levická).

Organizers from the Slovak National Corpus Department of the Ľ. Štúr Institute of Linguistics of the Slovak Academy of Sciences received upon the first call more than 90 registrations. Within the set deadline they received 44 articles presenting the latest results in the above-mentioned areas which were carefully reviewed by the programme committee members who recommended as many as 31 papers for publishing in the proceedings. We would like to express our gratitude to all reviewers who helped to improve the quality of the published papers and the conference itself.

Special thanks goes to the Editorial Board and Chief Editor of the Journal of Linguistics for providing the possibility to publish the articles on scientific research topics without a long tradition in the Slovak linguistics. Approaches and findings of the corpus and computer-oriented researches in the field of grammar, lexicology, terminology and speech analysis can contribute to the existing methodological approaches aiming at better understanding and description of the language, as well as at extending the current knowledge.

The event includes 2 plenary talks and 32 presentations by 69 authors – out of which 18 were prepared by a single author, while 16 presentations resulted from the co-authorship. From among 11 participating countries, the vast majority of authors come from the Czech Republic (32), followed by the Slovak Republic (14) and Russia (8), other countries are represented by fewer authors (Georgia, Croatia, Finland, Austria, Ukraine, Germany, Poland, Sweden). We wish all the participants of Slovko 2017 a successful conference, useful and enriching discussions as well as numerous opportunities for networking leading to future cooperation.

The following, 10th jubilee edition of Slovko – planned to be held in 2019 – will be focusing, in its third thematic area, on topics that appear already in this edition: dynamics of language and grammar changes in the contemporary language based on corpus data.

Mária Šimková
Translated by Adriána Žáková

GEORGIAN DIALECT CORPUS: LINGUISTIC AND ENCYCLOPEDIC INFORMATION IN ONLINE DICTIONARIES[1]

MARINA BERIDZE – DAVID NADARAIA – LIA BAKURADZE
Arnold Chikobava Institute of Linguistics, Tbilisi State University, Georgia

BERIDZE, Marina – NADARAIA, David – BAKURADZE, Lia: Georgian Dialect Corpus: Linguistic and Encyclopedic Information in Online Dictionaries. Journal of Linguistics, 2017, Vol. 68, No 2, pp. 109 – 121.

Abstract: The Georgian Dialect Corpus (GDC) has been created within the framework of the project "Linguistic Portrait of Georgia". It was the first attempt to create a structured corpus of Georgian dialects. The work of this project includes building the technical framework for a corpus, collecting the corpus (text) data of Georgian dialects including the lexicographic data (dictionaries), their linguistic processing, digitizing, developing annotation framework, making decision on the morphosyntactic annotation. Currently, the Georgian Dialect Corpus is a platform consisting of the dialect corpus, the text library, the lexicographical database/online dialect dictionaries. For the purposes of developing the lexicographical database and dialect dictionaries, we have created a new program – the Lexicographic Editor. It allows us to structure and improve the dictionaries with multiple linguistic and lexicographic information. The lexicographic concept of the GDC has been developed taking into consideration linguistic and social features of the Georgian dialects.

Keywords: corpus linguistics, corpus lexicography, dialect corpora

1 INTRODUCTION

The languages spoken in Georgia generally belong to the Kartvelian languages, also called South Caucasian languages, or Iberian languages, a family of languages including Georgian proper, Svan, and Zan (further split into Mingrelian and Laz) as well as dialects of Georgian. The Georgian Dialect Corpus covers 17 Georgian dialects out of which 3 are spoken outside the country. They are Fereydanian in Iran, Ingiloan in Azerbaijan and so-called "Chveneburebi" in Turkey. Also, the Laz language is mostly spoken in Turkey. Only two Laz speaking villages do exist in Georgia.

The linguistic research of the Georgian dialects dates back to the 1920s, however there were several preliminary works and descriptions of dialect vocabulary in the 19th century.

The Georgian dialects are classified according to their ethnogeography, region and particular linguistic features. The varied and diverse vocabulary plays an important role in identifying a particular dialect or sub-dialect.

In the 20th century, there were several large migrations in Georgia, such as the massive migrations during the Soviet period; environmental migrations; migrations

[1] This work was supported by Shota Rustaveli National Science Foundation (SRNSF) [grant number 217008].

DE GRUYTER OPEN
DOI 10.1515/jazcas-2017-0022

as a result of war conflicts in Abkhazia and Shida Kartli. All of these factors caused the distortion of the linguistic boundaries in these areas.

In the following sections we will present the lexicographic database, representation of linguistic annotation.

2 ONLINE DICTIONARIES IN GDC

2.1 The Lexicographic Database of the GDC

The lexicographic database of the GDC is a separate section of the corpus. This platform has functions such as collecting, processing and converting lexicographic data into dictionaries.

The lexicographic database includes:

- Digitized online dictionaries from earlier printed dictionaries
- Various lexicographic data collected from lexicographic fieldwork
- Lexicographic data published by various authors
- Lexicographic data extracted from the existing linguistic and/or ethnographic studies.

The lexicographic database grows continually, with new texts being added over time. The database covers over 10 dialects and lists about 60 000 entries. This database has been developed based on the traditional lexicographic principles and methods. The research team will follow this methods in compiling comprehensive online dictionary of other Georgian dialects. Overall, four online dictionaries has been published so far. The published dictionaries are: Fereydanian, Ingiloan, “Chveneburebi”, and Laz dictionaries. New dictionaries with corresponding lexicographic data will be added to the corpus interface.

2.2 Georgian Dialect Lexicography and the Lexicographic Principles of GDC

The Georgian dialect vocabulary is represented in the sources as follows:

- Georgian monolingual and multilingual dictionaries
- Dialect dictionaries
- Data from monographs, research outputs, publications.

The quality and the content of the Georgian dialect dictionaries varies greatly. There are several comprehensive, academic dictionaries, such as dictionary of Ingiloan dialect [7] Kartlian dialect [8], Imeretian dialect [9], [10], Gachechiladze [10], Khevsurian dialect [11], Tushetian dialect [12], Adjarian dialect [13] and etc.

In some cases, there are several dictionaries available for one dialect. These dictionaries can vary in size, content and lexicographic principle, such as the dictionaries for Gurian, Imeretian and Ingiloan dialects.

There are not sufficient lexicographic data for some Georgian dialects. In this case, lexicographic information about those dialects has been collected from various monographs, publications, manuscripts, for example, for Lechkhumian, Pshavian and Mtiuletian dialects. Additionally, the comprehensive dialect dictionary comprising all Georgian dialects is also available [14].

In general, the existing dialect dictionaries do not follow a single lexicographic principle. In particular, the dictionary entries are presented differently in various dictionaries, the structure of the entries and linguistic and encyclopedic information in each dictionaries varies massively.

The Dialect Vocabulary in the Dialect Dictionaries Represents the following:
- Dialect vocabulary used in the standard Georgian language
- Vocabulary that differs in word-form and meaning, collocation and/or phrases
- Vocabulary common to standard Georgian language, but having different meaning in dialects
- Some Archaic vocabulary that are only preserved in dialects
- Some common vocabulary for both dialects and standard Georgian language,
- Some foreign borrowings
- Morphemes and modal elements that are specific to a particular dialect
- Some proper nouns [15].

In general, the Georgian dialect dictionaries do not include (or very rarely) proper names, surnames, toponyms, regular phonetic variations, foreign borrowings, that has developed their own dialect forms.

As discussed above, the dictionary entries are structured differently in each dialect dictionary. There is no correspondence between the macro- and micro-structure of the dictionaries. In addition, there is no coherence in illustrating the lexicographic information. Some dictionaries present several phonetic variants as separate dictionary entries, or describe several grammatical variations. Thus, each dictionary uses its principle for presenting linguistic, encyclopedic information in the dictionaries.

The lexicographic concept and principle, we are using in our research, aims to reach the following goals:
- to edit, unify and re-structure the existing dictionaries,
- to add new dictionary entries from the dialect corpus (data collected from linguistic field works),
- to improve the dictionary by adding dictionary examples, linguistic and encyclopedic information.

3 LINGUISTIC INFORMATION IN GDC

3.1 The Lexicographic Editor of GDC

The lexicographic editor of GDC is a platform that allows documenting and re-structuring lexicographic information.

The lexicographic information in the editor can be integrated by directly adding texts or uploading Excel spreadsheets. The editor is easy to use. It does not require any special knowledge of a particular program(s).

The lexicographic editor has several functions and fields/tabs as follows:

1. Saving – there are several saving tabs, where information of a lexicographic source is being saved. This information allows us to create a new dictionary entry, to edit or correct it. As each dictionary entry has several lexicographic sources, there are several tabs accordingly.

2. to add a lemma;
3. to write up the description for the dictionary entry;
4. to add encyclopedic information;
5. to add information about foreign borrowings;
6. to add foreign borrowings with their corresponding spelling in original language;
7. to add information about the structure of the dictionary entry (enclitic, composition, compound grammatical forms, simple collocation etc.);
8. to add lexical information about the dictionary entry (neutral, specialised etc.);
9. to add information about the semantic group;
10. to add information about lexicographic sources;
11. to add a grammatical marker for the lemma;
12. to add the grammatical variation(s) of the lemma;
13. to add the derivation variation(s) of the lemma;
14. to add the phonetic variation(s) of the lemma;
15. to indicate synonymous words;
16. to link to other lexical item;
17. to add dictionary example;
18. to provide translation for the dictionary example or to add additional comment;
19. to add the source for the dictionary example.

As shown above, the lexicographic editor allows us to compile a dictionary entry, to edit it, make some changes there, and publish it.

All stages are carried out separately: processing, editing and publishing. At the processing stage, only two tabs are active, these are Lemma and Save tabs.

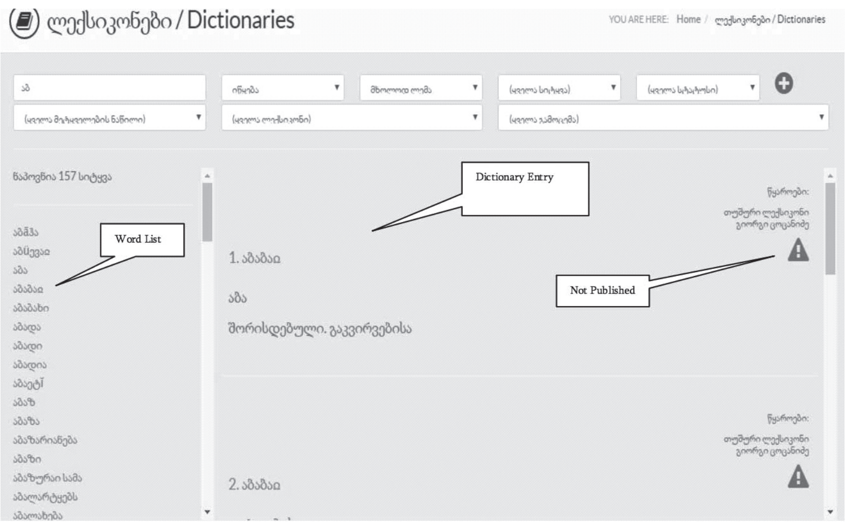
The first step is adding lemma with the corresponding lexicographic sources. After that, existing lexicographic information can be re-structured according to the specific fields. The final step is publishing the entry, when it becomes available to broader dictionary users.

The unpublished (unedited) dictionary entry can be searched in the lexicographical database with the information of the lexicographic source. However, the dictionary entry is not complete as it does not have additional linguistic and encyclopedic information. Also, lemma and its variants are not marked according to the GDC principles (Fig.1).

For the user purposes, unedited and edited dictionary entries can be distinguished by marking them with special symbols. Cf. Fig.1 and Fig.2.

3.2 Lemmas in GDC

In the GDC lemmas represent the following: a single word, collocation, multi-component units that are considered as an individual word. The collocations may include unlimited number of words, sometimes it can be a whole sentence (if a phrase). In the corpus platform, in the lemma field, a single word or collocation are written without indicating the variations.



ლექსიკონები / Dictionaries

YOU ARE HERE: Home / ლექსიკონები / Dictionaries

აბ | იწყება ▾ | მხოლოდ ლემა ▾ | (ყველა სიტყვა) ▾ | (ყველა სტატუსი) ▾

(ყველა მეტყველების ნაწილი) ▾ | (ყველა ლექსიკონი) ▾ | (ყველა გამოცემა) ▾

ნაპოვნია 157 სიტყვა

აბშა
აბüევაჲ
აბა
აბაბაჲ
აბაბახი
აბადა
აბადი
აბადია
აბაეტÏ
აბაზ
აბაზა
აბაზარიანება
აბაზი
აბაზური სამა
აბალარტყებს
აბალახება

1. აბაბაჲ

აბა

შორისდებული. გაკვირვებისა

წყაროები:

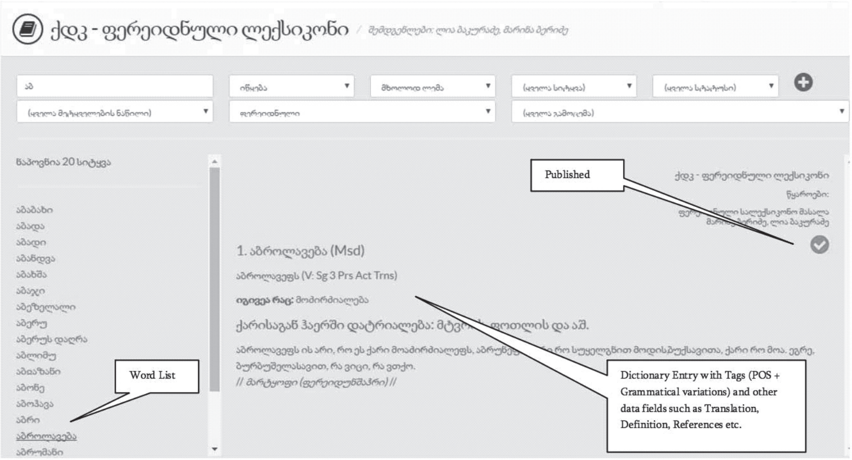
თუშური ლექსიკონი
გიორგი ცოცანიძე

2. აბაბაჲ

წყაროები:

თუშური ლექსიკონი
გიორგი ცოცანიძე

Fig.1. The progress of the Dictionary Entry Compilation



ქდკ - ფერეიდნული ლექსიკონი / შემდგენლები: ლია ბაკურაძე, მარინა ბერიძე

აბ | იწყება ▾ | მხოლოდ ლემა ▾ | (ყველა სიტყვა) ▾ | (ყველა სტატუსი) ▾

(ყველა მეტყველების ნაწილი) ▾ | ფერეიდნული ▾ | (ყველა გამოცემა) ▾

ნაპოვნია 20 სიტყვა

აბაბახი
აბადა
აბადი
აბანდვა
აბახშა
აბაჯი
აბეზელალი
აბერუ
აბერუს დაღრა
აბლიმუ
აბჲაზანი
აბონე
აბოჰავა
აბრი
აბროლავება
აბრუმანი

ქდკ - ფერეიდნული ლექსიკონი

წყაროები:

ფერეიდნული სალექსიკონო მასალა
მარინა ბერიძე, ლია ბაკურაძე

1. აბროლავება (Msd)

აბროლავეფს (V: Sg 3 Prs Act Trns)

იგივეა რაც: მოძირძიალება

ქარისაგან ჰაერში დატრიალება: მტვრის, ფოთლის და აშ.

აბროლავეფს ის არი, რო ეს ქარი მოაძირძიალეფს, აბრუნებს, რო სუყელგნით მოდისბუქსავითა, ქარი რო მოა. ეგრე, ბურბუშელასავით, რა ვიცი, რა ვთქო.

// მარტყოფი (ფერეიდუნშაჰრი) //

Published
Word List
Dictionary Entry with Tags (POS + Grammatical variations) and other data fields such as Translation, Definition, References etc.

Fig.2. The Published Dictionary Entry in GDC

A lemma for nominals (noun, adjective, pronoun and numeral) are: nominative case, singular.

Lemmas for the verb are presented as follows: 3rd person of subject, resent tense, Singular. However in some cases (taking into consideration the lexicographic source) a Masdar can function as the lemma.

The dictionaries also include other parts-of-speech: adverbs, particles, conjugations, interjections, also some root or inflectional morphemes. These parts-of speech, unlike nominals and verbs, are rather easy to deal with in POS-tagging, as they do not change their forms.

Unlike traditional dictionaries, the dictionaries in the GDC include different types of proper names, such as ethnonyms, toponyms, hydronyms, anthroponyms, zoonyms, oeconyms etc., and foreign borrowings.

3.3 The Dictionary wordlist. A Lemma and Word.

The dictionary wordlist is a list of words that are represented in the corpus with corresponding lexicographic information. The wordlist is compiled as follows:

A lemma for a particular word is added in the lemma tab:

1. the lemma is linked to its phonetic variant(s), where the number of phonetic variant(s) is not limited.
2. the lemma is linked to its grammatical/derivational variant(s), where the number of variant(s) is not limited.

In addition, these variation(s), both lemmas and grammatical/derivational variation(s) have corresponding grammatical markers.

POS-tags are selected from the existing tagset that is based on hierarchical structure. The first line in this hierarchical structure is a list of parts-of-speech, also markers for the masdar, participle and other elements, such as the nominal root, the verbal root, the preverb etc. The second line of the hierarchy is a description of some basic grammatical features, also some derivational features.

Lemma – tagged with the first hierarchical marker.

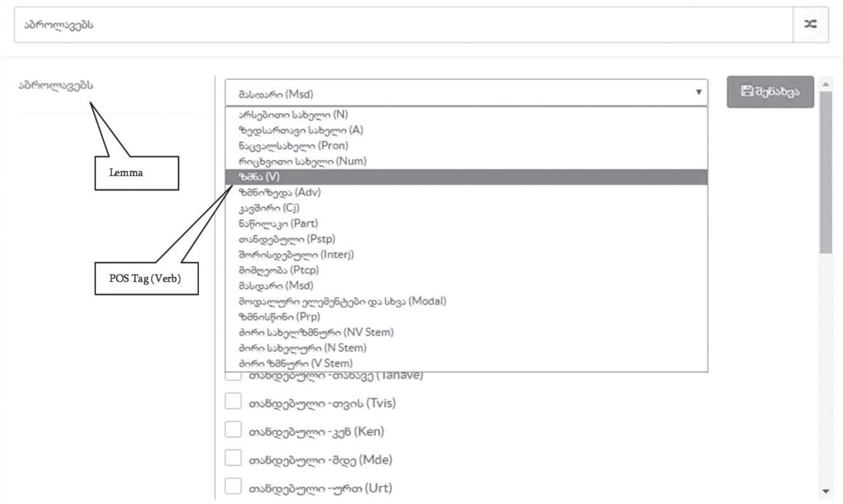
Phonetic variant(s) – not tagged as the grammatical status of a phonetic variant is identical to lemma.

Grammatical/Derivational variant(s) – tagged according to its grammatical features.

Therefore, lemmas in the online dictionary are presented with several variants/ forms within one dictionary entry.

The query interface of the online dictionaries allows:

- to enter a search lemma and a word
- to search for individual words with their associated part-of-speech tags.



აბროლავებს

აბროლავებს

მასდარი (Msd)

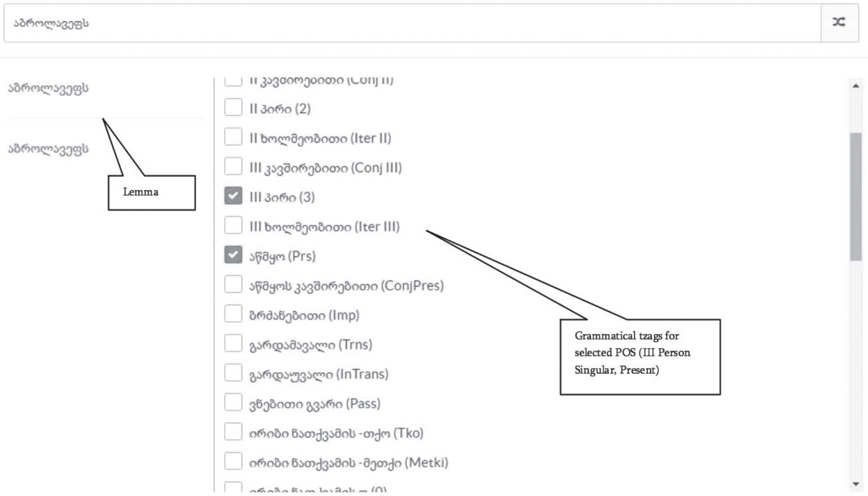
Fig. 3. Marking the Variants _ Example 1

Figure 3 shows the first stage of annotating grammatical and derivational variants, where the selected word-form has POS-tag assigned:

Lemma: abrolaveba (Msd) _ ‘to cause spinning by wind’

Grammatical variant: abrolavebs (Verb _V)

After the mark-up, the information about the verbal categories are added, For instance, Person of Agreement, Number of agreement, Screeve etc. e.g.: abrolavebs _V Prs Sg 3



აბროლავეფს

აბროლავეფს

აბროლავეფს

- □ II კავშირებითი (Conj II)
- □ II პირი (2)
- □ II ხოლმეობითი (Iter II)
- □ III კავშირებითი (Conj III)
- ☑ III პირი (3)
- □ III ხოლმეობითი (Iter III)
- ☑ აწმყო (Prs)
- □ აწმყოს კავშირებითი (ConjPres)
- □ ბრძანებითი (Imp)
- □ გარდამავალი (Trns)
- □ გარდაუვალი (InTrans)
- □ ვნებითი გვარი (Pass)
- □ ირიბი ნათქვამის -თქო (Tko)
- □ ირიბი ნათქვამის -მეთქი (Metki)

Lemma
Grammatical tzags for selected POS (III Person Singular, Present)

Fig. 4. Marking the Variants _ Example 2

3.4 Why to Include Separate Forms in the Dictionaries?

We have decided to include other variation(s) in the dictionaries for several reason. Firstly, following the dictionary principles of the GDC, a single word or collocation is represented in the lemma field. Secondly, it provides a comprehensive information about the word with their various phonetic variations. Additionally, it has somehow solved the problem of representing verbs in the dictionaries. In particular, there is no infinitive in Georgian and the masdar cannot represent the verb with its grammatical information in most cases. The Georgian monolingual dictionary [16] has established a rule for representing the verb in the dictionaries, where the verb is represented in 3rd person of subject, present tense, singular.

Our approach in representing a verb in the online dictionary has the following principles: no more than one word in the lemma tab, but the entry allows linking additional information such as phonetic and grammatical/derivational variations.

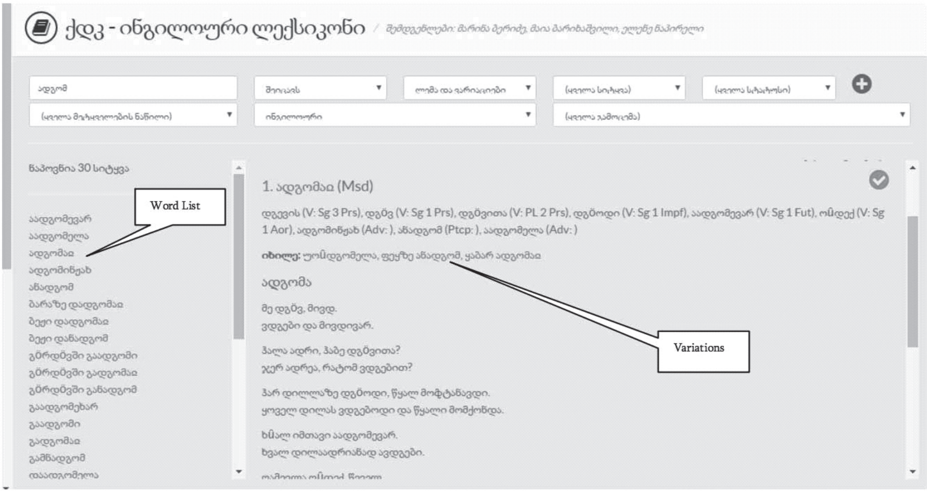
As an example, we will use the word adgomaj “stand up” from the Ingiloan dictionary. As for the grammatical and derivational variations, the words has eight forms. These forms are extracted from dictionary examples or texts from seven lexicographic sources. These variants are for example:

dgevis (V: SG 3 PRS) _ ‘h/she stands up’; dgövi (V: SG 1 PRS) _ ‘I stand up’; dgövita (V: PL 2 PRS) _ ‘we stand up’; dgöodi (V: SG 1 IMPF)‘I was standing up’; aadgomevar (V: SG 1 FUT)- ‘I will stand up’; ovdek (V: SG 1 AOR)_ ‘I stood up’;

adgomi žanx (ADV) _ ‘time to stand up’; anadgom (PTCP) ‘standing’; aadgomela (PTCP) _ ‘to stand up’.

Based on the lexicographic sources, we made decision to introduce Masdar form as a dictionary entry (lemma) – adgomaj _ MSD. As for the grammatical variations, present tense, 3rd person of subject forms are also included – (dgevis _ V: SG 3 PRS).

One of the main principles in introducing a new variant to the word is that it should be well represented in various contexts in the lexicographic sources. As mentioned above, all these variations are included in the dictionary wordlist. Each variant can be used as a search word linked to the dictionary headword. See Fig.5.



აადგომევარ
აადგომელა
ადგომაჲ
ადგომინჴახ
ანადგომ
ბარაზე დადგომაჲ
ბეჟი დადგომაჲ
ბეჟი დანადგომ
გო̈რდო̈ვში გაადგომი
გო̈რდო̈ვში გადგომაჲ
გო̈რდო̈ვში განადგომ
გაადგომეხარ
გაადგომი
გადგომაჲ
გამნადგომ
დაადგომელა

1. ადგომაჲ (Msd)

დგევის (V: Sg 3 Prs), დგო̈ვ (V: Sg 1 Prs), დგო̈ვითა (V: PL 2 Prs), დგო̈ოდი (V: Sg 1 Impf), აადგომევარ (V: Sg 1 Fut), ოჩდექ (V: Sg 1 Aor), ადგომინჴახ (Adv:), ანადგომ (Ptcp:), აადგომელა (Adv:)

იხილე: უოჩდგომელა, ფეხზე ანადგომ, ყაბარ ადგომაჲ

ადგომა

მე დგო̈ვ, მივდ.
ვდგები და მივდივარ.

ჰალა ადრი, ჰაბე დგო̈ვითა?
ჯერ ადრეა, რატომ ვდგებით?

ჰარ დილლაზე დგო̈ოდი, წყალ მოჶტანავდი.
ყოველ დილას ვდგებოდი და წყალი მომქონდა.

ხჩალ იმთავი აადგომევარ.
ხვალ დილაადრიანად ავდგები.

Fig. 5. Corpus Query Interface: Lemma and Variations

3.5 Definitions in GDC

Definitions in GDC are distributed in three different ways as follows:

- Simple variants (equivalents)
- Definition proper
- Encyclopedic information.

In the part Simple variants, one or several direct variants of dialect forms are introduced. In Definitions, information about the word that has a different meaning from the standard Georgian is given. The part Encyclopedic information includes different types of encyclopedic information that is entirely different from the definition.

It also contains other lexicographic information – the online dictionary has the function of adding lexicographic sources or dictionary examples. It allows to add unlimited number of dictionary examples and indicate our comments or translations.

The lexicographic sources can be selected from the reference database with predefined standard metadata. This function of the editor helps in improving the original dictionaries by adding the dictionary examples, indicating other existing lexicographic sources. The process implies combining several lexicographic sources, including fieldwork data. The lexicographic sources are linked to a particular dictionary entry, but not to the whole dictionary. These enables the users to narrow down the query by indicating a particular lexicographic source or sources.

3.6 Additional Linguistic Information for the Dictionary Entry

As described above, the dictionary also offers information on the foreign borrowings and transliteration. This can be displayed in two ways. First, from the language selection tab by choosing the language. Second, the corresponding word in the original script can be chosen in a separate tab. This information is relevant in terms of sociolinguistics and ethnolinguistics, especially for the Georgian dialects that are spoken outside Georgia. These dialects have developed in foreign environment having no contact with other Georgian languages or dialects. As a result, there are many borrowings in these dialects.

Therefore, it is very relevant for our research to mark foreign borrowing in the Georgian dialects and indicate their Georgian (dialect) synonyms. It is an additional function for these online dictionaries. The dictionary editor can assign such properties as lexical/semantic groups and word structure.

A sub-section of the word (lemma) tagset is a lexical tagset, where both simple and compound words can be classified according to its lexical type. In lexical type, we cover the following: general vocabulary, literary vocabulary and specialized vocabulary. The literary vocabulary includes proverbs, expressions etc. The terms in specialised vocabulary have indicated the relevant field. A special attention is paid to the multi-element collocations.

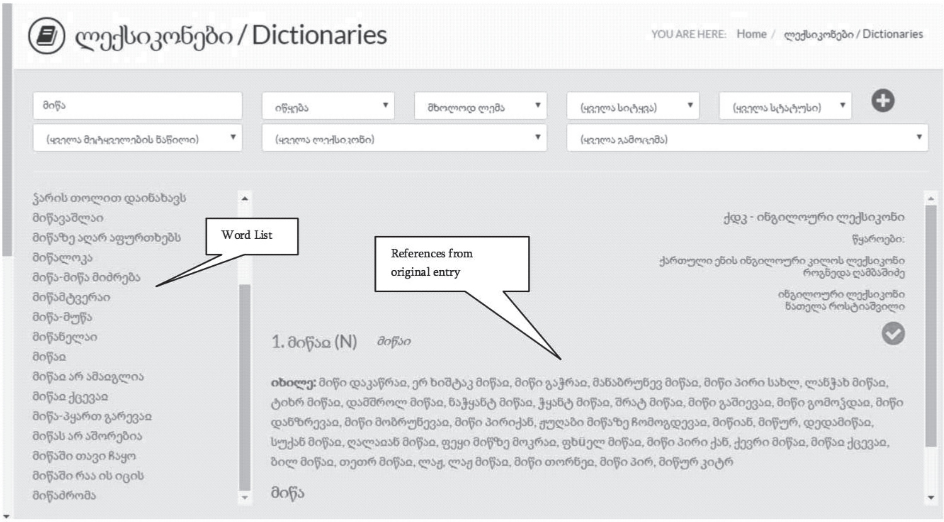
Internal links. The synonymous words will be interlinked in the dictionary. The primary links connect synonymous words both within one dialect dictionary and in other dialect dictionaries.

3.7 Single Term Equals Single Sense

In the dialect dictionaries, word senses including ambiguous, figurative, specialized etc. are presented as a separate dictionary entry. For example, the dictionary entry mic'ai ('soil'/ 'land') in Ingiloan dialect dictionary is realized in 33 different ways. In our dictionary, all these sense are presented as separate dictionary entries that are connected with one another with the internal link.

Thus, the dictionary entry from the comprehensive lexicographic sources [7] are represented with a rather short description in our lexicographic database (mic'ai ('soil'/ 'land'). Also, the dictionary entry contains two examples. The rest of information from the lexicographic sources is distributed among 33 new dictionary entries compiled from these sources. In the new dictionary, the entries, lexicographic sources and examples are provided.

In some cases, information about the alternative sources is also added, such as information about other dictionaries, corpora, authors etc. In some cases an extra description about the entry provided when necessary. All new dictionary entries are interlinked with the original dictionary list (Fig. 7).



ლექსიკონები / Dictionaries

YOU ARE HERE: Home / ლექსიკონები / Dictionaries

მიწა

იწყება

მხოლოდ ლემა

(ყველა სიტყვა)

(ყველა სტატუსი)

(ყველა მეტყველების ნაწილი)

(ყველა ლექსიკონი)

(ყველა გამოცემა)

ჯარის თოლით დაინახავს
მიწავაშლაი
მიწაზე აღარ აფურთხებს
მიწალოკა
მიწა-მიწა მიძრება
მიწამტვერაი
მიწა-მუწა
მიწანელაი
მიწაჲ
მიწაჲ არ ამაჲგლია
მიწაჲ ქცევაჲ
მიწა-პყართ გარევაჲ
მიწას არ აშორებია
მიწაში თავი ჩაყო
მიწაში რაა ის იცის
მიწაძრომა

ქდკ - ინგილოური ლექსიკონი

წყაროები:

ქართული ენის ინგილოური კილოს ლექსიკონი
როგნედა ღამბაშიძე

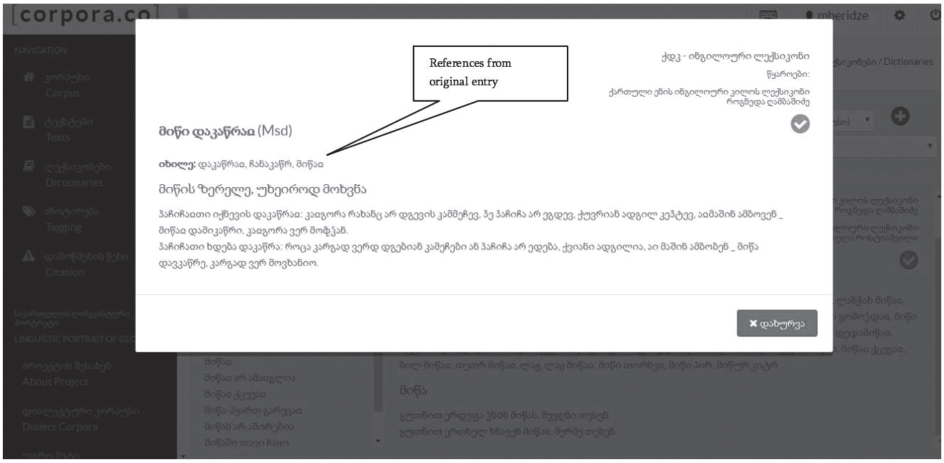
ინგილოური ლექსიკონი
ნათელა როსტიაშვილი

1. მიწაჲ (N) *მიწაი*

იხილე: მიწი დაკაწრაჲ, ერ ხიშტაკ მიწაჲ, მიწი გაჭრაჲ, მანაბრუნევ მიწაჲ, მიწი პირი სახლ, ლანჭახ მიწაჲ, ტიხრ მიწაჲ, დამშროლ მიწაჲ, ნაჭყანტ მიწაჲ, ჭყანტ მიწაჲ, შრატ მიწაჲ, მიწი გაშიევაჲ, მიწი გომოჴდაჲ, მიწი დანზრევაჲ, მიწი მობრუნევაჲ, მიწი პირიქან, ჟუღაბი მიწაზე ჩომოგდევაჲ, მიწიან, მიწურ, დედამიწაჲ, სუქან მიწაჲ, ღალაჲან მიწაჲ, ფეყი მიწზე მოკრაჲ, ფხüელ მიწაჲ, მიწი პირი ქან, ქევრი მიწაჲ, მიწაჲ ქცევაჲ, ბილ მიწაჲ, თეთრ მიწაჲ, ლაჟ, ლაჟ მიწაჲ, მიწი თორნეჲ, მიწი პირ, მიწურ კიტრ

მიწა

Fig. 6. Internal Links in the Dictionaries – Example 1



ქდკ - ინგილოური ლექსიკონი

წყაროები:

ქართული ენის ინგილოური კილოს ლექსიკონი
როგნედა ღამბაშიძე

მიწი დაკაწრაჲ (Msd)

იხილე: დაკაწრაჲ, ჩანაკაწრ, მიწაჲ

მიწის ზერელე, უხეიროდ მოხვნა

ჰაჩიჩაჲთი იქნევის დაკაწრაჲ: კაჲგორა რახანც არ დგევის კამეჩევ, ჰე ჰაჩიჩა არ ეგდევ, ქუვრიან ადგილ კეჭტევ, აჲმაშინ ამბოვენ _ მიწაჲ დამიკაწრი, კაჲგორა ვერ მოჭჴან.
ჰაჩიჩათი ხდება დაკაწრა: როცა კარგად ვერდ დგებიან კამეჩები ან ჰაჩიჩა არ ედება, ქვიანი ადგილია, აი მაშინ ამბობენ _ მიწა დავკაწრე, კარგად ვერ მოვხანით.

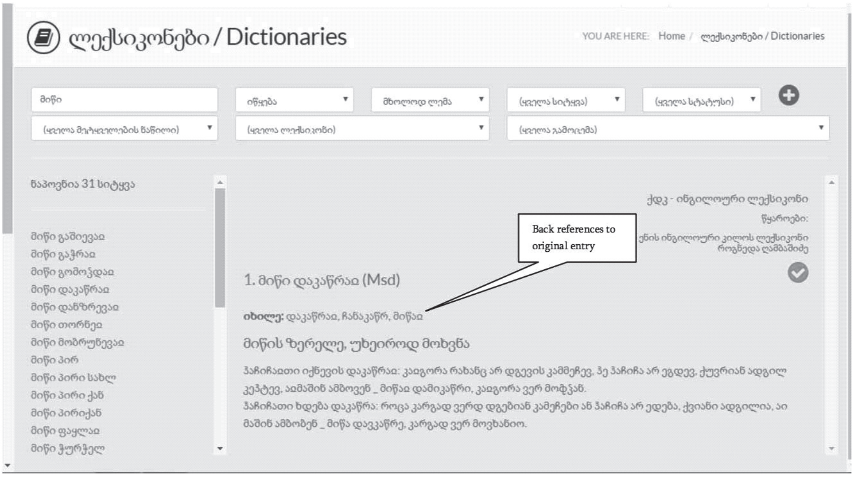
✖ დახურვა

Fig. 7. Internal Links in the Dictionaries – Example 2

Figure 8 shows the link from the original dictionary entry to the new dictionary entry – miçi daḳaçraj ‘to plow soil unfruitfully’. In this example, a new dictionary entry provides the description of the lexicographic sources and the dictionary examples.

The internal links are bidirectional, Figure 8 shows the links of the new dictionary entry with the original entry, also with other dictionary units:

miçi daḳaçraj> miçaj
miçi daḳaçraj>daḳaçraj



ლექსიკონები / Dictionaries

YOU ARE HERE: Home / ლექსიკონები / Dictionaries

მიწი | იწყება ▾ | მხოლოდ ლემა ▾ | (ყველა სიტყვა) ▾ | (ყველა სტატუსი) ▾

(ყველა მეტყველების ნაწილი) ▾ | (ყველა ლექსიკონი) ▾ | (ყველა გამოცემა) ▾

ნაპოვნია 31 სიტყვა

მიწი გაშიევაჲ
მიწი გაჭრაჲ
მიწი გომოჭდაჲ
მიწი დაკაწრაჲ
მიწი დანზრევაჲ
მიწი თორნეჲ
მიწი მობრუნევაჲ
მიწი პირ
მიწი პირი სახლ
მიწი პირი ქან
მიწი პირიქან
მიწი ფაყლაჲ
მიწი ჭურჭელ

ქდკ - ინგილოური ლექსიკონი

წყაროები:

ენის ინგილოური კილოს ლექსიკონი
როგნედა ღამბაშიძე

Back references to original entry

1. მიწი დაკაწრაჲ (Msd)

იხილე: დაკაწრაჲ, ჩანკაწრ, მიწაჲ

მიწის ზერელე, უხეიროდ მოხვნა

ჰაჩიჩაჲთი იქნევის დაკაწრაჲ: კაჲგორა რახანც არ დგევის კამეჩევ, ჰე ჰაჩიჩა არ ეგდევ, ქუვრიან ადგილ კეჰტევ, აჲმაშინ ამბოვენ _ მიწაჲ დამიკაწრი, კაჲგორა ვერ მოჭჯან.
ჰაჩიჩათი ხდება დაკაწრა: როცა კარგად ვერდ დგებიან კამეჩები ან ჰაჩიჩა არ ედება, ქვიანი ადგილია, აი მაშინ ამბობენ _ მიწა დავკაწრე, კარგად ვერ მოვხანით.

Fig. 8. Internal Links in the Dictionaries – Example 3

3.8 Morphosyntactic and Semantic Markers

Morphosyntactic and Semantic Markers for the GDC have been developed taking into consideration the existing standards (TEI, EAGLES, The Leipzig Glossing Rules) and the Georgian National Corpus (GNC: `http://gnc.gov.ge`). Due to the specific features of the Georgian dialects, the markers were partially harmonised and modified, the additional markers were introduced when necessary.

Particularly, the most difficult tasks are related to the Georgian dialects that are spoken outside Georgia, such as the Fereydanian dialect in Iran. This dialect has been under direct influence of the Persian language having no contact with any Georgian languages or dialects for about four centuries. The new steps for cultural reintegration started by the end of 20th century. There are some new developments in the dialect and to capture these new or specific features, we have introduced new markers, as follows:

Pseudo _ incorrect forms from literary Georgian that are attested in the corpus data.

New _ Lexical parallel forms due the influence of the Georgian language.

Morphosyntactic and semantic annotation in the dictionary is carried out according to two hierarchical lists: In the first hierarchy, there are markers for lemma only, whereas in the second one, we have tags for morphological features. All these tags are selected and added to the relevant lexical unit and indicated in a separate tab.

3.9 The Dictionary Query Functions

The interface of the online dictionaries in the GDC allows for:

- Search for word, or a part of word
- Search for lemma or lemma variations
- Search for foreign borrowings, with indicating a specific language
- Search through a particular dialect
- Search according to a particular part-of-speech

- Search according to a particular grammatical features
- Search according to the status of the dictionary entry, e.g. in-progress etc.
- Search for completed dictionary entry
- Search according to the lexicographic source.

4 FUTURE PROSPECTS

The research team is currently working on the new project (funded by the Shota Rustaveli National Science Foundation) on the Lexicographic Database for Georgian dialects. The project aims to develop a comprehensive lexicographic database for the Georgian dialects and to build a dialect platform with the user interface. In this platform, we plan to integrate the corpus and lexicographic data of the Georgian dialects. This research will also examine the cartographic visualization of linguistic diversity of the Georgian dialects.

One of the plans of our research team is also to improve morphosyntactic and semantic annotation in the corpus. This will allow us to build different types of dictionaries, for example dictionaries only nominals, verbs, foreign borrowing dictionaries, phrase/collocation dictionaries etc.

The Georgian Dialect Corpus and its lexicographic database represents both synchronic and diachronic aspects of the dialects. We are currently developing the database about migrations through cartographic visualization. These will allow us to analyse linguistic data taking into account the linguistic area and migration.

The Georgian dialect corpus and visualization of linguistic diversity through cartography in the corpus will be one of the main linguistic resources. It can be introduced in teaching modules at universities. Also, it can be used in linguistic and interdisciplinary research.

References

[1] Beridze, M. et al. (2009). The Corpus of Georgian Dialects. In *Proceedings of the NLP, Corpus Linguistics, Corpus Based Grammar Research. Fifth International Conference Smolenice*, pages 25–35, Jazykovedný ústav Ľ. Štúra SAV, Bratislava, Slovakia.

[2] Beridze, M. et al. (2015). The Georgian Dialect Corpus: Problems and prospects. In *Proceedings of the conference on Historical Corpora Challenges and Perspectives*, pages 323–333, Frankfurt, Germany.

[3] Beridze, M. et al. (2011). Dictionary as a textual component of Corpus (Georgian Dialect Corpus). In *Proceedings of the conference on corpus linguistic*s, pages 92–97, St. Petersburg, Russia.

[4] Beridze, M. et al. (2014). Lexicographical concept of Georgian Dialect Corpus and problems of morphological analysis. In *Proceedings of the conference on Applied Linguistics in Science and Education*, pages 91–94, St. Peterburg, Russia.

[5] Beridze, M., Lortkipanidze, L., and Nadaraia, D. (2015). Dialect Dictionaries in the Georgian Dialect Corpus. In *Logic, Language, and Computation. 10th International Tbilisi Symposium on Logic, Language, and Computation, TbiLLC 2013*, pages 82–96, Springer, Berlin – Heidelberg, Germany.

[6] Beridze, M. et al. (2016). Lexicographic Potential of the Georgian Dialect Corpus. In *Proceedings of the XVII EURALEX International Congress, Lexicography and Linguistic Diversity*, pages 300–309, Ivane Lavakhishvili Tbilisi State University, Tbilisi, Georgia.

[7] Gambashidze, R. (1988). *Dictionary of Ingiloan dialect of Georgian Language*. Ganatleba, Tbilisi.

[8] Meskhishvili, M. et al. (1981). *Dictionary of Kartlian Diallect*. Metsniereba, Tbilisi.

[9] Dzotsenidze, K. (1974). *Upper Imeretian Dictionary*. Ivane Javakhishvili Tbilisi State University, Tbilisi.

[10] Gachechiladze, P. (1976). *Lexical material of the Imeretian dialect*. Metsniereba, Tbilisi.

[11] Chincharauli, Al. (2005). *Khevsurian Dictionary*. Kartuli Ena, Tbilisi.

[12] Tsotsanidze, G. (2012). *Dictionary of Tushian Dialect*. Sulakauri Publishing, Tbilisi.

[13] Nizharadze, S. (1975). *Adjarian dialect*. Sabchota Adjara, Batumi.

[14] Glonti, Al. (1975). *Dictionary of Georgian dialects*. Ganatleba, Tbilisi.

[15] Martirosov, A. (1985). The Main Issues of the Study of Georgian Dialect Vocabulary and Compilation of Dictionaries. *Ibero-Caucasian linguistics*, XXIII:139–148.

[16] *Explanatory dictionary of the Georgian language*. (1950–1964). 8 vols. Georgian Academy of Sciences [in Georgian]. Georgian Academy of Sciences, Tbilisi.

[17] Georgian National Corpus. Accessible at: `http://gnc.gov.ge.`

[18] Georgian Dialect Corpus. Accessible at: `http://corpora.co.`

MODELING SEMANTIC DISTANCE IN THE PATTERN DICTIONARY OF ENGLISH VERBS

SILVIE CINKOVÁ – ZDENĚK HLÁVKA
Faculty of Mathematics and Physic, Charles University, Prague, Czech Republic

CINKOVÁ, Silvie – HLÁVKA, Zdeněk: Modeling Semantic Distance in the Pattern Dictionary of English Verbs. Journal of Linguistics, 2017, Vol. 68, No 2, pp. 122 – 135.

Abstract: We explore human judgments on how well individual patterns of 29 target verbs from the Pattern Dictionary of English Verbs describe their random KWICs. We focus on cases where more than one pattern is judged as highly appropriate for a given KWIC and seek to estimate the effect of event participants (arguments) being denotatively similar in two patterns, considering all pair combinations in a given lemma. We compare this effect to the effect of several contextual features of the KWICs, the effect of paired PDEV implicatures implying each other, and the effect of belonging to a given lemma. We show that the lemma effect is still stronger than any feature going across lemmas we have examined so far, so that each verb appears to be a little universe in its own right.

Keywords: usage patterns, lexicography, verbs, CPA, semantics, word embeddings, WSD, graded decisions, corpus, English, annotation

1 INTRODUCTION

Since many verbs are perceived as highly polysemous, their senses are both difficult to determine when building a lexicon entry and to distinguish in context when performing Word Sense Disambiguation (WSD). An alternative to verb senses is *usage patterns* coined by Hanks in the *Pattern Dictionary of English Verbs* (PDEV, [1], Fig. 1). Previous studies ([2], [3]) have shown that PDEV represents a valuable lexical resource for WSD, in that annotators reach good interannotator agreement despite the semantically fine-grained microstructure of PDEV.

Recently, we created a data set annotated with graded decisions (VPSGradeUp, cf. Section 2.2) from PDEV to investigate features suspected of blurring distinctions between the patterns [4]. We have been preliminarily considering features related to the KWICs independently of the lexicon design, such as *finiteness*, *argument opacity*, and *factuality*[1] of the target verb on the one hand, and those related to the lexicographical design of PDEV, such as *textual entailment* between PDEV *implicatures* within a lemma or *denotative similarity of the verb arguments* on the other hand.

This paper focuses on the denotative similarity of the verb arguments. We have attempted to approach it in a quantitative way by modeling it as *semantic distance* between the *corresponding syntactic slots* in pattern definitions in a PDEV lemma (henceforth *colempats*), as comprehensively described in Section 6. We compare all colempats pairwise, examining their scores in the graded decision annotation (see

[1] For explanation of terms used in this section, kindly refer to Section 2.

DOI 10.1515/jazcas-2017-0023

Section 2.2) with respect to how much they compete to become the most appropriate pattern, as well as the semantic distance between their subjects, objects, and adverbials. To quantify the comparisons, we have introduced a measure of *rivalry* for each pair of colempats. Rivalry increases, the more *appropriate* both colempats are considered for a given KWIC and the more similar their *appropriateness* scores are (see Sections 4.1 and 4.2). We have observed a significant association between high *rivalry* in paired colempats and their corresponding arguments being labeled with denotatively similar semantic labels (henceforth *semlabels*, see Section 2.1).

2 RELATED WORK AND IMPORTANT TERMS

2.1 Pattern Dictionary of English Verbs (PDEV)s

This section not only gives a brief description of PDEV, butalso introduces key terms that will be used throughout this paper.

PDEV's core idea is that a verb has no meaning in isolation; instead, it has a *meaning potential*, whose diverse components and their combinations are activated by contexts. To capture the meaning potential of a verb, the PDEV lexicographer manually clusters random KWICs into a set of prototypical usage patterns, considering the semantic and morphosyntactic similarity alike (Fig. 1).

Each PDEV *pattern* contains a *pattern definition* and an *implicature* to explain or paraphrase its meaning. Both are shaped as finite clause templates where important syntactic slots are populated with *semantic type* labels, alternatively with a set of collocates (*lexical set*), and, complementary to both, optional *semantic roles*. This paper merges them all under the umbrella term *semlabels*.

2.2 Graded Decisions: an Alternative WSD Setup

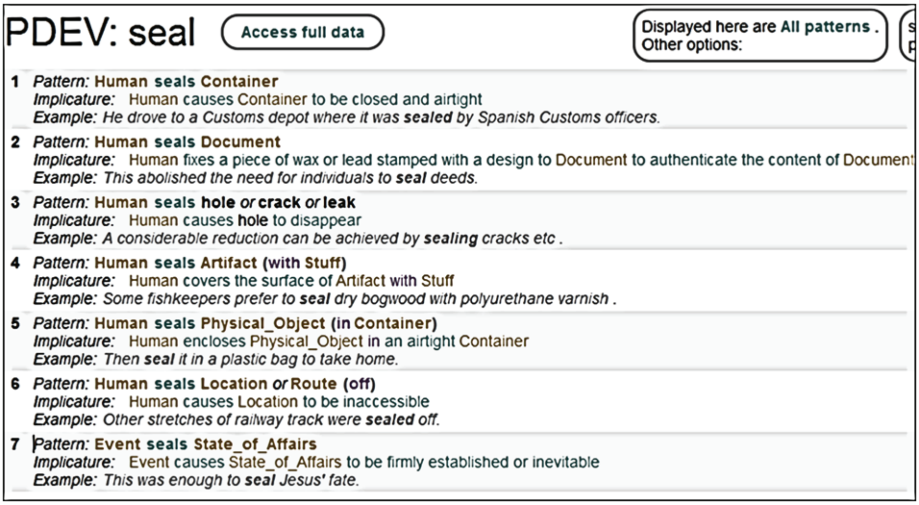
The graded-decision data set used in this paper draws on Erk et al. [5], who experimented with the WSD setup: instead of assigning a single sense to a given context of the key word, the annotators indicated on a Likert scale[2] how well each sense matched a given KWIC, allowing for ties. The data set has two subsets: WSim with graded decisions on matching relations between WordNet synsets and KWICs of 11 selected key word lemmas; and USim with graded decisions on how well two different words in two different KWICs paraphrase each other. Both displayed very good annotator correlation, suggesting graded decisions be a sensible alternative to the traditional WSD setup.

2.3 Verb Finiteness

Finiteness is a morphosyntactic category associated with verbs. Virtually all verbs appear in finite as well as infinite forms when used in context. A finite verb form is such a verb form that expresses person and number. Languages differ in whether these categories are expressed morphologically (e.g. by affixes or stem vowel changes) or syntactically (obligatorily complemented with a noun/pronoun

[2] Likert scale is a psychometric scale used in opinion surveys. It enables the respondents to scale their agreement/disagreement with a given opinion.

expressing these categories explicitly). Finite forms are typically all indicative and conditional forms, as well as some imperative forms, e.g. *reads, are reading, (they) read, čtu, gehst, allons!*. Infinite forms are infinitives *(to read, to have read, to be heard, to have been heard)* and participles along with gerunds and *supines (reading, known, deleted, försvunnit)*. The grammars of many languages know diverse other finite as well as infinite verb forms. Infinite forms typically allow more argument omissions than finite forms: *to go to town* vs. **went to town*. This suggests that descriptions of events rendered by infinite verb forms may be more vague, and, in terms of annotation, more prone to match several different patterns/senses at the same time. Verb finiteness is easy to determine, and therefore it was only annotated by one annotator in our data set.



PDEV: seal Access full data Displayed here are All patterns .
Other options:

1 *Pattern:* Human seals Container
Implicature: Human causes Container to be closed and airtight
Example: He drove to a Customs depot where it was ***sealed*** *by Spanish Customs officers.*

2 *Pattern:* Human seals Document
Implicature: Human fixes a piece of wax or lead stamped with a design to Document to authenticate the content of Document
Example: This abolished the need for individuals to ***seal*** *deeds.*

3 *Pattern:* Human seals hole *or* crack *or* leak
Implicature: Human causes hole to disappear
Example: A considerable reduction can be achieved by ***sealing*** *cracks etc .*

4 *Pattern:* Human seals Artifact (with Stuff)
Implicature: Human covers the surface of Artifact with Stuff
Example: Some fishkeepers prefer to ***seal*** *dry bogwood with polyurethane varnish .*

5 *Pattern:* Human seals Physical_Object (in Container)
Implicature: Human encloses Physical_Object in an airtight Container
Example: Then ***seal*** *it in a plastic bag to take home.*

6 *Pattern:* Human seals Location *or* Route (off)
Implicature: Human causes Location to be inaccessible
Example: Other stretches of railway track were ***sealed*** *off.*

7 *Pattern:* Event seals State_of_Affairs
Implicature: Event causes State_of_Affairs to be firmly established or inevitable
Example: This was enough to ***seal*** *Jesus' fate.*

Fig. 1. PDEV entry

2.4 Argument Opacity

Argument opacity typically, but not necessarily, relates to verb finiteness. By argument opacity we mean how many arguments relevant for disambiguation of the target verb are either omitted in the context (e.g. subject in infinitive) or ambiguous or vague. Ambiguous and vague arguments are often arguments expressed by personal pronouns that refer to entities mentioned distantly from the target verb, sometimes even not directly, but by longer chains of pronouns (so-called *coreference* or *anaphora chains*), or arguments expressed by indefinite or negative pronouns. Some examples of opaque verb contexts:

The Greater London Council was ABOLISHED in 1986. (Who abolished it?)

The company's ability to adapt to new opportunities and capitalize on them depends on its capacity to share information and involve everyone in the organization in a systemwide search for ways to improve, ADJUST, adapt, and upgrade. (Who exactly adjusts what?)

This feature, too, was annotated by a single annotator in our data set. The categories were “opaque subject”, “opaque object”, “opaque arguments” (i.e.; more slots at the same time, including adverbials) or empty field.

2.5 Textual Entailment Between PDEV Implicatures

Textual Entailment, mentioned in the Introduction, is the key notion of Recognizing Textual Entailment (RTE), a computational-linguistic discipline coined by Dagan et al. [6]. The task of RTE is to determine, “given two text fragments, whether the meaning of one text can be inferred (entailed) from another text. More concretely, the applied notion of textual entailment is defined as a directional relationship between pairs of text expressions, denoted by T the entailing ‘text’ and by H the entailed ‘hypothesis’. We say that T entails H if, typically, a human reading T would infer that H is most probably true”. So, for instance, the text *Norway’s most famous painting, ‘The Scream’ by Edvard Munch, was recovered yesterday, almost three months after it was stolen from an Oslo museum* entails the hypothesis *Edvard Munch painted ‘The Scream’* [6]. We have pursued a double RTE annotation of pairs of PDEV implicatures, measured the interannotator agreement, and investigated its effect on pattern distinctions [7].

2.6 Word Embeddings

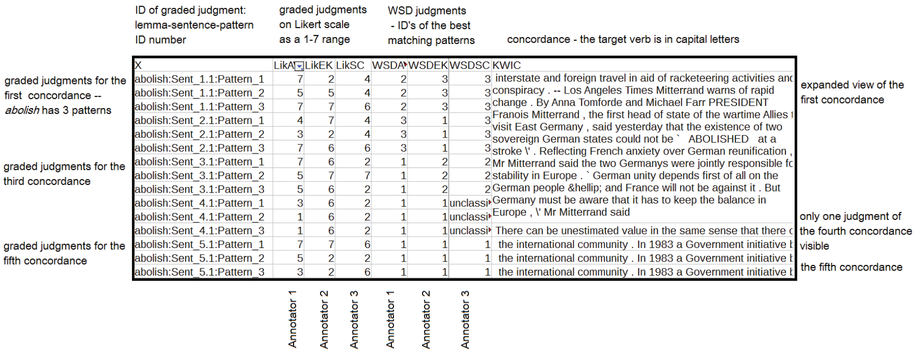
To compute the semantic distance between semlabels, we used *word embeddings*. Word embeddings are vector representations of the individual words in a corpus. Each vector dimension represents a word from that corpus. The more similarly words are distributed in the corpus, the more similar the directions of their respective vectors are; that is, the mutual semantic similarity of words is quantified by the (cosine) similarity of their vector representations. Cosine similarity renders the correlation coefficient between these two vectors, ranging from -1 to 1. *Word embeddings* is, loosely speaking, a quantitative expression of the Firthian “knowing the word by the company it keeps”. The embeddings used in this paper are based on Word2Vec [8], a neural network trained to reconstruct linguistic contexts of words. We use its implementation for R, *text2vec* [9].

3 GRADED DECISIONS ON VERB USAGE PATTERNS

3.1 VPS-GradeUp

The VPS-GradeUp data set draws on Erk’s experiments with paraphrases (USim). It consists of both graded-decision and classic-WSD annotation of 29 randomly selected PDEV lemmas: *seal, sail, distinguish, adjust, cancel, need, approve, conceive, act, pack, embrace, see, abolish, advance cure, plan, manage, execute, answer, bid, point, cultivate, praise, talk, urge, last, hire, prescribe*, and *murder*. Each lemma comes with 50 KWICs processed by three annotators in parallel.

In the graded-decision part, the annotators judged for each pattern how well it described a given KWIC, on a Likert scale. In the WSD part, each KWIC was assigned one best-matching pattern. The entire data set contains WSD judgments on 1 450 KWICs, corresponding to 11 400 graded decisions. A more detailed description of VPS-GradeUp is given by Baisa et al. [4]. Fig. 2 presents the most essential annotation elements.



X	LikA	LikEK	LikSC	WSDA	WSDEK	WSDSC	KWIC
abolish:Sent_1.1:Pattern_1	7	2	4	2	3	3	interstate and foreign travel in aid of racketeering activities anc
abolish:Sent_1.1:Pattern_2	5	5	4	2	3	3	conspiracy . -- Los Angeles Times Mitterrand warns of rapid
abolish:Sent_1.1:Pattern_3	7	7	6	2	3	3	change . By Anna Tomforde and Michael Farr PRESIDENT
abolish:Sent_2.1:Pattern_1	4	7	4	3	1	3	Franois Mitterrand , the first head of state of the wartime Allies t
abolish:Sent_2.1:Pattern_2	3	2	4	3	1	3	visit East Germany , said yesterday that the existence of two
abolish:Sent_2.1:Pattern_3	7	6	6	3	1	3	sovereign German states could not be ` ABOLISHED at a
abolish:Sent_3.1:Pattern_1	7	6	2	1	2	2	stroke \' . Reflecting French anxiety over German reunification ,
abolish:Sent_3.1:Pattern_2	5	7	7	1	2	2	Mr Mitterrand said the two Germanys were jointly responsible fc
abolish:Sent_3.1:Pattern_3	5	6	2	1	2	2	stability in Europe . ` German unity depends first of all on the German people &hellip; and France will not be against it . But
abolish:Sent_4.1:Pattern_1	3	6	2	1	1	unclassi	Germany must be aware that it has to keep the balance in
abolish:Sent_4.1:Pattern_2	1	6	2	1	1	unclassi	Europe , \' Mr Mitterrand said
abolish:Sent_4.1:Pattern_3	1	6	2	1	1	unclassi	There can be unestimated value in the same sense that there c
abolish:Sent_5.1:Pattern_1	7	7	6	1	1	1	the international community . In 1983 a Government initiative b
abolish:Sent_5.1:Pattern_2	5	2	2	1	1	1	the international community . In 1983 a Government initiative b
abolish:Sent_5.1:Pattern_3	3	2	6	1	1	1	the international community . In 1983 a Government initiative b

Fig. 2. Human judgments in Graded-Decisions and WSD tasks. Each line contains one graded judgment by all annotators. The WSD judgments repeat as many times as there are patterns to judge for each KWIC

4 MEASURES OF APPROPRIATENESS AND RIVALRY

4.1 Appropriateness

The appropriateness of a pattern for a given KWIC line is based on the triple of annotation judgments and conflates their sum and standard deviation in this formula:

$$Appropriateness = \sum(x) - \frac{sd(x)}{3.5}$$

The function returns values in the range of 3 to 21. The 3.5 coefficient is roughly the maximum standard deviation (sd) possible with three judgments ranging from 1 to 7. (The x value must be a natural number ranging from 1 to 7 and the sum must be the sum of exactly 3 such x.) Appropriateness reflects both the mean and the dispersion of the judgments, unlike mere mean or median.

4.2 Rivalry

To compare the competition between PDEV patterns in pairs, we have introduced *rivalry*. Rivalry always concerns the appropriateness rates for a pair of patterns of one lemma (*colempats*, see Section 1), being computed for all pairs. Rivalry increases with the *appropriateness* of each colempat (Section 4.1) and with decreasing difference between the appropriateness values in the given colempat pair: the higher the rivalry, the more the two patterns compete for becoming selected as the best match in the WSD annotation. The rivalry function is simple:

$$Rivalry = max(appr_{pair}) - max((appr_{pair}) - min(appr_{pair})) = min(appr_{pair})$$

Under $appr_{pair}$ we understand two computed appropriateness values of patterns in a colempat pair: $max(appr_{pair}$ and ”; $min(appr_{pair}$. They represent the higher and the lower appropriateness, respectively. Hence, rivalry is defined as the difference between the higher appropriateness value and the difference between that and the lower appropriateness value, which boils down to the lower appropriateness.

It is to be emphasized that rivalry is always computed *on a given KWIC*. Hence we cannot tell e.g. the rivalry between *abandon_1* and *abandon_3* in general, but we get one rivalry value of this colempat pair for each of the 50 KWICs.

Measuring rivalry is interesting, even though we have not yet abstracted from individual KWICs; it enables us to quickly and consistently identify cases of pattern overlap for further analysis of both the design of the patterns and of contextual features in the KWICs affected.

5 LEMMA-WISE CLUSTER ANALYSIS OF PATTERNS AND CONCORDANCES ACCORDING TO APPROPRIATENESS

The appropriateness function allows us a first visual overview of the individual lemmas regarding how well and how many colempats matched each individual KWIC. Using a standard clustering algorithm, we created a heatmap diagram for each lemma. Fig. 4 shows heatmaps for six selected lemmas. The heatmaps revealed striking differences between the individual lemmas, leaving the impression that each verb behaves in its own way. Nonetheless, we hazarded a coarse division into three groups:

Unproblematic lemmas, such as *murder* and *hire*, represent verbs whose PDEV patterns are distinct. For *murder*, the data contained only usages covered by Pattern 1. Two KWICs were not covered well by any pattern, although Pattern 1 marginally matched in the second last KWIC (white). In case of *hire*, the PDEV patterns were also distinct, with all three even occurring.

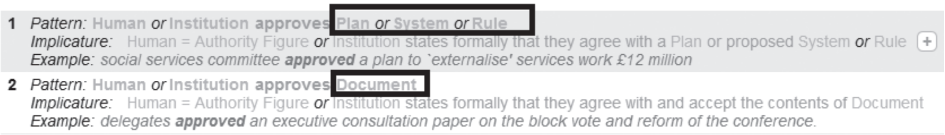
Lemmas with competing patterns. The lemmas *approve* and *last* reveal high pattern rivalry – an indication that something either in the entry design or in the contexts makes sharp distinction impossible and is potentially harmful for the WSD-interannotator agreement outcome. We carried out a manual analysis of these cases, which yielded a pool of possible features that increase rivalry, on one of which we focus in this paper.

Lemmas with many KWICs uncaptured by the current PDEV patterns. For instance, the lemmas *pack* and *seal* contain a non-trivial proportion of KWICs in which all patterns display low appropriateness by the current PDEV patterns. This suggests a problem in the entry. Two preliminary explanations are that either the KWICs contain unknown usages and that the entry ought to be complemented with new pattern(s), or that a number of patterns contain a relevant interpretation aspect, but not the ones perceived as important.

6 SEMANTIC DISTANCE BETWEEN SEMLABELS IN PDEV

6.1 Selectional Preferences of Verb Uses Modeled by Synslots and Semlabels

The manual inspection of clusters (Section 5) revealed that KWICs often contain objects that happen to match two semantic types at the same time, due to regular polysemy [10] or semantic coercion [11]. These semantic types can be the only aspect in which two pattern definitions differ. When, moreover, the implicatures of these two pattern definitions *entail* one another (see Section 2.5), the two patterns (colempats) become intuitively hard to distinguish, as is repeatedly illustrated by high rivalry scores of such colempat pairs. In our sample, this happened most prominently in colempats *cancel_2/cancel_6 and approve_2/approve_1*. (see Fig.3).



1 *Pattern:* Human *or* Institution approves Plan *or* System *or* Rule
Implicature: Human = Authority Figure *or* Institution states formally that they agree with a Plan or proposed System *or* Rule +
Example: *social services committee* ***approved*** *a plan to `externalise' services work £12 million*

2 *Pattern:* Human *or* Institution approves Document
Implicature: Human = Authority Figure *or* Institution states formally that they agree with and accept the contents of Document
Example: *delegates* ***approved*** *an executive consultation paper on the block vote and reform of the conference.*

Fig. 3. PDEV entry of the verb *approve*. Both the two colempats are transitive, and their subjects are populated with the semantic types Human and Institution. Their pattern definitions differ only in the semantic types of their direct objects (Plan/System/Rule vs. Document). Whenever a KWIC contains a noun such as *contract*, *agreement*, and *treaty* as its direct object, these two colempats become indistinguishable, since Plan/System/Rule are more often than not semantically associated with Document, and the implicatures entail each other, too

Therefore, when considering factors potentially increasing rivalry, the similarity of selectional preferences is among the most prominent suspects. We introduce a measure of *semantic distance*, which operationalizes the similarity of selectional preferences to quantify its effect on colempat rivalry.

The concept of *semantic distance between corresponding synslots in a colempat pair* is based on the following understanding of the PDEV entry, as already sketched out in Section 2.1: each PDEV entry consists of patterns. Patterns consist of pattern definitions and implicatures. Each pattern definition consists of a main predicate constituted by the target verb (entry lemma) and a number of arguments (syntactic slots, *synslots,* with different syntactic functions), which are populated by *semlabels*. The lemma-pattern combination (e.g. *abandon_1*) is called *lempat*. Lempats with the same lemma are called *colempats*. A *colempat pair* is a pair thereof. Verb arguments are called *synslots*. Semlabels populating each synslot along with the syntactic functions of the synslots represent the selectional preferences of the target verb in a given pattern. Selectional preferences in verbs describe "knowledge about possible and plausible fillers for a predicate's argument positions" [12, p. 723]. Any clustering of verb occurrences according to their semantic similarity is bound to rely on the selectional preferences, and PDEV entries capture them explicitly. Based on the Distributional Hypothesis [13], we can assume that the more similar the selectional preferences of two colempats, the more semantically related the colempats are. Consequently, more similar colempats should display higher rivalry than less similar colempats.

6.2 Semlabels in Synslots and Corresponding Synslot Pairs

When modeling the selectional preferences of the target verb in colempats by semlabels in synslots, the corresponding synslots were easily extracted by their syntactic labels provided in PDEV. The actual problem was the *semantic distance* between a pair of corresponding synslots in itself: synslots are often populated with more than one semlabel (cf. Object in Pattern 6, Fig. 1), whereas we needed a one-number summary. First we computed a semantic-distance score for *each possible pair of semlabels* (more in Section 6.3) and mapped these scores on observed pairs of semlabels populating the pairs of corresponding synslots. For instance, in a pair of colempats where the subject of Colempat A was populated by HUMAN and ANIMAL, while that of Colempat B by INSTITUTION, LAND, and ANIMAL, we used the semantic-distance scores for the following semlabel pairs: HUMAN-INSTITUTION, HUMAN-LAND, HUMAN-ANIMAL, ANIMAL-INSTITUTION, ANIMAL-LAND, and ANIMAL- ANIMAL. From these 6 scores, we needed a one-number summary to render the semantic distance between the entire corresponding synslots, which we computed as described in Section 6.4.

6.3 Computing the Semantic Distance Between Individual Semlabels

To determine the semantic distance between the individual semlabels used in PDEV, we exploited the Distributional Hypothesis [13] by transforming a corpus of pattern definitions and implicatures from the entire PDEV into a vector space with each token represented by a vector, whose dimensions reflected the co-occurrence with other tokens in the space. To render tokens as vectors, we used *tex2vec* [9], an implementation of word embeddings [8] for R. Although *text2vec* offers a large general-language vector space, we preferred to train our own on the text of PDEV patterns, since we were interested in the distributional similarity of the words (e.g. HUMAN) *used as PDEV labels* rather than in their regular usage.

Before training the corpus, we cleaned the PDEV data to capture all semlabels and minimize their variants. Markup variation was less a problem in the approx. 200 semantic types listed in the PDEV documentation than in lexical sets, semantic roles, and diverse grammatical markers. First, we cleared the text of punctuation, converted all tokens to lower case, and erased numerical indices (e.g. HUMAN 1). We also had the corpus lemmatized using MorphoDiTa [14]. Then we extracted all multi-word semlabels (e.g. HUMAN GROUP) and collapsed their strings into one token with underscores (obtaining e.g. *human_group*). This cleaned and lemmatized corpus was put into text2vec to obtain vector representations for each token.

When computing the vector space of PDEV with *text2vec*, we set a minimum frequency threshold to 5 and limited the number of vector dimensions to 50. Having obtained a 50-dimension vector for each token, we computed their pairwise similarities with cosine. Cosine ranges between -1 and 1 and expresses the angle between the two vectors. The more the angle differs from 0, the more dissimilar the two vectors are. We transformed the results into a range between 0 and 1 by adding 1 to each cosine value and dividing the result by 2. The resulting structure was a distance matrix for almost all pairs of semlabels (for exceptions see Section 6.4).

At this point, we had a mutual-similarity score for nearly each possible pair of PDEV semlabels, and thus we could establish the mutual similarity of all semlabel pairs within a corresponding synslot pair (cf. the example in Section 6.2 – we would have 6 cosine similarity scores to characterize the relation between the subjects of the Colempat A & Colempat B pair).

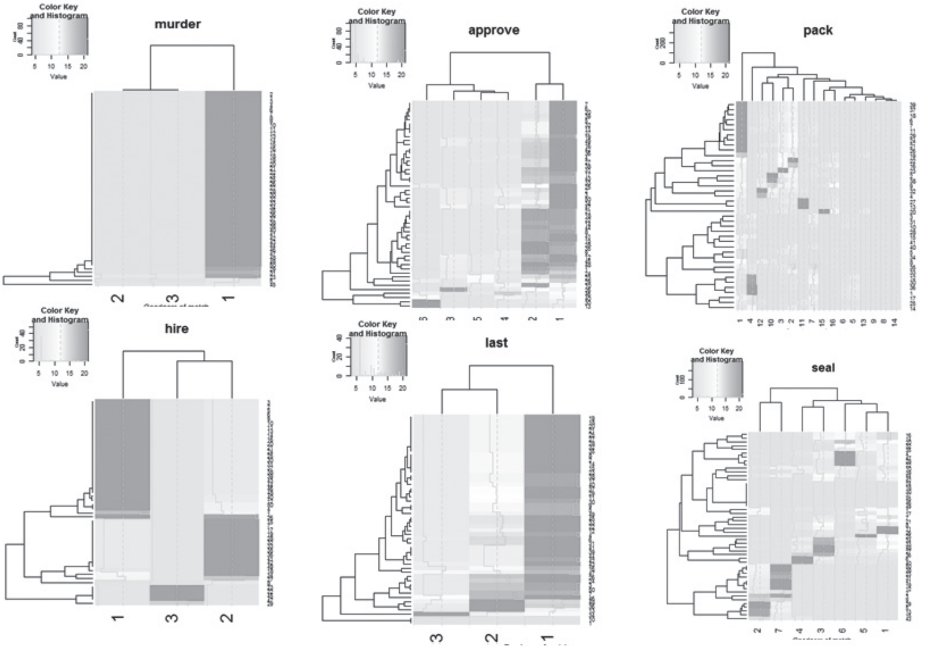


Fig. 4. Heatmaps of selected lemmas. The lines and columns represent the KWICs and the PDEV patterns, respectively. The color appropriateness scores range from turquoise (low appropriateness) to purple (high appropriateness). Each line is indexed with a unique KWIC ID on the right and is therefore easily tracked back in the corpus for a more detailed analysis

6.4 Determining the Semantic Distance Between Entire Synslots

At that point, we needed a one-number summary for each pair of corresponding synslots in a pair of colempats to be derived from the cosine similarities of individual semlabels. We defined it as the *Hausdorff distance* between the first corresponding synslot and the second corresponding synslot. The Hausdorff distance is commonly used to model the distance between two subsets of a metric space, in our case two synslots, if we pretend that the semlabels, that each of them contains, are points in the metric space. To compute the Hausdorff distance between Colempat A and Colempat B, we took one semlabel of Colempat A after another; for each we computed its distance to each semlabel in Colempat B and saved the shortest distance. From these distances we took the maximum. For this computation we adapted the *HD* function from the *polydect* R package [15] to immediately process distances between points across the subsets instead of deriving them from coordinates of each point as the original function would have done. Before we had to solve three minor issues:

1. Transformation of the individual cosine similarities to render *distance* rather than *similarity* to conform to the concept of Hausdorff distance; that is, have high score mean *dissimilarity*;
2. Model cases of only one colempat having the given synslot;
3. Dealing with semlabels for which we had not obtained a cosine similarity score.

Transformation of similarity into distance. We subtracted the cosine similarity value from 1 to get a number that would *decrease* with growing similarity instead of increasing. The resulting structure was a distance matrix for semlabel pairs, required by the adapted *HD* function.

Synslot mismatch between colempats. In many pairs, the inventory of synslots was different for each colempat; e.g. one colempat had a direct object while the other not. We modeled the syntactic mismatch as a *tenfold of the maximum distance observed in the matrix*. This is an approximation of the common annotator experience that differences in argument structure usually help distinguish patterns at the first sight, while distinctions based on semlabels in corresponding synslots are often blurred – especially when each synslot is populated by several semlabels and some of them are present in both, or when they are very general (e.g. PHYSICAL OBJECT), or if it is difficult to find a reasonable hyperonym (one-word and not too general) for them.

Missing cosine similarity/distance for a pair of semlabels. The automatic cleaning of the PDEV corpus before training the vector space by *text2vec* (Section 6.3) captured virtually all PDEV semantic types, but even so the vector space missed many members of lexical sets and semantic roles (see Section 2.1), because they were too rare to pass the frequency threshold. We set the cosine similarity/distance of pairs consisting of one or two unmatched semlabels to the *mean cosine similarity/distance observed in the data set*.

We computed the Hausdorff distances for each syntactic function of synslots separately. As a last step, we computed their sum for each colempat pair. This number rendered the semantic distance between the entire colempats.

7 ASSOCIATION BETWEEN RIVALRY AND SEMANTIC DISTANCE OF CORRESPONDING SYNTACTIC SLOTS

As the first approximation, we have run the Pearson correlation test with rivalry and the semantic distance between colempats (i.e. the sum of the semantic distances of all observed synslots in the colempat pair). The test detected a statistically significant negative correlation (p-value ≈ 0.005). This conforms to our intuition: the larger the semantic distance, the smaller the rivalry between patterns. However, the effect is very small (95% confidence interval between -0.002 and -0.004).

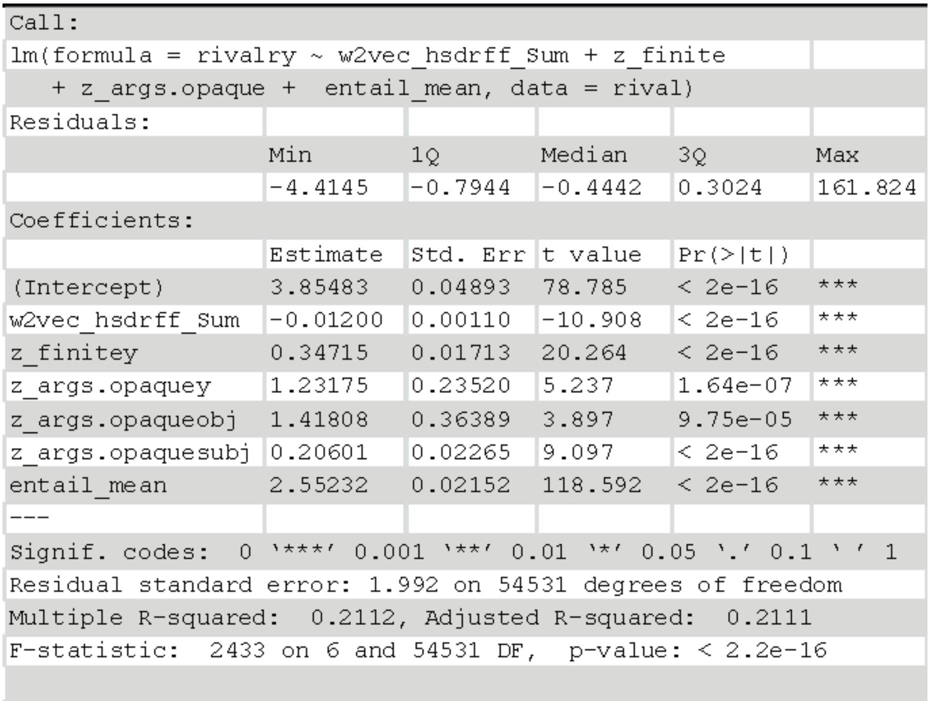
We also examined possible correlations between rivalry and the semantic distance in the individual syntactic functions of synslot pairs (i.e. subjects, objects, etc.), but the results were not statistically significant for any of them.

8 DISCUSSION AND FUTURE WORK

We have observed a statistically significant but tiny effect of the semantic distance of colempats on pattern rivalry – a far smaller one than we had expected. The annotator intuition certainly suggests that, when considering a verb as an event with participants, the cognitive characteristics of event participants are important for the perception of the event itself, and we may believe that we are able to compare how similar or different the typical participants of two events are, but there is no straightforward way to operationalize this perception. We decided to rely on *text2vec* as a robust state-of-the art device for lexical semantic analysis, which goes beyond simple distributional similarity yielded by binary distance matrices, but we naturally do not know what exactly happens behind the scenes of the neural network. Then, even if a vector representation had been the most adequate approach, we might have lost much of the actual effect size by observing the *vectors of labels* in the pattern definitions instead of *vectors of the actual words* populating the KWICs in the corresponding synslots. *Labels* were easily extracted from PDEV, while the extraction of *words* populating relevant synslots would have been much more complex, involving coreference resolution, dealing with coordinated elements, extensive manual checks, etc. Also, our approach to mismatches in the argument structure between patterns has been rather crude. In fact, our approach assigns the tenfold of the largest semantic distance to all cases where the Hausdorff distance could not be computed, which includes a synslot present in one but absent in the other colempat as well as a synslot missing in both and thus clearly blurs an important difference which could have added to the effect.

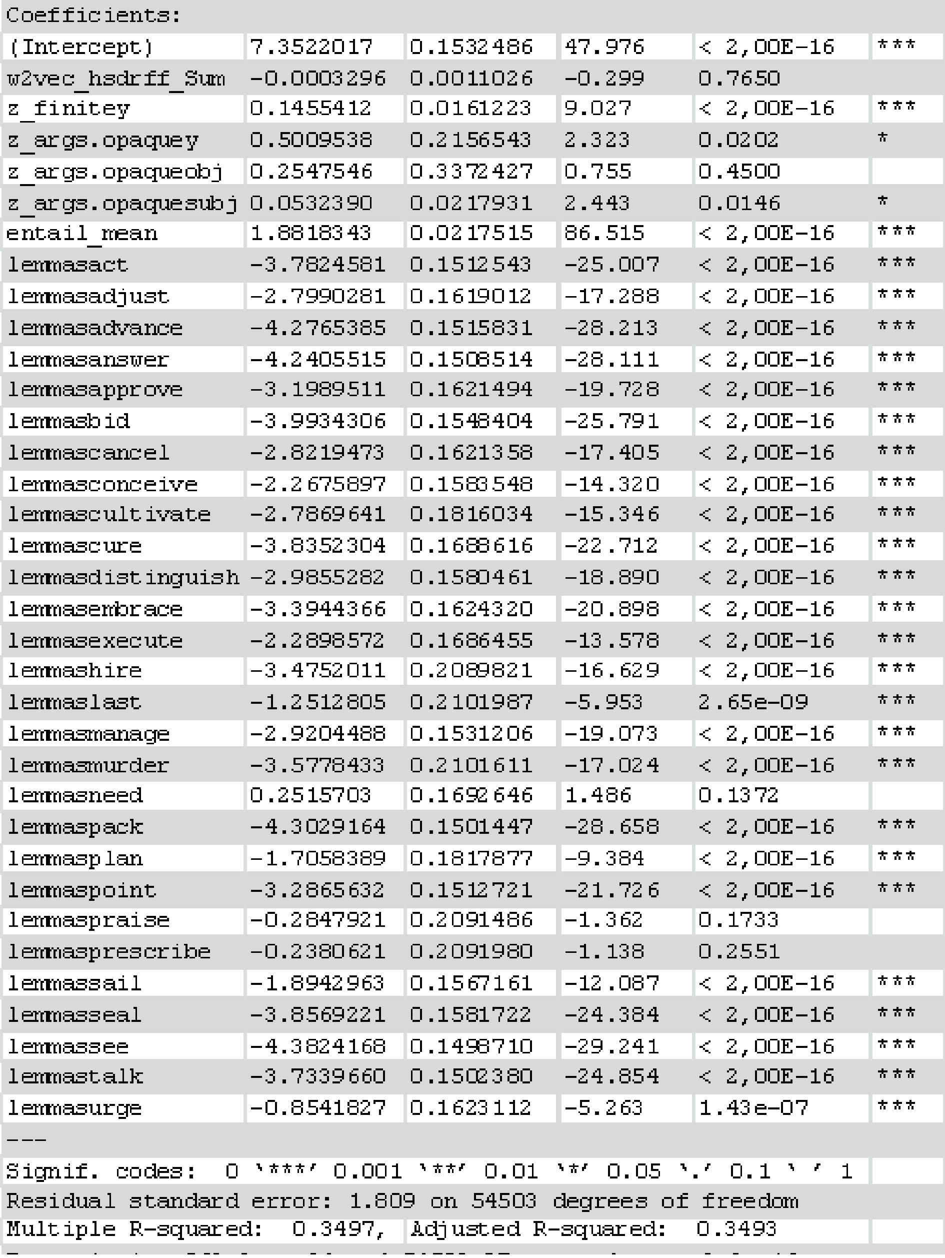
Apart from the semantic distance, we have been preliminarily examining other features suspect of increasing rivalry, such as the explicite presence/absence of relevant arguments (*argument opacity*) and *finiteness* of the target verb in the KWICs, along with a very preliminary manual annotation of *mutual implications of paired pattern implicatures* (only 2 annotators, 0.45 Cohen's κ). A statistically significant linear model predicting rivalry (Table 1) finds all these predictors significant, but the semantic distance focused in this paper turns out weakest. Interestingly, verb finiteness (promising more explicit contexts) does not help to distinguish between patterns, but in fact increases rivalry. Considering the argument opacity, the model grows slightly more powerful when the opacity of subjects and objects is singled out, with opaque object being the most rivalry increasing predictor from the opacity family (coeff. 1.42). The most effective rivalry increaser turns out to be implicatures implying each other, raising each rivalry unit by 2.55 (to the extent we can believe averaged triple human judgments on entailment). We have also been considering the factuality of the target predicates (for which we have used verb finiteness here as a primitive proxy), but a pilot annotation has yielded poor interannotator agreement, making results based on such data highly speculative.

All the aforementioned predictors are apparently not general enough to beat the effects of individual lemmas, as Table 2 reveals: most lemmas are significant, have high coefficients, and increase the predictive power of the model (cf. R-squared in both): despite efforts to find universal features, each verb appears to remain a little universe in its own right.



```
Call:
lm(formula = rivalry ~ w2vec_hsdrff_Sum + z_finite
   + z_args.opaque +  entail_mean, data = rival)
Residuals:
                  Min       1Q       Median    3Q       Max
                  -4.4145   -0.7944  -0.4442   0.3024   161.824
Coefficients:
                  Estimate  Std. Err t value   Pr(>|t|)
(Intercept)       3.85483   0.04893  78.785    < 2e-16  ***
w2vec_hsdrff_Sum  -0.01200  0.00110  -10.908   < 2e-16  ***
z_finitey         0.34715   0.01713  20.264    < 2e-16  ***
z_args.opaquey    1.23175   0.23520  5.237     1.64e-07 ***
z_args.opaqueobj  1.41808   0.36389  3.897     9.75e-05 ***
z_args.opaquesubj 0.20601   0.02265  9.097     < 2e-16  ***
entail_mean       2.55232   0.02152  118.592   < 2e-16  ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 1.992 on 54531 degrees of freedom
Multiple R-squared:  0.2112, Adjusted R-squared:  0.2111
F-statistic:  2433 on 6 and 54531 DF,  p-value: < 2.2e-16
```

Tab. 1. Linear model predicting rivalry from semantic distance, verb finiteness, argument opacity and mutual implications between paired implicatures.



Coefficients:					
(Intercept)	7.3522017	0.1532486	47.976	< 2,00E-16	***
w2vec_hsdrff_Sum	-0.0003296	0.0011026	-0.299	0.7650	
z_finitey	0.1455412	0.0161223	9.027	< 2,00E-16	***
z_args.opaquey	0.5009538	0.2156543	2.323	0.0202	*
z_args.opaqueobj	0.2547546	0.3372427	0.755	0.4500	
z_args.opaquesubj	0.0532390	0.0217931	2.443	0.0146	*
entail_mean	1.8818343	0.0217515	86.515	< 2,00E-16	***
lemmasact	-3.7824581	0.1512543	-25.007	< 2,00E-16	***
lemmasadjust	-2.7990281	0.1619012	-17.288	< 2,00E-16	***
lemmasadvance	-4.2765385	0.1515831	-28.213	< 2,00E-16	***
lemmasanswer	-4.2405515	0.1508514	-28.111	< 2,00E-16	***
lemmasapprove	-3.1989511	0.1621494	-19.728	< 2,00E-16	***
lemmasbid	-3.9934306	0.1548404	-25.791	< 2,00E-16	***
lemmascancel	-2.8219473	0.1621358	-17.405	< 2,00E-16	***
lemmasconceive	-2.2675897	0.1583548	-14.320	< 2,00E-16	***
lemmascultivate	-2.7869641	0.1816034	-15.346	< 2,00E-16	***
lemmascure	-3.8352304	0.1688616	-22.712	< 2,00E-16	***
lemmasdistinguish	-2.9855282	0.1580461	-18.890	< 2,00E-16	***
lemmasembrace	-3.3944366	0.1624320	-20.898	< 2,00E-16	***
lemmasexecute	-2.2898572	0.1686455	-13.578	< 2,00E-16	***
lemmashire	-3.4752011	0.2089821	-16.629	< 2,00E-16	***
lemmaslast	-1.2512805	0.2101987	-5.953	2.65e-09	***
lemmasmanage	-2.9204488	0.1531206	-19.073	< 2,00E-16	***
lemmasmurder	-3.5778433	0.2101611	-17.024	< 2,00E-16	***
lemmasneed	0.2515703	0.1692646	1.486	0.1372	
lemmaspack	-4.3029164	0.1501447	-28.658	< 2,00E-16	***
lemmasplan	-1.7058389	0.1817877	-9.384	< 2,00E-16	***
lemmaspoint	-3.2865632	0.1512721	-21.726	< 2,00E-16	***
lemmaspraise	-0.2847921	0.2091486	-1.362	0.1733	
lemmasprescribe	-0.2380621	0.2091980	-1.138	0.2551	
lemmassail	-1.8942963	0.1567161	-12.087	< 2,00E-16	***
lemmasseal	-3.8569221	0.1581722	-24.384	< 2,00E-16	***
lemmassee	-4.3824168	0.1498710	-29.241	< 2,00E-16	***
lemmastalk	-3.7339660	0.1502380	-24.854	< 2,00E-16	***
lemmasurge	-0.8541827	0.1623112	-5.263	1.43e-07	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					
Residual standard error: 1.809 on 54503 degrees of freedom					
Multiple R-squared: 0.3497,		Adjusted R-squared: 0.3493			

Tab. 2. Linear model from Tab. 1 enriched with lemmas as predictors

ACKNOWLEDGEMENTS

This work was supported by the Czech Science Foundation Grant No. GA 15-20031S and by the LINDAT/CLARIN project No. LM2015071 of the MEYS CR. All computation has been performed in R [16].

References

[1] Hanks, P. (2000). *Pattern Dictionary of English Verbs*. UK. Accessible at: `http://pdev.org.uk.`

[2] Cinková, S., Holub, M., Rambousek, A., and Smejkalová, L. (2012). A database of semantic clusters of verb usages. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, pages 3176–3183, European Language Resources Association, Istanbul, Turkey.

[3] Cinkova, S., Krejčová, E., Vernerová, A., and Baisa, V. (2016). Graded and Word-Sense-Disambiguation Decisions in Corpus Pattern Analysis: a Pilot Study. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 23–28, European Language Resources Association (ELRA), Paris, France.

[4] Baisa, V., Cinková, S., Krejčová, E., and Vernerová, A. (2016). VPS-GradeUp: Graded Decisions on Usage Patterns. In *LREC 2016 Proceedings*, Portorož, Slovenia.

[5] Erk, K., McCarthy, D., and Gaylord, N. (2009). Investigations on Word Senses and Word Usages. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 10–18, Association for Computational Linguistics, Suntec, Singapore.

[6] Dagan, I., Dolan, B., Magnini, B., and Roth, D. (2010). Recognizing textual entailment: Rational, evaluation and approaches – Erratum. *Natural Language Engineering*, 16(1):105.

[7] Cinková, S. and Vernerová, A. (to appear). Are Annotators'Word-Sense-Disambiguation Decisions Affected by Textual Entailment between Lexicon Glosses? *ITAT 2017*.

[8] Mikolov, T., tau Yih, W., and Zweig, G. (2013). Linguistic Regularities in Continuous Space Word Representations. In *Proceedings of NAACL-HLT*, pages 746–751, The Association for Computational Linguistics, Atlanta, Georgia.

[9] Selivanov, D. (2016). text2vec: Modern Text Mining Framework for R. Accessible at: `https://CRAN.R-project.org/package=text2vec.`

[10] Martínez Alonso, H., Sandford Pedersen, B., and Bel, N. (2013). Annotation of regular polysemy and underspecification. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 725–730, Association for Computational Linguistics, Sofia, Bulgaria.

[11] Pustejovsky, J., Rumshisky, A., Moszkowicz, J., and Batiukova, O. (2009). Glml: Annotating argument selection and coercion. In *IWSD-8: Eighth International Conference on Computational Semantics*, pages 169–180, Tilburg, Netherlands.

[12] Erk, K., Padó, S., and Padó, U. (2010). A Flexible, Corpus-Driven Model of Regular and Inverse Selectional Preferences. *Computational Linguistics*, 36(1):723–763. Accessible at: `https://doi.org/10.1162/coli_a_00017.`

[13] Harris, Z. S. (1970). Papers in structural and transformational linguistics. Reidel.

[14] Straka, M., Hajič, J., and Straková, J. (2016). UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. In Calzolari, N., Choukri, K., Declerck, T., Grobelnik, M., Maegaard, B., Mariani, J., et al., editors, *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, pages 4290–4297, European Language Resources Association, Paris, France.

[15] Zhihua, S. (2008). R Polydect. Accessible at: `https://github.com/cran/polydect/blob/master/R/HD.R`, retrieved 2017-01-22.

[16] R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Accessible at: `https://www.R-project.org/.`

GOLDEN RULE OF MORPHOLOGY AND VARIANTS OF WORDFORMS

JAROSLAVA HLAVÁČOVÁ
Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic

HLAVÁČOVÁ, Jaroslava: Golden Rule of Morphology and Variants of Wordforms. Journal of Linguistics, 2017, Vol. 68, No 2, pp. 136 – 144.

Abstract: In many languages, some words can be written in several ways. We call them variants. Values of all their morphological categories are identical, which leads to an identical morphological tag. Together with the identical lemma, we have two or more wordforms with the same morphological description. This ambiguity may cause problems in various NLP applications. There are two types of variants – those affecting the whole paradigm (global variants) and those affecting only wordforms sharing some combinations of morphological values (inflectional variants). In the paper, we propose means how to tag all wordforms, including their variants, unambiguously. We call this requirement "Golden rule of morphology". The paper deals mainly with Czech, but the ideas can be applied to other languages as well.

Keywords: morphology, global variants, inflectional variants, multiple lemma, Golden rule of morphology

1 TERMINOLOGY

As there are quite a lot of different approaches to some basic linguistic terms, let us at the beginning make clear about the terminology.

Wordform[1] is every string of letters that forms a normal word of a language. English examples: *get, gets, sisters, where*, Czech examples *dostal, dostaneš, sestrám, kam*.

Lemma is basic wordform. It is often used in dictionaries as a head word. Lemmas of examples from the preceding paragraph are: *get, get, sister, where*, the Czech ones: *dostat, dostat, sestra, kam*. From the lemma, individual wordforms can be created by means of inflection.

Paradigm is a set of wordforms that can be created by means of inflection from their basic wordform (lemma). There can be more than one (variant of a) lemma included in one paradigm (i.e. *color, colour* – see below).

Examples: wordforms belonging to the lemma *get*, namely *get, gets, got, gotten, getting* form one paradigm. Its representative is the lemma *get*. Another paradigm is the set of wordforms *color, colors, colour, colours*, with two lemma variants: *color* and *colour*. Czech example is presented in Table 1.

Variants are those wordforms that belong to the same paradigm and values of all their morphological categories are identical.

[1] It is often written as two words — word form. We have chosen this spelling as it avoids confusion with homographic reading when speaking about word forming.

DOI 10.1515/jazcas-2017-0024

Example: the pair of wordforms *got, gotten* are variants of past participle of the paradigm from the previous paragraph, represented by the lemma *get*.

Morphological category is a morphological property of words, for instance gender, tense, case.

Morphological value is a value of a morphological category. For instance, there are two values for the morphological category of number (singular, plural), seven values of the morphological category case in Czech.

Every wordform belongs to a paradigm that is represented by a lemma. We can also say that the wordform belongs to the lemma, or is derived from the lemma. Then, zero, one or more wordforms can be derived from a lemma, with a given set of morphological values.[2] In this paper, we will deal with the case of more than one wordforms for a given lemma and set of morphological values.

Maximal set of morphological values is the set that is sufficient for morphological description of a single wordform of a given lemma. What belongs to the maximal set of morphological values, usually depends on part-of-speech classificattion of the given lemma. For example number, gender, case, degree and status of negation are needed to describe a particular wordform of an adjective lemma in Czech.[3]

Morphological tag is a code – maximal set of the morphological values of a given wordform.[4]

2 GOLDEN RULE OF MORPHOLOGY

Given a lemma and a maximal set of morphological values, no more than one wordform should exist, belonging to that lemma and having those morphological values:

lemma + morphological tag → single wordform

This requirement is called "Golden rule of morphology" (see also [7], [8]) and is essential especially for generation of wordforms. If there were more than one wordform, a generator (automatic as well as human one) would not know, which variant to choose. Moreover, the variants can differ in a style or other non-morphological characteristics, and their replacement could be inappropriate in a given context.

Another reason for accepting the Golden rule is an unambiguous identification of wordforms in morphological (and other) dictionaries. Then, we can use the pair <lemma, morphological tag> as a unique identifier for each wordform.

[2] The alternative of no wordforms arises when the set of morphological values is not appropriate for the given lemma, for example no verb can be derived with a given value of the morphological category case.

[3] Example: maximal set of morphological categories for description of the wordform nehezkou (not pretty – instrumental), belonging to the lemma hezký (pretty), is number (sing), gender (fem), case (instr), degree (1) and negation (N). Avoiding any of them, more than one wordform would result – for instance without specifying the category of negation, there would be hezkou and nehezkou.

[4] There are several types of morphological tags, but their specific appearance is not important, if they contain all the morphological information needed for unique wordform description. That is the reason why we present no specific suggestions for tagging the new features.

2.1 Violation of the Golden Rule – Example

Let us have a look at the paradigm that is represented by lemma *okénko, 'small window'* (diminutive of *okno, 'window'*). This lemma has two more variants in Czech, namely *okýnko* and *vokýnko.* Every variant has 10 different wordforms, both standard and colloquial. In Table 1 we can see all of them. Each line of the table represents variants for the same combination of morphological values. Notice especially the last two lines of the table, which are doubled. There we have six different wordforms for one morphological tag. It means that one maximal set of morphological values (one morphological tag) describes six different wordforms.

Morphology (Case & Number)	Wordforms		
nom sg/acc sg	*okénko*	*okýnko*	*vokýnko*
gen sg/nom pl/acc pl	*okénka*	*okýnka*	*vokýnka*
dat sg/loc sg	*okénku*	*okýnku*	*vokýnku*
instr sg	*okénkem*	*okýnkem*	*vokýnkem*
gen pl	*okének*	*okýnek*	*vokýnek*
dat pl	*okénkům*	*okýnkům*	*vokýnkům*
loc pl	*okénkách* *okéncích*	*okýnkách* *okýncích*	*vokýnkách* *vokýncích*
instr pl	*okénky* *okénk<u>ama</u>*	*okýnky* *okýnk<u>ama</u>*	*vokýnky* *vokýnk<u>ama</u>*

Tab. 1. Paradigm *okénko, okýnko, vokýnko*

In the table, the first two columns under the title Wordforms include two standard variants, the third (greyish) column is colloquial. Moreover, all columns contain a wordform that is also colloquial, due to its colloquial ending, namely -*ama* (underlined in the Table 1). Thus, the lower rightmost wordform is colloquial twice – once due to its inclusion under a colloquial basic form, secondly due to its colloquial ending.

In our example, the Golden rule of morphology does not hold true. Even if we declared each of the three columns a separate paradigm, it would not hold true because of the two lower lines. Moreover, the three basic wordforms really are variants and they should belong to the same paradigm. We need other means how to distinguish all the variants and ensure the validity of the Golden rule of morphology.

3 TYPOLOGY OF VARIANTS

Following the observation from the example, we define two types of morphological variants – one affecting the whole paradigm and the second one affecting only a specific combination of morphological values. The former one is called **global**, the latter one **inflectional**.

Inflectional variants are those variants that relate only to some wordforms of a paradigm defined by a special combination of morphological values.

Global variants are those variants that relate to all wordforms of a paradigm, and always in the same way.

In accordance with the variant types we define two new morphological categories which describe them. Before we formally define their values, let us have a look at their nature.

3.1 Inflectional Variants

In Czech, the majority of inflectional variants differ in endings. There are patterns that include the inflectional variants for particular morphological values. In that sense we can say that they are mostly systematic. Examples of inflectional variants are in Table 2. The upper part contains systematic inflectional variants, in the rest there are more specific variants, that affect only those lemmas mentioned, or, as in the last line, a small set of similar words, in this case some verbs of movement.

Morphology	Czech variants	Czech lemma	English translation
loc pl	*hradu / hradě*	*hrad*	*(in the) castle*
loc pl	*lesu / lese*	*les*	*(in the) forest*
nom pl	*soudcové / soudci*	*soudce*	*(the) judges*
1st person pl	*mažeme / mažem*	*mazat*	*we spread*
1st person pl	*jdeme / deme / jdem / dem*	*jít*	*we are walking*
comparativ	*bílejší / bělejší*	*bílý*	*more white*
imperativ	*běž / poběž*	*běžet*	*run!*(imperativ)

Tab. 2. Examples of Czech inflectional variants

3.2 Global Variants

This type of variants is often not morphological, but orthographic. However, concerning automatic natural language processing, there is no difference between the two. The major point is that the pairs of variants have different spelling, for whatever reason. Thus, we do not distinguish between morphological and orthographic (or even other, e.g. etymological, stylistic) types.

Global variants can also be systematic, but the system always affects all wordforms belonging to a lemma. Examples of the systematic global variants are in Table 3.

Global variants always differ in one or more letters, no matter if at the beginning, in the middle, or at the end of the lemma. There can be more letter changes within a single word.

Description	Czech variants (lemmas)	English translation
protetic *v-*	*okno / vokno*	*window*
-ismus/-izmus	*feminismus / feminizmus*	*feminism*
generally *-s-/-z-*	*kurs / kurz*	*course*
-t-/-th-	*Atény / Athény*	*Athens*

Tab. 3. Examples of Czech global variants

We have already mentioned the possibility of not distinguishing the global variants. We can assign an individual lemma to both (all of) variants. Thus, we could have two lemmas, e.g. *okno, vokno* (*window*), with their own separate paradigms. However, there are words, especially foreign names with more spellings. For instance *Afghanistan* has 8 different spellings occurring in Czech texts[5]. It is reasonable to subsume them all under a single paradigm.

For linguists – corpus users – it is very convenient. If a corpus manager is designed and configured appropriately, they need not remember all the possible variants, but put only one of them into a query, and the resulting concordances will contain all of the possible variants. At the same time, it must be naturally possible to exclude some of them, if the user wants so, but this can be done by means of the query language, e.g. regular expressions.

The way how to join the variants together, while allowing their separate tagging to distinguish them, is described in the next section.

For computational linguists, it is also useful to have the variants labelled, because it is often necessary to automatically select one from more possibilities. If the variants were not described individually, there would be no clue for a selection of one of them. The tools even would not "know" that there are more possibilities.

4 HOW TO TAG VARIANTS

4.1 Present State

Examples from English

In English, there is almost no existence of inflectional variants. There are a few exceptions with two wordforms for past participle, for instance the verb *to get*, with two possible wordforms *got* and *gotten*, or past tense and past participle (*hanged* and *hung* for the lemma *to hang*).

There are quite a lot of global variants, especially due to wide area where English is spoken. Each region can have its specific variants. Well-known differences are between the British and American spellings. Probably the most common type of global variant for English is the type *ou-o,* as in pairs *colour/color*, *labour/labor*.

English tagsets, as far as we know, do not take care about variants. For instance, the both global variants *color* and *colour* mentioned in the previous paragraph, have the identical tag in the British National Corpus [2], namely NN1 for singular nouns and VVI for infinitive form of verbs. The same can be stated about inflectional variants for past participle *got* and *gotten* (both has morphological tag VVN), or *hung* and *hanged* (both VVN as past participles, or VVD as past tense).

Czech Case

In Czech, there are two main morphological tagsets (Prague tagset see [3], Brno tagset see [4]), both taking care about variants, but none of them being entirely satisfactory and consistent. The major point is that neither of them distinguishes between the two types, inflectional and global. They both use a single slot in the morphological tag for their description.

[5] Afghánistán, Afgánistán, Afganistán, Afghanistán, Afghanistan, Afganistan, Afghánistan, Afgánistan

As we have seen, there can be more variants for one combination of morphological values within a given paradigm, some of them differing inflectionally, some globally, and others in both ways. One category of variants is thus not enough. There have to be two of them.

Another problem is values of the two new morphological categories. Both present Czech morphological systems make distinction among styles of the variants. There are variants equipollent, archaic or bookish, colloquial, dialectical, to name just the most common ones. In other words, the values of the variants have an evaluative nature.

There is a number of studies about styles of variants for Czech. However, there is often little agreement, e.g. whether a variant is (still) colloquial or whether it has (already) penetrated into the standard vocabulary, or vice versa, whether a variant is (still) standard, or whether it is (already) archaic. It results in an inconsistent description. Thus, the list of variant values should be stated objectively and without evaluating ambitions.

4.2 Values of the Morphological Categories Describing Variants

As stated above, our new proposition is to avoid any evaluation. The values of both morphological categories, Global and Inflectional variant, should be strictly formal. The main reason for their introduction was only their distinction. Then we add them to the morphological tag in order to ensure the validity of the Golden rule of morphology.

The proposed values of the variant categories are based on differences between pairs of variants. Thus, we define the opposite values long and short (according to long and short vowels, e.g. *é/e*), hard and soft (according to hard and soft consonants, e.g. *s/š*), etc. Table 4 lists the most common types of global variants together with examples and values of the morphological category Global variant. Inflectional variants have similar set of values. Values of variants that are not common (see the example of *Afghanistan*) are tagged with numbers. If needed, the set of values might be enlarged.

5 LEMMATIZATION AND VARIANTS

We have solved the problem of variants, but created another one: Which of the possible global variants should be the representative of the whole paradigm? In other words: What is lemma of a paradigm with global variants? Which properties are essential for its selection?

We present several requirements that seem naturally and reasonably at first glance. It should be neutral, it must not be archaic nor colloquial. The problem is that the styles are not (and cannot) be defined exactly, they are changing and there is no agreement, as we have already mentioned. Moreover, there are often two, or even more equipollent variants. There are also variants that do not have a neutral counterpart. Ultimately, it was the reason why we do not use these concepts for their tagging.

It should be the most frequent (most common). According to our linguistic intuition, the basic variant should be that one, which is more common, but this characteristic is also very difficult to detect. Of course, we can use the frequency, or better said, the average reduced frequency (see [5]) calculated from a referential corpus (Czech National Corpus for Czech — see [1], for instance), but it often happens, that the corpus does not contain some, or even none of the variants, or the difference between their (average reduced) frequencies is negligible. There is no entirely representative corpus in which we could trust with respect to frequency or commonness. We could find more requirements for such a representative, but none of them is entirely neutral. There is another solution – multiple lemma.

Type	Example	Values of the morphological category Global variant
o-vo	*okno --- vokno*	0 --- v
ý-ej	*mýdlo --- mejdlo*	0 --- j
z-s	*klauzule --- klausule*	z --- s
t-th	*tema --- thema*	0 --- h
é-í	*kolébka --- kolíbka*	e --- i
é-ý	*okénko --- okýnko*	e --- y
á-e	*originální --- originelní*	a --- e
á-a	*Abrahám --- Abraham*	long --- short
é-e	*acetylén --- acetylen*	
ó-o	*salón --- salon*	
ý-y	*apetýt --- apetyt*	
í-i	*alexandrín --- alexandrin*	
ů-u	*přezůvky --- přezuvky*	
ú-u	*Plútarchos --- Plutarchos*	
s-š	*student --- študent*	hard --- soft
t-ť	*vlaštovka --- vlašťovka*	
n-ň	*šnůra --- šňůra*	
d-ď	*dolík --- ďolík*	
e-ě	*Bardejov --- Bardějov*	
z-ž	*zbrzďování --- zbržďování*	

Tab. 4. List of most common types of global variants

5.1 Multiple Lemma

Every paradigm can have not only one representative, but as many as there are global variants of its lemma. In other words, the lemma of a paradigm with global variants is a set of lemmas (see also [6], [7]). Then, if a corpus user asks for a lemma in his/her query, he/she needs not to care about a "basic" global variant, but can use whichever lemma, that come under the desired paradigm. They even need not to know all possible lemmas that could belong under the paradigm.

Multiple lemma		*okénko*	*okýnko*	*vokýnko*
Code of Global variant		e0	y0	yv
Morphology		Wordforms		
Case & Number	Inflectional Variant			
nom sg/acc sg		*okénko*	*okýnko*	*vokýnko*
gen sg/nom pl/acc pl		*okénka*	*okýnka*	*vokýnka*
dat sg/loc sg		*okénku*	*okýnku*	*vokýnku*
instr sg		*okénkem*	*okýnkem*	*vokýnkem*
gen pl		*okének*	*okýnek*	*vokýnek*
dat pl		*okénkům*	*okýnkům*	*vokýnkům*
loc pl	a	*okénkách*	*okýnkách*	*vokýnkách*
loc pl	i	*okéncích*	*okýncích*	*vokýncích*
instr pl	y	*okénky*	*okýnky*	*vokýnky*
instr pl	m	*okénkama*	*okýnkama*	*vokýnkama*

Tab. 5. The example with the multiple lemma {*okénko, okýnko, vokýnko*}. Every wordform has distinguished Global variant (columns) and Inflectional one, where necessary (4 bottom lines). The global variant "ev" (*vokénko*) is not included, though theoretically possible.

If a single global variant is desired, it has to be selected from the set by adding another condition to the query. There are two possibilities:

– to specify spelling of the lemma or wordform, or
– to specify the type of the global variant.

Software tools used for searching the corpus (corpus managers) are able to deal with multiple values of attributes. Let us present our new suggestions on the example that we have used as the introduction into the problem of variants. The lemma representing the whole paradigm presented in Table 1 is the set {*okénko*, *okýnko, vokýnko*}. If we wanted to search for all occurrences of this multiple lemma in corpus, we need not to write our query using regular expression like [lemma="v?ok[éý]nko"][6].

We can just state any of the three lemmas and will get what we wanted. The Table 5 presents the example from the Table 1 with all the wordforms distinguished by means of global and inflectional variants. We do not intentionally specify the shape of the morphological tags, because there are more ways how to code the information about the variants. There are several types of morphological tags, even for Czech, and each system can subsume the new morphological categories differently. One of possible suggestions can be found in [7].

6 FINAL REMARKS

We have proposed how to deal with variants of wordforms and lemmas. The main reason, why to distinguish them, is the effort to support the Golden rule of morphology, which ensures an unambiguous description of each wordform of a language. Without proper tagging variants it could not hold, which would cause problems in various fields of natural language processing – generating text, machine translation, indexing wordforms, to name just a few.

[6] Theoretically, this regular expression search also for possibly non-existing, absurd lemma *vokénko*.

There are two types of variants, inflective and global. They should be treated as two different morphological categories, as they may be combined in many ways. The existing systems do not distinguish between them, which causes a violation of the Golden rule of morphology.

The existence of global variants leads to a multiple lemma – set of all global lemma variants. This concept is more general and objective than choosing one representative from the set of variant lemmas, as there is no entirely objective criterion for that.

However, there is no general way how to deal with the variants, each application has to choose its own way. There can be an application where the preference is given to the variant with the substring *-t-* over the variant with the (oldish one) *-th-* for modern translations. With a strict description of all wordforms, especially their variants, such a preference is easy to implement.

Although we decided not to tag the variants with any semantic or stylistic labels, such as emotional, colloquial, archaic etc., it might be useful to do so. The main reason why we do not include any type of evaluation into the morphological description is that there is no exact rule how to define individual values of such labels. Even experts are not able to agree on objective criteria. Moreover, the meaning of those labels changes in time. However, if it was needed, the formal non-evaluative tagging of variants enable to make decisions tailored to various special and detailed requirements. Without a strict unambiguous description of each wordform, there would be not possible to make extensions mentioned above.

ACKNOWLEDGEMENTS

The work on this paper was supported by the grant number 16-18177S of the Grant Agency of the Czech republic (GAČR) and *LM2015044* of the Ministry of Education, Youth and Sports.

References

[1] Czech National Corpus: Accessible at: `http://ucnk.ff.cuni.cz/`.

[2] British National Corpus. Accessible at: `http://www.natcorp.ox.ac.uk/`.

[3] Hajič, J. (2004). *Disambiguation of Rich Inflection*. (Computational Morphology of Czech). Karolinum, Praha.

[4] Brno morphological analyzer ajka. Accessible at: `http://nlp.fi.muni.cz/projekty/ajka/index.htm`.

[5] Savický, P. and Hlaváčová, J. (2002). Measures of Word Commonness. *Journal of Quantitative Linguistics*, 9(3):215–231.

[6] Hlaváčová, J. (2011). Problém variantních tvarů slov při automatickém zpracování jazyka. In *Information Technologies – Applications and Theory*, pages 75–78, Univerzita Pavla Jozefa Šafárika v Košiciach, Slovakia.

[7] Hlaváčová, J. (2009). *Formalizace systému české morfologie s ohledem na automatické zpracování českých textů*. Ph.D. thesis, FF UK.

[8] Hlaváčová, J. (2008). Pravopisné varianty a morfologická anotace korpusů. In *Grammar & Corpora / Gramatika a korpus 2007*, pages 161–168, Academia, Praha, Czech Republic.

MORPHOLOGICAL DISAMBIGUATION OF MULTIWORD EXPRESSIONS AND ITS IMPACT ON THE DISAMBIGUATION OF THEIR ENVIRONMENT IN A SENTENCE[1]

MILENA HNÁTKOVÁ – VLADIMÍR PETKEVIČ
Faculty of Arts, Charles University, Prague, Czech Republic

HNÁTKOVÁ, Milena – PETKEVIČ, Vladimír: Morphological Disambiguation of Multiword Expressions and Its Impact on the Disambiguation of Their Environment in a Sentence. Journal of Linguistics, 2017, Vol. 68, No 2, pp. 145 – 155.

Abstract: This study concerns the impact of the collocation/phraseme disambiguation component within the complex system of the rule-based morphological disambiguation of Czech. This system constitutes one of the two main disambiguation subsystems that are responsible for the morphological disambiguation of the corpora of synchronic Czech within the Czech National Corpus project. We will show that although the part of texts constituted by collocations/phrasemes (generally multiword expressions – MWEs) is relatively small and consequently the errorfree morphological disambiguation of MWEs covers only a small portion of textual material, such perfectly disambiguated fragments in sentences help to improve the disambiguation of the rest, non-MWE part of sentences.

Keywords: multiword expressions, lexical database, morphological analysis, morphological ambiguity, morphological disambiguation, process of disambiguation, Czech National Corpus

1 INTRODUCTION

The series of corpora of synchronic Czech within the Czech National Corpus, viz. SYN2005, SYN2010, SYN2013PUB, SYN2015, versions of SYN,[2] are morphologically disambiguated by a complex process in which two main components cooperate: the rule-based disambiguation system called *LanGr* ([5], [6], [3], [4], [7], [8], [9]) and the stochastic tagger called *Featurama* (`https://sourceforge.net/projects/featurama/`). This hybrid disambiguation system is activated immediately after morphological analysis: individual morphological homographs are subject to the disambiguation of

(i) lemmas, and
(ii) morphological tags, including part-of-speech tagging.

2 THE DISAMBIGUATION PROCESS

The first disambiguation component, the LanGr system, consists of ca. 2 600 handcrafted linguistic rules that are

[1] This paper was financially supported by the grant Between Lexicon and Grammar (Mezi slovníkem a gramatikou – Grant Agency of the Czech Republic, reg. no. 16-07473S).

[2] `http://korpus.cz`

DOI 10.1515/jazcas-2017-0025

(a) developed on the basis of linguistic introspection and checked on corpus data, and also
(b) non-automatically inferred from corpus data.

Linguistic rules are written in a special programming language and their performance consists in the context-based gradual deletion of incorrect lemmas and tags assigned to individual tokens. First, the *LanGr* system processes the output of morphological analysis which assigns every token all of its tags and lemmas; the recall of morphological analysis is currently 99.25%. As the morphological analyzer assigns all tokens all of its lemmas and tags regardless of the context, the tokens are assigned the highest amount of incorrect tags, i.e. the precision is lowest possible on disambiguation input. The disambiguation consists in keeping the best possible recall (close to 100%) and in gradually increasing precision by removing lemmas and tags that are incorrect in the given context.

Disambiguation rules are contained in two main groups:
a) safe rules organized in two subgroups: Safe0 containing entirely safe rules and Safe1 containing slightly less safe rules
b) heuristic rules (Heu).

An input sentence is gradually more and more disambiguated by the rules' application until – ideally – a full disambiguation is achieved, i.e. each token is assigned the only correct lemma and tag. If the rule-based tagger is unable to entirely delete all inappropriate tags and lemmas in the input sentence, the remaining incorrect ones are removed by the second disambiguation component: stochastic tagger Featurama.

The process of the rule-based morphological disambiguation also involves the collocational module *Phras* ([1], [2]), identifying and properly disambiguating multiword expressions (MWEs). Thus, the following modules take part in the disambiguation process:
(i) LanGr tagger based on manually written rules;
(ii) Phras module using a lexical database of maximally disambiguated MWEs;
(iii) parameterizable stochastic tagger, currently Featurama.

The cooperation of the modules consists in the following sequence of operations applied to a sentence:

1733853790 **1st step**: The output of morphological analysis is processed by entirely safe rules (Safe0 group). The rules gradually disambiguate the sentence, i.e. the number of incorrect tags decreases. The process continues till there is nothing to disambiguate, i.e. till the rules in recurrent cycles exhaust their disambiguation capacity.

1733853790 **2nd step**: *Phras* module is invoked: it identifies MWEs in the sentence and performs disambiguation of their components as much as possible.

1733853790 **3rd step**: The set of safe rules Safe0 is reapplied. After these rules finish their job, i.e. they are not able to disambiguate any more, the 1733853790 4th step follows.

1733853790 **4th step**: Three sets of rules, i.e. Safe0, Safe1 and the set of heuristic rules Heu, are applied in cycles to disambiguate the sentence till they cannot disambiguate any more;

1733853790 **5th step**: The remaining incorrect tags intact up to now by the *LanGr* system are removed by the stochastic tagger Featurama and a postprocessing phase (see below).

Table 1 presents a quantitative contribution (in %) of each subsystem within the entire morphological analysis and disambiguation system, where the subsystems are as follows:

Morph – morphological analysis

Safe0 – safe rules: Safe0 (cf. 1st and 1733853790 3rd step above)

Phras – phraseme module Phras processing MWEs (cf. 1733853790 2nd step above)

SSH – the sets of rules Safe0, Safe1 and Heu applied together (cf. 1733853790 4th step above)

Tagger – stochastic tagger Featurama (cf. 1733853790 5th step above)

Post – postprocessing phase (verbal aspect added; possible reinterpretation of controversial part-of-speech annotation, e.g. adverb/particle; finalization of named entities processing...).

after	% of all tokens / incrementally	% of all words / incrementally
Morph	22.07	25.88
Safe0	31.46 / 53,53	36.88 / 62.76
Phras	**0.69 / 54.22**	**0.80 / 63.56**
Safe0	**2.17 / 56.39**	**2.54 / 66.10**
SSH	6.55 / 62.94	7.68 / 73.78
Tagger	22.13 / 85.07	25.94 / 99.72
Post	0.23 / 85.30	0.27 / 99.99

Tab. 1. Contribution of individual subsystems to the entire disambiguation of the texts contained in the sample Newton corpus of journalistic texts (the size in tokens including punctuation marks: 1 735 482 098; words: 1 480 369 445). The contribution is measured merely by the number of achieved unambiguous tags assigned to words after each phase of processing, the quality of disambiguation (recall and precision in the strict sense of the word) is not accounted for here.

In the middle column, the ratio of fully disambiguated tags of words in % after each phase of processing is presented with respect to all tokens (= all corpus positions including punctuation). In the right column, the ratio with respect to word forms only (i.e. without punctuation) is shown. Thus, morphological analysis identifies 22.07% of all tokens and 25.88% of all words as morphologically unambiguous word forms. The Safe0 rule group is able to disambiguate further 31.46% words that were ambiguous after morphological analysis etc. till all words (= 85.30% of all tokens) are unambiguously disambiguated (the rest of the tokens, i.e. 14.70%, is constituted by punctuation tokens). The figures in the right column have the same meaning as in the middle column but they are counted with respect to words only.

Table 2 shows the average number of tags assigned to tokens (words + punctuation marks) after each stage of processing. The figures in the second column mean that the average number of tags assigned to tokens by morphological analysis is almost 11; if punctuation is not taken into account the average number is 12.28, and if only ambiguous words are considered, the average number is more than 16. Table 2 demonstrates the paramount importance of the Safe0 set of rules that is able to decrease the average number of tags assigned by morphological analysis to tokens, words and ambiguous words to 2.81, 3.08 and 5.80, respectively.

after	All tokens counted	only words counted	only ambiguous words counted
Morph	10.62	12.28	16.21
Safe0	2.81	3.08	5.80
Phras	**2.76**	**3.03**	**5.75**
Safe0	**2.54**	**2.77**	**5.40**
SSH	1.92	2.04	4.17

Tab. 2. Average number of tags per word form achieved in the same annotated Newton corpus

Table 1 and Table 2 present, in fact, the measure of precision in a very coarse way since only the ratio of deleted tags in % is shown without taking into account whether only incorrect tags were deleted.

In Table 3 we present the recall after each processing step.

after	recall
Morph	99.25%
Safe0	99.09%
Phras	**99.07%**
SSH	98.82%

Tab. 3. The recall of (i) the morphological analyzer (Morph), (ii) the safe rules (Safe0), (iii) the MWE module (Phras), (iv) Safe0+Safe1+Heu(ristic) rules (SSH)

We see that the recall decreases very slightly: Safe0 rules make only 0.16% errors, the error rate of the MWE module is only 0.02%. The entire rule-base disambiguation system decreases recall after morphological analysis by only 0.43% (99.25 – 98.82).

The accuracy (recall + precision) of the entire disambiguation system, i.e. including the Featurama tagger and the postprocessing phase Post, is ca. 95.1%.

It is to be noted that the disambiguation system described above is not used in syntactic parsing. Stochastic parsers applied to Czech reduce morphological ambiguity to a large extent but the recall they achieve in morphological disambiguation proper is always lower (ca. 93%) than the recall of the system just depicted.

3 THE PHRAS MODULE

Now we will focus on the Phras module disambiguating MWEs in more detail. In Table 1 we see that it contributes to the overall disambiguation success rate only marginally (0.69%). However, if Phras is applied, i.e. if a sentence contains a MWE that is contained in the MWE lexical database exploited by Phras, it paves the way for the Safe0 rules that are able to remove further 2.17% tags thus allowing for further disambiguation. The average number of tags assigned to tokens, words and ambiguous words decreases by 0.05% (cf. Table 2) after the Phras module is invoked.

The performance of the Phras module will be demonstrated on examples (taken from sentences contained primarily in the SYN2015 corpus) showing how Phras
(i) disambiguates MWEs themselves (par. 3.1),
(ii) contributes to the disambiguation of the environment of MWEs in a sentence (par. 3.2).

3.1 Disambiguation of MWEs

Phras exploits the lexical database of fully or partially disambiguated MWEs. The fixed part of these expressions is fully disambiguated, the variable inflectional part is disambiguated only partially, but as much as possible. We will present two motivating examples demonstrating part-of-speech and case disambiguation.

Example 1
In the MWE
(1) *brány pekla*
gates$_{\text{Noun-Npl.Fem/Apl.Fem/Vpl.Fem}}$ of_hell$_{\text{Noun.Gsg.Neut}}$
there is a word form *brány* 'gates' 1733853794 that is, morphologically, part-of-speech ambiguous – it is:
(i) genitive singular (Gsg), or nominative/accusative/vocative plural (Npl/Apl/Vpl)[3] of the feminine noun *brána* 'gate'
(ii) passive participle in feminine plural / masculine inanimate plural of the verb *brát* 'take'.
In (1), the form *brány* is, however, a part-of-speech unambiguous feminine noun in plural and three cases: Npl/Apl/Vpl since the entirely unambiguous morphological interpretation depends on a textual context.
The other word in (1), *pekla*, is also part-of-speech ambiguous since it is:
(i) Gsg/Npl/Apl/Vpl of the neuter noun *peklo* 'hell'
(ii) past participle in feminine singular / neuter plural of the verb *péci* 'bake'.
In (1), the word form *pekla* is unambiguous: Gsg of the neuter noun *peklo.*

Example 2
In the MWE lexical database entry for the MWE,
(2) *ekonomický růst*
economic$_{\text{Adj-Nsg.MascInan/Asg.MascInan}}$ growth$_{\text{Noun-Nsg.MascInan/Asg.MascInan}}$
'economic growth'
the form *ekonomický* 'economic' is a part-of-speech unambiguous adjective in Nsg/Asg masculine inanimate; the part-of-speech ambiguous form *růst* 'growth / to grow' is disambiguated as a masculine inanimate noun in Nsg/Asg ('growth'), rather than the infinitive of the verb ('to grow'). Moreover, this database entry contains information that both forms agree in number, gender and case.
The Phras module is also very helpful in disambiguating proverbs and other sentential idioms as is shown in the following example.

[3] The remaining cases in the declension system of Czech are: dative (D), locative (L) and instrumental (I).

Example 3

In the process of morphological disambiguation, the proverb:

(3) *Komu není rady, tomu není pomoci*.

To_whom is_not advice$_{\mathrm{Gsg.Fem}}$, to_that$_{\mathrm{DsgMasc}}$ is_not help$_{\mathrm{Gsg.Fem}}$.

'There are none so deaf as those who will not hear'

contained in the MWE lexical database is first processed by the Safe0 rules. They cannot cope with two nouns in the genitive of negation (constructions with the genitive of negation are rare in modern Czech, being associated only with a limited set of nouns), namely *rady*$_{\mathrm{Gsg.Fem}}$ 'advice' and *pomoci*$_{\mathrm{Gsg.Fem}}$ 'help', because the word *rady* and *pomoci* can also be a form of the masculine animate noun *rada* 'counsellor' and the infinitival form of the verb *pomoci* 'to help', respectively. Moreover, the form *tomu* 'to_that' is not only dative singular (Dsg) masculine form of the pronoun *ten* 'that'*,* but also Dsg neuter form of the pronoun *to* 'it'. The collocational module resolves all these ambiguities and entirely disambiguates the proverb.

3.2 Disambiguation of MWEs' Context

On several examples, we will show how disambiguation of MWEs performed by the Phras module can improve disambiguation of their sentential context. These randomly chosen examples are to elucidate the main objectives of the disambiguation of MWEs and of their content: part-of-speech disambiguation, primarily deciding between nouns, verbs and adjectives, and case disambiguation (concerning nouns, adjectives, pronouns and numerals), which is the most difficult subtask of the whole disambiguation process.

Example 4

In sentence:

(4) *Asadův režim* ***nenese odpovědnost*** *za použití zbraní hromadného ničení.*

Asad régime$_{\mathrm{Noun.Nsg.MascInan}}$ **not_bears responsibility**$_{\mathrm{Noun.Asg.Fem}}$ for exploitation of_ weapons of_mass destruction.

'Asad régime **does not bear responsibility** for the exploitation of the weapons of mass destruction.'

Phras identifies the pair *nenese odpovědnost* ('not_bears responsibility') of the MWE ***nést odpovědnost*** 'bear responsibility' and disambiguates its components. The unambiguous present 3rd person singular negative form *nenese* of the transitive verb *nést* 'bear' poses no disambiguation problem, but the feminine noun *odpovědnost* 'responsibility' can morphologically be Nsg/Asg. The masculine inanimate noun *režim* 'régime' is case ambiguous in the same way. As the disambiguated MWE is contained in the MWE lexical database, Phras unequivocally disambiguates *odpovědnost* in (4) as Asg. The general rules cannot solve, on the basis of sole syntax, the classical disambiguation problem in Czech consisting in the disambiguation of the pattern:

Noun1$_{\mathrm{Nom/Acc}}$ Verb$_{\mathrm{Trans.Pres.3rd.Sg}}$ Noun2$_{\mathrm{Nom/Acc}}$

where either Noun1 and Noun2 is in the nominative and accusative case, respectively, or vice versa.

As Phras disambiguates *odpovědnost* as Asg, it fundamentally helps to disambiguate the sentence: as *odpovědnost* is in Asg, the noun *režim* 'régime' cannot be in non-prepositional Asg (the valency of the verb *nést* does not admit two accusative objects and, moreover, the noun *odpovědnost* cannot head an accusative nominal phrase having the syntactic function of adverbial) and that is why it is in Nsg. After such a correct disambiguation, it is then, e.g., no problem for a parser of Czech to assign proper syntactic functions to the nominal phrase *Asadův režim* 'Asad régime' (= subject) and to the nominal phrase *odpovědnost* 'responsibility' (= object). Thus the rest of the sentence is also influenced: there are no non-prepositional nouns as objects in accusative[4] in the sentence. In particular, the word *ničení* 'destruction' cannot be in non-prepositional accusative. The importance of the disambiguation of the MWE *nést odpovědnost* is thus clearly demonstrated. There are many such support verb (verbo-nominal) constructions in Czech as *nést odpovědnost* and the more such constructions are contained in the MWE lexical database, the better ad more accurate Phras is[5]. The disambiguation of such support verb constructions is of paramount importance especially in cases where some of the collocation components are not only case ambiguous (as in *odpovědnost*) but even part-of-speech ambiguous: e.g. the MWE *nabýt dojmu*$_{\text{Noun.Gsg.MascInan}}$ 'get an impression' contains the form *dojmu* that is morphologically Gsg/Dsg/Lsg of the masculine inanimate noun *dojem* 'impression', or 1st person singular present tense of the verb *dojmout* 'impress'; the verbo-nominal construction *má*$_{\text{Trans.Pres.3rd.Sg}}$ *štěstí* (lit. 'has happiness' '(s)he is lucky') contains the part-of-speech ambiguous word *má* (morphologically either 3$^{\text{rd}}$ person singular present tense of the verb *mít* 'have', or Nsg.Fem/Npl.Neut/Apl.Neut/Vpl. Neut of the possessive pronoun *můj* 'my'; the verbo-nominal construction *svalit vinu*$_{\text{Noun.Asg.Fem}}$ 'throw the blame (for something on someone)' contains the part-of-speech ambiguous word *vinu* (morphologically, the form *vinu* is either Asg of the feminine noun *vina* 'guilt', or 1733853796 1$^{\text{st}}$ person singular present tense of the reflexive verb *vinout se*, 'to wind') etc. If such words were erroneously part-of-speech disambiguated, undoubtedly the disambiguation of other words in sentences containing such MWEs would be badly affected.

Example 5

In sentence:
(5) *Manželé přijedou na plzeňské* ***hlavní nádraží*** *parním vlakem.*
Married_couple will_arrive to$_{\text{Prep.Acc}}$ Pilsen$_{\text{Adj.Asg.Neut}}$ main$_{\text{Adj.Asg.Neut}}$ railway_station$_{\text{Noun.Asg.Neut}}$ by steam engine.
'The married couple will arrive at Pilsen main railway station by steam engine.'
there is a frequent collocation *hlavní nádraží* 'main railway station', where the word *hlavní* 'main' is an adjective agreeing with the noun *nádraží* 'railway station' in number (singular/plural), gender (neuter) and case (nominative/accusative/vocative). However, it can also be a form of the feminine noun *hlaveň* 'barrel' in Isg/Gpl. If the

[4] Generally, nominal phrases can also head adverbials of time (duration) or regard in accusative as their attributes, but the set of head nouns governing such adverbials is limited.

[5] Some examples of support verb constructions: *vynést rozsudek* 'pronounce judgement', *upírat zrak* 'fix one's eyes on someone', *nabýt dojmu* 'get an impression', *mít štěstí* 'be lucky', *mít smysl* 'have sense', *mít pocit* 'have a feeling', *mít právo* 'be entitled to', *mít naději* 'have a hope', *dávat přednost* 'have preference (for something)', *svalit vinu* 'throw the blame (for something) on someone', *učinit rozhodnutí* 'make a decision'...)

disambiguation system chose this nominal interpretation of the word *hlavní* rather than the adjectival one, the disambiguation of the context would be wrong: the prepositional phrase (PP) ***na plzeňské hlavní nádraží*** (lit. 'to Pilsen main railway station') would be incorrectly split into three parts:

(i) PP *na*$_{\text{Prep-Acc/Loc}}$ *plzeňské*$_{\text{Adj}}$ 'to Pilsen'[6]

(ii) feminine noun *hlavní* 'barrel' that is morphologically in Isg/Gpl and does not agree with the PP *na plzeňské* in case since the preposition *na* generally requires accusative/locative and the adjective *plzeňské* morphologically is, i.a., Gsg.Fem/Dsg.Fem/Lsg.Fem/Npl.Fem/Apl.Fem/Vpl.Fem/Nsg.Neut/Asg.Neut/Vsg.Neut

(iii) neuter noun *nádraží* 'railway station' that is morphologically Nsg/Gsg/Dsg/Asg/Vsg/Lsg/Npl/Gpl/Apl/Vpl.

Thus, the rule-based system would not know how to disambiguate the ambiguous PP *na plzeňské.* Moreover, the case and number of the neuter noun *nádraží* could hardly be identified. If, on the contrary, Phras disambiguates *hlavní nádraží* as a collocation, i.e. *hlavní* as an adjective 'main' coforming a nominal phrase with the noun *nádraží* (the adjective *hlavní* agrees with the noun *nádraží* in number, gender and case), the subsequently applied rules can assume that the sequence ***na plzeňské hlavní nádraží*** complies with the PP pattern:

Prep$_{\text{Acc}}$ Adj$_{\text{Asg.Neut}}$ Adj$_{\text{Asg.Neut}}$ Noun$_{\text{Asg.Neut}}$

This means that the rules can recognize the sequence as one PP in Asg.[7]

Example 6

In sentence:
(6) *Prioritou je* ***zajištění odbytu***.
Priority is$_{\text{Verb.Pres.3rd.Sg}}$ securing$_{\text{Noun.Nsg.Neut}}$ of_sales$_{\text{Noun.Gsg.MascInan}}$.
'The priority is the securing of sales.'
there are three part-of-speech ambiguous words:

(a) *je* is:
  (i) 3rd person singular present tense form of the verb *být* 'be'
  (ii) Asg of the 3rd person neuter personal pronoun *ono* 'it'
  (iii) Apl of the 3rd personal pronoun (all genders) *oni/ony/ona* 'they';
(b) *zajištění* is:
  (i) Npl.MascAnim/Vpl.MascAnim form of the adjective *zajištěný* 'secured'
  (ii) Nsg/Gsg/Dsg/Asg/Vsg/Lsg… of the deverbal neuter noun *zajištění* 'securing';
(c) *odbytu* is:
  (i) Gsg/Dsg/Lsg of the masculine inanimate noun *odbyt* 'sales'
  (ii) passive participle of the transitive verb *odbýt* 'do sloppily' in Asg.Fem.

General disambiguation rules can hardly correctly disambiguate the nominal phrase (NP) *zajištění odbytu* 'securing of sales' as well as its immediate context: the

[6] For simplicity reasons, we omit the interpretation of the word form *plzeňské* as a (deadjectival) noun.

[7] They can, moreover, disambiguate *plzeňské* as an adjective rather than as a deadjectival noun (univerbization: *plzeňské pivo* 'Pilsner beer' → *plzeňské* 'Pilsner').

word *je*. The Phras module identifies the pair *zajištění odbytu* as a collocation where *zajištění* is disambiguated as a noun rather than as an adjective, and *odbytu* as a noun in Gsg rather than as a verbal passive form. For the subsequent rules it will then be much easier to identify the entire NP *zajištění odbytu* 'securing of sales' as an NP where *zajištění* is in Nsg, and also *je* as a verbal predicate in singular rather than as a personal pronoun.

Example 7

The sentence:
(7) *Nasypala dovnitř* ***prací prášek***.
She poured inside$_{\mathrm{Adv}}$ washing$_{\mathrm{Adj.Asg.MascInan}}$ powder$_{\mathrm{Noun.Asg.MascInan}}$.
'She poured inside the washing powder.'
is difficult to disambiguate since there are two part-of-speech ambiguous forms (*dovnitř* 'inside / into', *prací* 'washing / of-works'), and especially in such structures the processing of collocations can be very helpful. The word *dovnitř* is either (i) a preposition ('into') requiring genitive, or (ii) an adverb ('inside'); the word *prací* is either (i) Isg/Gpl form of the feminine noun *práce* 'work', or (ii) a very number-case ambiguous form of the soft adjective (the root ending in *-í*) *prací* 'washing'. The rules could incorrectly identify the form *prací* as Noun.Gpl.Fem 'work', and the word form *dovnitř* as a preposition taking genitive and thus the pair *dovnitř prací* could be identified as a genitive prepositional phrase with the meaning 'into the works'. If, on the contrary, Phras correctly identifies the pair *prací prášek* as a noun phrase ('washing powder') contained in the lexical database where *prací* is an adjective agreeing with the noun *prášek* in number (= singular), gender (= masculine inanimate) and case (= nominative/accusative), the subsequent rules will exclude *dovnitř* as a preposition taking genitive, thus interpreting *dovnitř* only as an adverb.

Example 8

The word *místo* is ambiguous between a noun 'place' and a preposition 'instead of'. A correct disambiguation of this very frequent word is crucial for the errorfree disambiguation of clauses where the word *místo* appears. The main disambiguation problem consists in that the noun *místo* often collocates with a NP in the genitive case and the preposition *místo* takes genitive, too. If typical collocations with the noun *místo* are contained in the MWE lexical database exploited by the Phras module, the disambiguation of sentences containing such collocations is much better. For instance, in sentence
(8) *Policie obrátila* ***místo činu*** *vzhůru nohama*.
Police$_{\mathrm{Noun.Nsg.Fem}}$ reversed place$_{\mathrm{Noun.Asg.Neut}}$ of_crime$_{\mathrm{Noun.Gsg.MascInan}}$ upwards with legs.
'The police put the scene of crime out of joint.'
the Phras module identifies *místo* as a noun rather than as a preposition since it uses a partially disambiguated collocation *místo činu* (lit. 'place of crime', 'scene of crime') contained in the lexical database: *místo*$_{\mathrm{Noun\text{-}Nsg.Neut/Asg.Neut/Vsg.Neut}}$ *činu*$_{\mathrm{Noun.Gsg.MascInan}}$, where the components have disambiguated morphological properties as indicated. In sentence (8), the noun *místo* is unambiguously in Asg since it does not agree with the feminine singular predicate *obrátila* 'reversed' in gender and therefore

it cannot be the subject in the nominative case. As *místo* is in accusative, *policie* ‘police’ cannot be in accusative (the verb *obrátila* cannot take two objects in accusative), it can only be in Nsg (correct), or Gsg/Vsg (incorrect), or Npl/Apl/Vpl (incorrect). If *místo* were erroneously identified as a preposition, the accusative reading of the form *policie* could not be syntactically excluded.

Most frequent right nominal collocations with the noun *místo* are as follows: *místo určení*[8] *‘*destination’, *místo činu ‘*scene of crime’, *místo nehody* ‘accident site’, *místo konání ‘*venue’, místo *narození* ‘place of birth’, *místo nálezu* ‘place of finding’, *místo spolujezdce* ‘passenger seat’, *místo odpočinku* ‘resting place’. Such collocations contained in the lexical database and exploited by the Phras module often help to disambiguate the context of these collocations.

Example 9

In sentence:
(9) *Řeč je o dobrovolných* ***dárcích krve***.
Talk is about voluntary$_{\text{Adj.Lpl.MascAnim}}$ donors$_{\text{Noun.Lpl.MascAnim}}$ of blood$_{\text{Noun.Gsg.Fem.}}$
‘Voluntary blood donors are being talked about.’
the Phras module identifies *dárcích* as Lpl of the masculine animate noun *dárce* ‘donor’ since it is a component of the MWE *dárce*$_{\text{Noun.MascAnim}}$ *krve*$_{\text{Noun.Gsg.Fem}}$ ‘blood donor’. In the nominal Lpl phrase *o dobrovolných dárcích krve* ‘about voluntary donors of blood’ the adjective *dobrovolných* ‘voluntary’ is also in Lpl.MascAnim (due to agreement with *dárcích* in number, gender and case). However, the form *dárcích* is, morphologically, also Lpl form of the masculine inanimate noun *dárek* ‘present’. Without knowing the existence of the MWE *dárce krve* the rules could erroneously disambiguate and lemmatize the form *dárcích* as a Lpl.MascInan form of *dárek*. If so, the form *dobrovolných* ‘voluntary’ would then be erroneously also disambiguated as Lpl.MascInan (due to agreement). Moreover, the morphologically ambiguous form *krve* is correctly disambiguated as Gsg.

4 CONCLUSION

In the paper, we have demonstrated the significance of MWEs’ morphological disambiguation – performed by a special Phras module on the basis of a lexical database containing (partially) disambiguated MWEs – for the successful disambiguation of the other, non-MWE parts of sentences containing MWEs, the disambiguation being performed by subsequent disambiguation rules. Further work will consist in improving the collaboration of the Phras module with the general rules as to the division of labour: which MWEs are to be processed by general rules and which should be included in the lexical database and processed by the Phras module. Furthermore, the database will constantly be enhanced.

[8] The neuter nominal forms *určení* (‘destination’) and *narození* (‘birth’) in bold are, generally, part-of-speech ambiguous: they are also adjectives (‘determined’ and ‘born’) in Npl.MascAnim/Vpl. MascAnim.

References

[1] Hnátková, M. (2006). Typy a povaha komponentů neslovesných frazémů z hlediska lexikálního obsazení (Types and nature of components of non-verbal phrasemes from the viewpoint of lexical elements). In *Kolokace*. *Studie z korpusové lingvistiky*, pages 142–167, Nakladatelství Lidové noviny – Ústav Českého národního korpusu, Praha.

[2] Hnátková, M. and Kopřivová, M. (2014). From Dictionary to Corpus. In Jesenšek, V. and Grzybek, P., editors, *Phraseology in Dictionaries and Corpora, ZORA 97*, pages 155–168, Maribor, Slovenia.

[3] Jelínek, T. (2008). Nové značkování v Českém národním korpusu (New annotation in the Czech National Corpus). *Naše řeč*, 91(1):13–20.

[4] Jelínek, T. and Petkevič, V. (2011). Systém jazykového značkování současné psané češtiny (The system of linguistic annotation of contemporary written Czech). In *Korpusová lingvistika Praha 2011, sv. 3: Gramatika a značkování korpusů*, pages 154–170, Nakladatelství Lidové noviny / Ústav Českého národního korpusu, Praha, Czech Republic.

[5] Květoň, P. (2006). *Rule-Based Morphological Disambiguation (Towards a Combination of Linguistic and Stochastic Methods)*. PhD thesis. MFF UK, Praha.

[6] Petkevič, V. (2006). Reliable Morphological Disambiguation of Czech: Rule-Based Approach is Necessary. In Šimková, M., editor, *Insight into the Slovak and Czech Corpus Linguistics*, pages 26–44, Veda (Publishing House of the Slovak Academy of Sciences & Ludovít Štúr Institute of Linguistics of the Slovak Academy of Sciences), Bratislava, Slovakia.

[7] Petkevič, V. (2014). Problémy automatické morfologické disambiguace češtiny. *Naše řeč*, 97(4–5):194–207.

[8] Petkevič, V. (2014). Ambiguity, language structures and corpora. In *La linguistique* (coordonné par Radimský, J. and Pešek, O., editors, Le Cercle linguistique de Prague – II. D'hier à aujourd'hui), vol. 50, 2014-2, pages 63–82, Presses Universitaires de France.

[9] Petkevič V. (2014). *Morfologická homonymie v současné češtině*. Nakladatelství Lidové noviny – Ústav Českého národního korpusu, Praha.

VALENCY POTENTIAL OF SLOVAK AND FRENCH VERBS IN CONTRAST

KATARÍNA CHOVANCOVÁ – LUCIA RÁČKOVÁ –
DAGMAR VESELÁ – MONIKA ZÁZRIVCOVÁ
The Faculty of Arts, Matej Bel University in Banská Bystrica, Slovakia

CHOVANCOVÁ, Katarína – RAČKOVÁ, Lucia – VESELÁ, Dagmar – ZÁZRIVCOVÁ, Monika: Valency Potential of Slovak and French Verbs in Contrast. Journal of Linguistics, 2017, Vol. 68, No 2, pp. 156 – 168.

Abstract: The paper presents results of synchronous contrastive study of fifteen most frequent Slovak full verbs and their French equivalents by the method of corpus analysis aimed at observation and comparison of their valency potential in relation to their semantic structure. The inventory of valency structures of Slovak verbs and their French equivalents shows not only differences, but also, to a great extent, identical semantic-syntactic connectivities. The main apport of the study lies in the contrastive research perspective and the interdisciplinary character on the crossroads of grammar, semantics, syntax, cognitive and corpus linguistics. Findings can be of use to linguists, terminologists, lexicographers, authors of textbooks and grammars, translators and interpreters, as well as to French-speaking learners of Slovak and Slovak students of French.

Keywords: linguistics, grammar, corpus, verb, valency

1 INTRODUCTION

The paper arises from contrastive research of valency of Slovak and French verbs carried out at Matej Bel University in Banská Bystrica in cooperation with the Ľudovít Štúr Institute of Linguistics of the Slovak Academy of Sciences within the research grant project VEGA – *Valenčné potencie slovies v kontraste*/Valency Potential of Verbs in Contrast (2014–2016). It sums up research objectives, methodology and results. It points out at specificities of contrastive analysis of valency structures of Slovak and French verbs and presents partial conclusions.

The research was based on the premises of general and contrastive linguistics. It focused on the verb as a crucial point of syntax and, specifically, on valency as one of its distinctive features. In linguistics, as well as in methodology of teaching foreign languages, verb valency has frequently been researched on. A contrastive linguistic approach linked to computer-based treatment of language is less frequent, still very enriching.

2 RESEARCH ON VALENCY POTENTIAL OF SLOVAK AND FRENCH VERBS IN CONTRAST

The research was centered on a synchronous contrastive study of valency and semantic structures of fifteen most frequent Slovak full verbs (according to the

DOI 10.1515/jazcas-2017-0026

frequency list generated from the monolingual corpus *sme2011*, a part of written corpora of the Slovak National Corpus) and their French equivalents in order to investigate and compare their valency potential in relation to their semantic structure in Slovak and French language. Among selected Slovak verbs, the following full polysemic units appeared (listed from the most frequent, with most frequent English equivalents): *povedať* (to tell), *hovoriť* (to speak), *dostať* (to get), *tvrdiť* (to affirm), *prísť* (to arrive), *hrať* (to play), *získať* (to get, to obtain), *platiť* (to pay), *myslieť* (to think), *rozhodnúť* (to decide), *vidieť* (to see), *stáť* (to stand), *čakať* (to wait), *nájsť* (to find)*, patriť* (to belong). All of these units are treated as independent head words in Ivanová et al. (2014). The verb *hovoriť* (to speak) stands together with its aspectual pair word *povedať* (to tell) in one entry, then it stands alone in a separate entry, as well. This treatment is identical to the one presented in [3].

Leaning on already existing inventory of possible valency structures of Slovak full verbs we constructed an inventory of valency structures of French equivalents of different meanings of these Slovak verbs in form of bilingual dictionary entries. Meanings of verbs are the part of cognitive systems of two typologically different languages: Slovak, predominantly fusional language, and French, an analytic one. Despite this difference, the languages show not only differences, but also identical syntactic and semantic connectivities.

Original verbal lexemes are taken from [3]. We share the concept of valency adopted by its authors, defined as the capacity of the verb to control a certain number of arguments, determining their formal and semantic features.

Identification of French equivalents of various meanings of Slovak polysemic verbs and contrastive-comparative analysis of valency structures in both languages have been part of our research. The contrastive study was based on several Slovak and French theoretical works, among others [8], [11] and [12]. Among French sources, there are [13], [1] and [7].

The inventory of valency structures of Slovak full verbs and their French equivalents and description of their valency properties with regard to identical, partially identical and different semantic and syntactic features in Slovak and French was based on specialized corpora *sme2011* and *LeMonde0.3*.

Sme2011 is a specialized monolingual written corpus containing press articles published in SME, the Slovak National daily newspaper, from January 1 to December 31, 2011. It was created as a selection of texts contained in *prim-6.1.public-all*, main part of the Slovak National Corpus. It contains 6 516 876 text units (tokens), out of which 5 409 453 word forms. The corpus, consisting in 409 509 sentences, is fully lemmatized and morphologically annotated. The automatic morphological annotation was done by *Morče* tool based on the morphological tagset used in the Slovak National Corpus.

Le Monde 0.3 is a specialized foreign-language written corpus containing press articles from Le Monde, French national daily newspaper, published from January 1 to December 31, 2011. It contains 21 969 159 text units (tokens) and consists of 829 092 sentence structures. It is fully lemmatized and morphologically annotated. The automatic morphological annotation was done by TreeTagger and it uses a free set of morphological tags for French language. *Sme2011* and *LeMonde0.3* can be browsed using *NoSketch Engine* tool of the Slovak National Corpus.

Research results were published as a bilingual dictionary of valency structures of Slovak and French verbs [14]. The publication is primarily destinated to students of French philology and translation studies, as well as scholars dealing with comparison of Slavic and Romance languages and with corpus linguistics.

3 EXAMPLE OF CONTRASTIVE PRESENTATION OF SLOVAK AND FRENCH VALENCY STRUCTURES

The basic entry unit of the dictionary *Valenčné potencie slovies v kontraste* [14] is represented by a Slovak full polysemic verb with specific aspectual characteristics whose semantic structure consists in 1 to n partial meanings (verbal lexemes). Idiomatic meanings, as defined in [3], were excluded. For each meaning, one or more semantically equivalent French lexical units are given. The structure of the bilingual dictionary entry has two parts – the Slovak one (starting point for comparison) and the French one. The Slovak part of the bilingual entry contains:

a) Slovak full verb to be analyzed,
b) partial meaning of the Slovak full verb no. 1 to n,
c) valency structure of the Slovak verb in form of a valency pattern*,*
d) a morphological-syntactic characteristic of left and right arguments appearing in the valency pattern and semantic roles of arguments of the verbal lexeme,
e) synonyms of the Slovak verbal lexeme,
f) description of the meaning of the Slovak verbal lexeme no. 1 to n*,*
g) examples of use of the valency structure of the Slovak verb taken from *sme2011*.

Points b) to f) are taken from [3].

The French part of the bilingual entry contains:

a) the French equivalent of the Slovak verbal lexeme no. 1 to n,
b) valency structure of the French equivalent,
c) description of the meaning of the French lexical unit semantically equivalent to the Slovak verbal lexeme (based on French monolingual dictionaries),
d) examples of use of the valency structure of French lexical units equivalent to Slovak verbal lexeme, taken from *Le Monde 0.3*.

In all, 20 bilingual entries were analyzed.

To give an example of the contrastive approach, we present the analysis of the second meaning of the polysemic verb *patriť* (to belong) and its French equivalents. According to [3], monoaspectual imperfective verb *patriť ndk* (to belong as an imperfective verb) is formed of five independent verbal lexemes (five meanings); four of them (non idiomatic ones) are treated in the dictionary. The partitive meaning *patriť 2* "byť členom, súčasťou niečoho" (to be part of something), is given below as an example of contrastive presentation of valency structures. This meaning can be expressed, in French, by two different lexemes, both of them semantically equivalent: *appartenir* (to belong) "faire partie organique d'un ensemble" (to be an organic part of a whole) and *rentrer* (to fit) *"*faire partie de, être contenu, inclus dans une classe, une catégorie" (to be a part of something, to be contained in something, to be included in a class or a category). The first equivalent, *appartenir*, requires a right-side argument fulfilling the syntactical role of indirect object, introduced by the only possible preposition *à*. The argument can be expressed by a subordinate clause introduced by *à ce que* (Slovak valency structure being expressed as **VŠ$_{\mathbf{slo}}$: S$_{\mathbf{n}}$ – VF – k S$_{\mathbf{d}}$/medzi S$_{\mathbf{a}}$**, the corresponding French valency structure being expressed as **VŠ$_{\mathbf{fra}}$: S$_{\mathbf{S}}$ – VF – à S$_{\mathbf{COI}}$/PS$_{\mathbf{à\ ce\ que}}$**). In the valency structure of the intransitive verb *rentrer*, with a broader meaning than *appartenir*, the right side of the valency structure is reduced – the indirect object is not a part of it anymore. The obligatory right-side argument is, in this case, the adverbial of direction **VŠ$_{\mathbf{fra}}$: S$_{\mathbf{S}}$ – VF – ADV$_{\mathbf{dir3}}$**.

VL/SLO/2: patriť

VŠ$_{\mathbf{slo}}$: S$_{\mathbf{n}}$ – VF – k S$_{\mathbf{d}}$/medzi S$_{\mathbf{a}}$

S$_{\mathbf{n}}$: [BO] STAT$_{part}$

k S$_{\mathbf{d}}$/medzi S$_{\mathbf{a}}$: [BO] TOTUM

SYN/SLO/VSSSKZ: zaraďovať sa[1]

DEF/VL/SLO/2: byť členom, súčasťou niečoho[2]

EXSLO/SME2011:

1. Majster gotických malieb z Okoličného ***patril*** k najlepším maliarom stredoeurópskej neskorej gotiky[3]. (SME 2011-01-04)
2. K jeho obľúbenej literatúre ***patrí*** Starec a more[4]. (SME 2011-01-10)
3. Artefakt starý 3400 rokov ***patrí*** medzi najznámejšie staroveké pamiatky[5]. (SME 2011-01-25)

→ EKV/FRA/VL/SLO/2: appartenir

VŠ$_{\mathbf{fra}}$: S$_{\mathbf{S}}$ – VF – à S$_{\mathbf{COI}}$/PS$_{\mathbf{à\ ce\ que}}$

DEF/FRA/TLF: faire partie organique d'un ensemble[6]

EXFRA/LEMONDE0.3:

1. Né en 1786, il ***appartient*** à une génération qui se détourne de la vogue du premier romantisme noir[7]. (Le Monde, 4 avril 2011)

[1] To range among something.

[2] To be a member, a part of something,

[3] The master of Gothic painting from Okoličné ***ranged among*** the best artists of late Gothic period in the Central Europe.

[4] The Old Man and the Sea ***ranged among*** his favourite books.

[5] A 3400-year-old artifact ranges among the most famous Ancient relics.

[6] To be a organic part of a whole.

[7] Born in 1786, he ***belongs*** to a generation which rejects the wave of the first romanticism noir.

2. Née le 20 octobre 1928 à Caudéran, Hélène Surgère ***appartenait*** à la bourgeoisie, qu'elle a fuie[8]. (Le Monde, 1^{er} avril 2011)
3. Ainsi, l'on apprend que les poufs animaliers ***appartiennent*** à ce que les deux artistes appellent avec tendresse le « design de compagnie »[9]. (Le Monde, 24 janvier 2011)

→ EKV/FRA/VL/SLO/2: rentrer

VŠ$_{\mathbf{fra}}$: S$_{\mathrm{S}}$ – VF – ADV$_{\mathbf{dir3}}$

DEF/FRA/CNRTL: faire partie de, être contenu, inclus dans une classe, une catégorie[10]

EXFRA/LEMONDE0.3:

1. Pour montrer sa bonne volonté, la BCE, elle, continuera d'accepter que les banques à qui elle prête de l'argent lui apportent en garantie de la dette portugaise, malgré le fait que celle-ci ne ***rentre*** plus dans les critères acceptés par l'institut monétaire[11]. (Le Monde, 11 juillet 2011)
2. Les grandes croix ***rentrent*** dans le trésor de la tradition, lequel est patrimoine national et n'a pas d'étiquette sociale[12]. (Le Monde, 15 juillet 2011)
3. Or je ne pense pas que faire décoller tous les avions **rentre** dans ces conditions[13]. (Le Monde, 24 décembre 2011)

The following table contains an overview of valency patterns corresponding to all meanings of the verb *patriť* (belong) and its French equivalents.

No.	**VŠslo**	**VŠfra**	**EKV/FRA**	**Degree of equivalence**
VL/SLO/1 byť vlastníctvom niekoho, prislúchať (to be a property of someone, to belong to someone)	S$_{\mathrm{n}}$ – VF – S$_{\mathrm{d}}$	S$_{\mathrm{S}}$ – VF – à S$_{\mathrm{COI}}$	appartenir (to belong)	1

[8] Born on 20th October 1928 in Caudéran, Hélène Surgère ***belonged*** to the bourgeoisie that she flew from.

[9] Thus, we learn that the poufs with animal patterns ***belonged*** to what the two artists gently called « company design ».

[10] To be a part of something, to be contained in something, to be included in a class or a category.

[11] To show its good intentions, the CEB will continue to accept that the banks to which it lends money bring as a guarantee a part of the Portuguese debt, despite the fact it does not ***fit*** the criteria accepted by the monetary financial institution.

[12] Big crosses ***belong*** to the treasure of tradition, which is national heritage and does not bear a social label.

[13] I don't think that making all the planes take off ***fits*** the conditions.

VL/SLO/2 byť členom, súčasťou niečoho (to be a member, a part of something)	S_n – VF – k S_d/ medzi S_a	S_S – VF – à S_{COI}/$PS_à$ ce que	appartenir (to belong)	3
		S_S – VF – ADV_{dir3}	rentrer (to fit)	3
VL/SLO/3 mať niekde náležité miesto (to have one's proper place)	S_n – VF – ADV_{dir3}	S_S – VF – ADV_{dir3}	appartenir (to belong)	1
		S_S – [$VF_{avoir/trouver}^{+}$ sa place] – ADV_{loc}	avoir / trouver sa place (to have/to find its place)	3 (+ change in the formal expression of the predicate)
VL/SLO/4 byť určený, týkať sa niečoho (to be destinated to something, to concern something)	S_n – VF – S_d – ADV_{kauz}	S_S – VF – à S_{COI} – ADV_{kauz}	revenir (to relate to, to go back to)	1

Tab. 1. Valency structures of *patriť* and its French equivalents

4 ASSESSING EQUIVALENCE OF VALENCY STRUCTURES

The contrastive analysis of valency structures of Slovak and French verbs lets us asses the degree of equivalence of valency patterns of verbal lexemes in both languages. The equivalence is understood as a variable, expressing a functional correspondence of compared elements. Its value can be scaled from "zero" through "partial" to "total".

When assessing equivalence of valency structures of verbs, it is necessary to take into consideration two aspects. The first of them is a potential ambivalence of the relation between the formal expression of relationships within the valency structure of the verb and the functional value of its arguments. In some cases, valency structures can contain functionally equivalent elements having different formal means of expression. This results from different typological features of Slovak and French, mainly from the existence of nominal flexion in Slovak and its corresponding expression by analytical structures in French. Therefore, while evaluating valency structures, syntactical function of valency arguments will be considered binding. Thus, Slovak valency pattern **S_n – VF – S_d** and French valency pattern **S_s – VF – à S_{COI}** will be seen as equivalent.

Another factor influencing contrastive assessment of Slovak and French valency structures is the potential character of some arguments. If some arguments in Slovak valency pattern are potential and the corresponding (functionally equivalent) arguments in the French valency structure are obligatory or vice versa, we define the degree of equivalence of these valency structures as partial. Thus, the pair of valency structures $\mathbf{S_n}$ **– VF –** $\mathbf{S_a}$ in Slovak and $\mathbf{S_s}$ **– VF – (**$\mathbf{S_{COD}}$**)** in French is seen as partially equivalent.

The comparison concerns primary, non-transformed valency structures of analyzed verbs. Contrastive assessment of transformed structures would be highly limited by systemic non-correspondence of their formal expression. This is the case of character and position of reflexive structures in French and Slovak, as well as systematics and semantics of impersonal verb forms entering active or passive structures. These specific questions require special attention.

In total, 165 pairs of valency patterns were compared. In some cases, only one French valency pattern corresponded to the original Slovak valency pattern related to a particular meaning of the Slovak verb. Elsewhere, a multilateral correspondence was observed: several valency patterns of one or more French verbs could be paired with the original Slovak valency pattern. Thus, the total number of pairs of valency patterns does not reflect the actual number of semantic equivalences between Slovak verbs and their French equivalents. For instance, the analysis of the Slovak verb *verb rozhodnúť 'decide' (finite)/rozhodovať (non-finite)* in its perfective and imperfective form) reveals the presence of 3 valency patterns corresponding to 3 different meanings of the verb. On the other hand, we identify 8 valency patterns corresponding to 19 different meanings of 5 different French verbs semantically equivalent to *rozhodnúť dk/rozhodovať ndk* (*décider* – to decide, *résoudre* – to resolve, *arbitrer* – to arbitrate, *trancher* – to cut, *statuer* – to state).

With regard to the above stated remarks, we set up following degrees of equivalence:

1 – total functional equivalence of valency and non valency components
2 – total functional equivalence of valency participants
3 – partial functional equivalence of valency participants
4 – zero functional equivalence of valency participants.

Given the predominantly obligatory character of left-side participant, the focus is on measuring the degree of functional equivalence of right-side participants.

Degree 1 indicates cases when Slovak and French valency structures contain an equal number of right-side participants and each obligatory or potential right-side participant in the Slovak valency structure finds, in the French valency structure, a corresponding participant with identical syntactic function (for instance $\mathbf{S_a}$ **–** $\mathbf{S_{COD}}$) and with identical obligatory or potential character. At the same time, Slovak and French valency structures contain an equal number of identical non-valency complementations.

Degree 2 is used for cases when Slovak and French valency structures contain an equal number of functionally corresponding right-side elements. These participants correspond in their syntactic function as well as in their obligatory or potential character.

Degree 3 is used for cases when:

a) Slovak and French valency structures do not contain an equal number of obligatory and/or potential participants, still there is at least one relation of functional correspondence between a right-side participant in the Slovak valency structure and a right-side participant in the French valency structure,

b) there are differences in the obligatory and/or potential character of functionally corresponding right-side participants.

Degree 4 indicates cases when it is not possible to identify any relation of functional equivalence between a participant in the Slovak valency structure and a participant in the French valency structure.

While assessing the equivalence of structures, variability of morphological expression of valency participants is not taken into account, i.e. $\mathbf{S_a/VV_{\check{z}e}}$ and $\mathbf{S_{COD}}$ are considered equivalent.

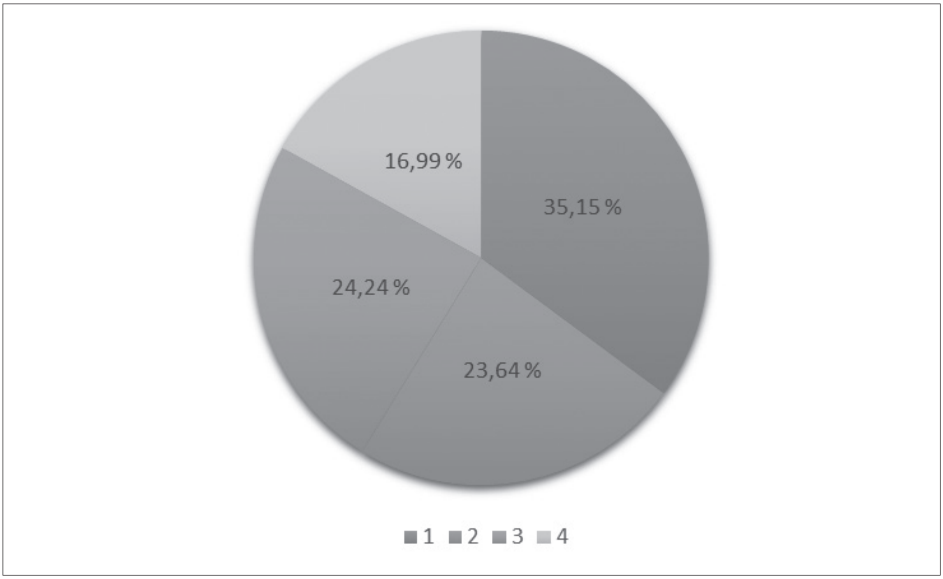


Fig. 1. Degrees of equivalence of Slovak and French verb valency structures

The proportional representation of various degrees of equivalence is shown in Figure 1. The largest group is marked as degree 1 of equivalence (35.15%), the least numerous group is characterized by zero functional equivalence – degree 4 (16.99%).

Degrees 1 and 2 have important ratios. The sizes of these groups indicate that the conventional idea about substantial differences in valency of verbs between French and Slovak – a belief which is often present in teaching/learning of French as a foreign language – does not necessarily reflect reality.

On the other hand, the cases where zero degree of equivalence (degree 4) of valency participants has been proved, may become the most important source of negative transfer in the situations of language contact in everyday communication, as well as in the process of language acquisition.

5 DEGREES OF EQUIVALENCE OF SLOVAK AND FRENCH VALENCY STRUCTURES

The corpus contains all 4 degrees of equivalence of valency structures of the Slovak and French verbs, illustrated as follows.

5.1 Total Equivalence of Valency and Non-valency Components

Degree 1 (total functional equivalence of valency and non-valency complementations) can be seen in the derived meaning of *vidieť/uvidieť 4* (to see) "stretať sa s niekým, dostávať sa do styku" (to meet someone, to get in touch with someone) and its French equivalent *voir* (to see) „fréquenter quelqu'un, le rencontrer lors d'une visite, dans le cadre de relations familiales ou sociales, rencontrer quelqu'un, se trouver par hasard en sa présence" (to see someone, to meet someone in a visit within a family of a society, to meet someone by accident):

(1) Bývam v jednom dome s Jožkom Stümpelom a Marcelom Hossom, ***vidíme sa*** takmer každý deň[14]. (SME 2011-09-23)

(2) Chaque fois que des jeunes gens se décident à aller rejoindre une manifestation, ils font leurs adieux à leurs proches comme s'ils les ***voyaient*** pour la dernière fois[15]. (Le Monde, 17 juin 2011)

The corresponding valency structures are the following:

VŠ$_{slo}$: S$_{n}$ – VF – S$_{a}$ – ADV$_{loc}$ – ADV$_{temp}$ – ADV$_{meas}$
VŠ$_{fra}$: S$_{S}$ – VF – S$_{COD}$ – ADV$_{loc}$ – ADV$_{temp}$ – ADV$_{meas}$

The assessment of the degree of equivalence was the same when valency structures contained only valency participants. This case can be illustrated by the primary meaning of *získať/získavať 1* (to get, to obtain) "dosiahnuť, nadobudnúť"" and its equivalent *obtenir* (to obtain) "parvenir à se faire accorder, à se faire donner ce que l'on veut avoir, ce que l'on demande" (to arrive to be awarded, to be given what one wants to have, what one demands) with the following valency structures: **VŠ$_{slo}$: S$_{n}$ – VF – S$_{a}$ – (od S$_{g}$)** and **VŠ$_{fra}$: S$_{S}$ – VF – S$_{COD}$ – (de S$_{COI}$)**:

(3) Niektoré spoločnosti však ***získali*** od NASA milióny dolárov a ich stroje už do kozmu lietajú[16]. (SME 2011-05-10)

(4) [...] l'activisme français a été décisif pour ***obtenir*** le feu vert de l'ONU[17]. (Le Monde, 29 mars 2011)

An identical degree of equivalence of valency structures can be seen in the primary meaning of *patriť 1* "byť vlastníctvom niekoho, prislúchať"" and its French equivalent *appartenir* "être la propriété de quelqu'un ; être le droit ou le privilège de quelqu'un" with valency structures **VŠ$_{slo}$: S$_{n}$ – VF – S$_{d}$** and **VŠ$_{fra}$: S$_{S}$ – VF – à S$_{COI}$**:

[14] I live in the same house as Jožko Stümpel and Marcel Hossa, we ***see*** each other almost each day.

[15] Every time the young people decide to go to a demonstration, they take a farewell of their closest as if they were ***saw*** them for the last time.

[16] Some societies ***have obtained*** millions dollars from NASA and their spacecraft fly to space already.

[17] [...] French activism was decisive in ***obtaining*** a consent from the UN.

(5) Budovy na predanom pozemku ***patria*** mestu a developer ich chce zadarmo. (SME 2011-01-03)

(6) Les deux hôtels ***appartiennent*** à l'homme d'affaires saoudien, le prince Al-Waleed. (Le Monde, 27 janvier 2011)

In both cases, the right-side participant is an indirect object expressed, in Slovak, without a preposition. In French, the preposition is needed. The syntactic function of participants was given priority before the means of their formal expression, as the form is often inluenced by factors of linguistic typology, i. e. $\mathbf{S_d}$ cannot have a non-prepositional French equivalent and it most likely corresponds to **à** $\mathbf{S_{COI}}$. The same approach was adopted elsewhere in case of differences in formal expression of components of valency structures caused by typological differences between the languages.

5.2 Total Functional Equivalence of Valency Participants

Degree 2 of equivalence was used to label the cases of total functional equivalence of valency participants when there is, at the same time, absence of equivalence in non-valency complementations. As an example, let us state *vidieť 2* "stať sa schopným vidieť", corresponding to the French verb *voir* "percevoir les objets du monde extérieur par l'intermédiaire des organes de la vue":

(7) A v noci ***vidíme*** šesťkrát lepšie ako vy! (SME 2011-03-03)
$\mathbf{VŠ_{slo}: S_n – VF –} \mathrm{ADV_{temp} – ADV_{meas}}$

(8) [...] je ne ***voyais*** déjà pas très bien de près à l'œil nu, maintenant c'est aussi de loin. (Le Monde, 10 mars 2011)
$\mathbf{VŠ_{fra}: S_S – VF –} \mathrm{ADV_{temp} – ADV_{meas} – ADV_{mod} – ADV_{instr}}$

Differences in the number and the character of adverbial complements can be seen in examples (7) and (8). Adverbials appear in the French valency structure ($\mathbf{VŠ_{fra}}$), however, the are missing from the Slovak one.

5.3 Partial Functional Equivalence of Valency Participants

Degree 3 of equivalence refers to situations when there is a difference in the obligatory or potential character of participants or when there is a functional equivalence of some of (but not all) valency participants on one or on the other side. The potentiality of valency participants indicates their possible absence in the surface structure. It means they are not necessarily realized; still, they are semantically binding and they belong among valency participants. They appear in brackets in valency patterns.

Changes in obligatory or potential character of valency participants can be observed in the verb *platiť/zaplatiť 2* (to pay) "dávať peniaze za nejakú hodnotu" (to give money for a value) and its French equivalent *payer* (to pay) "verser une somme d'argent, pour s'acquitter de ce qui est dû ou pour acheter quelque chose ; s'acquitter, par un versement, de ce qui est dû" (to give a sum of money to spend what is due or to buy something; to give out what is due). The difference consists in the position of the right-side participant with the syntactic function of direct object. Corpus data let us conclude that, in French, this participant is of potential nature.

(9) Viac ľudí bude ***platiť*** aj odvody, do vrecka budú mať hlbšie motoristi a ľudia si budú musieť vybrať medzi zamestnaním a predčasnou penziou[18]. (SME 2011-01-03)
$\mathbf{V\check{S}_{slo}: S_n - VF - S_a - (S_d) -} \mathrm{ADV_{mot} - ADV_{remp} - ADV_{instr}}$

(10) Il est d'autant plus important que 80 % des passagers sont anglais, ***paient*** en livres, quand en face, nous réglons nos dépenses en euros[19]. (Le Monde, 25 janvier 2011)
$\mathbf{V\check{S}_{fra}: S_S - VF - (S_{COD}) - (\grave{a}\ S_{COI}) -} \mathrm{ADV_{mot} - ADV_{meas} - ADV_{instr}}$

Partial functional equivalence can be illustrated by the Slovak verb *patriť 2* (to belong) "byť členom, súčasťou niečoho" (to be a member, a part of something) and its French equivalent *rentrer* (to go into, to fit) "faire partie de, être contenu, inclus dans une classe, une catégorie" (to be a part of something, to be contained in something, to be included in a class or a category). *Rentrer* is an intrasitive verb, semantically broader than the other equivalent of the same Slovak verb, *appartenir* (to belong) "faire partie organique d'un ensemble" (to be an organic part of something). The right side of the valency structure of *rentrer* is reduced: unlike in the Slovak valency structure, the indirect object is no longer present. Instead of it, we observe the presence of an obligatory adverbial expressing direction $\mathbf{ADV_{dir3}}$.

(11) Majster gotických malieb z Okoličného ***patril*** k najlepším maliarom stredoeurópskej neskorej gotiky[20]. (SME 2011-01-04)
$\mathbf{V\check{S}_{slo}: S_n - VF - k\ S_d/medzi\ S_a}$

(12) Or je ne pense pas que faire décoller tous les avions ***rentre*** dans ces conditions[21]. (Le Monde, 24 décembre 2011)
$\mathbf{V\check{S}_{fra} : S_S - VF - ADV_{dir3}}$

Functional equivalence of left-side participants is preserved.

5.4 Zero Functional Equivalence of Valency Participants

Degree 4 of equivalence relates to cases which do not fit any of the degrees described above, i. e. it is not possible to identify any relation of equivalence between right-side participants in Slovak and French valency structures. The Slovak verb *čakať* (to

[18] More people will ***pay*** fund contributions, it will become more difficult for car owners and people will have to choose between work and early retirement.

[19] It is important that 80 % passengers are English and ***pay*** in pounds and, on the opposite corner, we pay our expenses in euro.

[20] The master of Gothic painting from Okoličné ***ranged among*** the best artists of late Gothic period in the Central Europe

[21] I don't think that making all the planes take off ***fits*** the conditions.

wait, to expect) expressing the meaning "predpokladať (často nepríjemnú) udalosť" (to await that something (unpleasant) happens) is an example. The Slovak valency structure contains a right-side participant $\mathbf{S_{a}/VV_{\text{že.}}}$ The valency structure of the French equivalent *s'attendre* (to expect) contains a right-side participant of a different kind, preceded by the preposition *à*. This preposition introduces an indirect object in the dative case $\mathbf{\text{à } S_{COI}/PS_{\text{à ce que}}/\text{à INF:}}$

(13) Brusel *čaká*, že naša ekonomika porastie na budúci rok o 1,1 percenta a česká o 0,7 percenta[22]. (SME 2011-11-11)
$\mathbf{V\check{S}_{slo}\text{: } S_{n} – VF – S_{a}/VV_{\text{že}}}$

(14) Le groupe américain Blackrock, qui a réalisé cette étude pour le compte de l'institut d'émission, *s'attend* à l'aggravation des pertes bancaires entre 2011 et 2013, sous l'effet conjugué de la politique d'austérité, des faillites de PME et du trou noir représenté par les prêts aux promoteurs immobiliers[23]. (Le Monde, 2011-04-02)
$\mathbf{V\check{S}_{fra} : S_{S} – VF_{refl} – \text{à } S_{COI}/PS_{\text{à ce que}}/\text{à INF}}$

Moreover, a difference in the expression of the predicate can be observed. The French valency structure, unlike the Slovak one, contains a reflexive verb.

6 CONCLUSION

The research presented has been primarily inspired, on the one hand, by our efforts to formulate a new conception of contextualized grammar of the Slovak language for the French-speaking public. On the other hand, it is connected with the study of interference phenomena between French as a source language (mother tongue) and Slovak as a target language (acquired language) which has shown a potential difficulty in acquiring valency structures, as possible sources of negative transfer.

Presented results prove that differences between valency structures in both analyzed languages concern only a certain part of the ensemble of valency structures. At the same time, it has been confirmed that there is a large number of valency structures demonstrating, in a contrastive perspective, a high degree of functional equivalence. When conceiving comparative-contrastive descriptions of grammatical systems, as well as in the process of didactic mediation and facilitation of language acquisition, it is desirable not to present valency of verbs as a necessarily problematic phenomenon, but to focus the attention on structures where differences are actually observed. We believe that a contrastive overview of valency structures of the Slovak and French verbs, as it is presented in our works, can be of good use when trying to accomplish this objective.

[22] Brussels ***expects*** our economy to grow by 1.1 percent and Czech economy by 0.7 percent next year.

[23] The American group Blackrock, who launched this research for the emission institution, ***expects*** bank losses to get worse between 2011 and 2013 under the joint effect of an austere policy, SME bankruptcies and a blackhole represented by loans in real estate business.

References

[1] Ducrot, O. and Schaeffer, J.-M. (1972). *Nouveau dictionnaire encyclopédique des sciences du langage*. Éditions du Seuil, Paris.

[2] Ivanová, M. (2006). *Valencia statických slovies*. Filozofická fakulta Prešovskej univerzity, Prešov.

[3] Ivanová, M., Sokolová, M., Kyseľová, M., and Perovská, V. (2014). *Valenčný slovník slovenských slovies na korpusovom základe.* Filozofická fakulta Prešovskej univerzity v Prešove, Prešov.

[4] Kačala, J. (1989). *Sloveso a sémantická štruktúra vety*. VEDA, Bratislava.

[5] Klimová, K. (2012). *Questioni di aspetto verbale: un confronto tra italiano e slovacco*. Aracne Editrice, Roma.

[6] Křečková, V. (2012). *Terminologie & Linguistique: construction des ensembles terminologiques bilingues (slovaque – français).* Fakulta humanitných vied, Univerzita Mateja Bela, Banská Bystrica.

[7] Neveu, F. (2011). *Dictionnaire des sciences du langage. 2^{e} édition revue et augmentée.* Armand Colin, Paris.

[8] Pauliny, E. (1943). *Štruktúra slovenského slovesa. Štúdia lexikálno-sémantická.* Slovenská akadémia vied a umení, Bratislava.

[9] Pognan, P. (2008). De la théorie à l'application: Vallex, une démarche exemplaire*. The Prague Bulletin of Mathematical Linguistics*, 89:97–106.

[10] Ráčková, L. (2016). Pragmatické aspekty slovenčiny cez prizmu jej frankofónneho používateľa. *Motus in verbo*, V(1):29–34.

[11] Ružička, J. (1968). Valencia slovies a intencia slovesného deja. *Jazykovedný časopis*, 19(1):50–56.

[12] Sokolová, M. (1993). *Sémantika slovesa a slovesný rod*. VEDA – Vydavateľstvo Slovenskej akadémie vied, Bratislava.

[13] Tesnière, L. (1959). *Éléments de syntaxe structurale*. Klincksieck, Paris.

[14] Zázrivcová, M. et al. (2017). *Valenčné potencie slovies v kontraste*. Belianum, Banská Bystrica.

MICROSYNTACTIC ANNOTATION OF CORPORA AND ITS USE IN COMPUTATIONAL LINGUISTICS TASKS[1]

LEONID IOMDIN
A. A. Kharkevich Institute for Information Transmission Problems,
Russian Academy of Sciences, Moscow, Russia

IOMDIN, Leonid: Microsyntactic Annotation of Corpora and Its Use in Computational Linguistics Tasks. Journal of Linguistics, 2017, Vol. 68, No 2, pp. 169 – 178.

Abstract: Microsyntax is a linguistic discipline dealing with idiomatic elements whose important properties are strongly related to syntax. In a way, these elements may be viewed as transitional entities between the lexicon and the grammar, which explains why they are often underrepresented in both of these resource types: the lexicographer fails to see such elements as full-fledged lexical units, while the grammarian finds them too specific to justify the creation of individual well-developed rules. As a result, such elements are poorly covered by linguistic models used in advanced modern computational linguistic tasks like high-quality machine translation or deep semantic analysis. A possible way to mend the situation and improve the coverage and adequate treatment of microsyntactic units in linguistic resources is to develop corpora with microsyntactic annotation, closely linked to specially designed lexicons. The paper shows how this task is solved in the deeply annotated corpus of Russian, SynTagRus.

Keywords: Text corpora, Russian syntactically tagged corpus SynTagRus, syntactic idioms, microsyntactic annotation, microsyntactic dictionary

1 INTRODUCTORY REMARKS

The theory of microsyntax has been developed by the author over the last 15 years (recent publications include [1], [2], [3], [4]). In this theory, which has much in common with construction grammar (see e.g. [5], [6], [7] and [8][2], two main groups of linguistic units are distinguished: lexically centered syntactic idioms and lexically unrestricted non-standard syntactic constructions.[3] Throughout the paper, I will be mostly concerned with these units, which will be referred to as microsyntactic units. Primarily, I will consider syntactic idioms.

[1] The author is grateful to the Russian Humanitarian Scientific Foundation for their support of this work with a grant (No. 15-04-00562). Special thanks also go to anonymous reviewers of the submitted version of the paper, who provided some valuable remarks.

[2] Interestingly, the last paper by P. Lauwers and N, Van Wettere introduces the term “microconstructionalization”, which is an additional evidence of the proximity (but not the identity!) of the approaches.

[3] In fact, some non-standard syntactic constructions are lexically constrained in the sense that they contain two or even more occurrences of the same word. Russian has a great variety of such constructions, each having unique syntactic peculiarities and subtle semantic features, as e.g. ***rabota rabotoj**, no nado otdoxnut' »* ‘work is work but one needs a rest’ or ***videt' ja ne videl**, no slyshal ob etom. »* ‘ I didn’t really see it but I heard about it’ (lit. ‘to see I saw not but heard about it’). Probably Russian has many more constructions with lexical repetitions than e.g. English (cf. a relatively full list of English tautological constructions in [9]).

DE GRUYTER OPEN
DOI 10.1515/jazcas-2017-0027

Microsyntactic units are poorly represented even in traditional linguistic resources, such as monolingual or bilingual dictionaries or descriptive grammars. The reason for this is obvious: syntactic idioms are difficult to attach to a particular lexical entry (so they are often just mentioned and briefly commented on in an entry for one of the words constituting the idiom), while non-standard constructions are too specific to find a place for themselves in general grammars. In computational linguistic resources, microsyntactic elements are even less visible (as are idiomatic entities in general). As a result, they are often disregarded in high-end computational linguistics tasks, such as deep semantic analysis, quality parsing, question answering, or machine translation – or, at best, treated with *ad hoc* solutions.[4]

The project outlined below is an attempt to improve the state of affairs at least partially. The idea is twofold: 1) to create a special dictionary of microsyntactic units of Russian, which should provide comprehensive information on such units and ensure their effective use in computational linguistics applications, and 2) to develop a text corpus which should incorporate annotation of such units. The former type of resource, the Microsyntactic dictionary of Russian, has been described in detail in [4] and [10]. In what follows, I will focus on the second goal, i.e. the development of the corpus with microsyntactic annotation, which, so far, has been only briefly reported in [4] and [11].

2 MICROSYNTACTIC ANNOTATION IN SYNTAGRUS

Rather than create a new corpus with microsyntactic annotation from scratch, we decided to enhance the existing SynTagRus corpus of Russian texts, developed by our Laboratory of computational linguistics at the A. A. Kharkevich Institute for Information Transmission Problems in Moscow. For the recent state-of-the-art of SynTagRus, see [12]. Even though this corpus is not too large (it now contains about 1 million word tokens), it has several layers of annotation, including markup for (1) morphology, (2) syntax (in the formalism of dependency trees), (3) lexical senses (for words whose ambiguity is reflected in the underlying Combinatorial dictionary of Russian),[5](4) parametric lexical functions (in the sense of Meaning Û Text by Igor Mel'čuk [14]), (5) certain types of ellipsis and, recently, (6) anaphoric relations: the latter are currently traceable beyond the sentence level so that the antecedents of pronouns can be found either in the same sentence or in a text fragment comprising two preceding sentences (see [15], [16].)

Microsyntactic tagging is thus the seventh layer of SynTagRus markup.

2.1 Purpose of Microsyntactic Tagging

What is the purpose of creating this markup? It is a commonly known fact that a corpus annotated for lexical senses of words is a valuable linguistic resource

[4] A typical *ad hoc* solution is representing a multiword microsyntactic element as a single word, e.g. represent the sequence *in fact* as an unsegmented unit, ignoring cases where it is not, as in *in fact checking* or where it is part of a longer set phrase like *in fact or spirit*.

[5] We also held experiments of supplying SynTagRus with semantic markup on the basis of Juri Apresjan's system of fundamental classification of predicates (see [13]), but this markup is not maintained now.

instrument in solving many sophisticated theoretical and practical tasks, including those associated with theoretical semantics, monolingual and bilingual lexicography, WSD, and deep semantic analysis. In many cases, microsyntactic elements are polysemous, so, in a way, microsyntactic markup is close to lexical sense annotation.

Text corpora annotated for senses of words are few for many languages, including Russian, and they are seldom large; see e.g. [17] for the Russian equivalent of the SemCor corpus annotated with WordNet word senses (see [18], [19] for details).

We may be disappointed with the fact that such corpora are scarce and small, but at least they exist for standard words and are available for researchers. However, there have been no corpus resources at all so far that could provide markup for syntactically challenging phraseological units, including, of course, microsyntactic units. This means that the reported resource is, in all probability, the first one of its kind.

It must be noted that, over the last couple of years, considerable time and effort has been devoted by corpus developers to annotate text corpora of a variety of languages for multiword expressions (MWE) (see e.g. a recent overview [20], with extensive bibliography, and a comprehensive paper [21] on corpus annotation with verbal MWEs – specifically, light verb constructions of various types). It may seem, at first glance, that our research exactly falls within MWE annotation framework. Yet our goal is more specific and, in a way, more ambitious: we focus on linguistic units that have considerable syntactic specificity and strive to present their internal syntactic arrangement and determine how these units are incorporated into the sentence structure.

As a matter of fact, microsyntactic markup of the corpus is not an easy task. On the one hand, it is difficult to discriminate between a microsyntactic element and an arbitrary sequence of words, which may even span over different syntactic chunks. On the other hand, there exist no ready lists of microsyntactic units that could be viewed as exhaustive, or even representative. The available phraseological dictionaries provide no good approximation: most of the traditional idioms present in such dictionaries have no distinctive syntactic properties and cannot be considered as microsyntactic units, while many such units do not appear in such dictionaries.

2.2 Two Strategies

To make up for this lack of initial data, we used two different tactics of tagging SynTagRus for microsyntactic elements:

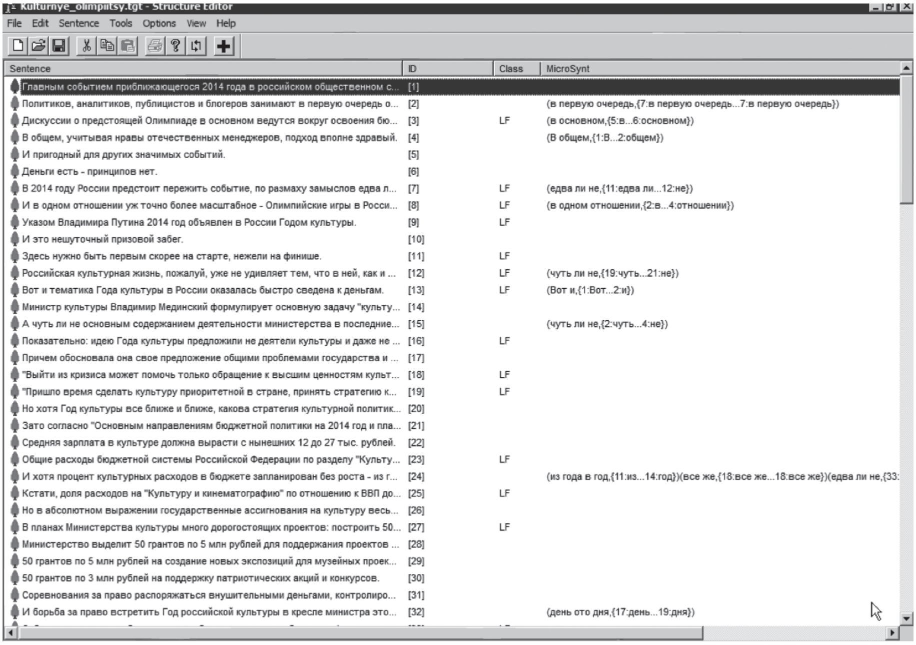
1) continuous analysis of whole individual texts, aimed at finding all candidates to microsyntactic elements within this text;
2) targeted search for linear strings and/or syntactic subtrees composed of such words about which we have had previous knowledge or reasonable conjecture that they form, or may form, microsyntactic units. To give a few examples, these are strings or subtrees like *vse ravno* ‘all the same’, *kak budto* ‘as though’, *kak by* ‘sort of’, *vse že* ‘yet’, *kak raz* ‘exactly, namely’, *kol’ skoro* ‘since; as long as’, *razve čto* ‘if only, except that’, *poka čto* ‘so far; for the time being’, *tol’ko liš’* ‘nothing but; as soon as’*, malo li* ‘one never knows; all sorts of

things', *vo čto by to ni stalo* 'at any cost; whatever happens', *ni razu* 'not once', *to i delo* 'over and over again', *čert znaet + interrogative word* 'devil knows (what, where,…)', *to i delo* 'ever so often', *to li delo* 'how much better', etc.[6]

Sure enough, in both cases only manual annotation of text for microsyntactic elements was possible: partial automation of microsyntactic elements could be done at the first stage of tagging in cases where strings of words constituting such elements had no gaps in between, with subsequent careful editing.

Using both tactics, we were able to obtain draft versions of microsyntactic markup of the corpus fragments, which were later subjected to thorough expert linguistic analysis, which revealed, among other things, that the number of microsyntactic elements occurring in the text is quite considerable. In a considerable number of texts, as many as a quarter of sentences were found to contain at least one microsyntactic element, so the microsyntactic markup turns out to be a frequent corpus feature.

Fig. 1 and Fig. 3 below are screenshots presenting the results of microsyntactic markup obtained by the two tactics. Fig. 1 shows the annotation for a fragment of a running journalistic text called *Kul'turnye olimpijtsy* 'Cultural Olympians'. The text, published by the popular Moscow *Novaya gazeta* newspaper in 2013, is a typical sample of SynTagRus material. It consists of 132 sentences, of which 33 (exactly 25%) were found to contain at least one microsyntactic element.



Sentence	ID	Class	MicroSynt
Главным событием приближающегося 2014 года в российском общественном с...	[1]		
Политиков, аналитиков, публицистов и блогеров занимают в первую очередь о...	[2]		(в первую очередь,{7:в первую очередь...7:в первую очередь})
Дискуссии о предстоящей Олимпиаде в основном ведутся вокруг освоения бю...	[3]	LF	(в основном,{5:в...6:основном})
В общем, учитывая нравы отечественных менеджеров, подход вполне здравый.	[4]		(В общем,{1:В...2:общем})
И пригодный для других значимых событий.	[5]		
Деньги есть - принципов нет.	[6]		
В 2014 году России предстоит пережить событие, по размаху замыслов едва л...	[7]	LF	(едва ли не,{11:едва ли...12:не})
И в одном отношении уж точно более масштабное - Олимпийские игры в Росси...	[8]	LF	(в одном отношении,{2:в...4:отношении})
Указом Владимира Путина 2014 год объявлен в России Годом культуры.	[9]	LF	
И это нешуточный призовой забег.	[10]		
Здесь нужно быть первым скорее на старте, нежели на финише.	[11]	LF	
Российская культурная жизнь, пожалуй, уже не удивляет тем, что в ней, как и ...	[12]	LF	(чуть ли не,{19:чуть...21:не})
Вот и тематика Года культуры в России оказалась быстро сведена к деньгам.	[13]	LF	(Вот и,{1:Вот...2:и})
Министр культуры Владимир Мединский формулирует основную задачу "культу...	[14]		
А чуть ли не основным содержанием деятельности министерства в последние...	[15]		(чуть ли не,{2:чуть...4:не})
Показательно: идею Года культуры предложили не деятели культуры и даже не ...	[16]	LF	
Причем обосновала она свое предложение общими проблемами государства и ...	[17]		
"Выйти из кризиса может помочь только обращение к высшим ценностям культ...	[18]	LF	
"Пришло время сделать культуру приоритетной в стране, принять стратегию к...	[19]	LF	
Но хотя Год культуры все ближе и ближе, какова стратегия культурной политик...	[20]		
Зато согласно "Основным направлениям бюджетной политики на 2014 год и пла...	[21]		
Средняя зарплата в культуре должна вырасти с нынешних 12 до 27 тыс. рублей.	[22]		
Общие расходы бюджетной системы Российской Федерации по разделу "Культу...	[23]	LF	
И хотя процент культурных расходов в бюджете запланирован без роста - из г...	[24]		(из года в год,{11:из...14:год})(все же,{18:все же...18:все же})(едва ли не,{33:
Кстати, доля расходов на "Культуру и кинематографию" по отношению к ВВП до...	[25]	LF	
Но в абсолютном выражении государственные ассигнования на культуру весь...	[26]		
В планах Министерства культуры много дорогостоящих проектов: построить 50...	[27]	LF	
Министерство выделит 50 грантов по 5 млн рублей для поддержания проектов ...	[28]		
50 грантов по 5 млн рублей на создание новых экспозиций для музейных проек...	[29]		
50 грантов по 3 млн рублей на поддержку патриотических акций и конкурсов.	[30]		
Соревнования за право распоряжаться внушительными деньгами, контролиро...	[31]		
И борьба за право встретить Год российской культуры в кресле министра это...	[32]		(день ото дня,{17:день...19:дня})

Fig. 1. Microsyntactic markup of a running text

[6] In order to avoid extended discussion, we list only one or two English equivalents for any microsyntactic units cited. Interestingly, in almost all of the above cases Russian microsyntactic units correspond to multiword English microsyntactic units which we use as glosses. It can thus be hypothesized that the number of microsyntactic phenomena and their typology in various languages may be quite comparable.

Currently, the markup looks as follows: a special field in the XML file representing the text cites the name of the microsyntactic element (in the case of syntactic idioms, it is normally a string of words, possibly with a figure attached to it if the syntactic idiom happens to be ambiguous) and the linear segment containing this element. For instance, a rather long sentence (24) of this text

I xotja procent kul'turnyx rasxodov v bjudžete zaplanirovan bez rosta - ***iz goda v god*** *1,5% – on* ***vse že*** *vdvoe vyshe, naprimer, čem procent rasxodov na fizkul'turu i sport, kotorye kažutsja nekotorym publicistam* ***edva li ne*** *glavnym prioritetom sovremennoj Rossii* 'And although the percentage of cultural expenditure in the budget is planned without growth – 1.5% from year to year – it is still twice as high, for example, than the percentage of spending on physical education and sports, which seem to some publicists to be almost the main priority of modern Russia'

contains three microsyntactic units (shown in boldface) – *iz goda v god* 'from year to year', *vse že* 'yet' and *edva li ne* 'almost'. In order to see how these units are incorporated into the syntactic structure, one needs to see the syntactic tree and identify the elements of the syntactic idioms as part of this tree.

Fig. 2 below shows a fragment of the syntactic tree for the above sentence with the first of the syntactic idioms discussed – *iz goda v god:*

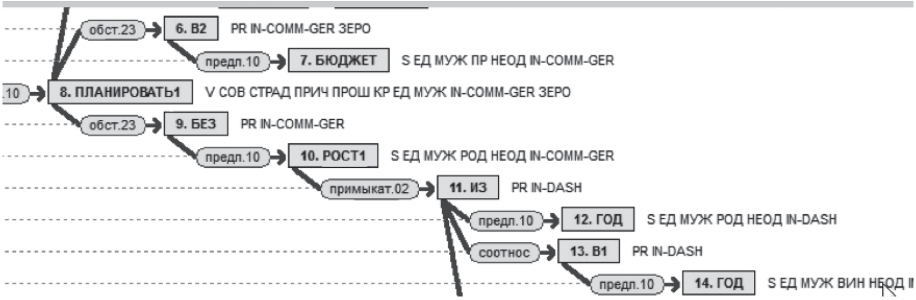


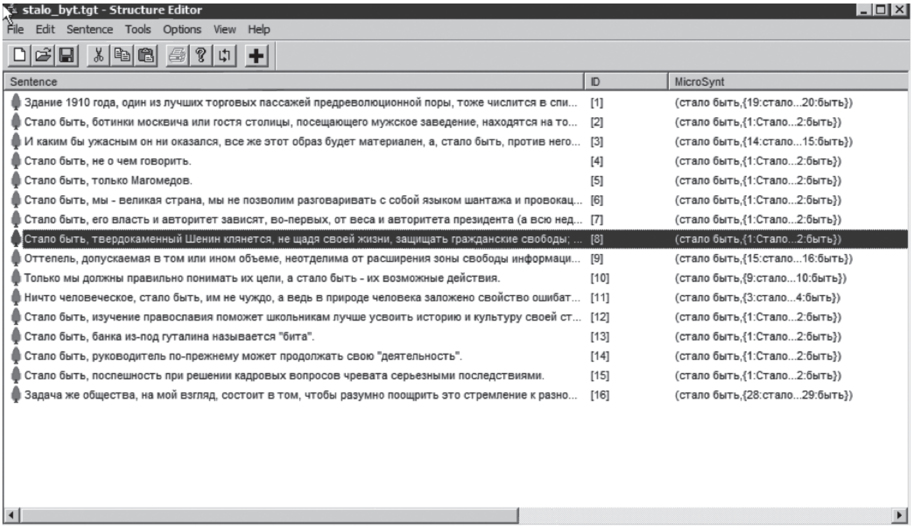
Fig. 2. A fragment of the syntactic tree structure containing a microsyntactic unit

It can be seen that the syntactic idiom occupies the nodes from 11 to 14, its local head, the preposition *iz* 'from', is dominated by the noun *rost* 'growth' and subordinates the noun *god* 'year' using the prepositional syntactic relation. The other prepositional phrase of this idiom (*v god* 'to year') is dominated by the first preposition with the correlative syntactic relation. So, the internal arrangement of the syntactic idiom within the structure has to be determined additionally: if the two prepositional phrases formed no such idiom, both prepositions would be most likely dominated in parallel by a verb or other predicate word.[7]

Fig. 3 below represents the second approach to microsyntactic annotation – the targeted search for possible microsyntactic units. In this case, we searched for sentences that are likely to contain a syntactic idiom *stalo byt'* 'hence, consequently'. The query for this unit (functioning as a parenthetical adverb despite being composed

[7] The syntactic representation of SynTagRus follows the conventions of the ETAP-3 parser (see [22], [23]), which in its turn heavily relies on the syntactic component of the Meaning ⇔ Text theory [14].

of two broad semantics verbs) was simple: find sentences with the wordform *stalo* (‘which is the neuter gender singular of the past tense of the verb *stat’* ‘begin’) followed by the wordform *byt’* (the infinitive of the verb *byt’* ‘be’).



Sentence	ID	MicroSynt
Здание 1910 года, один из лучших торговых пассажей предреволюционной поры, тоже числится в спи...	[1]	(стало быть,{19:стало...20:быть})
Стало быть, ботинки москвича или гостя столицы, посещающего мужское заведение, находятся на то...	[2]	(стало быть,{1:Стало...2:быть})
И каким бы ужасным он ни оказался, все же этот образ будет материален, а, стало быть, против него...	[3]	(стало быть,{14:стало...15:быть})
Стало быть, не о чем говорить.	[4]	(стало быть,{1:Стало...2:быть})
Стало быть, только Магомедов.	[5]	(стало быть,{1:Стало...2:быть})
Стало быть, мы - великая страна, мы не позволим разговаривать с собой языком шантажа и провокац...	[6]	(стало быть,{1:Стало...2:быть})
Стало быть, его власть и авторитет зависят, во-первых, от веса и авторитета президента (а всю нед...	[7]	(стало быть,{1:Стало...2:быть})
Стало быть, твердокаменный Шенин клянется, не щадя своей жизни, защищать гражданские свободы; ...	[8]	(стало быть,{1:Стало...2:быть})
Оттепель, допускаемая в том или ином объеме, неотделима от расширения зоны свободы информаци...	[9]	(стало быть,{15:стало...16:быть})
Только мы должны правильно понимать их цели, а стало быть - их возможные действия.	[10]	(стало быть,{9:стало...10:быть})
Ничто человеческое, стало быть, им не чуждо, а ведь в природе человека заложено свойство ошибат...	[11]	(стало быть,{3:стало...4:быть})
Стало быть, изучение православия поможет школьникам лучше усвоить историю и культуру своей ст...	[12]	(стало быть,{1:Стало...2:быть})
Стало быть, банка из-под гуталина называется "бита".	[13]	(стало быть,{1:Стало...2:быть})
Стало быть, руководитель по-прежнему может продолжать свою "деятельность".	[14]	(стало быть,{1:Стало...2:быть})
Стало быть, поспешность при решении кадровых вопросов чревата серьезными последствиями.	[15]	(стало быть,{1:Стало...2:быть})
Задача же общества, на мой взгляд, состоит в том, чтобы разумно поощрить это стремление к разно...	[16]	(стало быть,{28:стало...29:быть})

Fig. 3. Microsyntactic Markup of SynTagRus sentences with the unit *stalo byt’*

As seen from the screenshot, all 16 sentences satisfying the search query were tagged for the unambiguous microsyntactic unit *stalo byt’*. This means that, within the corpus, no sentences could be found in which the string *stalo byt’* meant something different (a random juxtaposition of the two wordforms, or a different phrase). It can be conjectured that this binary unit is very stable in the language, effectively excluding other lexical competitors. This hypothesis is easily confirmed by a search for the same string in a much larger corpus (the Russian National Corpus at `www.ruscorpora.ru`): we could find, using rather sophisticated contexts, only a very few sentences in which this string proved to be unrelated to our syntactic idiom. One such sentence, *No kogda by ni žil, nado vo čto by to ni stalo byt’ čestnym čelovekom* (Venedikt Erofeev) ‘Whenever one lives one needs by any means to be an honest man’ happened to have a phrase boundary between *stalo* (which, amusingly, was part of a different syntactic idiom – *vo čto by to ni stalo* ‘at whatever cost, by any means’) and *byt’*. Actually, all other occurrences of the string in the large corpus followed the same pattern as found in SynTagRus.

3 FIRST RESULTS

Even though regular microsyntactic tagging of the SynTagRus corpus was started only a few months ago, a number of linguistically interesting results could already be found.

1. Despite the fact that SynTagRus has a relatively small size, it proved to be quite representative of microsyntactic phenomena. Most microsyntactic elements tagged according to the second tactics of preliminary search for promising occurren-

ces could actually be detected (although some of them could naturally be represented by several examples only).

2. The extent of ambiguity of microsyntactic elements was found to vary significantly from one unit to the other.

Some elements proved to be quite homogenous. In addition to the case considered above (with *stalo byt'*), another microsyntactic unit, *kak byt' (s chem-libo)* 'what to do (about something)' shared the same property of being (almost) unambiguous, and never occurring in extraneous contexts in the SynTagRus corpus (in fact, it requires a lot of linguistic inventiveness to find relevant examples of *kak byt'* falling outside of the syntactic idiom considered (see [10] for more detail).

At the same time, other microsyntactic units proved to be highly ambiguous within the corpus. Moreover, words constituting them occurred in contexts totally unrelated to any of the unit's senses, providing many false positives during the markup. An illustrative example is the ramified set of microsyntactic units *kak by*, which had a number of senses and generated a host of false positives (see the screenshot of Fig. 4 below).

On the one hand, there is a microsyntactic unit which we will refer to as *kak by 1* ≈'sort of': this is a discourse particle with the semantics of comparison or uncertainty, as in sentence (97) from the screenshot in Fig.4: *Takim obrazom, nastupalo kak by ravnovesie* 'Thus, a kind of balance was established'.

On the other hand, there is an entirely different microsyntactic unit, the conjunction *kak by 2,* which is only used as a strongly governed word with many predicates sharing the semantics of apprehension, such as the verbs *bojat'sja* 'to be afraid'*, opasat'sja* 'to fear'*, ispugat'sja* 'to be scared', *sledit'* 'to make sure'*,* the nouns *bojazn', strax, opasenie* 'fear', and the predicative adverbs *strashno, bojazno* 'fearful', as in sentence (109) from the same screenshot: *Potom ja zamatyvalas' šal'yu i uxodila ne oboračivayas', boyas', kak by mne ne predložili deneg za niščiy vzgljad* (I.Grekova) 'Then I wrapped myself in a shawl and left without turning around, being afraid that I would be offered money for my beggarly look'.

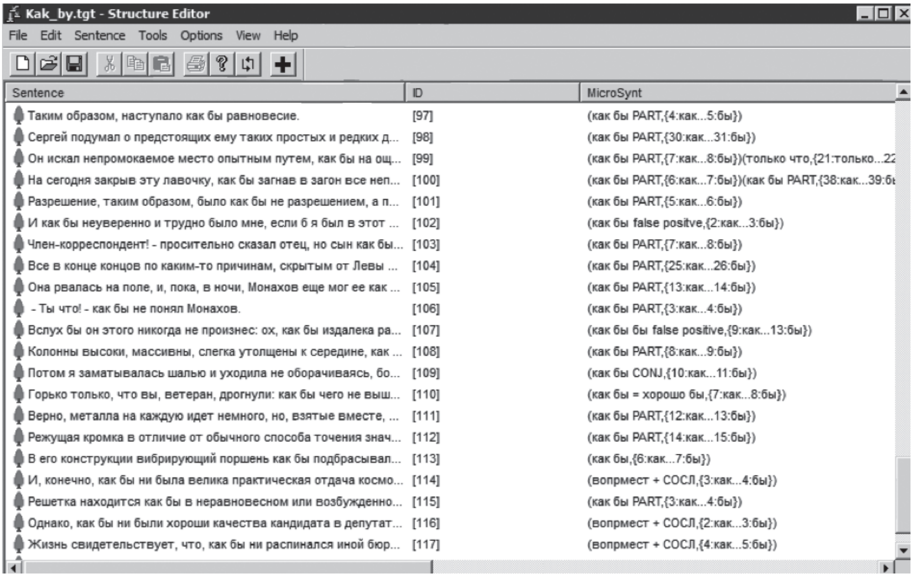
Yet another syntactic idiom composed of *kak* and *by* is a modal sentential adverb that implicitly expresses the speaker's wish – *kak by 3*. It is represented in such corpus sentences as *Kak by v kamennyj vek ne skatit'sja* 'It would be good not to slide back into the stone age' or *Kak by obojtis' bez etogo, ostaviv samuju sut'* [A.Bitov] 'I wish we could manage without it, leaving only the most crucial thing'.

In addition to these senses (plus a set of microsyntactc units which are longer than *kak by* and have to be distinguished from the above units), SynTagRus has a number of sentences that do not involve microsyntactic units formed with *kak by* despite the fact that physically this string is present. In particular, some sentences contain the construction with the emphatic particle *ni*: *Kak by nam ni xotelos' povysit' kačestvo školnogo obrazovanija, na eto potrebuetsja ešče mnogo let* 'However much we want to improve the quality of school education, this will require many years yet'. We believe that in such cases a good solution is to leave the sentence marked-up, introducing a "false positive" tag. Such a solution may seem a controversial one as it is not routinely applied in corpus annotation. I believe, however, that it may be very helpful not only as a provisional step at preparatory stages of corpus annotation but as a clear indication of the fact that the respective string does not form an idiomatic unit and represents a free juxtaposition of words otherwise belonging to such a unit. It may be viewed as a sort of negative linguistic material (in the sense of the Russian scholar Lev Shcherba), which can provide interesting linguistic insight for the grammarian and the lexicographer alike.

3. Normally, SynTagRus is representative enough of the most syntactic idioms having the same "lemma" name. However, to be sure that we have not missed anything, additional search is recommended for really ambiguous entities. For the *kak by* host of idioms, we were able to find one more interesting microsyntactic idiom formed with *kak* and *by* beyond the material of the corpus. It can be illustrated by a sentence present in the Russian National Corpus:

— *Kak by ne burja moskovskaja sobiraetsja, – pokrutil golovoj storož i povernul s pogosta von*. [B.Evseev]. 'Isn't it the case that the Moscow tempest is approaching? – The watchman twisted his head and went away from the cemetery'.

The meaning of this idiom (*kak by 4*) can be explained as follows: 'There are signs that the Moscow tempest is approaching, which is undesirable'. Importantly, in such cases a semantically void negation must be present – just like in the case with *kak by 2*.



Sentence	ID	MicroSynt
Таким образом, наступало как бы равновесие.	[97]	(как бы PART,{4:как...5:бы})
Сергей подумал о предстоящих ему таких простых и редких д...	[98]	(как бы PART,{30:как...31:бы})
Он искал непромокаемое место опытным путем, как бы на ощ...	[99]	(как бы PART,{7:как...8:бы})(только что,{21:только...22
На сегодня закрыв эту лавочку, как бы загнав в загон все неп...	[100]	(как бы PART,{6:как...7:бы})(как бы PART,{38:как...39:бı
Разрешение, таким образом, было как бы не разрешением, а п...	[101]	(как бы PART,{5:как...6:бы})
И как бы неуверенно и трудно было мне, если б я был в этот ...	[102]	(как бы false positve,{2:как...3:бы})
Член-корреспондент! - просительно сказал отец, но сын как бы...	[103]	(как бы PART,{7:как...8:бы})
Все в конце концов по каким-то причинам, скрытым от Левы ...	[104]	(как бы PART,{25:как...26:бы})
Она рвалась на поле, и, пока, в ночи, Монахов еще мог ее как ...	[105]	(как бы PART,{13:как...14:бы})
- Ты что! - как бы не понял Монахов.	[106]	(как бы PART,{3:как...4:бы})
Вслух бы он этого никогда не произнес: ох, как бы издалека ра...	[107]	(как бы бы false positive,{9:как...13:бы})
Колонны высоки, массивны, слегка утолщены к середине, как ...	[108]	(как бы PART,{8:как...9:бы})
Потом я заматывалась шалью и уходила не оборачиваясь, бо...	[109]	(как бы CONJ,{10:как...11:бы})
Горько только, что вы, ветеран, дрогнули: как бы чего не выш...	[110]	(как бы = хорошо бы,{7:как...8:бы})
Верно, металла на каждую идет немного, но, взятые вместе, ...	[111]	(как бы PART,{12:как...13:бы})
Режущая кромка в отличие от обычного способа точения знач...	[112]	(как бы PART,{14:как...15:бы})
В его конструкции вибрирующий поршень как бы подбрасывал...	[113]	(как бы,{6:как...7:бы})
И, конечно, как бы ни была велика практическая отдача космо...	[114]	(вопрмест + СОСЛ,{3:как...4:бы})
Решетка находится как бы в неравновесном или возбужденно...	[115]	(как бы PART,{3:как...4:бы})
Однако, как бы ни были хороши качества кандидата в депутат...	[116]	(вопрмест + СОСЛ,{2:как...3:бы})
Жизнь свидетельствует, что, как бы ни распинался иной бюр...	[117]	(вопрмест + СОСЛ,{4:как...5:бы})

Fig. 4. Microsyntactic markup of a SynTagRus fragment containing sentences with the string *kak by*

4. The material of syntactic idioms present in SynTagRus provides us with valuable data on linear variations of these idioms, their syntactic structure, their obligatory and optional valencies, and most importantly, their unique semantic features, which should be thoroughly accounted for in the resources like Microsyntactic dictionary. We intend to use this opportunity to the fullest extent possible.

References

[1] Iomdin, L. L. (2013). Nekotorye mikrosintaksičeskie konstruktsii v russkom jazyke s učastiem slova *čto* v kačestve sostavnogo elementa. [Certain microsyntactic constructions in Russian which contain the word *čto* as a constituent element.] *Južnoslovenski filolog*, LXIX:137–147. [In Russian.]

[2] Iomdin, L. L. (2014). Xorošo menja tam ne bylo: sintaksis i semantika odnogo klassa russkix razgovornyx konstruktsij. [Good thing I wasn't there: syntax and semantics of a class of Russian colloquial constructions.] In *Grammaticalization and lexicalization in the Slavic languages. Proceedings from the 36th meeting of the commission on the grammatical structure of the Slavic languages of the International committee of Slavists*, pages 423–436, Verlag Otto Sagner, München/Berlin/Washington D.C. [In Russian.]

[3] Iomdin, L. L. (2015). Konstruktsii mikrosintaksisa, obrazovannye russkoj leksemoj *raz*. [Construction of microsyntax built by the Russian word *raz*.] *SLAVIA, časopis pro slovanskou filologii*, 84(3):291–306. [In Russian.]

[4] Iomdin, L. (2016). Microsyntactic Phenomena as a Computational Linguistics Issue. In *Grammar and Lexicon: Interactions and Interfaces. Proceedings of the Workshop*, pages 8–18, Osaka, Japan. Accesible at: `http://aclweb.org/anthology/W/W16/W16-38.pdf.`

[5] Fillmore, Ch. (1988). The Mechanisms of Construction Grammar. In *Proceedings of the Fourteenth Annual Meeting of the Berkeley Linguistics Society*, pages 35–55.

[6] Goldberg, A. (1995). *Constructions: A Construction Grammar Approach to Argument Structure*. University of Chicago Press, Chicago.

[7] Rakhilina, E. V., editor (2010). Lingvistika konstruktsij. [The linguistics of constructions.] Azbukovnik Publishers, Moscow. [In Russian.]

[8] Lauwers, P. and Wettere, van N. (2017). La Micro-constructionnalisation En Tandem: La Copularisation De Tourner/virer. *Langue française*, 194(2):85–103.

[9] Rhodes, R. (2009). Tautological constructions in English … and beyond. Presented to the Syntax and Semantics Circle, UCB. Accessible at: `http://linguistics.berkeley.edu/~russellrhodes/pdfs/syntax_circle_taut_qp.pdf.`

[10] Iomdin, L. (2017). Kak nam byt' s konstruktsijami tipa *kak byt*? [What to do about constructions like *what to do*?] *Computational Linguistics and Intellectual Technologies. Dialogue 2017*, 16 (23)(2):150–161. [In Russian, Engl. Abstract.]

[11] Marakasova, A. A. and Iomdin, L. L. (2016). Mikrosintaksičeskaja razmetka v korpuse russkix tekstov SynTagRus [Microsyntactic tagging in the SynTagRus corpus of Russian texts.] In *Informacionnye texnologii i sistemy 2016 (ITiS'2016). Sbornik trudov 40-oj meždisciplinarnoj školy--konferencii IPPI RAN*, pages 445–449, Repino, Saint Petersburg, Russia. [In Russian.] Accessible at: `http://itas2016.iitp.ru/pdf/1570285171.pdf .`

[12] Dyachenko, P. V., Iomdin, L. L., Lazursky, A. V., Mityushin, L. G., Podlesskaya, Yu, O., Sizov, V. G., Frolova, T. I., and Tsinman, L. L. (2015). Sovremennoe sostojanie gluboko annotirovannogo korpusa tekstov russkogo jazyka (SynTagRus). [The current state of the deeply annotated corpus of Russian texts (SynTagRus).] In *Nacional'nyj korpus russkogo jazyka. 10 let proektu. Trudy Instituta russkogo jazyka im. V.V. Vinogradova.* M, Vol. 6, pages 272–299. [In Russian.]

[13] Apresjan, Ju., D., Iomdin, L. L., Sannikov, A. V., and Sizov, V. G. (2004). Semantičeskaja razmetka v gluboko annotirovannom korpuse russkogo jazyka. [Semantic Tagging in a deeply annotated corpus of Russian.] In *Trudy mezhdunarodnoj konferencii «Korpusnaja lingvistika – 2004»,* pages 41–54, Izd-vo Sankt-Peterburgskogo universiteta, Saint Petersburg, Russia. [In Russian.]

[14] Mel'čuk, I. A. (1974). *Opyt teorii lingvističeskix modelej «Smysl Û Tekst»*. [An experience of creating the theory of linguistic models of the Meaning Û Text type.] Nauka Publishers, Moscow. [In Russian.]

[15] Inshakova, E. S. (2016). Razrešenie sintaksičeskoj mestoimennoj anafory v sisteme «ETAP-3». [Resolution of syntactic pronominal anaphora in the ETAP-3 system.] In *Informacionnye texnologii i sistemy 2016 (ITiS'2016). Sbornik trudov 40-oj meždisciplinarnoj školy-konferencii IPPI RAN*, pages 420–429, Repino, Saint Petersburg, Russia. [In Russian.] Accessible at: `http://itas2016.iitp.ru/pdf/1570282678.pdf.`

[16] Marakasova, A. A. (2016). Avtomatičeskoe razrešenie anafory v russkom tekste: slučaj nulevogo sub”ekta. [Automatic resolution og anaphora in a Russian text: the case of a zero subject.] In *Informacionnye texnologii i sistemy 2016 (ITiS’2016). Sbornik trudov 40-oj meždisciplinarnoj školy-konferencii IPPI RAN*, pages 431–436, Repino, Saint Petersburg, Russia. [In Russian.] Accessible at: `http://itas2016.iitp.ru/pdf/1570285121.pdf.`

[17] Dikonov, V. G. and Poritski, V. V. (2014). A Virtual Russian Sense Tagged Corpus and Catching Errors In A Russian Û Semantic Pivot Dictionary. *Computational Linguistics and Intellectual Technologies. Dialogue 2014*, 13(20):128–137.

[18] Mihalcea, R. (1998). SemCor semantically tagged corpus, SenseEval 2 & 3 data in SemCor format. Accessible at: `http://www.cse.unt.edu/~rada/downloads.html.`

[19] Petrolito, T. and Bond, F. (2014). A survey of WordNet Annotated Corpora. In *Proceedings of the Seventh Global WordNet Conference*, pages 236–243, Tartu, Estonia.

[20] Rosén, V., Smedt, K. de, Smørdal Losnegaard, G., Bejček, E., Savary, A. and Osenova, P. (2016). MWEs in Treebanks: From Survey to Guidelines. In *Proceedings, LREC 2016, Tenth International Conference on Language Resources and Evaluation*, pages 2323–2330, Portorož, Slovenia.

[21] Savary, A., Sangati, F., Candito, M. et al. (2017). The PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 31–47, Valencia, Spain.

[22] Apresjan, Ju. D., Boguslavsky, I. M., Iomdin, L. L., Lazursky, A. V., Mitjushin, L. G., Sannikov, V. Z., and Tsinman, L. L. (1992). Lingvističeskij processor dlja složnyx informacionnyx sistem. [A linguistic processor for complex information systems.] Nauka Publishers, Moscow. [In Russian.]

[23] Apresjan, Ju. D., Boguslavsky, I. M., Iomdin, L. L., and Sannikov, V. Z. (2010). Teoretičeskie problemy russkogo sintaksisa: Vzaimodejstvie grammatiki i slovarja. [Theoretical Issues of Russian Syntax: Interaction of the Grammar and the Lexicon.] In Apresjan, Ju. D., editor, *Jazyki slavjanskix kul’tur*. Moscow. [In Russian.]

CLITIC CLIMBING, FINITENESS AND THE RAISING-CONTROL DISTINCTION. A CORPUS–BASED STUDY[1]

EDYTA JURKIEWICZ-ROHRBACHER [1,2] – BJÖRN HANSEN[1]
– ZRINKA KOLAKOVIĆ[1]
[1] Universität Regensburg, Germany
[2] University of Helsinki, Finland

JURKIEWICZ-ROHRBACHER, Edyta – HANSEN, Björn – KOLAKOVIĆ, Zrinka: Clitic Climbing, Finiteness and the Raising-Control Distinction. A Corpus–Based Study. Journal of Linguistics, 2017, Vol. 68, No 2, pp. 179 – 190.

Abstract: In the paper, we discuss the phenomenon of clitic climbing out of finite da$_2$-complements in contemporary Serbian. Scholars' opinions on the acceptability and occurrence of this construction, based on a handful of self-made examples, vary considerably. Expanding on the assumption that the correctness of the phenomenon has often been denied due to its rareness we employ large corpora to examine the problem. We focus on possible constraints arising from the syntactic properties of clause-embedding predicates.

Keywords: srWac, constraints on clitic climbing, semifinite complements, syntax, Serbian

1 INTRODUCTION

Following [16, p. 162], under the term CLITIC CLIMBING (CC) we understand "constructions in which the clitic is associated with a verb complex in a subordinate clause but is actually pronounced in constructions with a higher predicate [...]", as illustrated in the Serbian example:

(1)	*Gde*	***nas***$_2$	*možete*$_1$	*naći*$_2$*?*
	where	us.ACC	can.2PRS	find.INF

‘Where can you find us?’ (srWaC v1.2)

In many languages, CC is only attested in complex clauses involving infinitives; cross-linguistically, CC out of complements with inflected verbs is a rare phenomenon.

In Serbian complement clauses the infinitive competes with the so-called *da*-complement, i.e. a verb marked for person and number which is introduced by an element usually treated as a complementizer as in (2):

[1] This study was carried out within the research project ‚Microvariation of the Pronominal and Auxiliary Clitics in Bosnian, Croatian and Serbian. Empirical Studies of Spoken Languages, Dialects and Heritage Languages‘ funded by the Deutsche Forschungsgemeinschaft (HA 2659/6-1, 2015-2018).

DE GRUYTER OPEN
DOI 10.1515/jazcas-2017-0028

(2)	(...) *na*	*celoj*	*toj*	*teritoriji*	*ne*	*možete*$_1$
	on	whole.LOC	that.LOC	territory.LOC	NEG	can.2PRS
	da	*nađete*$_2$	*500*	*stanovnika.*		
	COMP	find.2PRS	500	inhabitants.GEN		(srWaC v1.2)
	‘On that whole territory you cannot find 500 inhabitants.’					

However, it remains unclear up to what extent and under what circumstances CC out of *da*-complements is possible. The present paper approaches this problem empirically. Section 2 refers to the discussion on CC out of *da*-complements in Serbian and Section 3 introduces the Raising-Control distinction. Section 4 presents the main sources of data used in this study while Section 5 explains the data collection process and the difficulties it poses. Section 6 describes the results in detail, and is followed by the final Section 7, which draws conclusions from the main results and offers a suggestion for future research.

2 THE *DA*-COMPLEMENT AND CC IN SERBIAN

As the research on the syntax of Bosnian, Croatian and Serbian is divided into descriptive empirical studies on the one hand, and works with a formal theoretical orientation on the other, it comes as no surprise that in the literature we find largely contradictory statements concerning CC out of *da*-complements.

S. Stjepanović [17, p. 174ff] argues that *da*-complements and infinitival clauses allow CC in a similar way. However, discussing examples of CC out of *da*-complements, S. Stjepanović [17, p. 201] writes imprecisely that those “are acceptable sentences, however, they are short of perfect”. Similarly, according to [5, p. 243], movement out of the finite complement is only “marginally possible”. In contrast, D. Ćavar and C. Wilder [2, p. 41] and W. Browne [1, p. 41] argue that CC out of finite complements is strictly impossible. Finally, Lj. Progovac [13, p. 146] admits that “some speakers of Serbian” do not accept CC in the presented contexts. All the above-mentioned authors rely exclusively on self-constructed examples.

An early empirical work concerning CC is [11][2], who assumed that the variation in clitic positioning is closely related to the (at that time) new and increasing tendency to replace the infinitive with ‘*da* + Present tense’. He claimed that although ekavian Serbian speakers, who had already almost completely replaced the infinitive with ‘*da* + Present tense’, preferred keeping the pronominal clitic directly after *da* (without CC), CC was common in journalistic texts published in Sarajevo [11].

It has long been known that *da*-complements do not behave in a uniform way. M. Ivić [7] proposes to distinguish two complement types headed by *da*: ‘mobile present tense’ and ‘immobile present tense’, the previous being regularly marked for tense and the latter not. This distinction goes back to [6], and was further elaborated on by W. Browne [1] and O. Mišeska-Tomić [12] who used the terms *da*$_1$- and *da*$_2$-complement. Based on [18], we assume that if CC is possible, this is so in the case of *da*$_2$-complements. One hypothetical reason why some scholars reject the possibility

[2] He does not use the term clitic climbing.

of CC out of *da*-complements is its extreme rarity in comparison to equivalent constructions without CC. In our paper, we address the following research question:

Q1: To what extent is clitic climbing out of *da*$_2$-complements in Serbian possible?

3 CC AND THE RAISING-CONTROL DISTINCTION

If CC out of *da*$_2$-complements is possible, the question arises which syntactic features enable or block climbing. To start with, we investigate the potential link between CC and the Raising-Control Distinction, usually held to be crucial to categorizing different types of sentences with complement clauses.

Due to lack of space, we will confine ourselves to some basic empirical observations discussed in various theoretical frameworks. Roughly speaking, in raising constructions the subject does not receive its thematic role directly from the matrix predicate but from the embedded predicate. In a control construction, in contrast, the matrix verb and the embedded verb each assign a subject thematic role; therefore, there are two syntactic arguments present: the surface subject and the non-overt infinitival null subject PRO. W. Davies and S. Dubinsky [4, pp. 4–8] propose relatively robust, cross-linguistically applicable tests to distinguish raising from control constructions: i) the argument of the matrix predicate takes over the theta role of the argument of the embedded predicate, ii) the argument of the matrix predicate takes over the selectional restrictions:

(3)	*Veliki*	*vizionar*	*može*	*da*	*donese*
	big.NOM	visionaire.NOM	Can.3PRS	COMP	bring.3PRS
	najvredniji	*trofej*	*u*	*Ukrajinu.*	
	mostvaluable.ACC	trophy.ACC	in	Ukraine.ACC	(srWaC v1.2)
	‘The great visionary can bring the most valuable trophy to Ukraine.’				Raising

(4)	*Velike*	*sile*	*pokušale*	*su*	*da* PRO
	big.NOM.F	forces.NOM.F	try.PTCP.PL.F	be.3PL	COMP
	spreče	*taj*	*ustanak.*		
	stop.3PRS	that.ACC	rebellion.ACC		(srWaC v1.2)
	‘Big forces tried to stop that rebellion.’				Control

and iii) passivization does not change the meaning of the sentences:

(3’)	*Najvredniji*	*trofej*	*može*	*da*	*bude*
	most valuable.NOM	trophy.NOM	can.3PRS	COMP	be.3SG
	donesen	*u*	*Ukrajinu.*		
	bring.PASS.NOM.M	in	Ukraine.ACC		
	‘The most valuable trophy can be brought to Ukraine (by the great visionary).’				

(4’)	**Taj*	*ustanak*	*pokušao*	*je*	*da*
	that.NOM.M	rebellion.NOM.M	try.PTCP.SG.M	be3SG	COMP

bude *sprečen.*
be.3SG stop.PASS.NOM.M
'Attempts were made to have that rebellion stopped (by big forces)'

A distinction should be made between subject and object control constructions. Whereas predicates that have only one individual argument besides the clausal argument are always subject control predicates, polyvalent predicates may show either subject or object control. According to [15, p. 412], verbs denoting directive speech acts (e.g. *zamoliti* 'to request') belong to the canonical class of object control predicates, and predicates that refer to commissive speech acts (e.g. *obećati* 'promise') are typical subject control predicates.

(5) *Dekan*$_X$ *je* *sve* *prisutne*$_Y$ *zamolio* *da* PRO$_Y$
dean.NOM be.3SG all.ACC present.ACC ask.PTCP.SG.M COMP
kažu *svoje* *utiske (...)*
say.3PRS their.ACC impressions.ACC (srWaC v1.2)
'The Dean kindly asked the attending members to share their impressions (...).'

The point of departure of our study is divergent statements on the link between CC and the Raising-Control dichotomy in Czech. According to M. Rezac [14], CC is allowed out of infinitival complements of raising, subject-control, and object-control verbs alike. U. Junghanns [8] basically agrees with M. Rezac [14], but raises doubt as to the acceptability of CC with object control. Based on this debate, we formulate our second research question:

Q2: Does clitic climbing out of *da*-complements in Serbian depend on verb type with respect to the Raising-Control Distinction?

In order to approach the two research questions we examine the behaviour of CLs in relation to the type of clause embedding predicate (CEP). The most suitable method of exploring our research questions appears to be a corpus-based approach.

4 SOURCE OF DATA

The lack of resources for most of the South Slavic languages was recognized by the group of linguists behind the Regional Linguistic Data Initiative. Namely, only few electronically stored corpora of contemporary, original Serbian texts are easily accessible for research.

Since other relevant criteria for our study are size, morphosyntactic annotation, type and variety of texts, the number of relevant sources drops drastically. srWaC [9] seems currently to be the most suitable source for studies on rarely occurring phenomena[3].

[3] The only alternative is offered by the Corpus of Contemporary Serbian Langauge (version SrpKor2013) developed by Miloš Utvić and Duško Vitas, but it is five times smaller. The quality of metainformation and availability of search options are additional reasons why we excluded this corpus from the current study.

srWaC is the biggest corpus of Serbian. The current version 1.2 is a web corpus collected from .rs top-level domains containing nearly 555 million tokens. The corpus is automatically annotated with diacritic restoration, morphosyntax and lemma layers. The accuracy of morphosyntactic tagger performance has been evaluated at 92.33%, while part-of-speech tagger accuracy reached 97.86% in the tests [10]. Some imperfections of taggers can be identified through frequency list analysis and compensated for in the query formula.

The corpus is available for download, but it is also accessible via an on-line interface, NoSketchEngine, which offers a more convenient way for the linguistic community to search the corpus structures in comparison to self-written scripts. The on-line version provides a Corpus Query Language-based concordancer as well as many useful tools such as filtering or frequency lists.

Next to size, available meta-information and accessibility, a great advantage of srWaC is data variety. Analysis of url domain lists shows that not only does srWac cover texts typically included in corpora of standard language such as literary, journalistic and administrative texts, and academic and popular scientific texts: it is also a valuable source of less formal language appearing in user-generated content such as comments and fora.

Although the Internet is often criticized for poor quality of texts, which covers numerous spelling errors, omission of diacritic signs and non-standard use of upper and lower case, it is also a source of authentic, spontaneously produced written texts. Spatially unrestricted access to the Internet additionally gives some prospects for the study of regional differences, which otherwise might remain undiscovered.

The main drawback of srWaC, that is the lack of control for text-types and authorship, has not yet been solved, so some caution must be applied with regard to linguistic variation.

Another problem arises from the heterogeneity of text-types represented by a single domain, which is particularly important for separating narrative texts. Complementary literary texts, however, can easily be obtained from InterCorp [3]. The most recent, ninth version contains a section with original Serbian texts. Although it contains only 563 782 words, it comprises eight contemporary literary positions (all written after 1960), all in the core part of InterCorp. This implies that the annotation process has been manually revised. The corpus is accessible through the Kontext interface, which uses the same search engine as NoSketchEngine. The tagsets are identical in both corpora. Therefore, the same queries can be applied to both corpora, and InterCorp can be treated as a complementary source of literary data for our study.

5 QUERIES FOR THE DA$_2$-COMPLEMENT

5.1 Query Design

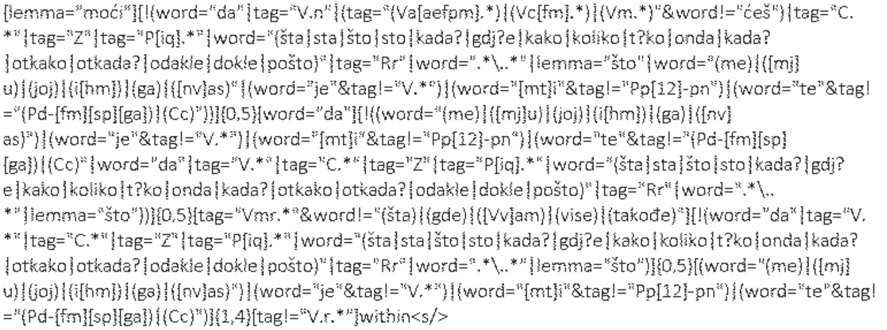
The corpus queries have to take into account four word-order patterns (CL—clitic, CEP—clause embedding predicate, *DA*—*da* complementizer, COMP—embedded finite complement):

1. CL CEP DA COMP
2. CEP CL DA COMP

3. CEP DA CL COMP
4. CEP DA COMP CL

1 & 2 are clear cases of CC, while in 3 & 4 no CC takes place. Each of the four elements in the pattern was defined in the query (see example query for pattern 3 in Figure 1) as a tag, word form or lemma, or a combination of these. This made it possible to exclude many ambiguities, such as the most often inaccurately lemmatized words. Instead of single CLs, we allowed clitic clusters comprising maximally four CLs.

In order to improve our recall, we allowed up to five empty positions between the core elements of the query. Thus, we were able to eliminate from those positions the core elements, as well as those expressions that refer to sentence clause crossing, i.e.: punctuation signs, accidentally attached dots, other verbs, conjunctions, participles and complementizers. We are, of course, aware that some markers can function within a sentence clause e.g. as connectors within a noun phrase (e.g. *Marko i Ana*), but due to a shortage in human resources, we could not afford excessive manual filtering.



```
[lemma="moći"][!(word="da"|tag="V.n"|(tag="(Va[aefpm].*)|(Vc[fm].*)|(Vm.*)"&word!="češ")|tag="C.
*"|tag="Z"|tag="P[iq].*"|word="(šta|sta|što|sto|kada?|gdj?e|kako|koliko|t?ko|onda|kada?
|otkako|otkada?|odakle|dokle|pošto)"|tag="Rr"|word=".*\..*"|lemma="što"|word="(me)|([mj]
u)|(joj)|(i[hm])|(ga)|([nv]as)"|(word="je"&tag!="V.*")|(word="[mt]i"&tag!="Pp[12]-pn")|(word="te"&tag!
="(Pd-[fm][sp][ga])|(Cc)")]{0,5}[word="da"][!((word="(me)|([mj]u)|(joj)|(i[hm])|(ga)|([nv]
as)")|(word="je"&tag!="V.*")|(word="[mt]i"&tag!="Pp[12]-pn")|(word="te"&tag!="(Pd-[fm][sp]
[ga])|(Cc)"|word="da"|tag="V.*"|tag="C.*"|tag="Z"|tag="P[iq].*"|word="(šta|sta|što|sto|kada?|gdj?
e|kako|koliko|t?ko|onda|kada?|otkako|otkada?|odakle|dokle|pošto)"|tag="Rr"|word=".*\..
*"|lemma="što"))]{0,5}[tag="Vmr.*"&word!="(šta)|(gde)|([Vv]am)|(vise)|(takođe)"][!(word="da"|tag="V.
*"|tag="C.*"|tag="Z"|tag="P[iq].*"|word="(šta|sta|što|sto|kada?|gdj?e|kako|koliko|t?ko|onda|kada?
|otkako|otkada?|odakle|dokle|pošto)"|tag="Rr"|word=".*\..*"|lemma="što")]{0,5}[(word="(me)|([mj]
u)|(joj)|(i[hm])|(ga)|([nv]as)")|(word="je"&tag!="V.*")|(word="[mt]i"&tag!="Pp[12]-pn")|(word="te"&tag!
="(Pd-[fm][sp][ga])|(Cc)")]{1,4}[tag!="V.r.*"]within<s/>
```

Fig. 1. Query example for Pattern 3

5.2 Choice of Tested Verbs

It goes without saying that due to imperfect annotation and search limitations, not all instances of the patterns in question could be retrieved. Apart from the well-acknowledged problems of recall, also the precision of searches caused problems.

While the size of InterCorp allowed for one general query per pattern, and filtering out wrong examples manually was possible, in srWaC the final formula brought us results exceeding our human capacities, but not yielding any viable results.

We narrowed the search in srWaC to a selection of non-reflexive CEPs from the list of 42 CEPs retrieved from InterCorp (five of each type plus two additional object-control predicates, as we expected smaller frequencies in this type), which had appeared more than once and which allowed only *da*$_2$-complements. Additionally, we had to take into account differences in frequencies of particular syntactic types of CEPs with the *da*$_2$-complement. Some raising verbs are particularly frequent, more so than for example object-control verbs. Therefore, we did not necessarily choose raising verbs with the highest frequencies, and we selected those belonging to different semantic types (modal and phasal verbs).

Finally, we tried to handle polyfunctionality in relation to syntactic type by eliminating verbs that due to their semantics can belong to different classes or appear with different *da*-complements. As such distinctions had not been annotated, they could not be easily distinguished on the query surface. This is why, for example, the verbs *(na)učiti* 'to learn/to teach', *znati* 'to know', *htjeti* 'to want/will', *morati* 'to must' and *trebati* 'to have to', *dati* 'to give' were excluded from the list of potential candidates.[4]

verb	**frequency of da$_2$ in InterCorp**	**estimated frequency of *da*$_2$ in srWaC**	**relative frequency of CC out of *da*$_2$ in srWac[4]**	**syntactic type**
moći 'can'	69	37526	0.0043	raising
nastaviti 'to continue'	2	1028	0.0029	raising
početi 'to start'	26	6546	0.0100	raising
prestati 'to stop'	5	1203	0.0116	raising
sm(j)eti 'to be allowed'	7	2027	0.0039	raising
nam(j)eravati 'to intend'	3	465	0.0021	subject control
nastojati 'to strive'	7	721	0.0027	subject control
pokušati 'to try'	7	4794	0.0047	subject control
um(j)eti 'to be able to'	8	1209	0.0016	subject control
usp(j)eti 'to succeed'	9	4331	0.0009	subject control

[4] Proportion of the estimated frequency of CC out of da$_2$-complements to the estimated frequency of all da$_2$-complements for the given CEP.

dozvoliti ‘to allow’	7	2528	0	object control
narediti ‘to order’	5	1174	0	object control
nat(j)erati ‘to force’	2	502	0	object control
zamoliti ‘to ask’	3	1584	0	object control
pustiti ‘to let’	3	534	0	object control
primorati ‘to force’	0	248	0	object control
pomoći ‘to help’	0	331	0	object control

Tab. 1. Selected CEPs

The size and precision of results still posed processing problems. As no gold standards have been broadly acknowledged we decided to follow some suggestions by S. Wallis [19], and accordingly the precision of queries was estimated through sampling. We took random samples of 100, which usually should give no more than a 10% margin of error at a confidence level of 95% regardless of the population size. We calculated the binominal probability confidence interval using the Clopper-Pearson exact method. We recalculated raw frequencies into estimated frequencies on the basis of the worst-case scenario of the obtained confidence intervals. These are used in the analyses in the next section.

6 RESULTS AND DISCUSSION

6.1 Constraints on CC from da$_2$-complements into the Matrix Clause in Serbian

Although part of our data is based on the worst-case scenario, our material provides empirical evidence that CC out of *da*$_2$-complements into matrix clauses is indeed possible, although it is most likely a marginal phenomenon. In addition to two examples of CC into raising predicate clauses obtained from the Intercorp subcorpus, our samples yielded 69 correct sentences with CC, from which we estimated a worst-case scenario of 286 CC cases in the whole examined population. The frequencies of CC normalized to the frequency of a *da*$_2$-complement for a particular verb are summarized as part of Table 1 as well as in Figure 2. Analysis of frequencies shows that CC out of *da*$_2$-complements occurs with verbs of varying frequencies. The Chi-square test of dependence between the syntactic type and clitic climbing yields a significant result (p = 7.948e-11).

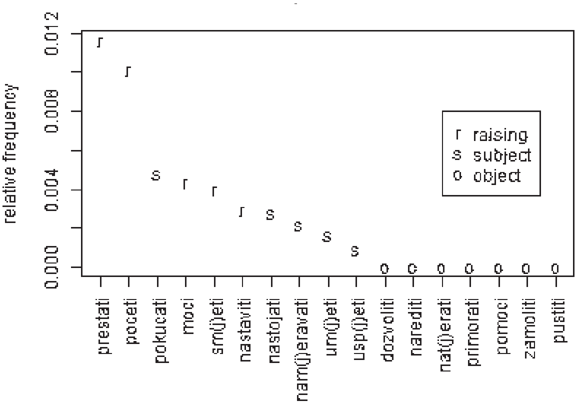


Fig. 2. Relative frequencies of CC for the retrieved CEPs

Figure 2 shows that the two phasal verbs *prestati* and *početi* have the highest relative frequency of CC out of *da*$_2$-complements, followed by *pokušati, moći, sm(j)eti* and *nastaviti*. An interesting finding is that object control CEPs do not seem to allow CC. We did not find a single example for the predicates we selected.

Further, cases in which the CL is placed to the right of the verb of the *da*-complement are extremely rare, albeit possible in all syntactic types. It is also very clear that regardless of the type of CEP, CLs tend to be placed directly after the *da* complementizer, the position which some scholars assumed to be the only possible and correct one (see [1, p. 41], [2, p. 41]).

Furthermore, in the case of CC, CLs tend to be left of the matrix verb, but can appear between the CEP and the *da* complementizer as well. If there are auxiliaries belonging to the CEP, climbed CLs can form clusters with them, as shown as in example (6). These examples disprove Todorović's claim that "if the matrix verb is in the past or future tense, whose auxiliary clitics carry the tense feature, no clitic climbing is allowed out of the subjunctive *da*-complement[5]" [18, p. 166].[6]

(6)	(...) *počeo*$_1$	***im***$_2$	*je*	*da*	*govori*$_2$	*o*
	start.PTCP.SG.M	them.DAT	be.3SG	COMP	speak.3PRS	about
	dolasku	*ove*	*grupe.*			
	arrival.LOC	this.GEN	group.GEN			(srWaC v1.2)

'(...) he began to speak to them about the arrival of this group.'

A reflexive CL can either climb with the pronominal, as in (7), or it can stay in the *da*$_2$-complement, as in (8).

(7)	*U*	*poslednje*	*vreme*	***mi***$_2$	***se***$_2$	*pocelo*$^{2}_{1}$
	in	past.ACC	time.ACC	me.DAT	REFL	start.PTCP.SG.N
	da	*desava*$_2$	*da*	*cujem*$_3$	(...)	
	COMP	happen.3PRS	COMP	hear.1PRS		(srWaC v1.2)

'Recently, it has started happening to me that I hear (...)'

[5] In her terminology, the subjunctive complement refers to *da*$_2$.

[6] As is known, many BCS Internet users do not use diacritics.

(8)	(...) *i*	*počelo*$_1$	***mi***$_2$	*je*	*da*	***se***$_2$
	and	start.PTCP.SG.N	me.DAT	be.3SG	COMP	REFL
	vrti$_2$	*u*	*glavi.*			
	spin.3PRS	in	head.LOC			(srWaC v1.2)

‘(...) and I started to feel dizzy.’

The fact that two CLs that were generated by the same verb do not have to climb together over *da*$_2$ was already observed by S. Stjepanović [17, p. 182]. Her examples, however, concern only two pronominal CLs and not the reflexive *se*. S. Stjepanović [17, p. 182] concludes that in the case of a split only a dative CL can climb, while an accusative CL stays in the *da*$_2$-complement. Additionally, we argue that if two CLs are generated in the *da*$_2$-complement and split, it is the pronominal that climbs, while the reflexive tends to stay in the *da*$_2$-complement. Moreover, it is worth mentioning that reflexive *se* did not climb with a pronominal CL if the matrix clause contained an auxiliary clitic. Since we did not find examples with three CLs (auxiliary, pronominal and reflexive) in a cluster, it seems that whenever there are three CLs in a sentence, the reflexive tends to stay in the *da*$_2$-clause.

Finally, it is worth mentioning that CC was not attested for the form *je* (acc.3sg.f). This needs further investigation, but could be due to error in tagging.

We also investigated embedded finite complements, but we did not observe any link between their semantic or syntactic properties and the inclination of their clitic pronominal complements’ to climb.

6.2 Diaphasic and Diatopic Variation

Regarding diaphasic variation, S. Marković [11] suggests that the phenomenon of CC out of *da*$_2$-complements most typically occurs in the journalistic register, but can also be found in literary texts. Our data confirm both statements. First, the subcorpus of InterCorp, consisting only of Serbian literary texts*,* provides two examples obviously belonging to the literary register. Second, 36 examples were published on Internet sites with predominantly journalistic texts. As regards diatopic variation, S. Marković [11] claimed that CC out of *da*$_2$-complements is typical of language use in Bosnia. In our sample, ijekavian spelling, which is typical of language use in Bosnia, Croatia and Montenegro, occurred in only 8 examples while ekavian spelling was used in the remaining 63 examples (including InterCorp).

7 CONCLUSIONS

In this paper, we addressed the syntactic mechanism of clitic climbing in the context of *da*$_2$-complements, which are characterized by the presence of a verb inflecting for person and number. This is an interesting topic because e.g. for Czech it is claimed that finite complements block CC. The point of departure of our study was the observation that there is a large disagreement as to the acceptability of CC out of *da*-complements. Whereas S. Stjepanović [17] allows the grammaticality of CC out of *da*-complements mainly within a unified formal theory of CC in BCS, other authors reject the grammaticality of this structure outright. We presented the results of a corpus-based study which had to overcome the various shortcomings of the available corpora of Serbian. We proposed solutions to enhance precision and recall by developing sophisticated CQL queries. Our data allow the following answers to our research questions given in Sections 2 and 3:

Q1: Serbian *da*$_2$-complements do marginally allow CC. In these cases, the climbed CL can form a cluster with the auxiliary CL of the matrix verb. We thus in principle agree with [17], but have to point out that we are dealing with a highly marginal construction. Examples did not support the occurrence of CC for all CL forms.

Q2: CC is possible in raising and in subject control contexts. It is, however, most probably blocked in the case of object control. This is in line with what has been claimed for Czech.

We found some further evidence for the following constraints: first, if two CLs are generated in a *da*$_2$-complement and split, it is the pronominal that climbs and the reflexive that stays in the complement; second, reflexive *se* does not climb if there is an auxiliary clitic in the matrix clause. This suggests that the pronominal CL and reflexive *se* behave differently, which leads to the conclusion that CC is not a unified syntactic mechanism. Finally, we were able to reject Todorović's hypothesis [18] that perfect or future auxiliaries block CC.

Since we are dealing with a rare phenomenon, which seems to be restricted not only syntactically, but also stylistically and regionally, the next study should involve native speakers who would judge the acceptability of such examples.

References

[1] Browne, W. (2003). Razlike u redu riječi u zavisnoj rečenici. *Wiener Slawistischer Almanach*, 57: 45–52.

[2] Ćavar, D. and Wilder, C. (1994). “Clitic third” in Croatian. *Linguistics in Potsdam*, 1:25–63.

[3] Čermák, F. and Rosen, A. (2012). The case of InterCorp, a multilingual parallel corpus. *International Journal of Corpus Linguistics*, 17(3):411–427.

[4] Davies, W. and Dubinsky, S. (2004). *The grammar of raising and control: A course in syntactic argumentation*. Blackwell, Malden.

[5] Franks, S. and King, T. H. (2000). *A Handbook of Slavic clitics*. Oxford University Press, Oxford.

[6] Gołąb, Z. (1964). The problem of verbal moods in Slavic languages. *International Journal of Slavic Linguistics and Poetics*, 8:1–36.

[7] Ivić, M. (1970). O upotrebi glagolskih vremena u zavisnoj rečenici: prezent u rečenici s veznikom *da*. *Zbornik za filologiju i lingvistiku*, 13(1):43–54.

[8] Junghanns, U. (2002). Clitic climbing im Tschechischen. *Linguistische Arbeitsberichte*, 80:57–90.

[9] Ljubešić, N. and Klubička, F. (2014). {bs,hr,sr}WaC — Web corpora of Bosnian, Croatian and Serbian. In *Proceedings of the 9th Web as Corpus Workshop* (WaC-9), pages 29–35, Gothenburg, Sweden.

[10] Ljubešić, N., Klubička, F., Agić Ž., and Jazbec I. (2016). New Inflectional Lexicons and Training Corpora for Improved Morphosyntactic Annotation of Croatian and Serbian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 4264–4270, ELRA, Paris, France.

[11] Marković, S. (1955). Položaj zamjeničke enklitike u vezi sa naporednom upotrebom infinitiva i prezenta sa svezicom *da*. *Naš jezik*, 1–2:33–40.

[12] Mišeska-Tomić, O. (2003). The Syntax of the Balkan Slavic future tenses. *Lingua*, 114: 517–549.

[13] Progovac, Lj. (2005). *A Syntax of Serbian: Clausal Architecture*. Slavica Publishers, Bloomington.

[14] Rezac, M. (2005). The syntax of clitic climbing in Czech. In *Clitics and affix combinations. Theoretical perspectives*, pages 103–140, Benjamins, Amsterdam, Nederlands.

[15] Stiebels, B. (2015). Control. In *Syntax – theory and analysis. An international handbook*, pages 412–446, De Gruyter, Berlin, Germany.

[16] Spencer, A. and Luís, A. R. (2012). *Clitics. An Introduction.* Cambridge University Press, Cambridge.

[17] Stjepanović, S. (2004). Clitic Climbing and Restructuring with "Finite Clause" and Infinitive Complements. *Journal of Slavic Linguistics*, 12(1):173–212.

[18] Todorović, N. (2012). *The Indicative and Subjunctive da-complements in Serbian: A Syntactic--Semantic Approach.* PhD Thesis, University of Illinois, Chicago.

[19] Wallis, S. (2014). Coping with imperfect data. Accessible at: `https://corplingstats.wordpress.com/2014/04/10/imperfect-data/`, retrieved 2017-01-31.

ON THE DEVELOPMENT OF AN INTERDISCIPLINARY ANNOTATION AND CLASSIFICATION SYSTEM FOR LANGUAGE VARIETIES – CHALLENGES AND SOLUTIONS

AGNES KIM[1] – LUDWIG M. BREUER[2]
[1] Department of Slavonic studies, University of Vienna, Austria
[2] Department of German studies, University of Vienna, Austria

KIM, Agnes – BREUER, Ludwig M.: On the Development of an Interdisciplinary Annotation and Classification System for Language Varieties – Challenges and Solutions. Journal of Linguistics, 2017, Vol. 68, No 2, pp. 191 – 207.

Abstract: The Special Research Programme (SFB) 'German in Austria: Variation – Contact – Perception' is a project financed by the Austrian Science Fund (FWF F60). Its nine project parts are collaboratively conducting research on the variation and change of the German language in Austria. The SFB explores the use and the subjective perception of the German language in Austria as well as its contact with other languages. Methodologically and theoretically, most SFB project parts are situated within variationist linguistics, others in contact linguistics and perceptionist linguistics. This paper gives an insight into the conception of a framework for the annotation and ultimately also classification of language varieties, which is being developed within the SFB. It outlines the requirements of the various project parts and reviews, whether and how standardised language codes (ISO 639) and language tags (following BCP 47) can be utilised for the annotation of language varieties in variationist linguistic projects.

Keywords: language varieties, dialects, language tags

1 INTRODUCTION AND CONTENT

The Special Research Programme (SFB) "German in Austria: Variation – Contact – Perception", funded by the Austrian Science Fund (FWF) is an interdisciplinary collaborative project, which conducts research on the variation and change of the German language in Austria. It consists of the three thematic pillars represented in its title, and thus explores the entire spectrum of language variation in German in Austria, the perception of German in Austria, and contact of German in Austria with other languages.

Six of the nine project parts are located at different departments of German linguistics at the Universities of Vienna, Salzburg, and Graz, as well as at the Austrian Academy of Sciences. Of the remaining, two project parts—those focusing on aspects of language contact—are situated at the Department of Slavonic studies at the University of Vienna. Additionally, the Centre of Translations Studies at the University of Vienna hosts the project part responsible for developing and implementing the Collaborative Online Research Platform "German in Austria". This platform will support the working process in the whole research cycle ranging from data querying, input, annotation and analysis to interactive online tools, which allow accessing the data in multiple ways. Furthermore, it will guarantee sustainable preservation of research both data and outcomes.

DE GRUYTER OPEN
DOI 10.1515/jazcas-2017-0029

Hence, the SFB aims at using and (if needed) enhancing existing (and standardised) annotation systems. In some cases (e.g., the tagging of specific syntactic phenomena), a completely new annotation scheme is necessary. Considering all possible annotation levels, the interdisciplinary orientation and the various (theoretical and empirical) approaches of the different project parts, a multidimensional and highly flexible annotation system is crucial to reconcile all these demands. Therefore, the SFB builds on a highly flexible annotation syntax with an underlying, equally strict description scheme.

In this paper, we focus on an annotation and classification framework of language varieties, which is supposed to serve as a basic description level for the language data gathered. Nonetheless, it should be kept in mind that it is a part of a larger annotation scheme, which also describes specific parts of the object language (e.g. part-of-speech-tagging).

In section 2 of this paper, we discuss several aspects of languages and their varieties that have to be considered in creating a custom-made annotation and classification framework of language varieties within the SFB. We discuss the different theoretical and methodological approaches of several project parts in order to outline their requirements of such a framework.

Section 3 discusses language codes and language tags with regard to their standardisation. We evaluate the benefits and difficulties of applying these standards within the SFB. Finally, section 4 proposes an according solution.

This paper gives a glimpse into the considerations of the working group responsible for designing and implementing a variety annotation and classification framework within the SFB.

2 REQUIREMENTS WITHIN THE SFB[1]

2.1 Task Cluster B: Variation

The three project parts within Task Cluster B focus on the dynamics of varieties of German in Austria in their linguistic and social structures. Methodologically and theoretically, they are situated within variationist linguistics. In order to answer their research questions, they collect language data of a large number of informants from all over Austria, from rural as well as urban areas. The elicited corpus will ultimately not only cover dialects from all over Austria but also other colloquial registers between (intended) dialects and the (intended) standard language (for the terminology see [1]). In addition to the data collected by the project parts themselves, comparative language

[1] A large-scale project like the SFB combines many aspects, which may be interesting for different kinds of linguistic projects. Thus, we would preferably describe the whole SFB in detail, considering theoretical and empirical approaches of its nine different project parts as well as the Collaborative Research Platform, which combines the project-internally orientated working infrastructure as well as – externally-orientated – means of dissemination and even elements of citizen science. However, this would go beyond the scope of this paper. Considering the main topic of this paper – the annotation and classification of languages and their varieties –, we focus on Task Cluster C. Since this Task Cluster does not only investigate German varieties, but conducts research into the contact of German in Austria with Slavic languages, too, it requires the whole scale of the presented annotation system. For closer information on the SFB in general, its goals and structure please refer to its homepage: `http://www.dioe.at/en`.

data such as linguistic descriptions are considered in order to trace the development of dialects in Austria over the course of the 20th century until the present. Among others, one outcome of this Task Cluster will be an online “speaking linguistic atlas”, in which audio samples are provided within an interactive geographic information system. Within the whole research platform, all data will be linked and interconnected to other data. In terms of (automatic) linking or filtering of these data, a standardised metadata set including normalised variety classifications is necessary [2].

2.2 Task Cluster C: Contact

Task Cluster C is concerned with the contact of German in Austria with other, particularly Slavic languages. It is orientated towards contact linguistics and research into multilingualism and links German with Slavic linguistics in an interdisciplinary fashion. In the first four years, both project parts within that Task Cluster employ a diachronic approach. Therefore, they deal with data types that are clearly distinct from the data elicited by the synchronically orientated project parts [3].

Project Part 5 analyses the context of language policies in several fields of action and tension, i.e., in administration, law, and especially in education. For this purpose, existing data from legislative texts, newspapers, archive materials, and other contemporary documents are being collected, connected, and analysed by methods from the critical discourse analysis. The Project Part’s central aim is to reconstruct the functional and metalinguistic dimensions of German in the multilingual Habsburg state and relate them to the conditions of language (and multilingualism) policies and planning in the Second Republic of Austria [4].

On the other hand, Project Part 6 focuses on linguistic contact phenomena, e.g. on all linguistic levels of German in Austria that have been explained by language contact with the Slavic languages. Its main goal is to give a comprehensive overview of alleged contact phenomena, provide a basic assessment of these and thus identify language myths associated with the contact with Slavic languages. The agglomeration of Vienna and its urbanolect will be of special interest. The Project Part particularly focuses on the exhaustive number of Slavic loanwords in German in Austria. For this purpose, linguistic and popular literature on the language contact phenomena in German in Austria [5] will be collected and processed.

Ultimately, Task Cluster C aims at establishing an *Information System on (historical) Multilingualism in Austria* (MiÖ) within the Collaborative Research Platform. This module will link and present quantitative data such as historical census data to qualitative data collected within the two project parts. It will further include a bibliography. The database shall make historical – and in a later stage – also present multilingualism in Austria and its linguistic, societal and historical conditions visible.

In order to ensure public searchability, we need to model and map historical and contemporary names and labels for languages and their varieties. These names’ political and contextual restrictions, connotations, as well as their change over time have to be considered in the model. Some of these names have already been well described, such as the German *Tschechisch* ‘Czech’ and *Böhmisch* ‘Bohemian’, their relation and development [6].

2.3 Task Cluster D: Perception

Task Cluster D deals with language attitudes and language perception with special regard to German in Austria. Project Part 10, for example, investigates language attitudes and perception within schools, i.e., of pupils and teachers. Of course, in that context not only the so called internal multilingualism, i.e., competence in both dialect and standard language, but also external multilingualism has to be considered.

Project Part 8, on the other hand, compares how standard varieties as well as other registers of German are conceptualised by adults living all over Austria. Therefore, laymen's names for varieties and registers play a prominent role. Again, given the need to classify the gathered qualitative data, the described annotation framework is crucial for this Task Cluster, too. In the semi-standardised interviews, laymen are, e.g., asked how they would call the varieties they use. To ensure comparable quantitative analyses of the variety names expressed, a standardised categorisation is vital [7].

2.4 Task Cluster E: Collaborative Online Research Platform

Task Cluster E develops and implements the Collaborative Online Research Platform of the whole SFB. This is supposed to be the main communication and research hub, as well as the platform for the dissemination of data and results. Thus, the platform does not only play a role amongst and within the various Project Parts, but also connects and presents the SFB and its results to the outside world.

All implemented tools shall provide means of machine-to-machine communication and thus interfaces to share the data with other tools (from other projects etc.). Therefore, Task Cluster E aims at a high interoperability of annotation schemes, corpora and the annotated data itself. Apart from this, the emphasis on addressing the non-academic public as well makes a lucidly comprehensible and explicitly described annotation framework indispensable [8].

2.5 Summary: Summary and Outline of a Model

As shown above, the project parts focus on various aspects of varieties of German in Austria and their variation. Therefore, they take differing angles of view onto them: Task Cluster B mainly considers varieties as its objects of interest, i.e., as its object languages. Task Cluster D, on the other hand, focuses on names of varieties, i.e., *glottonyms* (or: *glossonyms*) and the concepts that speakers connect to them. Within Task Cluster C, both dimensions are relevant depending on the project part. In addition, due to the focus on language contact these projects require a model of genetic language and variety affiliation to ensure the searchability of the data during the working process and on the Research Platform. As Task Cluster E does not focus on the object level, but rather on the more technical and standardisation level it will be neglected in the following discussion.

Resulting from the requirements named above, we propose three modules within our variety annotation and classification system (see Fig. 1).

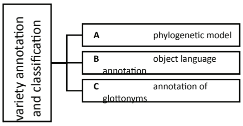

variety annotation and classification
A phylogenetic model
B object language annotation
C annotation of glottonyms

Fig. 1. Outline of a system for the annotation and classification of varieties

The modules have different functions within the working process and therefore differing statuses within the system: Module A provides a model for the phylogenetic affiliation of languages and their varieties. It serves as an auxiliary construction to ensure the searchability of the data. We therefore prefer a pragmatically orientated model within module A and consciously accept the simplifications that will have to be made in order to be able to model various language families and groups. Besides that, module A will be designed and provided as a closed model by the responsible working group. Of course, changes may be requested, but generally the module should not be changed on a regular basis during the working process.

Module B and C, on the other hand, are regarded to be working instruments, which can and should be adapted by the project parts according to their needs. The responsible working group provides the technical and content framework, as well as documentation and adaption guidelines. These modules serve as a comprehensive annotation framework for varieties and their names within the SFB. Possibly, we shall be able to develop classification systems for varieties of German in Austria, as well as for glottonyms based on the detailed analysis provided by several project parts at the end of the SFB.

Once more we will focus on the proposed modules, their content and technical modelling in section 4. However, we first will revise international systems and standards for language coding and language tagging. They are relevant, because one explicit goal of Task Cluster E in building the Online Research Platform is to develop best practise examples for handling variationist linguistic data. Especially this research area has recently led to a big amount of data. Consequently, the need for presenting these data online and connecting data from different sources has risen (cf. [9], [10]). Therefore, the development of best practise examples and new standardised annotation systems is vital for various variationist linguistic projects. Mutual interests lie in the connection of the different datasets gathered in different regions (*horizontal variation*), situations (*vertical variation*) and periods (*diachronic variation*). This requires a flexible, multidimensional, and thus complex but, given the large data sets, easy to use annotation system.

3 LANGUAGE CODES AND LANGUAGE TAGS

When it comes to identifiers of languages or language varieties, language codes need to be distinguished from language tags, even though the latter often refer and make use of the first. *Language codes* are alphabetical, numerical or alphanumerical identifiers, which uniquely refer to a certain language or language variety. The entities that the codes refer to are seen as rather well-defined and comparable according to the underlying language definition.

Language tags, on the other hand, allow for specifying deviations from default values of a given language in a certain text written or spoken in the according language. Therefore, they account for certain degrees of variation in language and are used like annotations rather than identifiers.

Below, we assess whether and how the most common standards and/or systems of language codes and tags may be used for linguistic projects in general and the SFB in special. In order to accomplish that task, we exemplarily compare whether and how varieties of German in Austria as well as the genetic affiliation of Czech could be identified and modelled by utilising the according code sets. As shown above, both perspectives are needed for our system.

When it comes to language codes, we focus on the ISO 639 standards, because only language tags, which make use of them, can be used in XML annotations, such as those following the TEI standards (after the `xml:lang` attribute) [11]. Therefore, we leave other coding systems such as the *Glottocodes* [12], [13] and the *Linguasphere codes* [14] aside.

3.1 Language Codes: The ISO 639 Standard Family

As of 2017, the ISO 639 standard family comprises five sub-standards (see Tab. 1). Four of them define alpha-2 or alpha-3 codes for "the representation of names of languages". Part 4 sets the principles of coding and provides application guidelines. The ISO 639 standard family is under the responsibility of ISO/TC 37 (ISO Technical Committee 37), which generally facilitates the standardization "of principles, methods and applications relating to terminology and other language and content resources in the contexts of multilingual communication and cultural diversity" [15]. The various ISO 639 sub-standards are rooted in quite divergent disciplines and projects, as we show below.

		FIRST RELEASE	**VALID VERSION**	**NUMBER OF SINGLE CODES**[2]
ISO 639-1	*Alpha-2 code*	1967	2002	204
ISO 639-2	*Alpha-3 code*	1998	1998[3]	506
ISO 639-3	*Alpha-3 code for comprehensive coverage of languages*	2007	2007[4]	7459
ISO 639-4	*General principles of coding of the representation of names of languages and related entities, and application guidelines*	2010	2010	
ISO 639-5	*Alpha-3 code for language families and groups*	2008	2008	115[5]

Tab. 1. ISO 639 standard family

In 2009, the ISO published a proposal for ISO 639-6, an 'Alpha-4 code for comprehensive coverage of language variants'. This standard was withdrawn in November 2014 and is not available anymore [16]. According to various sources [17], [18], it was to be based on the *Linguasphere Register of the World's Languages and Speech Communitie*s [14].

[2] The given numbers represent the authors' count based on code of lists that were retrieved from the respective registration authorities' websites in March 2017: [20] (ISO 639-1 and ISO 639-2), [21] (ISO 639-3) and [22] (ISO 639-5).

[3] According to [20], the code list itself has been updated on 18 March 2014 for the last time. The last change is dated to 21 Nov 2012.

[4] The ISO 639-3 codes have had their latest update on 17 Feb 2017 [21], i.e., four days before the 20th edition of the *Ethnologue* was published on 21 Feb 2017.

[5] The last update of an ISO 639-5 element took place on 2 Nov 2013 [22].

Interestingly, in 2016 the same subcommittee that bears responsibility for ISO 639 (ISO/TC 37/SC 2 – *Terminographical and lexicographical working methods*) initiated a new project for the standardisation of the "Identification and description of language varieties" (ISO/AWI 21636) [19]. Unfortunately, there is no further information on this project publically available.

3.1.1 ISO 639-1 and ISO 639-2

Kamusella [17] provides an embedding of the emergence of the ISO 639-1 and ISO 639-2 standards into socio-cultural developments of the 20th century [17, p. 62ff.]. Generally, he associates the processes of standardisation or uniformisation with "modernity", i.e., "the [international] spread of various technologies and cultural practices" [17, p. 59], which also require shared terminologies.

The ISO 639-1 list of alpha-2 codes was compiled for a primary use in terminology, too. It is maintained by the *International Centre for Terminology* (Infoterm) in Austria[6] and includes identifiers for "the most developed languages of the world, having specialized vocabulary and terminology" [23]. Therefore, languages need to fulfil a list of detailed criteria in order to be assigned an ISO 639-1 code.

ISO 639-2, on the other hand, is primarily rooted in bibliography: As the alpha-2 codes proved insufficient to identify a large number of publication languages, the ISO developed an alpha-3 code set based on the *MARC Code List for Languages*, a standard created by the US Library of Congress [24]. This institution was also made the registration authority for the ISO 639-2 standard.

Both the ISO 639-1 and the ISO 639-2 substandard require languages to meet detailed criteria in order to be assigned an own code (see Table 2[7]). As ISO 639-1 is seen as a subset of ISO 639-2, a language needs to fulfil both the criteria for ISO 639-2 and the more specific ones for ISO 639-1 in order to be assigned an alpha-2 code. Table 2 can be read this way, as we first present the requirements for ISO 639-2 and only then the ones for ISO 639-1. Generally, a single language code is provided for languages which are written in multiple orthographies and scripts. Dialects should be represented by the "same language code as that used for the language". According to [25], the difference between a dialect and language is to be decided "on a case-by-case basis".[8][9]

[6] `http://infoterm.info`, retrieved 2017-03-30.

[7] The information presented in Table 2 were retrieved from [25]. Kamusella [17] presents the related list, too. In comparison to the list, we slightly regroup the information in order to enlarge comparability.

[8] Documents such as "specialized texts, such as college or university textbooks, technical documentation manuals, specialized journals, subject-field related books, etc." [25].

[9] "E.g. technical dictionaries, specialized glossaries, vocabularies, etc. in printed or electronic form" [25].

	ISO 639-2	ISO 639-1
DOCUMENTATION	● one agency holds **50 different documents** (not limited to text) in the language or ● five agencies hold a total of 50 different documents in the language	● a significant body of existing documents[8] **written** in **specialized languages** ● a number of existing **terminologies** in various subject fields[9]
RECOMMENDATION		● recommendation of a specialized authority[10] ● support by one or more official bodies
NUMBER OF SPEAKERS		is considered
STATUS		recognized in one or more countries

Tab. 2. Requirements for ISO 639-1 and ISO 639-2

As can be seen from these requirements, both code sets have in common that the underlying language definition is a sociological one. ISO 639-1 basically provides codes for standard languages like `de` for *German* or `cs` for *Czech*. Neither aspects of language variation nor of their affiliation can be covered.

On the other hand, ISO 639-2 roughly lists what could be called *Ausbau* languages according to Kloss [26]. Therefore, in addition to `ger/deu`[11] for *(Standard) German*, there is an own code `gsw` for *Swiss German, Alemannic, Alsatic* (in German only: *Schweitzerdeutsch*).

In contrast to ISO 639-1, ISO 639-2 also provides the possibility to assign a collective alpha-3 code, if the requirements concerning the documentation of a language is not fulfilled [25]. Such collective codes thus identify groups of languages that could be used to model the genetic affiliation of languages, e.g., `ine` for *Indo-European languages* or `sla` for *Slavic languages*. However, the all-together 55 collective codes[12] are of course not sufficient to model linguistic affiliation across several Central European languages as will be required within the SFB.

Next to the individual language level and the language group level, there also exists a diachronic level in the ISO 639-2 code set: Diachronic varieties such as `gmh` *Middle High German* or `goh` *Old High German* can be identified by these alpha-2 codes. These multiple layers and the fact, that they are not clearly distinguished from each other clearly, points to the roots of the ISO 639-2 codes in bibliography.

3.1.2 ISO 639-3

What if an individual language did not meet the criteria for ISO 639-1 and ISO 639-2 and somebody still wanted it to be registered in the ISO 639 standard family? These

[10] Such as "a standards organization, governmental body, linguistic institution, or cultural organization" [25].

[11] 21 languages have alternative codes either for bibliographic use or for use in terminology. In these cases, the bibliographic code is listed first [20].

[12] These numbers are provided by Kamusella [17], who counted 484 codes within ISO 639-2 in 2011. In 2017, we counted 506 items, which means that the number of collective codes might be slightly higher, too.

languages may be "candidates for inclusion in ISO 639-3", as the Library of Congress suggests [27].

The history of the third part of ISO 639 is thoroughly and critically analysed by Kamusella [17]. Generally, it is closely associated with the Summer Institute of Linguistics, now: SIL International, a missionary linguistic organization, which is quite well known for its main publication, the *Ethnologue* [28]. It aims at giving a comprehensive overview of all languages spoken worldwide. Basically, the ISO adapted the identification codes for individual languages that were introduced in the 10th edition of the *Ethnologue* in 1984 [28] as the third part of ISO 639.

In his presentations [30], [31], Gary Simons, currently Chief Research Officer at SIL International[13], frequently cites Einar Haugen [32], who distinguishes a structural and a functional view with regard to the distinction of languages from dialects. The structural view describes "the language itself", whereas the functional view focuses on "its social uses in communication". Similar distinctions can be found in other seminar works of early sociolinguistics, such as Kloss [26], who distinguishes a sociological and a philological view as early as 1929 [33].

Simons associates the structural use of the terms *language* and *dialect* with the one "most commonly held by linguists" [30], thereby legitimating the approach supported by SIL International. Furthermore, he states that, following the premises of variationist linguistics, that "languages are not static objects", a language identifier in ISO 639-3 "denotes some range of language varieties" [34]. The main criterion for the distinction of different languages is their intelligibility (comp. Klosses *Abstand* languages [26]); the ethnolinguistic identity of a group of speakers is only considered in the second place [34].

This leads to a "Bible translation-based overcounting of languages imposed from outside", as Kamusella puts it [17, p. 76]. He assumes that SIL would count up to 40 different languages within the area in Central Europe, where German is spoken [17]. This estimation is quite realistic, if we consider that the *Ethnologue* [28] lists five Germanic languages within Austria (see below).

For modelling language variation of German in Austria, ISO 639-3 provides the code `bar` for *Bavarian* but no equivalent code for Alemannic varieties, which are spoken in Vorarlberg. The code `gsw`, which identifies "*Swiss German, Alemannic, Alsatic*" in ISO 639-2, does only refer to "*Swiss German*" in ISO 639-3. In 2011, a request[14] was made to register a code, preferably `aeg`, for "*Alemannic*". The code was supposed to have macrolanguage status and cover the individual languages `gct` *Colonia Tovar*, `gsw` *Swiss German*, `swg` *Swabian*, and `wae` *Walser*, which were already registered in ISO 639-3. The change request was rejected, because *Alemannic* would not meet the requirements for macrolanguages, as the individual languages listed above would not collectively be referred to as *Alemannic* in any contexts [35].

[13] `https://www.sil.org/biography/gary-f-simons`, retrieved 2017-03-16.

[14] The code request was made by Clemens-Valentin Kientzle, who seems to have been a student at the University of Freiburg, Switzerland at that time and had a leading position in the development of an Alemannic Wikipedia in 2011 (`http://www.freiburger-nachrichten.ch/kanton/sie-schreiben-wie-ihnen-der-schnabel-gewachsen-ist-aus-freude-am-dialekt`, retrieved 2017-03-29). The latter fact may have been a motiviation for the code request.

The existence of a code for Bavarian with the status of an individual language, a status, which could also be questioned, but no equivalent one for Alemannic of course makes it impossible to model, at least the dialectal groups of German in Austria.

As the ISO 639-3 standards and the *Ethnologue* [28] are closely related, it is interesting to take a look at its latest edition. The *Ethnologue* lists several varieties of German in Austria, which would be treated as such in a framework of German variationist linguistics, as individual languages (see Table 3)[15]. Interestingly, the *Ethnologue* uses `gsw` in order to refer to *Alemannic* in general.

It is obvious that from a variationist linguistic point of view, this list and its mapping to certain regions is simplistic, incoherent and based on questionable facts. The underlying *Abstand* paradigm implies distinct languages where a variationist perspective would be more appropriate. Thereby, it renders phenomena of vertical variation within the standard-dialect spectrum invisible.

ISO 639-3 does not assign any collective codes. Therefore, it is not possible to model linguistic affiliation with this code set.

code	***name***	***region***
`gsw`	Alemannic	Vorarlberg
`bar`	Bavarian	Lower Austria, Salzburg, Burgenland, Carinthia, Styria
`deu`	Standard German	Vorarlberg[16]
`swg`	Swabian	Tyrol, around the town of Ruette
`wae`	Walser	Tyrol, Paznauntal area

Tab. 3. Varieties of German in Austria according to the *Ethnologue*

3.1.3 ISO 639-5

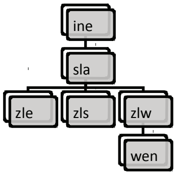
In 2009 the ISO published a fifth part of the ISO 639 standard family. It provides 'codes for language families and groups', some of which were already included in ISO 639-2. According to the Library of Congress, which maintains ISO 639-5 as well, these codes are intended to "support the overall language coding" and do not "provide a scientific classification of the languages of the world" [36].

For modelling the linguistic affiliation of Slavic languages in general, ISO 639-5 currently provides the codes listed in Table 4. As can be seen, for a basic model of linguistic affiliation within the Slavic languages the codes can be used quite accurately. However, if there exists a code for the Sorbian languages `wen`, it would be favourable to also have codes for the Czech-Slovak languages, Lechitic languages, and so forth.

[15] See `https://www.ethnologue.com/country/AT/languages`, retrieved 2017-03-15.

[16] On the relevant map, Standard German is linked to the cities of Vienna, Graz and Linz. On the other hand, it does not assign Standard German to Vorarlberg (see `https://www.ethnologue.com/map/AT`, retrieved 2017-03-15). Thus, the *Ethnologue* even contradicts itself.

[17] The codes that are novel in ISO 639-5 and had not already been part of ISO 639-2 are marked with an asterisk (*).

code[17]	name	ISO 639:5 hierarchy
`ine`	Indo-European languages	
`sla`	Slavic languages	
*`zle`**	East Slavic languages	
*`zls`**	South Slavic languages	
*`zlw`**	West Slavic languages	
`wen`	Sorbian languages	

Tab. 4. Modelling the Slavic languages with ISO 639-5 codes

Generally, we conclude that static lists of language codes, no matter whether designed in library contexts or linguistic enterprises, do hardly account for aspects of language variation. Still, if language data shall be annotated according to machine-readable standards (such as XML) in order to be processed by several applications, ISO 639 codes or language tags according to BCP 47 [37] have to be used.

3.2 Language Tags According to BCP 47

As already emphasised above, in comparison to language codes, language tags allow for annotating certain degrees of variation. In their context language variations can best be described as deviations from default settings. Language tags that can be used in XML annotations, e.g., following the `xml:lang` attribute, have to be designed according to BCP (Best Current Practise) 47 [37], a document issued by the IETF (Internet Engineering Task Force). This organisation's ambitious mission is to "make the Internet work better by producing high quality, relevant technical documents that influence the way people design, use, and manage the Internet" [38]. In contrast to the ISO, IETF relies on free community participation and is organised by the non-profit ISOC (Internet Society).

BCP 47 was issued in September 2009. According to this document, a language tag has the following structure, in which the individual subtags (e.g., `language` or `script`) need to be used in the given order.

```
language-extlang-script-region-variant-extension-privateuse
```

Except for the `language` subtag, the positions do not need to be specified and can be left empty. Some values even have to be suppressed with a certain `language` subtag, e.g., the script must not be specified, if a German text is written in Latin (see Fig. 2). If, on the other hand, it was written in Cyrillic, the script would have to be specified (`de-cyrl`). The W3-Consortium also advises to "keep the tag as short as possible" and thus encourages to leave out redundant subtags [39].

The individual subtags may only have certain values. Valid subtags are registered in the *IANA language subtag registry*[18]; the registration process for new subtags is described in BCP 47. Some subtag values are generally associated with ISO standards (see Table 5).

[18] `http://www.iana.org/assignments/language-subtag-registry/language-subtag-registry`, retrieved 2017-03-21, see Fig. 2 for an example. A subtag search tool is provided on `https://r12a.github.io/app-subtags/`, retrieved 2017-03-21.

```
%%
Type: language
Subtag: de
Description: German
Added: 2005-10-16
Suppress-Script: Latn
%%
```

Fig. 2. Entry for the language subtag “German” in the IANA *language subtag registry*

subtag	ISO standards
language	ISO 639 *Codes for the representation of names of languages*, preferably ISO 639:1
extlang[19]	ISO 639 *Codes for the representation of names of languages*, especially ISO 639:3
script	ISO 15924 *Codes for the representation of names of scripts*
region	ISO 3166 *Country codes*

Tab. 5. ISO-Standards in language tags according to BCP 47

The `variant` subtag may only carry values registered in the IANA *language subtag registry*. Each of these values is tied to a specific language and can therefore only be used in combination with a certain primary `language` subtag. There are two values for the `variant` subtag registered, which can be combined with the `language` subtag `de` (German, see Fig 3.) and none for `cs` (Czech).

```
%%
Type: variant
Subtag: 1901
Description: Traditional German orthography
Added: 2005-10-16
Prefix: de
%%
Type: variant
Subtag: 1996
Description: German orthography of 1996
Added: 2005-10-16
Prefix: de
%%
```

Fig. 3. Variant subtags for German in the IANA *language subtag registry*

If more specifications are needed, there are two possibilities: `extension` subtags are singletons that can be registered with IANA by organisations. Following these singletons, the according organisations themselves may define more subtags and their values.

[19] Extended language subtags, i.e., extlang subtags, are used to identify languages that are closely linked or seen as a variant of another language due to some reasons. Some variants of pluricentric languages such as Arabic can be described in that way, if there are ISO 639-3 codes for their single variants. A language tag consisting of a language subtag and an extlang subtag for Gulf Arabic would thus be ar-afb. But, it could also and should be referred to with a primary language subtag only (afb) [39].

`Private-use` sequences work similarly. They always begin with the singleton `-x-`, which is followed by subtags that are privately agreed on within a certain community. The W3-Consortium advises to use them "with great care", as they are "only meaningful within private agreements and cannot be used interoperability across the Web" [39]. Unfortunately, they seem to be the only solution for scientific projects with a variationist linguistic focus such as the SFB "German in Austria", because variation in language cannot be sufficiently annotated in XML-documents with the basic language tag syntax.

Still, we propose a language tag consisting of a primary language subtag and a region subtag to generically refer to the object languages that the SFB "German in Austria" is interested in, i.e., the whole spectrum of varieties of German used in Austria. This tag will serve as the basis for further specifications as long as our system has the status of a working annotation.

`de-AT`

In the long run, we could, of course, register an `extension` subtag, but such a system should be agreed upon as a community standard within at least German variationist linguistics and needs to be well designed and pretested.

4 SOLUTIONS FOR THE SFB: A SYSTEM IN PROGRESS

4.1 Module A: Modelling the Genetic Affiliation of Languages

As mentioned above, we consider module A as an auxiliary construction and therefore prefer pragmatic solutions and accept simplifications. Hence, we transfer a phylogenetic tree model into a relational database (see Table 6 and 7). Thereby, we make groups and categorisation levels more explicit than in the graphic representation of the tree model, in which the generations of different branches can only be 'seen' implicitly. Concerning the content, we first of all need to agree on a harmonisation of different tree models for several languages used in Central Europe. Secondly, we will have to define, what the name of each language family or group designates in an underlying ontology.

Currently, we have agreed on using and adapting the *Composite model* for the Indo-European languages developed by the MultiTree project [40] for the levels above individual languages, i.e., for language families and groups. Within module A, names for individual languages such as "German" or "Czech" do not refer to codified standard languages but to variety bundles, which are commonly addressed (and/or constructed) as "German" or "Czech". We also model levels of dialectal groups, such as Upper German with its subnodes Bavarian and Alemannic, because these levels are not considered research objects within the SFB.

Tables 6 and 7 exemplarily show, how the linguistic affiliation of Czech and Slovak would be modelled. In the *belongs_to* column, Table 6 refers to its own *ID* column; in the *type* column, it refers to Table 7. Note that type 3 is not assigned to any variety in Table 6. That level is needed to model the affiliation of German based on a simplified model adapted from MultiTree [40], which is depicted in Fig. 4. This figure also represents the type of visualisation underlying Tables 6 and 7, with the greyish bars corresponding to the variety types in Table 7 and the single boxes to the varieties in Table 6.

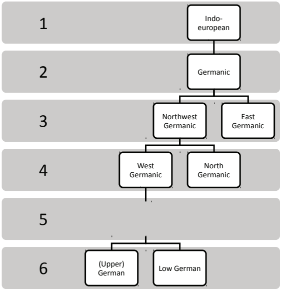


Fig. 4. Tree model for the Germanic languages with focus on German

ID	*variety_name*	*type*	*belongs_to*
1	Indo-European	1	
2	Slavic	2	1
3	East Slavic	4	2
4	West Slavic	4	2
5	South Slavic	4	2
6	Lechitic languages	5	4
7	Sorbian languages	5	4
8	Czech-Slovak languages	5	4
9	Czech	6	8
10	Slovak	6	8

Tab. 6. Variety table from module A for the Slavic languages with focus on Czech and Slovak

ID	**type_name**
1	language family
2	language group
3	subgroup 1
4	subgroup 2
5	subgroup 3
6	individual languages

Tab. 7. Variety type table from module A

4.2 Module B: Object Language Annotation System

The objective of module B is to provide an annotation framework for several dimensions of language variation of German in Austria. It shall enable corpus linguistic analyses, but should not impose a pre-defined classification upon the data. Table 8 shows the dimensions that it will have to account for. Furthermore, we indicate, which factors might specify these dimensions in the corpus of data collected within the SFB, as well as in other, external linguistic sources such as linguistic literature or other language resources.

Dimension	**factors within the corpus**	**factors in other linguistic sources**
***vertical variation** on the standard-dialect axis*	intended register	according classification
	code switch or style shift	
horizontal** (diatopic) **variation	place of recording	place of reference
***diachronic** variation*	time of recording	time of reference
***idiolectal** dimension of variation*	informant	author

Tab. 8. Dimensions of language variation to be considered in module B.

These factors could be transferred into a `private-use` language tag sequence that would have to be specified as belonging to module B by the singleton `-b-`.

```
de-AT-x-b-place-time-intend_register-register_shift-person
```

The `place` and `person` subtags will take their values from the place and person specific parts of the SFB database. Especially the values for subtags, which carry information on vertical variation, will be defined during the working process. The developing working group provides the technical framework, as well as the guidelines for adding and documenting language subtags and their values.

4.3 Module C: Annotation System for the Names of Varieties

The names of varieties and the connotative and/or evaluative meaning they develop depending on their *context*, their *reference* and the *kind of reference*, will be important for some SFB project parts, too. We understand *context* as the kind of text the glottonym appears in. Whether it is an interview conducted by the SFB or a 19th century legal text will clearly make a difference in its meaning. *Reference* is the language or variety, the glottonym refers to. It may specify this language or variety by attaching it to a certain place (e.g., *Viennese*), a certain person (i.e., an idiolect), or by embedding it in time. Furthermore, a glottonym may carry information on the register, which the reference language belongs to. The *kind of reference* expresses, whether the glottonym refers to the person using it and his/her way of speaking (*self-classification*) or whether he/she uses it to describe somebody else's speech (*hetero-classification*).

In XML, a relevant language tag would not follow the `xml:lang` attribute, because this attribute may only specify the object language, i.e., the language a source is written or spoken in. On the other hand, it would follow a `lang` attribute, which allows for the specification of language names, e.g., according to the TEI P5-guidelines [11]. Such a language tag would have the form:

```
language-x-c-place-time-register-person-context-reference_kind
```

The `language`, `place`, `time`, `register` and `person` subtag annotate the reference. Again, the singleton `-c-` indicates the module, in the context of which the tag needs to be interpreted.

5 CONCLUSIONS

This article has provided a glimpse into the development of a custom-made annotation framework for language varieties, which evolves from and shall be used within a collaborative linguistic project. It will serve as a working annotation and ultimately also enable querying the corpus of German in Austria, which is compiled by the eponymous SFB. On the long run, it shall also enable classification of varieties of German in Austria and provide a best practise example that might initiate the definition of community standards. Potentially, this best practise example can be extended to a new standard in terms of variationist linguistic variety annotation, if accepted and adopted by the community.

References

[1] Glauninger, M. M. (2012). Zur Metasozioseminose des ›Wienerischen. Aspekte einer funktionalen Sprachvariationstheorie. *Zeitschrift für Literaturwissenschaft und Linguistik*, 42(2):110–118.

[2] DiÖ (2017). Task-Cluster B: Variation. *DiÖ-Online*. Accessible at: `https://dioe.at/en/projects/task-cluster-b-variation/`, retrieved 2017-03-29.

[3] DiÖ (2017). Task Cluster C: Contact. *DiÖ-Online*. Accessible at: `https://dioe.at/en/projects/task-cluster-c-contact/`, retrieved 2017-03-29.

[4] DiÖ (2017). PP05: German in the context of the other languages of the Habsburg state (19th century) and the Second Austrian Republic. In *DiÖ-Online*. Accessible at: `https://dioe.at/en/projects/task-cluster-c-contact/pp05/`, retrieved2017-03-29.

[5] DiÖ (2017). PP06: German and the Slavic languages in Austria: Aspects of language contact. In *DiÖ-Online*. Accessible at: `https://dioe.at/en/projects/task-cluster-c-contact/pp06/`, retrieved 2017-03-29.

[6] Berger, Tilman (2007). Böhmisch oder Tschechisch? Der Streit über die adäquate Benennung der Landessprache der böhmischen Länder zu Anfang des 20. Jahrhunderts. In Nekula, M., Fleischman, I., and Greule, A., editors, Franz Kafka im sprachnationalen Kontext seiner Zeit. Sprache und nationale Identität in öffentlichen Institutionen der böhmischen Länder, pages 167–182, Böhlau, Köln – Weimar – Wien, Germany.

[7] DiÖ (2017): Task-Cluster D: Perception. In: *DiÖ-Online*. Accessible at: `https://dioe.at/en/projects/task-cluster-d-perception/`, retrieved 2017-03-29.

[8] DiÖ (2017). Task-Cluster E: Collaborative Online Research Platform. *DiÖ-Online*. Accessible at: `https://dioe.at/en/projects/task-cluster-e-research-platform/`, retrieved 2017-03-29.

[9] SYHD (2016). Syhd.info. Accessible at: `http://www.syhd.info/startseite/`, retrieved 2017-03-30.

[10] Schmidt, J. E., Herrgen, J., and Kehrein, R., editors (2008ff.) Regionalsprache.de (REDE). Forschungszentrum Deutscher Sprachatlas, Marburg. Accessible at: `https://regionalsprache.de/`, retrieved 2017-03-30.

[11] Text Encoding Inititative (2016). *P5:* Guidelines for Electronic Text Encoding and Interchange. Accessible at: `http://www.tei-c.org/release/doc/tei-p5-doc/en/html/`, retrieved 2017-03-29.

[12] Hammarström, H., Forkel, R., and Haspelmath, M. (2017). Glottolog 3.0. Accessible at: `http://glottolog.org`, retrieved 2017-03-29.

[13] Haspelmath, M. (2013). Can language identity be standardized? On Morey et al.'s critique of ISO 639-3. In *Diversity linguistics comment. Language structures throughout the world*. Accessible at: `http://dlc.hypotheses.org/610`, retrieved 2017-03-29.

[14] Dalby, D. (2012). The Linguasphere Register of the World's Languages and Speech Communities. First online reprint. Linguasphere press. Accessible at: `http://www.linguasphere.info/lcontao/bienvenue-welcome.html`, retrieved 2017-03-29.

[15] ISO (2017). ISO/TC 37. Terminology and other language and content resources. Accessible at: `https://www.iso.org/committee/48104.html`, retrieved 2017-03-28.

[16] ISO (2017). ISO 639-6:2009. Codes for the representation of names of languages – Part 6: Alpha-4 code for comprehensive coverage of language variants. Accessible at: `https://www.iso.org/standard/43380.html`, retrieved 2017-03-17.

[17] Kamusella, T. (2012). The global regime of language recognition. *International Journal for the Sociology of Language*, 218:59–86.

[18] Dalby, D., Gillam, L., Cox, Ch., and Garside, D. (2004). Standards for language codes: Developing ISO 639. In *Proceedings of the LREC 2004. Forth International Conference on Language resources and evaluation.* Accessible at: `http://www.lrec-conf.org/proceedings/lrec2004/pdf/327.pdf`, retrieved 2017-03-30.

[19] ISO (2017). ISO/AWI 21636. Identification and description of language varieties. Accessible at: `https://www.iso.org/standard/71300.html?browse=tc`, retrieved 2017-03-28.

[20] Library of Congress (2017). ISO 639-2 Codes for the Representation of Names of Languages. Accessible at: http://www.loc.gov/standards/iso639-2/php/code_list.php, retrieved 2017-03-17.

[21] SIL International (2017). ISO 639-3 Downloads. Accessible at: http://www-01.sil.org/iso639-3/download.asp, retrieved 2017-03-17.

[22] Library of Congress (2017). ISO 639-5 Codes for the Representation of Names of Languages. Part 5: Alpha-3 code for language families and groups. Accessible at: http://www.loc.gov/standards/iso639-5/id.php, retrieved 2017-03-17.

[23] Library of Congress (2017). ISO 639-2. Frequently Asked Questions (FAQ). Accessible at: http://www.loc.gov/standards/iso639-2/faq.html, retrieved 2017-03-17.

[24] Library of Congress (2017). Development of ISO 639-2. Accessible at: http://www.loc.gov/standards/iso639-2/develop.html, retrieved 2017-03-17.

[25] ISO 639 Joint Advisory Committee (2000): Working principles for ISO 639 maintenance (ISO 639/JAC N3R). Accessible at: http://www.loc.gov/standards/iso639-2/iso-639jac_n3r.html, retrieved 2017-03-28.

[26] Kloss, H. (1978). *Die Entwicklung neuer germanischer Kultursprachen seit 1800*. Pädagogischer Verlag Schwann, Düsseldorf.

[27] Library of Congress (2017). Criteria for ISO 639-2. Accessible at: http://www.loc.gov/standards/iso639-2/criteria2.html, retrieved 2017-03-17.

[28] Simons, G. and Fenning, Ch. D. (2017). Ethnologue: Languages of the World. 20th edition, SIL International, Dallas, Texas. Accessible at: http://www.ethnologue.com, retrieved 2017-03-29.

[29] Ethnologue (2017). History of the Ethnologue. Accessible at: https://www.ethnologue.com/about/history-ethnologue, retrieved 2017-03-30.

[30] Simons, Gary (2014). Terminology and language aspects in language coding. Presented at the *TKE 2014 Workshop: Language Codes at the Crossroads*. Berlin, Germany, 21 June 2014. Accessible at: https://tke2014.coreon.com/slides/2014_06_21_104_1030_Simons.pdf, retrieved 2017-03-16.

[31] Simons, G. (2013). ISO 639-3. Where are we and how did we get here? Presented at the *Workshop on Identifying Codes for Languages.* Newcastle, Australia, 9 February 2013. Accessible at: http://www-01.sil.org/~simonsg/local/ISO%20639-3.pdf, retrieved 2017-03-16.

[32] Haugen, E. (1966). Dialect, Language, Nation. In *American Anthropologist, New Series*, 6(4):922–935.

[33] Kloss, H. (1929). *Nebensprachen. Eine sprachpolitische Studie über die Beziehungen eng verwandter Sprachgemeinschaften*. Braumüller, Wien.

[34] SIL International (2017). Scope of denotation for language identifiers. Accessible at: http://www-01.sil.org/iso639-3/scope.asp, retrieved 2017-03-29.

[35] SIL International (2017). Change request documentation for: 2011-180. Accessible at: http://www-01.sil.org/iso639-3/chg_detail.asp?id=2011-180&lang=aeg, retrieved 2017-03-21.

[36] Library of Congress (2017). Codes for the Representation of Names of Languages. Part 5: Alpha-3 code for language families and groups. Introduction. Accessible at: http://www.loc.gov/standards/iso639-5/langhome5.html#intro, retrieved 2017-03-17.

[37] Phillips, A. and David, M. (2009). BCP 47. Tags for Identifying Languages. Accessible at: http://www.rfc-editor.org/rfc/bcp/bcp47.txt, retrieved 2017-03-29.

[38] IETF (2017). Mission Statement. Accessible at: https://www.ietf.org/about/mission.html, retrieved 2017-03-29.

[39] W3C (2017). Language tags in HTML and XML. Accessible at: https://www.w3.org/International/articles/language-tags/index.en#overview, retrieved 2017-03-29.

[40] MultiTree (2013). Indo-European: Composite. In *MultiTree: A digital library of language relationships*. Institute for Language Information and Technology, Ypsilanti, MI. Accessible at: http://multitree.org/, retrieved 2017-03-20.

POSSIBLE BUT NOT PROBABLE: A QUANTITATIVE ANALYSIS OF VALENCY BEHAVIOUR OF CZECH NOUNS IN THE PRAGUE DEPENDENCY TREEBANK

VERONIKA KOLÁŘOVÁ[1] – ANNA VERNEROVÁ[1] – JANA KLÍMOVÁ[1] – JAN KOLÁŘ[2]
[1] Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic
[2] Institute of Mathematics, Czech Academy of Sciences, Prague, Czech Republic

KOLÁŘOVÁ, Veronika – VERNEROVÁ, Anna – KLÍMOVÁ, Jana – KOLÁŘ, Jan: Possible but not Probable: A Quantitative Analysis of Valency Behaviour of Czech Nouns in the Prague Dependency Treebank. Journal of Linguistics, 2017, Vol. 68, No 2, pp. 208 –218.

Abstract: In order to optimize corpus searches for valency lexicon production, we analyse the relative frequencies of different combinations of valency complementations of Czech deverbal nouns in the Prague Dependency Treebank, considering differences between productively and non-productively derived nouns and their semantic class. We also classify combinations of forms of participants according to their frequency.

Keywords: valency, valency lexicon, Czech nouns, Word Sketch, corpus, Prague Dependency Treebank, quantitative analysis

1 INTRODUCTION

The practice of illustrating words in monolingual dictionaries with quotes goes at least as far back as Samuel Johnson's dictionary of English published in 1755. Until the advent of electronic corpora, quotes had to be collected and organised manually, mostly in the form of slips containing one quotation illustrating a selected keyword. Because of the scale of evidence that needed to be collected, volunteers from the public were involved in the major lexicographic projects such as the *Oxford English Dictionary* [22] (starting in 1858) or *Příruční slovník jazyka českého* [20] (starting after 1905). In the 1960's, first electronic corpora became available for linguistic research, and the first lexicographic project, relying fully on corpus data was started at the University of Birmingham in late 1970's, leading to the publication of the Collins COBUILD English Language Dictionary in 1987. Since then, reliance on corpus evidence is becoming the norm, especially in monolingual lexicography (COBUILD [4], Longman [14]). Due to the rapid growth of the size of the available corpora, the question of the modern-day lexicographer is not "How do I collect enough evidence?", but rather "How do I make sense of the vast amounts of evidence available to me? What tools can I use to discover patterns in the data and what can help me select appropriate quotes for the dictionary?", e.g, SketchEngine [9], GDEX [10], Pralex [13], DeepDict [3].

This paper is concerned with work on NomVallex, a corpus-based valency lexicon of Czech deverbal nouns (Section 3). Thus, the patterns, that we want to find in the data, concern the valency complementations of nouns, their forms and their possible and common combinations (Section 2). Sentences, in which the nouns

DOI 10.1515/jazcas-2017-0030

appear with several different complementations, are of particular interest due to their syntactic complexity; they are extracted from Czech corpora both automatically and manually (Section 4). Information about the preferred vs. alternative/supplementary forms and a rough indication of the frequency of forms (distinguishing common and rare forms) may be obtained by a quantitative analysis of the Prague Dependency Treebank (Section 5). This analysis may then be used to optimize the search for valency complementations and their combinations using Word Sketches (Section 6).

2 CLASSIFICATION OF ADNOMINAL FORMS AND THEIR COMBINATIONS

Czech is a highly inflectional language; valency complementations of a word are primarily distinguished by their morphological category of case. The forms expressing the individual complementations play a central role in the description of the valency behaviour of Czech nouns. The (im)possibility to use certain forms is even indicative of a meaning shift [12].

Adnominal forms often undergo both systemic and non-systemic changes when compared to forms of valency complementations of base verbs (the changes are also called typical and special shifts in surface forms of participants, see [12]). However, several constraints on adnominal forms and their combinations hold true; the major ones can be formulated as follows:

- an adnominal form is not allowed to have the form of a prepositionless accusative (with the exception of the free modification of duration, i.e. for how long, e.g. *čtení hodinu* ‘reading for an hour’);
- while prepositionless genitive and possessive forms (pronouns or adjectives) usually alternate as expressions of the same complementation, it is not possible to simultaneously express an agent by a prepositionless genitive and a different semantic role by a possessive form (e.g. **jejich bití chlapců* ‘their$_{\text{possessive}}$ beating of boys$_{\text{genitive}}$’ cannot express the verbal construction ‘boys are beating them’).

These constraints, semantic shifts and the syntactic complexity of a noun group in general have an impact on the number of possible expressions of nominal modifications which is often higher than the number of corresponding verbal ones. Various forms of valency complementations are available to enable nouns to form grammatical constructions with two or more of their complementations expressed. Typically, an adnominal complementation can be expressed by at least two forms (variants).

However, we can see differences between the respective forms (variants) and their combinations in their usage. Some forms of complementations are theoretically possible and grammatical, yet they only function as alternative or supplementary forms and they are very rare[1]. The case of complementation combinations is similar; some combinations of complementations and their forms are preferred, other combinations are only alternative or supplementary. A classification of adnominal forms is presented in Table 1.

[1] Herbst [7, p. xl] works with the following frequency scale: rare, frequent and very frequent.

Aspects of classification	Types of forms	
Grammar	grammatical	ungrammatical
Type of changes (shifts)	systemic	non-systemic
Preference in usage	preferred	alternative or supplementary
Frequency	frequent	rare

Tab. 1. Classification of adnominal forms

3 THE DEVELOPMENT OF NOMVALLEX

The corpus-based[2] valency lexicon of Czech nouns called NomVallex is a project building upon the theory of valency developed within the Functional Generative Description (FGD; [21]) and extending two existing valency lexicons developed within this tradition, Vallex (a valency lexicon of Czech verbs; [15]; [16]) and PDT-Vallex[3] (containing valency patterns of verbs, nouns, adjectives and adverbs as they occurred in the PDT-corpora; [6]; [24]). Vallex provides semantic class membership [8] and valency patterns for all meanings (i.e. lexical units) of verbs included and it was the reason why Vallex was chosen as the base for the NomVallex project. Valency properties of nouns included in NomVallex are captured in the form of a valency frame for each meaning (lexical unit), and an enumeration of combinations of adnominal complementations representing various valency patterns, as extracted from Czech corpora [11].

The valency theory for the theoretical framework of the FGD has been detailed in numerous studies addressing especially valency of verbs [19] and nouns [18]; [12]. The following types of complementations may fill in the individual slots of the valency frames of verbs:

– inner participants or arguments that can be obligatory or optional: Actor (ACT), Patient (PAT), Addressee (ADDR), Effect (EFF), Origin (ORIG) (e.g., *Vláda*$_{\text{.ACT}}$ *omezila těžbu*$_{\text{.PAT}}$ *uranu ze současných 950 tun*$_{\text{.ORIG}}$ *na 500 tun*$_{\text{.EFF}}$ *ročně* ‘The government$_{\text{.ACT}}$ restricted uranium mining$_{\text{.PAT}}$ from the current 950 tonnes$_{\text{.ORIG}}$ to 500 tonnes$_{\text{.EFF}}$ per year’);
– obligatory free modifications or adjuncts, especially those with the meaning of direction (e.g., *přijet někam*$_{\text{.DIR3}}$ ‘to arrive somewhere’) or location (e.g., *přebývat někde*$_{\text{.LOC}}$ ‘to dwell somewhere’) and manner (e.g., *chovat se dobře*$_{\text{.MANN}}$ ‘to behave well’).

The same inventory of valency complementations is assumed for deverbal nouns denoting an action. The inventory of valency complementations of non-deverbal nouns and deverbal nouns undergoing substantial shifts in their meaning is supplemented with some more modifications, especially with a special nominal participant Material (MAT; e.g., *skupina lidí*$_{\text{.MAT}}$ ‘group of people’, *jedno balení*

[2] Another approach to valency of Czech nouns (so-called corpus-driven approach) was applied by Čermáková (2009).

[3] http://hdl.handle.net/11858/00-097C-0000-0023-4338-F

másla$_{\text{.MAT}}$ 'one package of butter') and a free modification Appurtenance (APP; e.g., *Petrovo*$_{\text{.APP}}$ *auto* 'Peter's car', *oddělení odbytu*$_{\text{.APP}}$ 'sales department').

3.1 Semantic Classes in NomVallex and a Preliminary List of Entries

Nouns representing five semantic classes are included in NomVallex, namely Communication (e.g. *odpověď* 'answer'), Exchange (e.g. *dodávka* 'delivery'), Contact (e.g. *dotyk* 'touch'), Mental action (e.g. *dojem* 'impression'), and Psychological state (e.g. *obava* 'fear'). The assignment of a semantic class is carried over from Vallex: a noun is supposed to be assigned the same semantic class as its base verb in Vallex, with the exception of nouns that undergo a change in meaning. On the basis of the list of verbs in Vallex (see Table 2 for numbers of lexical units representing particular semantic classes), a preliminary list of noun entries was created. Within these semantic classes, we aim to provide valency patterns of all types of Czech nouns with a meaning denoting an action or an abstract result of an action. These nouns are either derived from verbs by productive means (suffixes *-(e)ní/tí*, as in *vykládání* 'explaining // unloading' or *pojetí* 'conception') or by non-productive means including the zero suffix (such as *vykládka* 'unloading', *výklad* 'explanation / interpretation'). The preliminary list of candidate entries to be included in NomVallex currently contains 1230 lemmas, cf. Table 3.

	Commu-nication	Exchange	Contact	Mental action	Psych. verbs	Total
Verbs in Vallex	428	182	125	338	143	1216

Tab. 2. Number of verbal lexical units in Vallex.

	Commu--nication	Exchange	Contact	Mental action	Psych. state	Total
Productively derived nouns	335	171	117	257	104	984
Non-productively derived nouns	110	38	14	56	28	246
Total	445	209	131	313	132	1230

Tab. 3. Number of lemmas of nouns included in the NomVallex preliminary list of entries

4 EXTRACTION OF VALENCY PATTERNS FROM CZECH CORPORA: METHODOLOGY

Searching for valency patterns of Czech nouns usually means searching for many various combinations of forms, including word order variants. We use the following Czech lemmatized and morphologically annotated corpora: the synchronic part of the Czech National Corpus (CNC)[4], the web corpus Araneum Bohemicum Maximum[5] [2] and corpora from the Prague Dependency Treebank Family,

[4] `http://korpus.cz/`

[5] `http://ucts.uniba.sk/aranea_about/index.html`

especially the Prague Dependency Treebank (PDT 3.0)[6]. The PDT 3.0 [1] contains Czech texts with complex and interlinked morphological (2 million words), syntactic (1.5 MW) and complex semantic annotation (0.8 MW).

Using the CNC and the Araneum corpus, valency patterns of Czech nouns are being extracted either with the help of Sketch Engine's Word Sketches [9], or by sophisticated CQL queries specified in the KonText application[7]. Searching through the PDT 3.0 is carried out by the tool called PML-TQ [23].

A manual syntactic annotation of the PDT 3.0 enables to carry out a precise quantitative analysis of all adnominal forms and their combinations (Section 5). However, as the corpus is rather small, some rare meanings and some non-systemic, supplementary or rare forms of complementations of nouns in their more frequent meanings do not occur in the data at all. In contrast, it is impossible to do a reliable quantitative analysis in the data of the CNC or the Araneum corpus unless a manual syntactic annotation is provided. On the other hand, these big corpora give evidence about assorted adnominal forms and their combinations that, although rare, should be captured in the valency lexicon.

While annotating individual lexemes from a manually prepared list of headwords (see Section 3), we suggest the following procedure:

1. To prepare a list of tentative lexical units by applying the systemic shifts to the units of the base verb; then to adjust this list to reflect meaning shifts and additional or missing lexical units;
2. To specify the most frequent forms and their combinations on the basis of the PDT-corpora data;
3. To use WordSketches with an extended Word Sketch Grammar to discover individual preferred forms of complementations of the lemma together with their most common/relevant lexical realisation; however, the Word Sketches are not sense disambiguated, so their output needs to be manually explored and the forms and examples added to the relevant lexical units as appropriate;
4. The statistical nature of Word Sketches makes them unsuitable for discovering alternative or supplementary forms; so we suggest to manually extend the list of possible realisations of each complementation with forms discovered by inspecting a sample of corpus concordances and/or by introspection and confirmed by manual search in the CNC data and the Araneum corpus;
5. To automatically create and run corpus searches for combinations of two or more expressed complementations, extracting concordances that could be used as dictionary examples;
6. To manually check the concordances extracted in step 5, selecting the appropriate examples.

In this paper, we refer to step 2 (Section 5) and explore the ways how steps 3 and 5 can be optimized in order to produce the least amount of output (Section 6) while providing the most useful evidence for the manual steps 4 and 6.

[6] http://ufal.mff.cuni.cz/pdt3.0

[7] http://wiki.korpus.cz/doku.php/en:pojmy:dotazovaci_jazyk

5 A QUANTITATIVE ANALYSIS OF COMBINATIONS OF PARTICIPANTS IN THE PDT 3.0

5.1 Annotation Scheme

The valency theory of the FGD was applied to the PDT-corpora data which resulted in a very complex and detailed annotation scheme [17]. Different meanings of words with valency that occur in the data are differentiated in PDT-Vallex. The annotation of valency consists of:

– determining and assigning a valency frame from PDT-Vallex;
– a lemma and a corresponding semantic role (ACT, PAT, ADDR, etc.) are assigned to the nodes for valency complementations expressed in the surface form of the sentence;
– obligatory valency complementations unexpressed on the surface are captured by an added (newly created) node with an artificial lemma (for example #PersPron), and the corresponding semantic role is also assigned.

First, we carried out a quantitative analysis focusing on relative frequencies of combinations of participants, modifying both productively and non-productively derived nouns in the PDT 3.0 (Section 5.2). Second, we present an analysis of combinations of adnominal forms for all productively derived nouns in PDT 3.0 (Section 5.3).

5.2 Relative Frequencies of Combinations of Adnominal Participants

Using the NomVallex preliminary list of entries (1230 lemmas, see Section 3), we searched through the PDT 3.0 for both productively and non-productively derived nouns representing the five selected semantic classes (Communication, Contact, Exchange, Mental action and Psychological state). 623 such lemmas occurred in the PDT 3.0 in a total of 8273 occurrences (see Table 4).

		Communication	Exchange	Contact	Mental action	Psych. state	Total
Productively derived nouns	Lemmas	145	94	30	107	29	405
	Occurrences	1552	699	128	1236	179	3794
Non-productively derived nouns	Lemmas	102	34	10	54	18	218
	Occurrences	2163	540	16	1256	504	4479
Total	Lemmas	247	128	40	161	47	623
	Occurrences	3715	1239	144	2492	683	8273

Tab. 4. Number of lemmas and occurrences of nouns found in the PDT 3.0

Nodes added for obligatory complementations that are not present on the surface layer of the sentence enable us to search also for the unexpressed elements and to differentiate expressed and unexpressed valency modifications in our searches. We carried out a quantitative analysis focusing on relative frequencies of combinations of adnominal participants (i.e., ACT, PAT, ADDR, EFF, and ORIG). Figures 1 and 2 show that the most frequent combination is the case when only the PAT is expressed (with the exception of non-productively derived nouns of Contact which represent the least frequent class and so the numbers may be influenced by their rare occurrence). The case when only ACT is expressed is the second most frequent combination, followed by the combinations ACT+PAT or PAT+ADDR, the latter of which is applicable only in the case of nouns that have ADDR in their valency frame. Interestingly, relative frequencies of the combination ACT+PAT are very low with nouns of Exchange and nouns of Contact. Relative frequencies of combinations of three participants – no more than 0.13% – are not shown in the Figures.

5.3 Relative Frequencies of Combinations of Forms of Adnominal Participants

Analysis of forms of complementations of all productively derived nouns in PDT 3.0 (not only the nouns representing the five selected semantic classes) strongly confirms the intuition that most nouns occurring with a valency complementation occur with a single complementation expressed in prepositionless genitive (*zvyšování ceny* 'increasing of the price'; relative percentage around 70%). Less common forms are listed in Table 5.

Concerning nouns with two complementations expressed, the most common combination is a complementation in genitive together with a complementation expressed by a prepositional group (*měření sil*$_{\text{.PAT}}$ *se Švédy*$_{\text{.ADDR}}$ 'pitting (one's) strength$_{\text{.PAT}}$ against the Swedes$_{\text{.ADDR}}$'; almost 3%).

As mentioned above, combinations of three expressed valency participants are rare; of these, the most common is noun in the genitive and two prepositional groups (*snížení investic*$_{\text{.PAT}}$ *z jedné miliardy*$_{\text{.ORIG}}$ *na 600 milionů*$_{\text{.EFF}}$ 'the fall of investments$_{\text{.PAT}}$ from 1000 million$_{\text{.ORIG}}$ to 600 million$_{\text{.EFF}}$'; 6 occurrences in PDT 3.0, making up for less than 0.1% of nouns with expressed valency participants).

The PDT 3.0 does not contain any instance of a noun modified by four expressed valency participants.

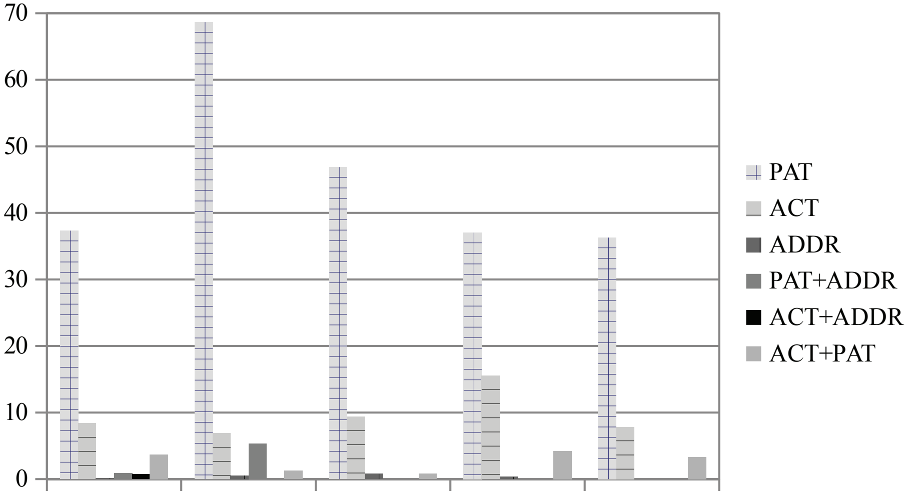


Fig. 1. Relative frequencies of selected combinations of participants modifying *productively* derived nouns in the PDT 3.0

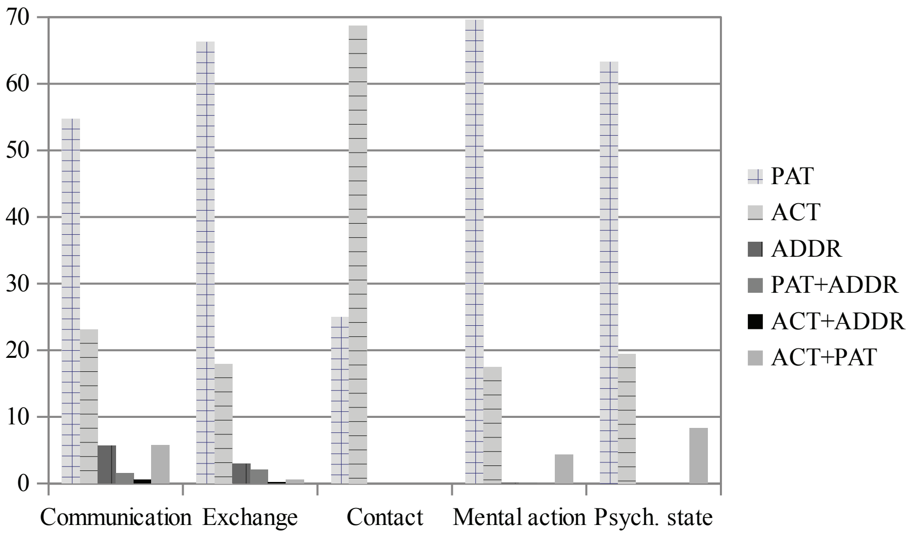


Fig. 2. Relative frequencies of selected combinations of participants modifying *non-productively* derived nouns in the PDT 3.0

Percentage	Expressed complementations	Example
around 70%	prepositionless genitive	*zvyšování ceny* 'increasing of the price'
almost 9%	prepositional groups	*srovnávání s čím* 'comparison with sth'
almost 6%	possessive forms	possessive adjectives: *Clintnův* 'Clinton's' possessive pronouns: *náš* 'ours'
almost 3%	prepositionless genitive + prepositional group	*měření sil*$_{\text{.PAT}}$ *se Švédy*$_{\text{.ADDR}}$ 'pitting (one's) strength$_{\text{.PAT}}$ against the Swedes$_{\text{.ADDR}}$'
about 1.5%	indeclinable noun	*vedení Oilers* 'the leadership of Oilers' acronyms: *rozdělení ČSFR* 'the division of ČSFR (i.e. Czechoslovakia)'
0.8%	content clause	*prohlášení, že ...* 'a declaration that …'
0.5%	infinitive	*oprávnění zastavit vozidlo* 'authorization to stop a vehicle'
< 0.1%	prepositionless genitive + prepositional group + prepositional group	*snížení investic*$_{\text{.PAT}}$ *z jedné miliardy*$_{\text{.ORIG}}$ *na 600 milionů*$_{\text{.EFF}}$ 'the fall of investments$_{\text{.PAT}}$ from 1000 million$_{\text{.ORIG}}$ to 600 million$_{\text{.EFF}}$'

Tab. 5. Combinations of forms expressing complementations of a single noun. Of the combinations with percentage below 1%, only those of particular interest are listed.

6 WORD SKETCH AND CORPUS SEARCH OPTIMIZATION

The fact. that a certain form is not common among adnominal complementations modifying a particular noun, does not necessarily mean that such complementation cannot be found by Word Sketches. Quite to the contrary, these forms may be specific to a limited group of nouns and thus forming more statistically significant collocations. For example, the dative complementations appear with less than 1% of noun instances in the examined PDT 3.0 data. However, using our extended SketchGrammar, a Word Sketch of the noun *předání* ‘delivery, handover’ contains several typical lexical realisations of the dative complementation: *exekutorovi* ‘to the executor’, *zdravotníkům* ‘to the paramedics’, *Číně* ‘to China’, *zákazníkovi* ‘to the customer’ etc.

On the other hand, we may try to judge the expected utility of WordSketches by comparing the output of the search for valency complementations only to the output of the search for any complementations of nouns, including the complementations that are not part of the valency frame (i.e. the free modifications). For example, there are twice as many occurrences of the combination of a complementation expressed as a noun in the genitive and complementation expressed by a prepositional group when we allow for non-valency complementations. In other words, if we created a Word Sketch rule for discovering this combination, we may expect it would trigger about half of the time on a combination such that at least one of the two complementations does not actually belong to valency frame of the noun.

Up to now, we have extended the Czech Word Sketch Grammar with some forms of adnominal collocations that are typical of valency, especially with collocations in prepositionless dative and instrumental. Upon further examination of the frequent forms, we plan to experiment with adding ternary relations capturing some of the most common combinations of two adnominal complementations.

7 CONCLUSION

The paper refers to the current work on NomVallex, the corpus-based valency lexicon of Czech nouns. We carried out a quantitative analysis of valency behaviour of Czech deverbal nouns in the PDT 3.0. Reflecting the difference between productively and non-productively derived nouns and their semantic class membership, we show that order of relative frequencies of combinations of adnominal participants is almost the same for all observed types of nouns. We also specify the most frequent and infrequent combinations of adnominal forms, which enables to optimize Word Sketches.

ACKNOWLEDGEMENTS

The research reported in the paper was supported by the Czech Science Foundation under the project GA16-02196S. This work has been using language resources developed, stored and distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2015071).

The last author was supported by grant no. RVO 67985840 of the Czech Academy of Sciences.

References

[1] Bejček, E. et al. (2013). Prague Dependency Treebank 3.0. Data/software, Univerzita Karlova v Praze, MFF, ÚFAL, Prague, Czech Republic.

[2] Benko, V. (2014). Aranea: Yet Another Family of (Comparable) Web Corpora. In Sojka, P. et al., editors, *TSD 2014*. LNAI 8655, pages 247–256, Springer International Publishing.

[3] Bick, E. (2009). DeepDict – A Graphical Corpus-based Dictionary of Word Relations. In Bick, E., Hagen, K., Müürisep, K., and Trosterud, T., editors, *Proceedings of the NODALIDA 2009 workshop Constraint Grammar and robust parsing*. NEALT Proceedings Series, Vol. 8, i-ii, Northern European Association for Language Technology (NEALT), Tartu University.

[4] *Collins COBUILD Advanced Learner's Dictionary* (2014).

[5] Čermáková, A. (2009). *Valence českých substantiv*. Lidové noviny, Praha.

[6] Hajič, J., Panevová, J., Urešová, Z., Bémová, A., Kolářová, V., and Pajas, P. (2003). PDT-VALLEX: Creating a Largecoverage Valency Lexicon for Treebank Annotation. In *Proceedings of The Second Workshop on Treebanks and Linguistic Theories*, pages 57–68, Växjö University Press, Sweden.

[7] Herbst, T., Heath, D., Roe, I. F., and Götz, D. (2004). A valency dictionary of English: a corpus--based analysis of the complementation patterns of English verbs, nouns, and adjectives. Walter de Gruyter, Berlin, Germany.

[8] Kettnerová, V., Lopatková, M., and Hrstková, K. (2008). Semantic Classes in Czech Valency Lexicon: Verbs of Communication and Verbs of Exchange. In *Lecture Notes in Computer Science, Vol. 5246, Proceedings of the 11th International Conference, TSD 2008*, pages 109–116, Springer, Berlin – Heidelberg, Germany.

[9] Kilgarriff, A. and Tugwell, D. (2001). WORD SKETCH: Extraction and display of significant collocations for lexicography. In *Proc Collocations workshop*, pages 32–38, ACL, Toulouse, France.

[10] Kilgarriff, A., Husák, M., McAdam, K., Rundell, M., and Rychlý, P. (2008). GDEX: Automatically finding good dictionary examples in a corpus. In *Proceedings of the 13th EURALEX International Congress*, pages 425–432, Spain.

[11] Klímová, J., Kolářová, V., and Vernerová, A. (2016). Towards a Corpus-based Valency Lexicon of Czech Nouns. In Kernerman, I. et al., editors, *Globalex 2016, Lexicographic Resources for Human Language Technology*, pages 1–7. Accessible at: `http://ailab.ijs.si/globalex/files/2016/06/LREC2016Workshop-GLOBALEX_Proceedings-v2.pdf.`

[12] Kolářová, V. (2014). Special valency behavior of Czech deverbal nouns. In Spevak, O., editor, *Noun Valency*, pages 19–60, John Benjamins, Amsterdam, Netherlands.

[13] Lexikální databáze Pralex: Accessible at: `http://lexiko.ujc.cas.cz/index.php?-page=23.`

[14] Longman Dictionary of Contemporary English. (2014). 6th Edition. Pearson.

[15] Lopatková, M., Kettnerová, V., Bejček, E., Vernerová, A., and Žabokrtský, Z. (2015). *VALLEX 3.0 – Valenční slovník českých sloves*. Charles University in Prague. Accessible at: `http://ufal.mff.cuni.cz/vallex/3.0/.`

[16] Lopatková, M., Kettnerová, V., Bejček, E., Vernerová, A., and Žabokrtský, Z. (2016). *Valenční slovník českých sloves VALLEX*. Karolinum, Praha.

[17] Mikulová, M. et al. (2006). Annotation on the tectogrammatical level in the Prague Dependency Treebank. Annotation manual. Technical Report TR-2006-30, ÚFAL MFF UK, Praha.

[18] Panevová, J. (2002). K valenci substantiv (s ohledem na jejich derivaci). In *Zbornik matice srpske za slavistiku*, vol. 61, pages 29–36.

[19] Panevová, J. (2014). Contribution of Valency to the Analysis of Language. In Spevak, O., editor, *Noun Valency*, pages 1–18, John Benjamins Publishing Company, Amsterdam, Netherlands.

[20] *Příruční slovník jazyka českého*. (1935–1957). Státní nakladatelství – Školní nakladatelství – Státní pedagogické nakladatelství, Praha.

[21] Sgall, P., Hajičová, E., and Panevová, J. (1986). *The Meaning of the Sentence in its Semantic and Pragmatic Aspects*. Reidel, Dordrecht.

[22] Stevenson A., editor (2010). *Oxford Dictionary of English*. Oxford University Press.

[23] Štěpánek, J. and Pajas, P. (2010). Querying Diverse Treebanks in a Uniform Way. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, pages 1828–1835, European Language Resources Association (ELRA), Valletta, Malta.

[24] Urešová, Z. (2012). Building the PDT-VALLEX valency lexicon. In *Proceedings of the fifth Corpus Linguistics Conference*, pages 1–18, University of Liverpool, Liverpool, UK.

NEW SPOKEN CORPORA OF CZECH: ORTOFON AND DIALEKT

ZUZANA KOMRSKOVÁ – MARIE KOPŘIVOVÁ – DAVID LUKEŠ
– PETRA POUKAROVÁ – HANA GOLÁŇOVÁ[1]
[1] Institute of the Czech National Corpus, Charles University, Prague, Czech Republic

KOMRSKOVÁ, Zuzana – KOPŘIVOVÁ, Marie – LUKEŠ, David – POUKAROVÁ, Petra – GOLÁŇOVÁ, Hana: New Spoken Corpora of Czech: ORTOFON and DIALEKT. Journal of Linguistics, 2017, Vol. 68, No 2, pp. 219 – 228.

Abstract: The paper introduces the ORTOFON corpus of spontaneous spoken Czech and the DIALEKT corpus of Czech dialects, their design principles and practical solutions adopted during data collection.

Keywords: dialectology, lemmatization, spoken corpus, tagging, transcription

1 INTRODUCTION

This paper introduces new spoken corpora prepared by the Institute of the Czech National Corpus (ICNC). The process of collecting recordings for the ORTOFON and DIALEKT corpora started in 2012 and both have finally been published on June 2, 2017. Both corpora are lemmatized and morphologically tagged.

The ICNC has a long tradition in creating spoken corpora. The first corpus of spoken Czech was the Prague Spoken Corpus (PSC) [5] whose recordings span the years 1988–1992 and were made in the Prague area only. Its follower – the ORAL series corpora[1] – focused on spontaneous spoken conversations of family members or friends from different parts of the Czech Republic, in the course of their natural, usual interactions (e.g. at home during a meal, in a restaurant, in the street). Except for the last corpus in the ORAL series (ORAL2013 [3]), these corpora (namely PSC, ORAL2006 [15], and ORAL2008 [23]) have been published only as transcripts, without the corresponding sound recordings. By contrast, ORAL2013 provides access to the actual recordings aligned with a one-tier transcript.

While the new ORTOFON corpus follows this tradition as far as the manner of data collection is concerned, the DIALEKT corpus is a new project line which focuses on monological spoken language showcasing traditional dialects. Both new corpora are based on a multi-tier transcription setup.

2 THE ORTOFON CORPUS

This new spoken corpus of spontaneous everyday communication has been published on June 2, 2017, following several months of final data selection and revision. The data

[1] The ORAL series corpora were integrated into the ORAL corpus with 6 361707 tokens. This corpus is lemmatized and morphologically tagged in the same way as the ORTOFON and DIALEKT corpora. More at [18].

DE GRUYTER OPEN
DOI 10.1515/jazcas-2017-0031

was collected during 2012–2017. In terms of linguistic annotation, it features lemmatization and morphological tagging (see section 4). The size of the final published corpus is 1 236 508 tokens. Like previous spoken corpora, the ORTOFON corpus is balanced with respect to several sociolinguistic categories.

The raw material consists of recordings of prototypical spoken language (Czech in our case) [7, p. 118], which is defined as informal and spontaneous conversations between people who know each other very well, situated in casual settings. The interactions take place in familiar environments (e.g. in private, among friends) and the situations are not experimentally induced. We only record adult speakers (18+ years old).

2.1 Metadata

Our external collaborators who record and transcribe the conversations were asked to provide a variety of information about each recording and each speaker. This information covers the two broad categories of "context-governed" and "demographic" details [4]. These enable the corpus user to restrict searches to specific types of extralinguistic context and to create subcorpora based on them. The goal is to capture as many of the factors which can possibly influence the conversation as possible.

The context-governed perspective covers general information about the recorded situation. There is a list of 12 pre-defined primary situation types, which distinguish the different possible settings in which the conversation could have taken place (for further details see [16], [17]). Another requirement is to enter the date, place, and corresponding geographical area of the recording location (the geographical areas are based on dialect areas which follow [1]). The collaborators are also asked to make a list of conversational topics and to fill them in. Apart from that, the relationship of speakers is indicated (one of partners, family, friends, acquaintances) and the total number of generations they represent (e.g. mother and daughter = two generations). There is also an assessment of the sound quality of the recording, which is useful for phonetic transcription. In the resulting corpus, the information related to the whole recording will be stored as per-document metadata.

The demographic perspective summarizes the speakers' characteristics; it is therefore mapped onto per-speaker metadata. In each recording, the speakers are numbered and cross-referenced with a speaker database. The database tracks the speakers' sociological characteristics, which include:

- gender
- age
- field and highest achieved level of education
- current and longest occupation
- childhood region and place of residence (until 15 years old), longest and current region and place of residence, and size of the corresponding administrative unit
- common speech defects.

2.2 Balancing the ORTOFON Corpus

The previous ORAL2008 and ORAL2013 corpora have been balanced according to three sociolinguistic variables: gender, age, and the highest achieved level of education.

Each variable was split into two levels (female × male, 18–34 years old × 35+ years old, non-tertiary × tertiary education) to avoid excessive fragmentation and to enable comparability with PSC. The balancing of the ORTOFON corpus is based on four sociolinguistic variables, namely the three previously mentioned ones and childhood region, which assumes ten dialect regions (see Fig. 1). The final corpus is trying to be representative (i.e. it includes speakers representing all possible combinations of the sociolinguistic variables, and as many different speakers as possible), and as balanced as possible (i.e. the proportions of all categories are roughly equal). Considering the target size of the corpus and the number of levels per the four variables, we get 1M / (2 × 2 × 2 × 10) = 12 500 tokens ideally for each combination, e.g. for female speakers 35+ y.o. with tertiary education from West Bohemia. We strove for a minimum of five different speakers per combination [9], which reduces the risk of a category being excessively tied to a single idiolect and maintains variability.[2]

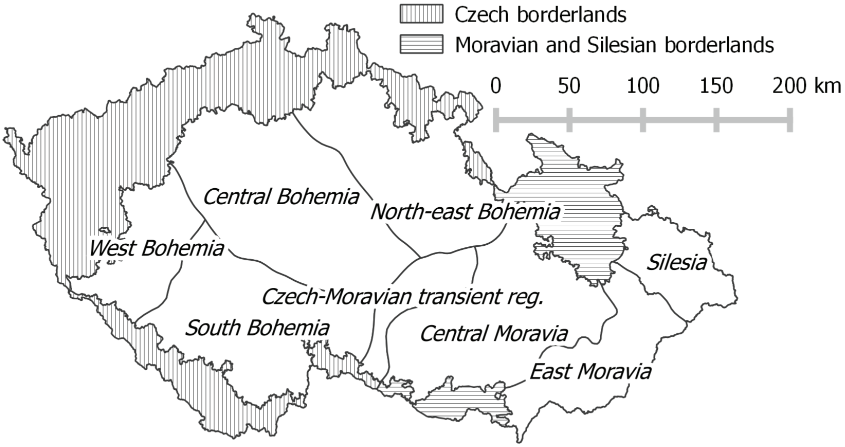


Fig. 1. Dialect regions in the ORTOFON corpus

The map shows all ten dialect regions. Their borders were determined according to several dialect studies (e.g. [14], [22]), so they have been slightly modified compared to ORAL2013.[3] While the previous ORAL series corpora only used the criterion of territory to a certain extent to make the data as representative as possible, ORTOFON treats the criterion of childhood territory on par with the other balancing variables.

2.3 Annotation Scheme

The main difference between the ORTOFON corpus and the ORAL series corpora is the multi-tier transcription. Every recording is transcribed using the ELAN[4] transcription software [21]. There are two main types of tiers (corresponding to

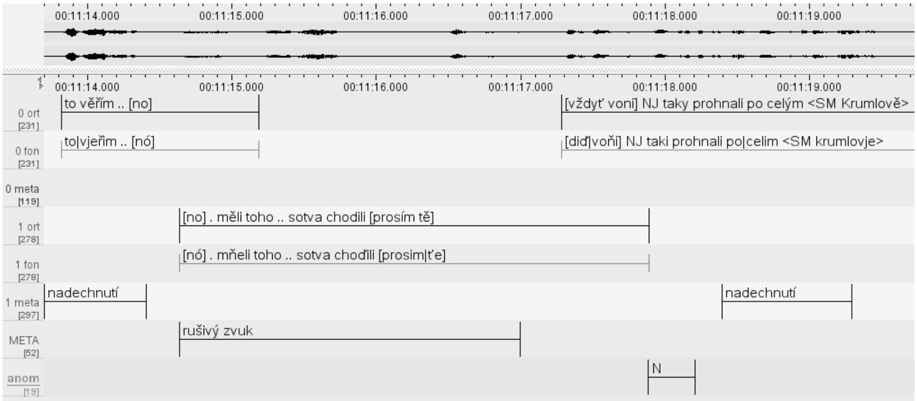
[2] More details at http://wiki.korpus.cz/doku.php/cnk:ortofon.

[3] The map is available at: https://wiki.korpus.cz/lib/exe/detail.php/cnk:o13.png.

[4] ELAN is being developed at the Max Planck Institute for Psycholinguistics, The Language Archive, Nijmegen, The Netherlands; URL: http://tla.mpi.nl/tools/tla-tools/elan/.

orthographic and phonetic transcription) and each speaker in every recording gets their own instance of both of them, which means that overlaps may be transcribed in parallel on the respective orthographic (and phonetic) tiers of the overlapping speakers (there are always the whole words in overlaps, the overlapping speech is marked by square brackets []). Speakers' turns are segmented into sub-units of a maximum length of 25 tokens for ease of parallel revision.

The transcription workflow proceeds stepwise from a basic orthographic transcription with annotation of metalinguistic information, through revisions, and eventually to phonetic transcription.



0 ort [231]: to věřím .. [no]
[vždyť voni] NJ taky prohnali po celým <SM Krumlově>

0 fon [231]: to|vjeřim .. [nó]
[diď|voňi] NJ taki prohnali po|celim <SM krumlovje>

0 meta [119]

1 ort [278]: [no] . měli toho .. sotva chodili [prosím tě]

1 fon [278]: [nó] . mňeli toho .. sotva choďili [prosim|ťe]

1 meta [297]: nadechnutí
nadechnutí

META [52]: rušivý zvuk

anom [19]: N

Fig. 2. Excerpt from a transcript for the ORTOFON corpus in the ELAN transcription program, showing the recording waveform at the top, with time-aligned orthographic, phonetic, and metalinguistic tiers for speaker 0 (0 ort, 0 fon, 0 meta) and speaker 1 (1 ort, 1 fon, 1 meta).

The multi-tier transcription shown in Fig. 2 illustrates the use of tiers: orthographic (ort), phonetic (fon), metalinguistic (meta, META), and anonymization (anom). The orthographic and phonetic tiers are reserved for speech transcription (see 2.3.1 and 2.3.2). Each speaker is further associated with their own metalinguistic tier (meta), which captures e.g. laughter or hiccups, i.e. paralinguistic sounds pertaining to a specific speaker, or pauses longer than two seconds. Additionally, there is another metalinguistic tier (META), only one instance per recording, whose purpose is to capture ambient sounds, e.g. phones ringing, dogs barking, or TV background noise. Both the meta and META tiers offer a list of pre-defined categories. Another layer (anom) is used for the anonymization of personal data, e.g. phone numbers, surnames, addresses. There is also a possibility to add another tier, the so-called JO tier, to capture the speech of a non-target speaker who disrupts the communication of target speakers, i.e. a waitress in a restaurant, or a child speaking to her mother. The anom and JO tiers are optional.

2.3.1 Orthographic Transcription

The starting point for annotation is the orthographic tier. It is optimized for a first quick transcription of the recording. Although the tier is named "orthographic", the transcription differs in some aspects from traditional written language. For instance, it captures dialectal features, e.g. variations in the endings for all types of conjugation and declension. Conversely and unlike the ORAL series, it preserves the quantity of vowels according to standard Czech, all consonants in consonant clusters (e.g. *já vždycky vím* instead of pronounced *já dycky vim*), and full form of formally reduced variants of words (e.g. *myslím dostal šestnáct* instead of *sim dostal šesnáz*).[5] In case of two (or more) possible variants of transcription of the word, we choose only one of them (*citron/citrón* > *citron*; *osum/osm* > *osm*; *benzin/benzín* > *benzin* etc.).

A very important requirement is to ensure that the transcription procedure is homogeneous across different recordings, which already span over four years. For this purpose, we worked out a detailed manual for all our collaborators where they can find examples and general rules for transcription.[6] This manual has been continuously updated with additional examples gleaned from the material.

The most important phenomena captured on the orthographic tier include:

- v- or h-prothesis: ***v**okno*, ***h**ulica*
- regional variants of vocalic changes: *ml**ý**n* - *ml**ej**n* - *ml**é**n*, *l**ou**ka* - *l**ú**ka* - *l**ó**ka*
- regional declension variants: *s mal**ejma** náklad**ama*** (instead of *s malými náklady*)
- regional conjugation variants: *maj**í*** - *maj* - *maj**ú*** - *maj**ó*** (3-PL-*mít*), *chc**u** říct* (instead of *chci říct*)
- shortened forms of the 3rd pers. sg. past participle normally ending in -l: *moh*, *spad*, *řek*

Another specificity is pausal punctuation, used also in the ORAL2013 corpus. In the ORTOFON corpus, the term "pause" became more accurate, i.e. at least 120 ms of silence or other nonverbal sounds, e.g. breath, cough, laugh. However, pauses shorter than 120 ms may be annotated under the looser concept of "prosodic boundary", which also covers prosodic segmentation phenomena not implemented by an actual interruption of the flow of speech, like tempo changes and intonation cues. The transcription distinguishes three types of pauses with different symbols:

- . on the ort layer for prosodic boundaries (including pauses up to 120 ms);
- .. on the ort layer for pauses from 120 ms to 2 s;
- a separate segment annotated as *dlouhá pauza* (*long pause*) on the meta layer for pauses longer than 2 s.

The orthographic layer captures the verbal and near-verbal content of the interaction including unfinished words, false starts, hesitations, response sounds, and overlaps (for details on the particular symbols used, see [16], [17]).

Paralinguistic and situational comments are mainly captured on the meta and META layers, but some of them are also present on the orthographic tiers. This occurs when they are tightly coupled to a particular segment of speech: either because they could affect voice quality, e.g. laughter, yawning, loudness, or because they convey additional information, e.g. speech in foreign language, recitation, singing. The tokens uttered with that concomitant feature are signalized by angle brackets <>, e.g. *ty máš <SM nápady>*.[7]

[5] There is a list of formally reduced variants which have been lexicalized and thus transcribed, e.g. *čéče* (but *čoveče* is transcribed as a full *člověče*), *páč* (instead of *poněvadž*).

[6] Accessible at: `https://wiki.korpus.cz.`

[7] <*SM* ...> marks laughter.

2.3.2 Phonetic Transcription

The phonetic tier is an innovation compared to the ORAL series corpora. It has its own rules, which allow us to capture real pronunciation using a simplified phonetic transcription. Although it does not aim to capture all phonetic variation (e.g. the scale of vowel reduction), it still offers basic pointers concerning variability in spontaneous speech. Standard alphabet characters, extended with a small set of specialized symbols, are used instead of the International Phonetic Alphabet (for details on this decision see [16], [17]).

The phonetic layer is closely integrated with the orthographic layer. Some orthographic words are merged into prosodic words (or stress groups) on the phonetic tier, but the space between them is not simply removed. Instead, it is replaced with the pipe | symbol, so as to preserve information about the location of the orthographic boundary and, by extension, a one-to-one correspondence between the tokens on the two tiers. This allows search query constraints to target both tiers simultaneously, providing the users with more control over their search results.

The phonetic layer captures the following phenomena (in the example pairs, the first half corresponds to the ort layer and the second to fon):

- some non-phonemic distinctions, e.g. labiodental [ɱ] or velar [ŋ]: *prosím vás* → *prosi**ɱ**|vás*, *tenkrát* → *te**ŋ**krát*
- assimilations of voicing: *kup mi to* → *ku**b**|mi|to*, *tvoje* → *t**f**oe*
- assimilations of place of articulation: *hodně* → *ho**ď**ňe* (see also examples under non-phonemic distinctions above)
- assimilations of manner of articulation: *od nás* → *o**n**|nás*
- shared phones, indicated via the underscore _ symbol: *dnes jsem se dobře vyspal* → *dne_|sem|se dobře vispal*
- epentheses and elisions: *zhasnout* → *zhas**t**nout*, *protože* → *bže*

3 THE DIALEKT CORPUS

This new corpus, published alongside ORTOFON, is our first attempt to build a collection of dialectal linguistic material compiled as a linguistic corpus. As far as we are aware, it is also the first dialectal corpus in the Czech Republic available through a user-friendly search interface, serving not only professional dialectologists but also the broader linguistics community, teachers and laypeople. Like the ORTOFON corpus, it is lemmatized and morphologically tagged.

The DIALEKT corpus differs from the ORTOFON in several characteristics. Firstly, it does not have a fixed size in tokens, it will be, hopefully, published regularly in versions with a growing amount of data.[8] The first version counts 128,289 tokens on the dialectological layer and 126,131 tokens on the orthographic layer. This is related to the second difference, that the corpus is not balanced, nor does it aim to be in the future. Thirdly, the material covers two broad stages of data collection: older data from the late 1950s up to the 1980s, which mostly comes from

[8] Creating a non-balanced, continuously growing version of the ORTOFON corpus, alongside the balanced one, is also under consideration.

the research effort which resulted in the Czech Linguistic Atlas [1], and new data since the 1990s [12]. This allows comparing the gradual loss of dialectal features in the respective dialectal regions. Additional differences concern the process of transcription (see 3.3).

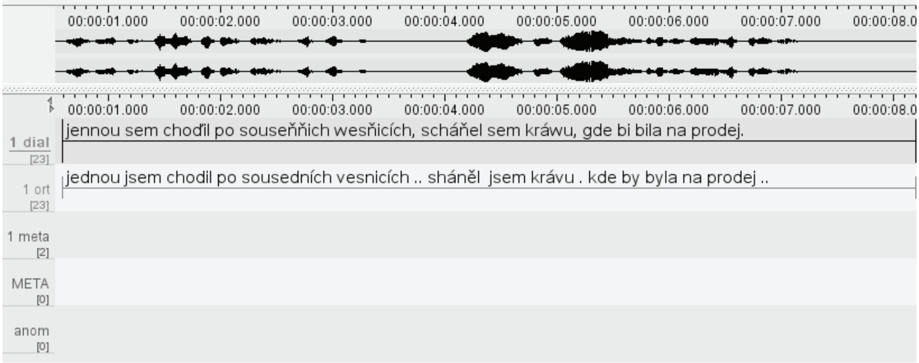
3.1 Metadata

Due to the two stages of data collection, the metadata about speakers and the whole recording were adapted. The dialectal speakers had to fulfil certain criteria: they had to have spent the great majority of their life in a single rural area without moving to another dialectal region, they had to be over 60 years old and not university educated. There were no limitations as far as their occupation, but some speakers (teachers for example) usually adjust their speech or care much more about dialectal features, which influences their spontaneity. Speakers tied to traditional rural professions were therefore given preference, which goes hand in hand with an interest in dialectal lexis.

Regional classification is, in contrast to the ORTOFON corpus, more detailed. The ten dialectal regions, which are the same for both corpora, are divided into smaller sub-areas with a specific type of a particular dialect, and those can subdivided even further, according to the traditional three-level hierarchy for classifying dialects (*nářeční oblast* > *nářeční typ* > *nářeční úsek*). The metadata also show which region belongs to which territory of the Czech Republic, i. e. Bohemia, Moravia, Silesia, and if the type of residence was town or country. Further details about the metadata are available in [12].

3.2 Annotation Scheme

The recordings for the DIALEKT corpus are transcribed according to a similar procedure as the ORTOFON corpus, using the same tools. The types of tiers are the same with one exception: there is a dialectological layer instead of the phonetic one, and it is considered as the primary one (the primary layer for the ORTOFON corpus is the orthographic one).



jennou sem choďil po souseňňich wesňicích, scháňel sem kráwu, gde bi bila na prodej.

jednou jsem chodil po sousedních vesnicích .. sháněl jsem krávu . kde by byla na prodej ..

Fig. 3. Excerpt from a transcript for the DIALEKT corpus in the ELAN transcription program

3.2.1 Orthographic Transcription

The main reason for multi-tier transcription of dialectal data was comparability with other spoken corpora in the CNC, especially the ORTOFON corpus, the facilitation of searching, and help for better lemmatization and tagging. But the richer variability in lexicon, morphology and phonology requires more aggressive standardization on the orthographic layer, which thus differs in some details from the corresponding one in the ORTOFON corpus.

The differences between the orthographic and dialectological tiers cover the following phenomena (the first word shows the transcription on the dialectological tier, the second its orthographic counterpart):

- v-prothesis is kept (***v**okno* > ***v**okno*), but h-prothesis is not (***h**erteple* > *erteple*)
- regional variants of vocalic changes are leveled on the orthographic tier: *k**ú**řilo s**a*** > *k**ou**řilo s**e***, *s**ej**tko* > *s**í**tko*
- regional variants of consonantic changes as well: *sva**r**eb* > *sva**t**eb*, *s**k**oval* > *sc**h**oval*, *kame**ň*** > *káme**n***

Other phenomena (e.g. vowel quantity, full form of consonant clusters and formally reduced variants, regional variants in declination and conjugation) are treated the same on the orthographic layers of both corpora.

3.2.2 Dialectological Transcription

The transcription rules for the dialectological layer are based on the usual conventions in the field of Czech dialectology.[9] This layer includes some specific symbols for dialectal vowels or consonants in order to capture the actual pronunciation, e.g. *vərch*, *býł*, *won*, *řezňičił*. In contrast, word boundaries are kept according the standard orthography and we use unrestricted syntactic punctuation, e.g. marking direct speech using quotes “”. Capital letters appear only at the beginning of proper names, like on the orthographic layer.

4 LEMMATIZATION AND TAGGING[10]

Even though the issue of lemmatization and tagging of spoken Czech has been discussed many times, practical attempts have been comparatively few, e.g. [9], [12], [13]. It is closely connected to the type of data (monologues, dialogues), and especially transcription rules, e.g. how the transcription is segmented, which type of punctuation is used, how much the transcript reflects real pronunciation etc. We decided to develop a pragmatically-minded custom solution based on existing and openly available tools, even though these are designed for written language.

The lemmatization and morphological tagging of both new spoken corpora were conducted according to the same process recently applied to the ORAL series [18]. We took the Czech morphological dictionary MorfFlex CZ [11] as a basis which has been manually and semi-automatic extended or cleaned according to the target register. The extensions refer mainly to register- and/or region-specific items, either full lexemes (lemmas *zbroják*, *škodárna*, *ikspéčka*) or inflectional variants (e.g. lemma *neděle* has two acc. sg. variants, *neděli* and *nedělu*), which were not

[9] We mostly follow the *Rules for the Scientific Transcription of Dialectological Records of Czech and Slovak* [8], but also take some inspiration from *Czech Dialectal Texts* [19] and the *Addenda to the Czech Linguistic Atlas* [5].

[10] For more information about lemmatization and tagging of the ORAL corpora see [18].

contained in the original morphological dictionary. Unsurprisingly, what makes lemmatization and tagging even remotely possible is the presence of an orthographic layer which is fairly close to standard language, at least in terms of transcribing the individual word forms.

5 CONCLUSION

Taken together, the ORTOFON and DIALEKT corpora allow users to research diachronic and diatopic variation in spoken Czech language through a convenient interface. Compared to previous spoken corpora built at the ICNC, they feature a more detailed annotation separated into several parallel layers accommodating speakers individually. The multi-tier transcription allows us to reserve one layer in both corpora for capturing pronunciation detail (be it from a phonetic – as in ORTOFON – or dialectological – as in DIALEKT – perspective), and another (called orthographic in both corpora) for general transcription. The orthographic layer serves as the basis for lemmatization and tagging of both spoken corpora.

This multi-tier transcription also presents challenges when indexing the corpora for querying with corpus tools which require a single authoritative tokenization of the text. A rigorous token-level alignment between the two tiers must be maintained at the transcription stage (as in the case of the ORTOFON corpus) or reconstructed (in the case of DIALEKT) in order to correctly link each token on the main layer with the corresponding token on the dependent layer.

A rich set of both context-dependent and demographic metadata provides additional perspectives on the collected material; especially the DIALEKT corpus provides useful information to researchers from related fields (sociologists, ethnographers, historians etc.). Both lines of data collection, as represented by the ORTOFON and DIALEKT corpora, will hopefully continue into the future.

ACKNOWLEDGEMENTS

This paper resulted from the implementation of the Czech National Corpus project (LM2015044) funded by the Ministry of Education, Youth and Sports of the Czech Republic within the framework of Large Research, Development and Innovation Infrastructures.

References

[1] Balhar, J. et al. (1992–2011). *Český jazykový atlas*. 6 sv. Academia, Praha.

[2] Balhar, J. et al. (2011). *Český jazykový atlas. Dodatky*. Academia, Praha.

[3] Benešová, L., Waclawičová, M., and Křen, M. (2013). ORAL2013: reprezentativní korpus neformální mluvené češtiny. ÚČNK FF UK, Praha. Accessible at: `http://korpus.cz`.

[4] Crowdy, S. (1993). Spoken Corpus Design and Transcription. *Literary and Linguistic Computing* 8(4):259–265.

[5] Čermák, F., Adamovičová, A., and Pešička, J. (2001). *PMK (Pražský mluvený korpus): přepisy nahrávek pražské mluvy z 90. let 20. století*. Ústav Českého národního korpusu FF UK, Praha. Accessible at: `http://www.korpus.cz`.

[6] Čermák, F. et al. (2007). *Frekvenční slovník mluvené češtiny*. Karolinum, Praha.

[7] Čermák, F. (2009). Spoken Corpora Design: Their Constitutive Parameters. *International Journal of Corpus Linguistics*, 14(1):113–123.

[8] Dialektologická komise České akademie věd a umění (1951). *Pravidla pro vědecký přepis dialektických zápisů českých a slovenských*. Česká akademie věd a umění, Praha.

[9] Feagin, C. (2002). Entering the community: Fieldwork. In Chambers, J. K., Trudgill, P., and Schilling-Estes, N., editors, *The Handbook of Language Variation and Change*, pages 20–39, Blackwell Publishing, Malden, MA.

[10] Goláňová, H., Kopřivová, M., Lukeš, D., and Štěpán, M. (2015). Kartografické a geografické zpracování dat z mluvených korpusů. *Korpus – gramatika – axiologie*, 11:42–54.

[11] Hajič, J. and Hlaváčová, J. (2013). MorfFlex CZ. Univerzita Karlova v Praze, MFF, ÚFAL, Praha.

[12] Hlaváčková, D. (2001). Korpus mluvené češtiny z brněnského prostředí a jeho morfologické značkování. *Slovo a slovesnost*, 62(1):62–70.

[13] Hlaváčková, D. and Osolsobě, K. (2008). Morfologické značkování mluvených korpusů, zkušenosti a otevřené otázky. In Kopřivová, M. and Waclawičová, M., editors, *Čeština v mluveném korpusu*, pages 105–114, Nakladatelství Lidové noviny / Ústav Českého národního korpusu, Praha, Czech Republic.

[14] Kloferová, S. (2000). *Mluva v severomoravském pohraničí*. Masarykova univerzita, Brno.

[15] Kopřivová, M. and Waclawičová, M. (2006). *ORAL2006: korpus neformální mluvené češtiny*. Ústav Českého národního korpusu FF UK, Praha. Accessible at: `http://www.korpus.cz`.

[16] Kopřivová, M., Goláňová, H., Klimešová, P., and Lukeš, D. (2014). Mapping Diatopic and Diachronic Variation in Spoken Czech: the ORTOFON and DIALEKT Corpora. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, pages 376–382, European Language Resources Association, Reykjavík, Iceland.

[17] Kopřivová, M., Goláňová, H., Klimešová, P., Komrsková, Z., and Lukeš, D. (2014). Multi-tier Transcription of Informal Spoken Czech: The ORTOFON Corpus Approach. In *Complex Visibles Out There*, pages 529–544, Univerzita Palackého v Olomouci, Olomouc, Czech Republic.

[18] Kopřivová, M., Komrsková, Z., Lukeš, D., and Poukarová, P. (2017). Korpus ORAL: sestavení, lemmatizace a morfologické značkování. *Korpus – gramatika – axiologie*, 15:47–67.

[19] Lamprecht, A. and Michálková, V., editors (1976). *České nářeční texty*. SPN, Praha.

[20] Lukeš, D., Klimešová, P., Komrsková, Z., and Kopřivová, M. (2015). Experimental tagging of the ORAL series corpora: Insights on using a stochastic tagger. In Král, P. and Matoušek, V., editors, *TSD 2015, LNAI 9302*, pages 342–350, Springer International Publishing.

[21] Sloetjes, H. and Wittenburg, P. (2008). Annotation by Category: ELAN and ISO DCR. In *LREC 2008: Sixth International Conference on Language Resources and Evaluation*, pages 816–820. Accessible at: `http://www.lrec-conf.org/proceedings/lrec2008/summaries/208.html`, retrieved 2017-07-31.

[22] Sochová, Z. (2001). *Lašská slovní zásoba*. Academia, Praha.

[23] Waclawičová, M., Kopřivová, M., Křen, M., and Válková, L. (2008). *ORAL2008: sociolingvisticky vyvážený korpus neformální mluvené češtiny*. Ústav Českého národního korpusu FF UK, Praha. Accessible at: `http://www.korpus.cz`.

WHAT DOES *ŽE JO* (AND *ŽE NE*) MEAN IN SPOKEN DIALOGUE

ZUZANA KOMRSKOVÁ
Faculty of Arts, Charles University, Prague, Czech Republic

KOMRSKOVÁ, Zuzana: What Does *že jo* (and *že ne*) Mean in Spoken Dialogue. Journal of Linguistics, 2017, Vol. 68, No 2, pp. 229 – 237.

Abstract: The goal of this paper is to examine the role of two collocations (*že jo* and *že ne*) in spoken dialogue. Both are said to be typical of spontaneous conversation and express a large scale of pragmatic functions, e.g. uncertainty of the speaker or a request for a backchannel. The examination of their positioning within the utterance in relation to the meaning of their close context helped us to identify the functions and to distinguish between cases which are simple co-occurrences of the conjunction *že* and the particle *jo/ne*, and those which are instances of the set phrase. The source material comes from the ORAL2013 and DIALOG corpora.

Keywords: DIALOG, spoken corpora, ORAL, co-occurrence

1 INTRODUCTION

Spoken communication offers many phenomena worthy of study which are closely related to speakers' pragmatic needs. Almost every spoken interaction is conducted with subjective goals and the speakers use various means, strategies, or methods to achieve them (perhaps with the exception of small talk, which is a purpose in and of itself; its participants follow a special goal: to be socially active). Pragmatics is a part of linguistics which focuses on the relationship between context and meaning, e.g. how language users are able to overcome ambiguity, how they understand each other with minimal use of verbal language or how they employ linguistic meanings with respect to social role, relationships etc.

The structure of spoken dialogue was first examined by Harvey Sacks and his fellow collaborators (e.g. [15]). Their methodology, called conversation analysis (CA), tries to uncover the system hidden in conversation routines, which e.g. allows the relatively regular changes of all speakers. Conversation analysis has been applied to several types of linguistic data, from spontaneous conversation to formal and moderated dialogues. Although the results of CA studies are interesting and inspiring, they are rather qualitative and based on small data samples, because the practitioners of CA are mainly interested in the details of every turn including situational circumstances. On the other hand, a traditional corpus study is quantitative, trying to identify patterns or special phenomena with the help of statistics or the functionality offered by corpus managers (e.g. collocation analysis, data sorting by different criteria). However, many research questions in linguistics would benefit from an integrated perspective, both qualitative and quantitative, on the examined phenomenon. This article focuses on spoken language captured in spoken corpora; therefore, it tries to take advantage of both methodologies as much as possible.

DE GRUYTER OPEN
DOI 10.1515/jazcas-2017-0032

Spoken conversation research has shown that the notion of ‘sentence’, which is one of the basic units in written language, is not ideal when considering spoken structure and syntax (cf. [7]). In addition to its problematic definition, two different persons tend not to agree on its boundaries in most cases.[1] There are many other proposals for a unit of spoken language, e.g. C-unit [10], AS-unit [5], but none of them has been widely accepted by the linguistics community. This article does not try to propose any new unit for speech analysis, but it considers one type of tag questions to be a marker of shorter syntactic structures within an utterance. The two Czech collocations (*že jo* and *že ne*) chosen for this article brought us to the broader topic of pragmatics in spoken language, especially to the meaning and status of collocations and their position within utterances, having at first sight little discernible structure.

2 THEORETICAL BACKGROUND

Tag questions or question tags (hereafter QT) are one of the characteristic devices used in spoken language. They may be considered as one type of questions, besides the yes/no and wh- questions (which is mostly a teaching perspective, e.g. [13]), or as an indirect form of question, transmuted into a statement with question particles or short set phrases appended, e.g. [3, p. 155].

In many European languages, these tags are realized by invariant forms, e.g. *nicht wahr?* in German, *n'est-ce pas?* in French, *že ano?* in Czech. The canonical English QTs are a reduced interrogative clause whose structure and lexical content is conditioned by the format of the anchor or host clause to which it is appended [6]. It consists of an auxiliary or a modal verb, a pronoun (or *there*), and an optional negative particle (typically the enclitic *n't*).

The function of QTs is mainly a request for more information, but they also perform a whole range of additional pragmatic functions. Hoffmann [6] provides a further classification, e.g. confirmatory tags, which express that the speaker is unsure about what s/he is saying; peremptory and aggressive tags, which are employed to close a discussion or to provoke and insult other speakers (cf. [1]); punctuational tags, which are employed to emphasize what the speaker says and do not expect any involvement or reply by the conversational partner. Rühlemann [14, p. 93] maintains that they simultaneously function as support markers or backchannels, as turn yielders, which Leech [10] shows in an analysis of Czech narrations. Rühlemann [14] states that one of their basic functions is also an invitation to co-construction and relates that to intonation. Lukásci [11] groups the various functions into two broader categories: epistemic modal, or affective. Epistemic modal tags express the speaker's uncertainty and thus are related to content, while affective tags show politeness. Sacks [15, p. 718] consider QTs a generally available exit technique for a turn.

Since QTs are characteristic for spoken communication, their description usually includes intonation patterns. Their classification is often clarified through

[1] This is one reason why transcriptions of spoken Czech captured in the ORAL2013 corpus [2] and its follower ORTOFON [9] feature pausal punctuation.

intonation, e.g. a rise indicates a real question, while a fall seeks confirmation [11]. Müllerová [12] found the tags with rising intonation to be much more frequent (90% in the BNC data) than those with falling intonation. The rising intonation may signalize syntactic co-construction of utterances (by another speaker) and, by extension the co-construction of meaning, when taking the turn. On the other hand, Müllerová [12] relates the intonation of QTs to the previous clause. If the previous clause or syntactic phrase is pronounced with falling/rising terminal intonation, then the same goes for the question tag.

In Czech, the most often mentioned QTs are the following: *viď, víš, (že) jo, (že) ne; see* e.g. [8], [12]. However, there are also several verb forms in the second person which serve as question tags, e.g. *chápeš, rozumíš, víš* etc. This article focuses on two Czech question tags: *že jo* and *že ne,*[2] whose meaning is summed up in [4] as an attitude stressing the uncertainty or truth of an assumption.

3 DATA OVERVIEW

Spoken data were extracted from the ORAL2013 [2] and DIALOG [16] corpora, which cover different types of spoken language. The first one captures the spontaneous conversations typically encountered within the family circle, among friends and relatives in general, in other words in such situations in which the speakers are not really self-conscious about the formal attributes of their speech. On the contrary, the DIALOG corpus contains publicly broadcast speech, collected from a variety of discussion programmes, from talk-shows to political debates. Both corpora include an orthographic transcription capturing all pronounced or unfinished words, hesitations, response noises etc. Ignoring several minor differences, like the amount of metadata or the transcription of overlaps and proper nouns, the comparison of both types of spoken language is possible without need for a complicated conversion between transcription systems.

Tab. 1 shows the frequencies of both tag questions within both corpora. The following search query was limited on the utterance of one speaker: [word="(?i)že"] [word="(?i)j[oó]|n[eé]"] within <sp/>. We decided against including an optional pause or hesitation between both parts of the collocation because there would be much more undesirable homonymy.

	ORAL2013 (i.p.m.)	DIALOG (i.p.m.)
že jo	57 057.77	945.73
že ne	1 016.51	96.42

Tab. 1. Frequency of both collocations within the ORAL2013 and DIALOG corpora

The difference between both corpora could be a result of the DIALOG corpus design, i.e. the different proportion of broadcast programmes. Tab. 2 introduces four

[2] The motivation for choosing these tags and not e.g. *že ano* was driven by their collocation strength measure in spoken corpora. Listing the closest right-sided collocates of the conjunction *že* according to the T-score, MI score and logDice, the type *jo forms the strongest pairing* within both corpora used for this article. The word *jo* may be seen as informal synonym for the response token *ano.* Their mutual antonym is the word *ne*, which forms a collocation with *že*.

genres[3] represented in this corpus, an annotation which was added externally at the Institute of the Czech National Corpus to provide a sub-classification of the programmes. The frequencies of collocations show that talk-show is the closest genre to intimate spontaneous conversation. As part of this ‘entertainment’ genre, besides many prepared questions, the host has to improvise in reaction to the answers.

	talk-show (ipm)	profile (ipm)	discussion (ipm)	debate (ipm)
že jo	3 316.88	1 977.19	847.03	65.14
že ne	114.38	62.24	96.64	99.19

Tab. 2. Frequency of both collocations within the genres of the DIALOG corpus

Looking closer at the titles with the highest frequency of the *že jo* collocation, the majority (27%) comes from the talk-show *Uvolněte se, prosím,* followed by the programmes *Krásný ztráty* (20%) and *Na plovárně* (15%), both included under the profile genre. The host of the programme *Uvolněte se, prosím* Jan Kraus has the highest frequency of this collocation (274 occurrences) among all speakers in the DIALOG corpus.

According to the metadata in the ORAL2013 corpus, the collocation *že jo* occurs more often in Bohemia than Moravia or Silesia. It confirms the observation by [10], that the Moravians and Silesians prefer to use shorter *že* or *ne.*

The amount of data, i.e. thousands of occurrences, precludes a manual analysis of all of them, especially as regards their function within the utterance. For the comparison of intimate and broadcast spoken language, manual analysis was performed on random samples of 115 occurrences. This number has been chosen for two reasons: the total amount of data (4 × 115) is manually manageable, and it is the absolute frequency of the less frequent collocation, i.e. *že ne,* in the DIALOG corpus. In almost every case, it is necessary to listen to a recording to determine function of the occurrence.

4 RESULTS

4.1 Functions of *že jo*

Although we were mainly interested in the pragmatic uses of *že jo*, the context shows other functions as well. The classification shown in Tab. 3 is derived from the data. The first category covers occurrences where the conjunction *že* is obligatory and introduces an object subordinate clause, e.g. *já si mysím že jo* (ORAL2013), *já vím že jo* (DIALOG). The second category describes the main pragmatic role of the QT and will be discussed in further detail under 4.3. The third category includes examples with the pragmatic function of (a request for) confirmation, clearly

[3] The genre of *talk-show* encompasses programmes which should mainly amuse. The genre called *profile* focuses on a single person (the guest) from several perspectives, e.g. a confrontation between the person’s perspective and that of his/her fans, the host etc. The difference between *discussion* and *debate* lies in the task of the participants. *Debates* are primarily political and their participants want to persuade others. In contrast, the participants of *discussions* are rather experts on a given topic who offer their professional opinion but do not need to persuade the viewers.

recognizable by intonation, e.g. sp1: *prášek ti zabere . když ho nejíš soustavně* sp2: *to je pravda* sp1: *že jo ?* (ORAL2013).

	ORAL2013	DIALOG
Valency	6	2
Question tag	103	109
Confirmation	6	3
Unclear	0	1

Tab. 3. Functions of the collocation *že jo* in both corpora within the samples of 115 occurrences

Tab. 3 confirms that the collocation *že jo* is mainly used as QT in both types of spoken data. Even though one third of the *že jo* instances comes from the *Uvolněte se, prosím* talk-show, the composition of a random sample from the DIALOG corpus spans all genres.

4.2 Functions of *že ne*

The analysis procedure was the same. In contrast to the previous collocation, a richer variety of functions was found. The higher frequency of non-pragmatic uses (valency and constituent negation) is due to the fact that the collocation *že ne* was chosen for its similar meaning to the previous one despite its lower frequency in both corpora.

The group of verbs with obligatory object (first group) mainly includes verbs of thinking and speaking, e.g. *myslet, představit si, doufat, obávat se, věřit, říkat, (po) tvrdit.* The identified lexicon was richer in the DIALOG corpus, most likely due to the speech of politicians. The second group will be discussed in the next section. The category of disagreement, the third group, consists of occurrences where the token *ne* was an answer or simple negation, e.g. *a on že ne* (ORAL2013)*, samozřejmě že ne* (DIALOG)*, rozhodne-li parlament že ne* (DIALOG). The negative was often intensified with *samozřejmě, právě, jistě.* The fourth group represents the cases of often emphatic, stand-alone negation applying to the next token, e.g. *kolikrát mu mistr říkal . že ne vítězství po boji ale před bojem* (ORAL2013), *rozdíly sou takové že ne každý kdo by si usmyslel by mohl vyjíždět* (DIALOG). Finally, the collocation *že ne* can also be part of another collocation: *ne že ne.*

	ORAL2013	DIALOG
Valency	72	66
Question tag	11	2
Disagreement	27	34
Constituent negation	3	12
Part of idiom *ne že ne*	0	1
Unclear	0	0

Tab. 4. Functions of the collocation *že ne* in both corpora within the samples of 115 occurrences

The analysis revealed interesting results in contrast to the collocation *že jo.* In the sample from DIALOG, almost one third of occurrences (classified among the first three most frequent groups) comes from the *Sedmička programme, a debate*. The higher incidence of QTs in the ORAL2013 sample may be explained through the higher uniformity of data, i.e. only intimate spontaneous conversation.

4.3 Detailed Analysis of Question Tags *že jo* and *že ne*

This part focuses only on those occurrences which were annotated as QTs; therefore, its results in the second part are related only to the samples.

First, we tried to find some context cues to generalize the characteristics, but no linguistic devices were identified as typical in the right- and left-sided closest context. Both corpora use the question mark as a signal of rising voice (or as a general signal of questions in ORAL2013), the DIALOG corpus even to indicate changes in intonation[4]. A co-occurrence of the collocation *že jo* and the question mark was found in 324 cases in the entire ORAL2013 corpus, i.e. only 4 % of all collocation occurrences, and a change of speaker immediately follows in 228 cases. The co-occurrence of *že ne* and ? was found in 38 cases in the entire corpus, but almost all of these cases present a question stressing a negation, e.g. sp1: *ty ale oni ty ségry nejsou . nemají stejného tatu* ***že ne ?*** sp2: *já mysim že jo* or simple emphasize the need for an answer: sp1: *ty seš u toho topení tebe zima mysim není že ? .* ***že ne ?*** sp2: *hmm ..* On the other hand, in the DIALOG corpus, intonation markers occupied a full third of *že jo* occurrences, though only 29 cases on the right side and 10 on the left side made it into the samples. Relying on the annotation of intonation, it seems that neither QT is pronounced with any distinctive characteristics.

Neither did the marking of pauses show any trend, although they occurred as expected before or after the collocation. In the sample from ORAL2013, pauses were most frequent on the ±1 position, but overall only in 31 cases on the right and 9 on the left side. Therefore, the next step was a more qualitative analysis focusing on function.

The term ‘question tag’ may be a little misleading, because it may associate to the token marking the sentence to be a question with the usual positioning of a tag, as a marker of something additional or extending, beyond that sentence, in a linear perspective after that sentence. Before delving into the issue of terminal position, we will deal with the role of QTs in dialogue.

The first group sums up the use of QTs for answering or rather confirming the main speaker’s statement. This strategy was very often used in talk-shows, e.g. Jan Kraus: *to znamená to vaření bylo trochu jiný* ***že jo*** Zdeněk Pohlreich: *to vaření bylo vo hodně jiný* (DIALOG), sp1: *nebudeš furt v lihu* ***že ne*** sp2: *ne jenom od rána* (ORAL2013). The next sample from ORAL2013 shows that the second speaker does not need an invitation in the form of QT, but reacts immediately and thus simultaneously with the QT, e.g. sp1: *už snad ho poprosili což jako neni se co divit [****že jo****]*[5] sp2: *[no neni]*(ORAL2013).

The second group consists of those QTs which are positioned after a summary of known information or generalisation and thus do not cause any verbal reaction on the part of the second speaker, e.g. *hrát vančurův dialog je samozřejmě daleko obtížnější* ***že jo*** (DIALOG), *no a směrem k Mrtvýmu moři to ubejvá ubejvá . a do Mrtvýho moře skoro už nic nepřitejká* ***že jo*** (ORAL2013). The QT can also introduce an additional expression or parenthesis. The position of the QT could also emphasize

[4] I.e. the question mark ? to indicate a high intonational rise, the comma , for a lower rise, and the full stop . for a fall.

[5] The square brackets [] mark the overlaps.

the following token/part, e.g. *a voni tam začali dávat židle* ***že jo*** *nahoru . jako klasicky ježky . a von chodil a sundával je zase dolů jo* . (ORAL2013)[6].

The QTs could be also used as connectives to change the topic or syntactic structure, e.g. *voni to mu to* ***že jo*** *to je taky proti vlhkosti ne todlencto* (ORAL2013). This use is closely connected with false starts or restarts, i.e. when the speaker does not prepare his/her speech or right words and produces many fillers, e.g. *to vypadá prostě fak to je to mmm . to vypadá . to prostě* ***že jo*** *to s* sám víš velmi dobře . že prostě todle vypadá dycky hrozně dobře . dyž si to takhle udělám že jo* (ORAL2013), *protože mm* ***že jo*** *pořád trochu se ňákym způsobem ta komunikace vázla* (DIALOG). The second use of *že jo* in the previous example could be included into the first group.

These three groups also show the position of QTs. In many of the examples mentioned (e.g. all examples in the first group), the position could be called terminal or final in relation to the complex syntactic unit, including verbal and nominal phrases. The intermediate or central position would consist of an insertion of QT between the head of a nominal phrase and its complement, e.g. adjective/pronoun and noun in *akorát ty vnitřní ty* ***že jo*** *rozměry nebudou takový* (ORAL2013), or adverbial attribute in *vy ste dokonce byl na stáži* ***že jo*** *ve spojených státech* (DIALOG). However, there are many occurrences where the QT introduces e.g. the sentence topic (from a functional sentence perspective point of view) or words which had previously in the sentence been substituted by pronouns, e.g. *co je v tom* ***že jo*** *v té středověké metalurgii se začínalo tím že se dělalo to dřevěné uhlí* (ORAL2013). This brings us back to the issue of sentences in spoken language.

Looking closer at utterances with several occurrences of QTs in examples (1) and (2), we believe that the distinction between the terminal and central position will help us identify the possible borderlines of syntactic units in speech.

(1): *[ale tak to ale tak to byla jen vyjímka]* ***že jo*** *že sem prostě šel plavat sem si mezitím uďál věci co sem potřeboval* ***že jo*** *ale normálně* ***že jo*** *dyž se pak budu chtít ráno vosprchovat* ***že jo*** *. tak tam tak je takováhle fronta* ***že jo*** *. zvláště dyž všichni vyrážíme ve stejnou dobu* . (ORAL2013)

(2): *a teď jako se tam takhle válel a teď říká super polohovací . a teď jak ta postel* ***že jo*** *se polohuje . no to je pecka . takhle tam hejbal tim zadkem* ***že jo*** *s tou postelí . pořád si šahal rukama do těch trenýrek . pořád se tam drbal na těch jako . koulích a kamarádka uplně v šoku* (ORAL2013)[7]

The occurrences of QTs in (1) are often followed by conjunctions, which introduce (relative) clauses. On the contrary in (2), the QTs are placed within the

[6] The first part of the utterance until the pause . is pronounced without hesitation or drawing a breath, as one complex unit, therefore we believe the parenthesis is rather the part *jako klasicky ježky* framed on both sides by pauses.

[7] The word *jako* occurs in a position similar to that of the collocation *že jo*, i.e. its first occurrence is in the initial cluster introducing the new interesting topic *a teď jako se tam takhle válel* and within the verb and its adverbial complement *drbal na těch jako . koulích;* the second occurrence of *jako* may be caused by shyness, an attempt to delay uttering the final word, or perhaps to find another, more appropriate one.

syntactic unit, in which they stress the following verb/prepositional phrase. The crucial step in finding the syntactic units is to find the finite verb. In any analysis of spoken language, the temporal aspect should not to be ignored, which means we need to account for how the speakers react to one another, e.g. in (3).

(3)
sp1: *jako kdybys šel tady [do Pančáku . přes poledne menu]* ..
sp2: *[to maj přes poledne jako menu no tak tady] máš*
sp1: *[tak to maš d* . dvojnásobně] ..* ***že jo*** *. než tady*
sp2: *[tak to maš d* . dvojnásobně]*

5 CONCLUSION

This article dealt with two Czech collocations, *že jo* and *že ne,* and their use in spoken language. Both collocations were analysed within two corpora of spoken Czech, namely ORAL2013 and DIALOG, which capture two different types of spoken data: informal spontaneous speech among friends and family members, and broadcast speech of hosts and their guests (celebrities, politicians etc.). The main goal of the analysis was to identify occurrences of both collocations in their specific use as question tags.

The analysis was conducted on four samples. Both samples of the collocation *že jo* showed predominant use as question tag, unlike both samples of *že ne,* where the co-occurrence of conjunction *že* and negative particle *ne* was detected rather than the collocation *že ne* per se. This finding was in accordance with the collocation measures.

The subsequent analysis tried to distinguish and define functions of the question tags. Comparing the results with [6], the confirmatory tag is the most frequent type. My data confirm the statement by [14] that question tags simultaneously function as backchannels. In addition, one of the basic functions was as a sentence topic marker, positioned directly before the part of the sentence to emphasize. Unlike the other functions, this does not respect the borderlines of syntactic phrases within the utterance. Although the question tags provide a useful division of speech, their position is closely related to their function and this should be always kept in mind, especially with respect to the broader debate about the spoken ‘sentence’. The distinctions between pragmatic and non-pragmatic (i.e. valency) use of both collocation could be provided within orthographic transcription (the simplest way could be to transcribe the pragmatic QT as a one word *žejo/žene*) which would be helpful for tagging as well.

ACKNOWLEDGEMENTS

The present article was written under the auspices of grant project No. 15-01116S (Syntax mluvené češtiny) [Syntax of spoken Czech]), supported by the Grant Agency of the Czech Republic.

References

[1] Algeo, J. (1990). It's a Myth, Innit? Politeness and the English Tag Question. In Ricks, Ch. and Michaels, L., editors, *The State of the Language*, pages 443–450, University of California Press, Berkeley, US.

[2] Benešová, L., Waclawičová, M., and Křen, M. (2013). ORAL2013: reprezentativní korpus neformální mluvené češtiny. ÚČNK FF UK, Praha. Accessible at: `http://korpus.cz`.

[3] Čermák, F. (2007). *Jazyk a jazykověda*. Karolinum, Praha.

[4] Čermák. F., Hronek, J., and Machač, J. (2009): *Slovník české frazeologie a idiomatiky 2, výrazy neslovesné*. Leda, Praha.

[5] Foster, P., Tonkyn, A., and Wigglesworth, G. (2000). Measuring spoken language: a unit for all reasons. *Applied Linguistics*, 21(3):354–375.

[6] Hoffmann, S. (2006). Tag questions in early and modern late English: historical description and theoretical implication. *Anglistik,* 17(2):35–55.

[7] Hoffmannová, J. and Zeman, J. (in print). Výzkum syntaxe mluvené češtiny: vstupní inventarizace problémů.

[8] Kolářová, I. (1997). *Vliv komunikace na významové posuny některých sloves*. UJEP, Ústí nad Labem.

[9] Kopřivová, M., Komrsková, Z., Lukeš, D., Poukarová, P., and Škarpová, M. (2017). *ORTOFON: korpus neformální mluvené češtiny s víceúrovňovým přepisem, verze 1 z 2. 6. 2017*. Ústav Českého národního korpusu FF UK, Praha. Accessible at: `http://www.korpus.cz`.

[10] Leech, G. (2000). Grammars of spoken English: new outcomes of corpus-oriented research. *Language Learning*, 50(4):675–724.

[11] Lukásci, Z. (2009). Language and gender. How question tags are classified and characterised in current EFL material. In Lugossy, R., Horváth, J., and Nikolov, M., editors, *UPRT 2008: Empirical studies in English applied linguistics*, pages 191–205, Lingua Franca Csoport, Pécs, Hungary.

[12] Müllerová, O. (2007). Postpozitivní *žejo*, *jo* ve vzpomínkovém vyprávění. In Hoffmannová, J. and Müllerová, O., editors, *Čeština v dialogu generací*, pages 291–297, Academia, Praha, Czech Republic.

[13] Quirk, R. and Greenbaum, S. (1993). *A university grammar of English* (27th ed.). Longman, Harlow.

[14] Rühlemann, C. (2007). *Conversation in context: a corpus-driven approach*. Continuum, London.

[15] Sacks, H., Schegloff, E., and Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4):696–735.

[16] Ústav pro jazyk český (2012). Korpus DIALOG 1.1. Ústav pro jazyk český, AV ČR, Praha. Accessible at: `http://ujc.dialogy.cz`.

GRAMMATICAL CHANGE TRENDS IN CONTEMPORARY CZECH NEWSPAPERS[1]

MICHAL KŘEN
Institute of the Czech National Corpus, Charles University, Prague, Czech Republic

KŘEN, Michal: Grammatical Change Trends in Contemporary Czech Newspapers. Journal of Linguistics, 2017, Vol. 68, No 2, pp. 238 – 248.

Abstract: The paper presents a corpus-driven method for the detection of recent grammatical change in contemporary Czech newspapers. It is based on a large and homogeneous material (825 million tokens of a single newspaper) that covers a 23-year time span. The task is operationalised into finding the most relevant frequency change manifested by selected subsets of the Czech tagset. The results show changing proportions of parts of speech, nominal cases etc. that indicate a shift towards more "verbal" language associated with increasing informality of the newspaper register.

Keywords: modern diachrony, language change, Czech, newspaper register, corpus composition

1 INTRODUCTION

The paper aims to investigate recent grammatical change that can be observed in contemporary Czech newspapers. It presents an automatic corpus-driven method used to detect morphological features that show the most considerable diachronic shift. Finally, the results as well as the limitations of such an approach are discussed.

The paper draws on previous research done mostly on English trying to detect recent language change [1], [3], [8], [9], [12], [13], [14]. Compared to them, this study can be characterised by the following:

- it is based on large and homogeneous data;
- morphological categories (rather than the often studied individual word forms) are investigated systematically and in a corpus-driven manner; this has been operationalised into finding the most relevant frequency changes manifested by selected subsets of the currently used Czech tagset;
- evaluation of the frequency differences is carried out using Mann-Kendall test and Theil-Sen estimator.

2 DATA

It is often emphasised that research aiming to discover recent language change should be based on large and homogeneous data covered by many data points [9], [14]. This has determined selection of SYN v4 as the base corpus [11]. With its 4.3

[1] This study was written within the programme Progres Q08 *Czech National Corpus* implemented at the Faculty of Arts, Charles University.

DOI 10.1515/jazcas-2017-0033

billion tokens (3.6 billion running words), it is the largest available traditional (as opposed to web-crawled) corpus of contemporary written Czech featuring reliable metadata and large homogeneous newspaper subcorpora. SYN v4 is uniformly processed, which includes text cleanup, de-duplication and other rather technical issues, as well as lemmatisation and morphological tagging [7], [10].

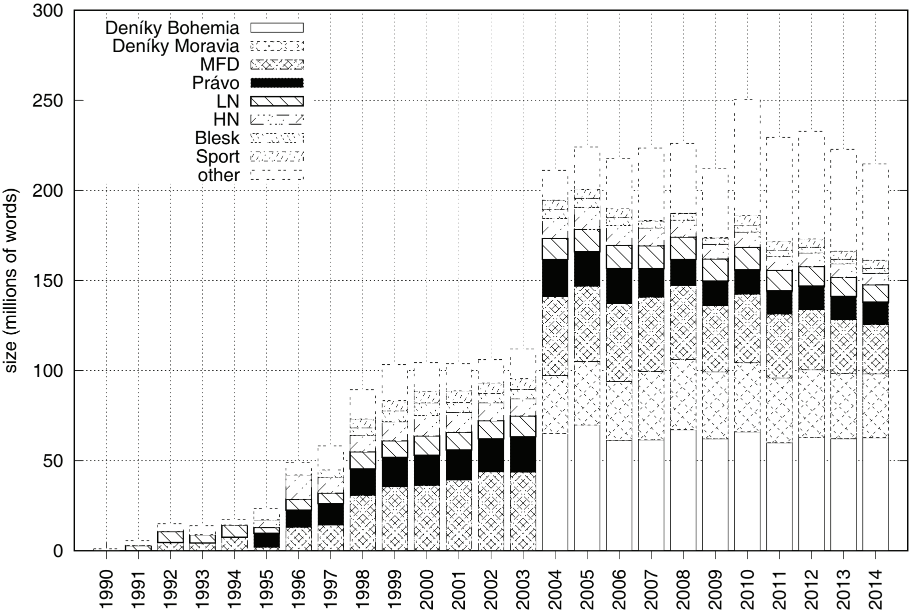


Fig. 1. Composition of the newspaper part of SYN v4

Composition of the newspaper part of SYN v4 is given in Figure 1.[2] As a rule, all the texts are incorporated in full, which means that there are always whole newspaper issues included in SYN v4. For the present study, a part of SYN v4 that contains only a single major national daily newspaper *Mladá fronta DNES* (MFD) was used. Its total size is 825 million tokens (687 million running words) and it covers the period 1992–2014, which means that it is the largest newspaper in terms of size and time span.

Subsequently, a virtual subcorpus of MFD was created for each year of the given period that was used for all the queries described in Section 3. It should be noted that some MFD issues are missing in SYN v4, especially from 1992–1995 (see Figure 2). However, this shortcoming should not distort the overall picture of the language used in MFD at the time and it is presumably outweighed by the greater number of data points available [14, p. 208].

[2] More information can be found at `https://wiki.korpus.cz/doku.php/en:cnk:syn:verze4`.

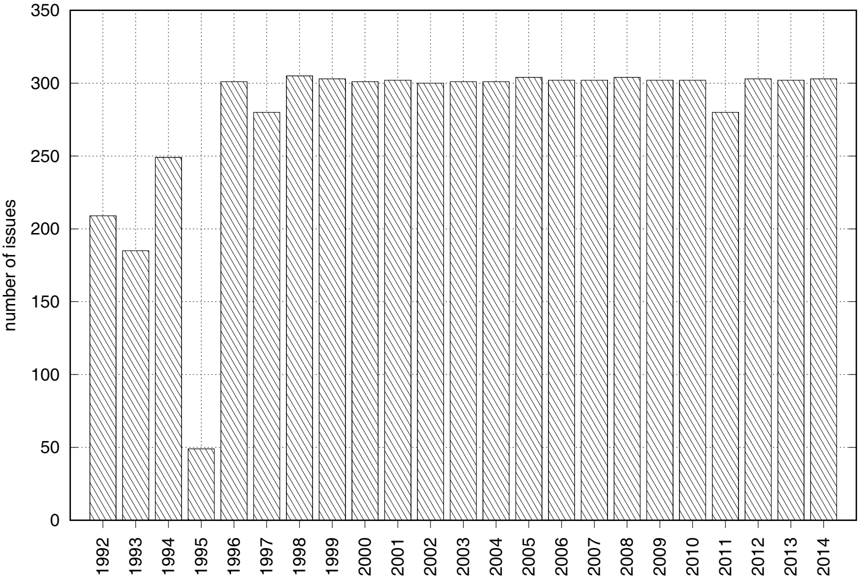


Fig. 2. Number of MFD issues per publication year in SYN v4

3 METHOD

3.1 Morphological Categories

The method is based on the Czech tagset currently used in the Czech National Corpus (CNC). The tagset draws on the original one developed by Jan Hajič [4] with some improvements and extensions.[3] Czech is a morphologically rich language and this is reflected also in the tagset: there are 4 351 different tags actually used in SYN v4. The tagset is positional, which means that each position in the tag (viewed as a string) represents a single morphosyntactic feature. For instance, one of the possible morphological tags for the word form *nejasnější* ('less clear') is `AAFS7----2N-----` which denotes the following features: adjective (A), regular adjective (A), feminine (F), singular (S), instrumental (7), comparative (2), negated form (N). For features not relevant for the given POS, '`-`' is used on the respective position.

The tags are very fine-grained, which means that their development over time would hardly show any convincing trend. Instead, it would be optimal to discover the trends for all possible tag combinations, and then to choose the most significant from among them. Although this procedure would guarantee that no relevant combination (in terms of the original tagset) is missed, it is also not feasible given the exponential size of the set of all possible subsets.

Therefore, instead of grouping the individual tags, *categories* were introduced that represent various morphological "dimensions", e.g. adjectives, feminine adjectives, adjectives in instrumental singular, negated adjectives, instrumental

[3] Detailed information available at `https://wiki.korpus.cz/doku.php/seznamy:tagy` (Czech only).

singular in general (without any POS restrictions), etc. The categories are defined by *regexps* (regular expressions) over the tagset that result from the expansion of variables of manually input *patterns*; the variables denote any possible value on a given position. The regexps are finally turned into the individual CQL *queries* for the Manatee query engine [15].

For instance, the pattern `${1}.*` with a variable on the first position was expanded into separate regexps for all possible values of POS, e.g. `N.*` for nouns, `A.*` for adjectives, `V.*` for verbs etc. The variables could also be freely combined to yield more complex sets of regexps for each pattern, for instance `${1}...${5}.*` with variables on the first and on the fifth position expanded into all combinations of POS and case, e.g. `N...1.*`, `N...2.*`, `N...3.*`, ... , `A...1.*`, ... , `A...7.*` etc. For some categories, CQL regular expressions were used directly (e.g. `P[PH5].*` for personal pronouns) and they were also combined with variables (e.g. `P[PH5]..${5}.*` for personal pronouns in every particular case).

The total number of input patterns was 161. After the expansion, they yielded 128 557 regexps that define the individual categories on various levels of granularity in different morphological dimensions. The regexps were turned into CQL queries simply by wrapping them into `[tag=”regexp”]`. Finally, all the CQL queries were run against the individual MFD subcorpora of the SYN v4 corpus using Manatee API. For every query, a *data row* of normalised frequencies was the result, showing the development of the particular category in MFD over time. As the pattern expansion mentioned above heavily overgenerates, many resulting queries gave no results, e.g. `[tag=”V...7.*”]` (verbs in instrumental). All such data rows were simply discarded and not included into the evaluation.

3.2 Evaluation

All non-zero data rows (4 628 in total) were used as an input into the statistical module that was employed to detect those with the most significant frequency development. Two methods have been used: Mann-Kendall test and Theil-Sen estimator.

Mann-Kendall is a non-parametric statistical test used to measure the correlation between ranks of two variables that is often used to assess the (upward or downward) monotonicity of the trend of the observed variable over time [6], [9]. Its values are in the <-1;1> range: 1 for perfect agreement (both variables increase/decrease simultaneously), -1 for perfect disagreement (the opposite of the above), 0 if there is no correlation observed. However, Mann-Kendall does not take into account the actual values as long as their rank order over the time remains the same. It also gives clear preference to smooth, monotonous frequency change which may not be the case in reality.

Therefore, it was supplemented by the Theil-Sen estimator that is also used for automatic detection of development trends in the Sketch Engine software (based on [5]). It is a robust linear regression method that computes a median slope of the overall trend. Theil-Sen overcomes local fluctuation in the observed trend and, at the same time, it naturally takes into account the actual frequency values (overall increase or decrease). This means that both methods complement each other.

4 Results

Evaluation of the results is complicated by the fact that most of the categories are multi-dimensional and interrelated with other ones. For instance, there is a slight increase observed for nouns in accusative, while there is a decrease for nouns in general and an increase for accusative (regardless of POS); the overall picture is presumably much more complex. At the same time, there are too many results to show as the space in this study is limited.

Therefore, two tables are presented, one for each method, with 15 items per table. The items have been selected as the most significant as detected by the given method while omitting near duplicates, e.g. `..FS2.*` vs. `..F.2.*` (‘F’ for feminine, ‘2’ for genitive), where the difference is only in the (un)specification of the number (‘S’ for singular).

For every item, the following is given:

- exact query in terms of the CNC tagset;
- value obtained by the respective method (Mann-Kendall or Theil-Sen);
- rank according to that method;
- relative frequencies (instances per million) in 1992 and 2014 (the first and the last year of the data row);
- overall increase or decrease (trend);
- characterisation of the query (category).

Query	value	rank	i.p.m. (1992)	i.p.m. (2014)	trend	category
[tag=“PDN.1.*“]	0.96	3	2604	5291	+	pronoun, demonstrative, neutral, nominative (mostly the form *to*)
[tag=“PH..4.*“]	0.95	8	610	1495	+	pronoun, personal in short form, accusative
[tag=“P5F.6.*“]	0.95	13	103	165	+	pronoun, personal after prep., feminine, locative (*ní*)
[tag=“P[PH5].*“]	0.94	25	7605	11848	+	pronoun, personal
[tag=“A...2.*“]	-0.94	32	31436	22036	-	adjective, genitive
[tag=“N...2.*“]	-0.94	34	88698	75080	-	noun, genitive
[tag=“PH.*“]	0.94	35	1476	3001	+	pronoun, personal in short form
[tag=“AG.P.*“]	-0.93	43	937	597	-	adjective, derived from present transgressive, plural
[tag=“..F.2.*“]	-0.93	46	49992	41048	-	feminine, genitive (any POS)
[tag=“P[PH567].*“]	0.93	56	23302	33128	+	pronoun, personal or reflexive
[tag=“P6..7.*“]	0.93	57	112	221	+	pronoun, reflexive in long form, instrumental (*sebou*)
[tag=“...S2.*“]	-0.93	58	85829	69032	-	singular, genitive (any POS)
[tag=“Vs.........P...P“]	-0.93	61	4477	1943	-	verb, passive participle, perfective
[tag=“P5.S4.*“]	0.92	69	161	336	+	pronoun, personal after prep., singular, accusative
[tag=“VB.....1F.*“]	0.92	71	319	582	+	verb, future tense, 1st person

Tab. 1. Mann-Kendall

Query	value	rank	i.p.m. (1992)	i.p.m. (2014)	trend	category
[tag=“N.*“]	-1316	1	313489	285839	-	noun
[tag=“....2.*“]	-1161	3	151214	128874	-	genitive (any POS)
[tag=“V.*“]	1146	4	115603	142042	+	verb
[tag=“...........A.*“]	1075	6	89508	116202	+	active voice
[tag=“VB.*“]	912	12	49506	66430	+	verb, present or future form
[tag=“V..............I“]	899	13	71988	90781	+	verb, imperfective
[tag=“........P.*“]	851	19	45479	61577	+	present tense
[tag=“A.*“]	-775	23	107316	91097	-	adjective
[tag=“P.*“]	727	26	60170	74222	+	pronoun
[tag=“....4.*“]	699	29	117906	135418	+	accusative (any POS)
[tag=“V..S.*“]	648	38	68746	84743	+	verb, singular
[tag=“.........1.*“]	-602	43	107619	95130	-	positive (comparison degree; adjectives and adverbs)
[tag=“.......3.*“]	591	44	49302	62770	+	3rd person (pronouns and verbs)
[tag=“N..S2.*“]	-562	48	60579	50022	-	noun, singular, genitive
[tag=“D.*“]	558	49	45980	57983	+	adverb

Tab. 2. Theil-Sen

To comment on the methods briefly, Theil-Sen tends to prefer more “global” categories (e.g. whole parts of speech), because they are more frequent. On the other hand, Mann-Kendall prefers more detailed categories (e.g. sub-part of speech combined with gender and/or case, sometimes even specific enough to single out a word form) that show monotonous development over time (by definition, monotonicity is the only evaluation criterion for Mann-Kendall).

It should be pointed out that the i.p.m. values given in both tables should be viewed as boundaries, as the i.p.m.’s for the individual years between 1992 and 2014 often develop evenly within this range (this is caused by the nature of the methods employed, especially the Mann-Kendall). Figure 3 and Figure 4 show examples of such development, one from each table, that depict one increasing and one decreasing trend. What can be observed is thus gradual, smooth and continuous frequency change of the individual grammatical categories.

In terms of the parts of speech, there is a steady increase observed for verbs, pronouns and adverbs that is complemented by the decrease of nouns and adjectives (Table 2). Even more significant change can be observed within the individual POS: Table 2 suggests that verbs in 3rd person, singular, present or future form,[4] imperfective, active voice are in the lead of the change (it is perhaps worth mentioning that all these categories constitute unmarked forms of a verb). Similarly, the increase of pronouns can be ascribed to demonstrative, personal and reflexive pronouns that often show strikingly monotonous trends (cf. Table 1 and Figure 3).

[4] Perfective verbs in Czech form future tense by their morphologically present form. This is reflected by the `VB.*` tag that denotes morphologically present verb forms.

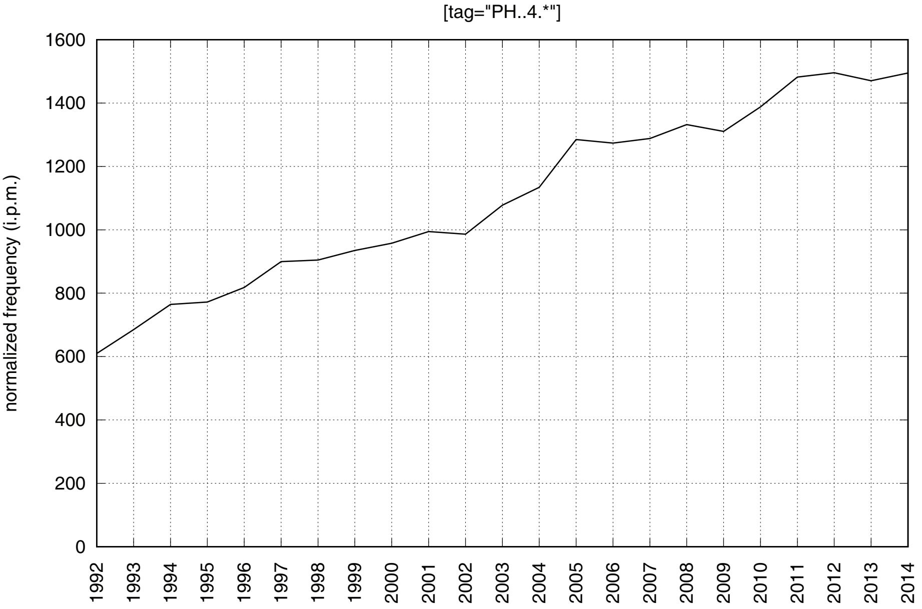


Fig. 3. Personal pronouns (short forms) in accusative

The detected trends can have various (and interrelated) causes: frequency change of the individual parts of speech suggests a shift towards more "verbal" language associated with a diversion from nominal expressions. This corresponds to the gradual move of the newspaper register towards fiction [2], [9], although one of the reasons is likely to be the growing proportion of leisure themes, interviews, weekend supplements etc. in the MFD subcorpora.

Numerous papers on recent language change also report increasing informality of the newspaper register [1], [3], [8], [9], [12], [13], [14]. This is seconded also in this study and illustrated by the decrease of rather formal expressions, namely passive participles and adjectives derived from the present transgressive (cf. Table 1).

As for the other morphological categories, there is an increase observed for accusative and decrease for genitive, regardless of POS. Given the gradual nature of this frequency shift, it is unlikely to be affected by disambiguation errors. However, one should be very cautious about its possible interpretation as an indication of a long-term typological change. Certainly more reasonable cause could simply be changing structure of MFD, perhaps also within its individual sub-registers, which will be discussed in the following paragraphs.

In order to investigate the composition of MFD and its possible influence on the presented results, the newspaper sections markup newly introduced in the SYN-series corpora was used. The information about the sections is available for all articles published in major newspapers (including MFD) since 2010. It is based on the original section titles taken over from the publishers that have subsequently been unified and classified into the sections available as a part of the newspaper article metadata (with a small percentage of section titles remaining as unclassified).

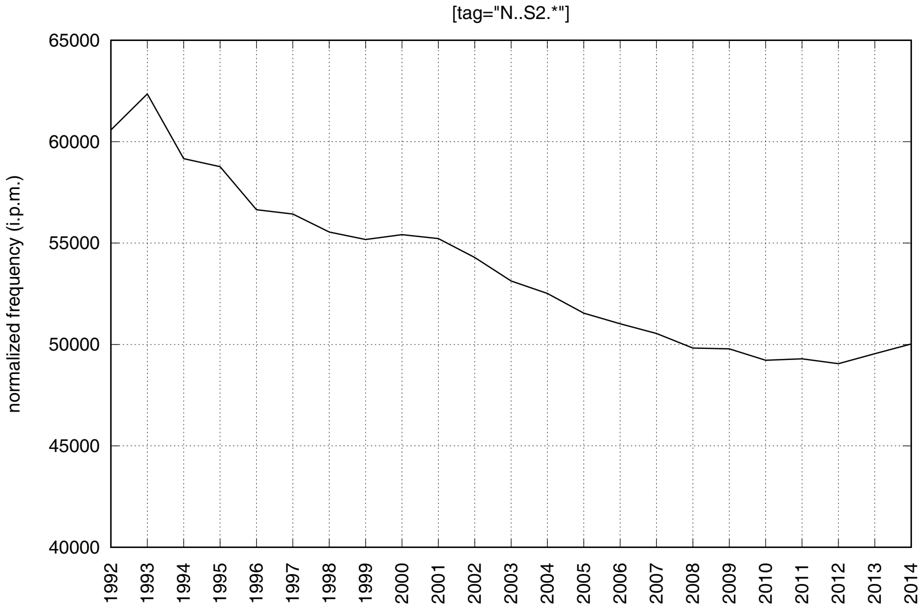


Fig. 4. Nouns in genitive singular

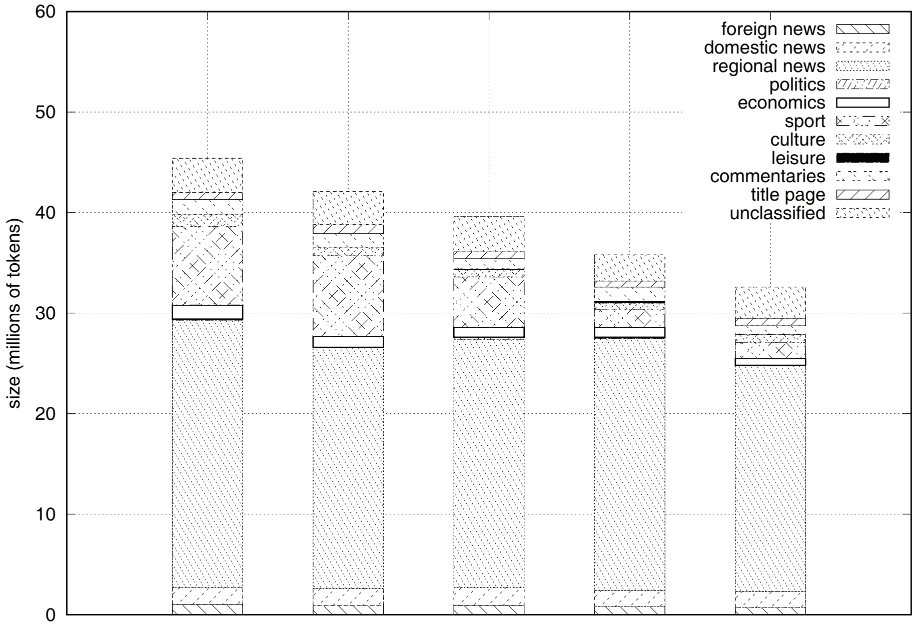


Fig. 5. Newspaper sections in MFD

Composition of MFD in terms of the individual sections in 2010–2014 is shown in Figure 5. There are two immediate observations to be made: first, the most prominent part of MFD are regional news, and second, the average size of an MFD issue is decreasing (as there is about 300 issues per year, see Figure 2).

The decreasing size of MFD in SYN v4 is the result of the decreasing volume of MFD texts received from the publisher's archive and available at the beginning of the corpus processing pipeline. As for the prevalence of regional news, the general policy (already mentioned above) is to include full texts or newspaper issues into SYN v4. However, MFD is sold in numerous regional versions that differ mostly in their regional news section. All the regional versions of MFD from the same day make up one issue on the input, which is then subject to de-duplication procedures on article level before its inclusion into SYN v4 [7, p. 160]. This affects mainly the non-regional articles typically removed as identical across the regional versions, while the bulk of the regional news remain as the prevailing part of MFD in SYN v4.

Although, there are no data on newspaper composition available for periods before 2010, it can be concluded that the results of this study are presumably not caused by the changing structure of MFD, but rather by gradual shifts within its individual sub-registers, most notably the regional news.

5 SUMMARY

The paper presented a corpus-driven method for detection of recent change in morphological categories that can be observed in contemporary Czech newspapers. The trends can be characterised as a shift towards more "verbal" language associated with increasing informality of the newspaper register.

The study certainly has its limitations. First, only categories that resulted from manually selected patterns have been considered in the evaluation. This means that some of them may have been left out, either by unintentional omission, or simply because the given combination was not considered potentially relevant.

Another limitation pertains to the morphological tagging that underlies the individual categories and that may not be ascribed only to the disambiguation accuracy. For instance, there has been incidentally discovered only a slight decrease of nominal (short) forms of adjectives (`AC.*`) during the examination of the results. This is in contradiction with their gradual replacement by their long counterparts in contemporary Czech. A key to the explanation is the word form *rád* ('glad'): from diachronic point of view, it is a nominal form of an adjective and it is also tagged as such. However, it is fossilised and contemporary Czech descriptions often treat it as an adverb. Since it is both very frequent and typical of informal language and topics, its increase almost compensates for the (quite significant) decrease of all other nominal forms. Figure 6 shows the resulting plot as a confluence of the two factors.

Last but not least, the study aimed at the detection of grammatical change trends in Czech newspapers. However, it analysed only one national daily newspaper (MFD) and came to the conclusion that the analysis is for the most part based on the regional news within MFD. This should not be seen as a shortcoming, as studies based on large and homogeneous data from restricted domain certainly have their value, although the results may not be as general as one would wish: "*The advantage of working with single source data ... is that, although the claims that can be made are necessarily limited, they are securely grounded*" [14, p. 216]. At the same time, the study has confirmed that the major challenge for research on recent language change is the data. CNC thus aims to continuously build corpora also from other domains to provide the research community with constantly growing material that could eventually bring corpus-derived insights into the nature of language change.

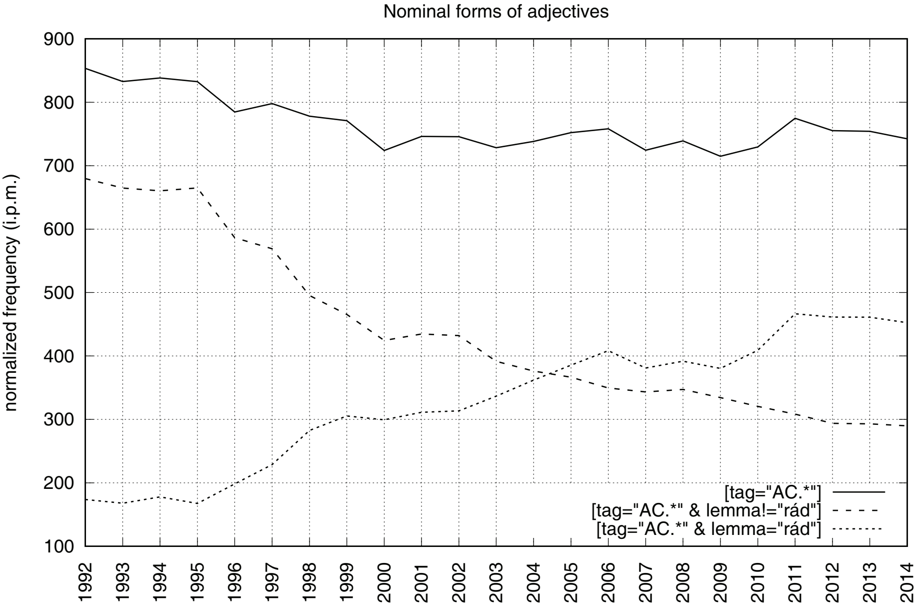


Fig. 6. *rád* vs. all other nominal forms of adjectives

References

[1] Baker, P. (2009). The BE06 Corpus of British English and recent language change. *International Journal of Corpus Linguistics*, 14(3):312–337.

[2] Bartoň, T., Cvrček, V., Čermák, F., Jelínek, T., and Petkevič, V. (2009). *Statistiky češtiny*. Nakladatelství Lidové noviny, Praha.

[3] Duguid, A. (2010). Newspaper discourse informalisation: a diachronic comparison from keywords. *Corpora*, 5(2):109–138.

[4] Hajič, J. (2004). *Disambiguation of Rich Inflection (Computational Morphology of Czech)*. Karolinum, Praha.

[5] Herman, O. and Kovář, V. (2013). Methods for Detection of Word Usage over Time. In *Seventh Workshop on Recent Advances in Slavonic Natural Language Processing, RASLAN 2013*, pages 79–85, Tribun EU, Brno, Czech Republic.

[6] Hilpert, M. and Gries, S. T. (2009). Assessing frequency changes in multistage diachronic corpora: Applications for historical corpus linguistics and the study of language acquisition. *Literary and Linguistic Computing*, 24(4):385–401.

[7] Hnátková, M. et al. (2014). The SYN-series corpora of written Czech. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014*, pages 160–164, ELRA, Reykjavík, Iceland.

[8] Hundt, M. and Mair, C. (1999). ’Agile’ and ’Uptight’ Genres: The Corpus-based Approach to Language Change in Progress. *International Journal of Corpus Linguistics*, 4(2):221–242.

[9] Křen, M. (2013). *Odraz jazykových změn v synchronních korpusech*. Nakladatelství Lidové noviny, Praha.

[10] Křen, M. et al. (2016). SYN2015: Representative Corpus of Contemporary Written Czech. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation, LREC 2016*, pages 2522–2528, ELRA, Portorož, Slovenia.

[11] Křen, M. et al. (2016). SYN corpus, version 4 from 16. 9. 2016. Ústav Českého národního korpusu FF UK, Praha. Accessible at: `http://www.korpus.cz.`

[12] Leech, G. (2004). Recent grammatical change in English: Data, description, theory. In *Advances in Corpus Linguistics. Papers from the 23rd International Conference on English Language Research on Computerized Corpora (ICAME 23)*, pages 61–81, Rodopi, Amsterdam, Netherlands.

[13] Mair, C., Hundt, M., Leech, G., and Smith, N. (2002). Short term diachronic shifts in part-of-speech frequencies: A comparison of the tagged LOB and F-LOB corpora. *International Journal of Corpus Linguistics*, 7(2):245–264.

[14] Millar, N. (2009). “Modal verbs in TIME: Frequency changes 1923–2006”. *International Journal of Corpus Linguistics*, 14(2):191–220.

[15] Rychlý, P. (2007). Manatee/Bonito – A Modular Corpus Manager. In *First Workshop on Recent Advances in Slavonic Natural Language Processing, RASLAN 2007*, pages 65–70, Brno, Czech Republic.

CORPUS-BASED SEMANTIC MODELS OF THE NOUN PHRASES CONTAINING WORDS WITH 'PERSON' MARKER

MARGARYTA LANGENBAKH
Taras Shevchenko National University of Kyiv, Ukraine

LANGENBAKH, Margaryta: Corpus-Based Semantic Models of the Noun Phrases Containing Words with 'Person' Marker. Journal of Linguistics, 2017, Vol. 68, No 2, pp. 249 – 257.

Abstract: The mechanism underlying constructing of lexically correct sequences of words is an object of attention both in theoretical and applied fields of linguistics. This paper reveals some aspects of modelling the patterns of semantic valence in noun phrases of NN (Noun+Noun) structure, one or both components of which contain the 'person' semantic tag. The research is based on the Corpus of Ukrainian and performed with the help of automatic language processing.

Keywords: natural language processing, semantic valence, noun phrases

1 COMPUTER-BASED REPRESENTATIONS OF SEMANTICS

The problems we discuss lie at the intersection of important theoretical (syntactical and lexical valence theory) and practical (semantic and informational analysis, text mining) studies. Last three decades in the computational linguistics brought a considerable amount of studies and projects representing different approaches to the meaning representation: the WordNet [17], FrameNet [3], Wikipedia-based annotation [20], Abstract Meaning Representation (AMR) [12], the method of propositions [20] etc. The WordNet project initiated the creation of numerous set of semantic dictionaries all over the world including Slavic languages that are represented by the Polish [8], Slovene [9], Croatian [19], Bulgarian [13] or Russian [4] projects. During the last decade some researches have been taken on the base of Ukrainian language [1], [14].

Active development of the corpus linguistics actualized a new field of study – the semantic annotation of the textual corpora. Now we have a semantic annotation for the Bulgarian [13], Polish [5], Russian [15] and Slovene [9] corpora. The semantic annotation for the Corpus of Ukrainian is being developed by the National Taras Shevchenko University of Kyiv [6]. In the next sections we discuss some theoretical and practical aspects of the taxonomy-based approach to the semantic annotation of the corpus by the example of noun phrases modelling.

2 THE THEORETICAL ASPECTS OF RESEARCH

Traditionally computer-based semantic analysis (as long as syntactic) is performed at the sentence level. But choosing the phrase as a unit of semantic processing has some advantages. Assigning the semantic roles to the sentence constituents is not a problem when we deal with monopropositional constructions. Polypropositional

DOI 10.1515/jazcas-2017-0034

sentences may contain a set of linked situations and the word included in more than one situation may have different roles in them. For example, in the sentence *Я не вважаю твого брата винуватцем сварки 'I don't think that your brother caused a quarrel',* the noun *брата 'brother'* is an object of one situation (to think him not to be a causer) and a subject of another (to cause a quarrel). Using a phrase as basic semantic construction may solve this problem.

The goal of our research was to investigate the mechanisms of semantic connection inside the noun phrases structure. In order to do this, we had to answer several questions.

The first question what taxonomic markers are needed to form the semantic models of noun phrases and if some fixed positions (head or adjunct).

Every natural language has some syntactical rules. In our article we tried to find out whether these rules are based on the grammar only or there are some lexical features that determine the words positions.

The second question was what combinations of the semantic tags correlate with the certain types of relations.

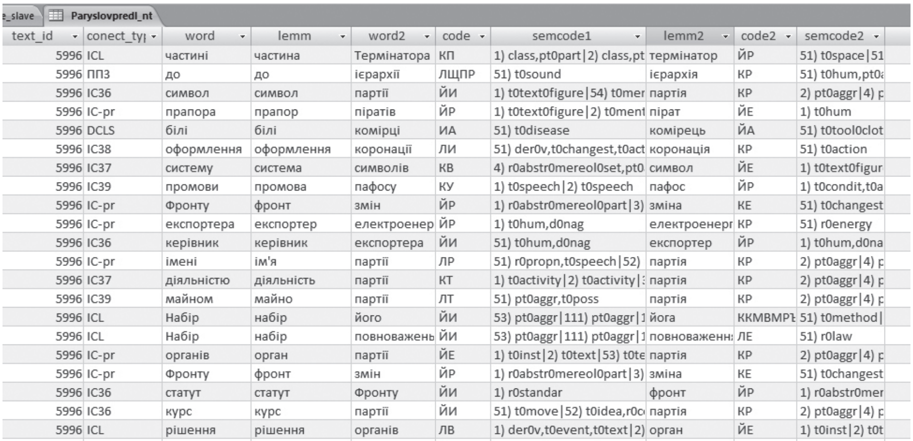
One more important question was whether the core meaning markers determine the relation or peripheral elements can also take part in this process. Many researchers of the lexical system emphasize on the different roles of semantic components of the word meaning. According to Y. Apresyan, "every lexical meaning has certain syntactical structure" [2, p. 9]. Examining the structure of lexeme, the linguists divide it into integral and distinctive semantic features [21, p. 78]. The first of them help to specify the semantic similarity between the words and the second – to distinguish the meanings of semantically similar lexemes. And one of our tasks was to find out if there is a correlation between the position of the semantic marker in the meaning structure and the word valence.

The last question we had to answer was what are the functions of the semantic model elements. E. Sapir proposed to differentiate the meanings into concrete (nominative, descriptive), derivational, relational and mixed concrete relational [11, p. 20]. The presence of relational markers in the lexemes gives them a potentiality to attach the other lexemes (the valence phenomenon [23], [13]). The functional structuring of the word semantics also plays an important role in the development of ontologies. For example, in the WordNet project the elements that have relational meanings are used to represent the semantic relations between the nominative elements [17]. In our research we tried to find out which of the meaning components determine the nature of syntagmatic links – relational that initiate the connection or nominative that define the nature of word lexical meaning.

3 TYPES OF RELATIONS AND THEIR SEMANTIC MARKERS IN THE NN PHRASES CONTAINING WORDS WITH 'PERSON' TAG

In our previous work we analysed the deep syntactic structure patterns of the noun phrases [16]. This research is a second attempt to investigate the mechanisms of the noun valence realization focused more on the lexical background of this phenomenon and performed with the help of machine semantic tagging.

We used the taxonomy-based semantic dictionary for automatic semantic analysis of the texts. The taxonomy consists of three general classes (Proper Names, General Names and Object Names), divided into a list of subclasses (178 taxons, the maximum depth of taxonomy – four levels). The dictionary was applied to the newspapers subcorpus (17 705 122 words) of the Corpus of Ukrainian (`http://mova.info/corpus.aspx`) processed by the AGAT NLP-system [7]. Then we automatically selected the phrases of the NN structure with their grammatical and semantic tags (Fig. 1). The query to the database returned 113 471 phrases consisted of 12 610 words[1].



text_id	conect_ty	word	lemm	word2	code	semcode1	lemm2	code2	semcode2
5996	ICL	частині	частина	Термінатора	КП	1) class,pt0part|2) class,pt	термінатор	ЙР	51) t0space|51
5996	ПП3	до	до	ієрархії	ЛЩПР	51) t0sound	ієрархія	КР	51) t0hum,pt0
5996	IC36	символ	символ	партії	ЙИ	1) t0text0figure|54) t0mer	партія	КР	2) pt0aggr|4) p
5996	IC-pr	прапора	прапор	піратів	ЙР	1) t0text0figure|2) t0ment	пірат	ЙЕ	1) t0hum
5996	DCLS	білі	білі	комірці	ИА	51) t0disease	комірець	ЙА	51) t0tool0clot
5996	IC38	оформлення	оформлення	коронації	ЛИ	51) der0v,t0changest,t0act	коронація	КР	51) t0action
5996	IC37	систему	система	символів	КВ	4) r0abstr0mereol0set,pt0	символ	ЙЕ	1) t0text0figur
5996	IC39	промови	промова	пафосу	КУ	1) t0speech|2) t0speech	пафос	ЙР	1) t0condit,t0a
5996	IC-pr	Фронту	фронт	змін	ЙР	1) r0abstr0mereol0part|3)	зміна	КЕ	51) t0changest
5996	IC-pr	експортера	експортер	електроенер	ЙР	1) t0hum,d0nag	електроенерг	КР	51) r0energy
5996	IC36	керівник	керівник	експортера	ЙИ	51) t0hum,d0nag	експортер	ЙР	1) t0hum,d0na
5996	IC-pr	імені	ім'я	партії	ЛР	51) r0propn,t0speech|52)	партія	КР	2) pt0aggr|4) p
5996	IC37	діяльністю	діяльність	партії	КТ	1) t0activity|2) t0activity|	партія	КР	2) pt0aggr|4) p
5996	IC39	майном	майно	партії	ЛТ	51) pt0aggr,t0poss	партія	КР	2) pt0aggr|4) p
5996	ICL	Набір	набір	його	ЙИ	53) pt0aggr|111) pt0aggr|1	йога	ККМВМРЪ	51) t0method|
5996	ICL	Набір	набір	повноважень	ЙИ	53) pt0aggr|111) pt0aggr|1	повноваженн	ЛЕ	51) r0law
5996	IC-pr	органів	орган	партії	ЙЕ	1) t0inst|2) t0text|53) t0te	партія	КР	2) pt0aggr|4) p
5996	IC-pr	Фронту	фронт	змін	ЙР	1) r0abstr0mereol0part|3)	зміна	КЕ	51) t0changest
5996	IC36	статут	статут	Фронту	ЙИ	1) r0standar	фронт	ЙР	1) r0abstr0mer
5996	IC36	курс	курс	партії	ЙИ	51) t0move|52) t0idea,r0c	партія	КР	2) pt0aggr|4) p
5996	ICL	рішення	рішення	органів	ЛВ	1) der0v,t0event,t0text|2)	орган	ЙЕ	1) t0inst|2) t0t

Fig. 1. Database structure

According to the results of our research, the relations in noun phrases containing words with the 'person' markers may be divided into three general types – actional, non-actional (different sorts of attribution) and part-whole relation (mereology). The actional relations describe the roles of participants in a situation frame. Such structures are the grammatical transforms of sentences (compare *виступ політика 'the speech of a politician'* and *Політик виступив 'The politician made a speech'*). The non-actional relations form the models which in some realizations are similar to the other type of noun phrases – constructions of the noun and adjective (*ввічливість працівника 'the politeness of the worker'* is comparable with the *ввічливий працівник 'the polite worker'*). But this type isn't completely identical to the noun-adjective attributive relations because some of non-actional phrases also may be transformed into predicative constructions, for example, in a case of possession (в*олодіння бізнесмена 'the possessions of a businessman' – бізнесмен володіє (чимось) 'the businessman possess (something)'*). These examples do not have such definitely actional semantics as the first general type of phrases we described previously, but the attribution characteristic in them has some specificity.

The part-whole relations embody the linguistic representation of mereology phenomenon. They are quite natural for the noun phrases because the 'part' and 'whole' concepts are usually associated with the objects and described by nouns.

[1] Because of the possible errors of language processing the real results may differ. The estimated error rate is 6–8%

All discussed types of relations can be specified by the semantic patterns they can be used in.

The actional relations are represented by such subtypes:

1. Subjective. These relations appear in the NN phrases that describe the connection between the action and its doer. The actor can possess any grammatical position in the phrase:

A) The position of the subjective grammatical head occurs in such patterns as <creation, person, nomen agentis + action> (*натхненник повстання ‘the inspirer of the revolt’*);
<person, nomen agentis + action> (*виконавець проекту ‘the executor of project’*) etc. The active mode of the person’s role is usually marked by the taxon ‘nomen agentis’ which describes the doer of certain action;

B) The position of the grammatical adjunct can be illustrated by such pattern as <interaction + person> (*поєдинок чемпіонів ‘the battle of champions’*).
Unlike the previous pattern, the adjunct does not obligatory have the ‘nomen agentis’ marker that can point at the lexical differences between the actions described by these two models.

2. Objective. This subtype is represented by the combination of a word with actional semantics and word with ‘person’ marker being a passive participant of the situation. The grammatical position of this element also may be different:

A) Head:
<person + action> pattern (*жертва насильства ‘the victim of the violence’*).

B) Adjunct:
<occupation, activity + person> (*лікування пацієнта ‘a treatment of the patient’*).

The difference between the lexical meanings of the words that name the actions in subjective and objective models can also produce a semantic variety of actional relations:

– process of creation: <action|process, creation + person> (*підготовка спеціалістів ‘a training of the professionals’*);

– contact, interaction: <action|process, contact + person, nomen agentis> (*учасник дебатів ‘the participant of debates’*);

– action or process which results in some changes: <action|process, change + person> *навчання дітей ‘teaching of children’*) etc.

3. Subject-object interaction. The noun phrases of this subtype contain the subjective noun that expresses the actor and, at the same time, points to the action, and a noun in an objective role. The ‘person’ marker can be found in the subjective, objective or both phrase components and the positions of the elements also may vary. The semantic variations are quite wide:

A) Creation of concrete or abstract objects:
<person, nomen agentis, process, creation + tool, furniture> (*виробник меблів ‘the manufacturer of furniture’*).

B) Transformation of an object, changing of its characteristics:
<person, nomen agentis, action|process, change + text> (*перекладач роману ‘the translator of the novel’*).

C) Supplying or consumption of products or services:
<person, nomen agentis + food> (*продавець картоплі ‘a potato seller’*) etc. Our taxonomy doesn’t have special tag for naming such class of activity, but these relations are quite regular and semantically different from the others, so we decided to mention them and suppose to add them to the new revised edition of taxons list.

D) Professional activity:
<person, nomen agentis, occupation + tool, transport> (*водій тролейбуса ‘a trolleybus driver’*).

E) Emotional or willing activity:
<person, nomen agentis, action|process, will + human qualities|mental sphere> (*захисник моралі ‘the defender of morality’*).

F) Participation or collaboration:
<person, nomen agentis, activity + event> (*учасник фестивалю ‘a participant of the festival’*).

G) Professional interaction:
<person, nomen agentis, occupation + person> (*радник директора ‘the counsellor of the chief manager’*).

In this subtype (as well as in the other actional constructions) the main lexical meaning of the head is a basis of relational subtype division. For example, in phrases which describe the professional activities the head-actor determines the sense of situation (*приборкувач* – a person who tame an animal) and the adjunct is a tool of concretization and bearer of passive valence marker (*змія ‘snake’* is an animal, so it can be an object of taming). The difference between the phrases *творець картини ‘the creator of the picture’* and *покупець картини ‘a buyer of the picture’* lies in the field of the head’s semantics: in the first case we have a person-creator and in the second – a person-consumer. This lexical difference points to the specific mode of the influence which the actor has on the object. The lexical meanings of the adjuncts, on the contrary, don’t play an essential role in this type of relations.

The non-actional relations usually are formed by the combination of personal noun and noun describing some characteristics. As well as in previous models, the components of non-actional phrases may appear in different positions: characteristic as the head and subject as adjunct (*самолюбство колег ‘the ambition of colleagues’*) or inverted variant (*людина честі ‘a man of honour’*), but the first pattern is more typical for Ukrainian.

This type of relations also integrates some different semantic variants – complementary, possessive and property attribution.

The property attribution can be found in the noun phrases which contain the word describing specific value of the certain property:

A) A feature of a human personality:
<human quality + person> (*велич постаті ‘a greatness of figure’*).

B) An emotional state:
<feeling + person> (*радість матері ‘a joy of mother’*).

The combination of subjective noun and abstract names of human qualities or activities (fields of knowledge or culture) gives us the complementary attribution (*доктор філософії ‘the philosophy doctor’*).

The similar situation we have in the case of possessive attributive semantics: *володар зброї 'the weapon holder'*. The possessive relations also can be divided into two subtypes:

A) Ownership:
<person, possession + tool, transport> (*власник автомобіля 'the car owner'*).

B) Occupancy:
<place, part of building + person, nomen agentis> (*кабінет секретаря 'a secretary's room'*).

As well as the subject-object interaction, the possessive relation of ownership is formed by the subjective noun that has possessive marker. The occupancy case isn't identified by the specific tag of our taxonomy. The choice of the adjunct doesn't rely on its specific lexical meaning but demands a specific relational potentiality (to be an object of possession).

The part-whole relations occur in the patterns, consisting of tags combinations in both phrase constituents. According to the set of markers we divided the patterns in two subtypes:

A) Body and its parts:
<part of the body + person> (*плече робітника 'the shoulder of the worker'*); this pattern includes the tag 'part' of the head and the tag 'person' of the adjunct.

B) Group of people or organization and its members:
<person + person, set of objects> (*член команди 'a team member'*); the head of this pattern has the 'person' taxon (it is not the only semantic component required for such type of relation, but our taxonomy doesn't have a tag for the meaning *'to be a representative of certain group'*), the adjunct must be marked as 'set of objects' or 'organization'.

Though the positions of relational markers bearers mainly are not strongly restricted, there are some exceptions when relational type clearly depends on the position of certain marker. For example, a model, built from the head representing the abstract name of class with the 'part' taxon and adjunct with the 'person' and 'set of objects' taxons expresses the 'part-whole' relation (*половина електорату 'a half of electorate'*). At the same time, the converse position of the 'set of objects' marker points to the complementary semantics of relation (*клас професіоналів 'a class of professionals'*). The 'nomen agentis' tag in the phrase head marks its active mode, identifies it as an actor whereas the main elements without this tag are more likely to be an object of an action or a bearer of some qualities. But there are rather rare, minor cases.

There are some situations in which we can't clearly define the type of relation. The first situation is caused by the grammatical homonymy. When the position of a head is occupied by the noun derived from the transitive verb and the adjunct is personal noun we usually are not able to say whether it is an object or subject of an action (*перевірка спеціаліста 'an examination of the professional'* may mean a work of specialist or an examination of his qualification level).

The second situation may be illustrated by the noun phrases, describing the personal relations (*друг сім'ї 'a friend of family'*). The interpretation of such patterns depends on the specific lexical meaning of the head: in some cases this combination forms the subjective relation (*голова родини ‘a head of family’*), in other – objective (*ставленик президента ‘the President’s protégé’*) or even syncretic types (*друг сім’ї* is a possessive model (*a friend of whom?*), and, from the other hand, subjective, that describes a pattern of behaviour).

Unfortunately, in this article we can’t present a full list of the patterns that we found but all of them fit into the three mentioned classes of relations. To sum up our surveillance we can make several conclusions. Firstly, the semantic relation in the noun phrases can be based on the markers of both words (e.g. subjective marker in the head and objective in adjunct) or only one of them. The duplication of certain marker in the phrase constituents rather means their connection abilities than forms the relation (except the case of semantic recursion such as *діти дітей ‘children of children’*).

Secondly, there are no strict rules concerning the quantity and positions of the markers that determine the relation. In some models they belong to the head, in some to the adjunct or even to both of them. We can only make an assumption that relation is more often based on the element that has stronger valence potential in its meaning (what agrees with the traditional valence theories). Usually they are the names of abstract classes, actors and attributes. Concerning the direction of semantic relation we can draw an analogy with syntax dependencies, divided into three classes – bilateral, unilateral and coordinate [22, p. 89]. The actional relations are usually bilateral because the actor and object of an action are obligatory participants of the situation. The non-actional relations may either be unilateral (in complementary phrases like *категорія читачів ‘a readers category’* the head *категорія ‘category’* defines the specific role of adjunct *читачів ‘readers’*) or coordinate (in the attributive phrases like *чесність політика ‘the honesty of a politician’* we can say that the grammatical head *чесність ‘honesty’* creates the complementary relation with the adjunct *політика ‘politician’* and the adjunct also induces a reverse relation of attribution). The part-whole relations are unilateral because the noun which names the ‘whole’ component of pattern is usually semantically independent.

Thirdly, there is no obvious correlation between the position of the semantic marker in the lexical meaning structure and its potency to establish a relation. In the other words, we cannot say that only heads or adjuncts regularly induce relations. The choice of the semantic markers (and a word correspondingly) needed to build a phrase is determined by the contextual and communicative requirements.

4 THE PROSPECTS OF USING THE TAXONOMY-BASED DICTIONARY IN THE NATURAL LANGUAGE PROCESSING

In the Section 3 we discussed the theoretical issues concerning the dictionary-based valence modelling in the noun phrases. But is the structure of our taxonomic dictionary suitable for automatic detecting the essential lexical features of the words in a context?

Grammatical and semantic annotation theoretically allows the computer to detect the connected words and build the structure of phrases and sentences. Also the machine can suggest the senses of these constructions according to the markers of their components and relational patterns. The semantic dictionary used in our research embodies the facet classification made by the example of the Russian National Corpora [15], that allows words to be included in several semantic classes and, consequently, to have different sets of markers [6]. It means that we can, treating words as structured semantic complexes, derive the relations between the words not from their integrated lexical meanings but from the certain elements of their semantics that actually provide the semantic connection.

Such approach has some deficiencies. Fixation of all variants of the homonymic and polysemic lexemes leads to high level of the ambiguity and so called informational noise produced by the minor or rare meanings. For example, the word *зміна ʻchange'* is interpreted by the computer as the 'person' because of its second meaning *'people working in shifts'* and, consequently, the phrase *активісти змін 'the activists of changes'* receives the model <person + person, set of objects> and may be treated as part-whole pattern. Regular wrong results are produced by the connotations of animal names: *використання собак 'the using of dogs', кількість свиней 'the quantity of pigs'* (because in Ukrainian we may use 'a dog' and 'a pig' as abusive words) etc. A majority of words (about 2/3 of dictionary list) has more than one meaning so it is the serious problem.

By applying relational patterns to the processed texts we can reduce the ambiguity to the certain extent. However, there are situations that need more complicated approach – examination of the style, genre, thematic of the text, contextual, statistic or stochastic information etc. Unfortunately, if the ambiguity level is very high such strategy may make the system too complicated.

There are two additional ways to reduce the ambiguity. The first one is to shorten the lexemes description omitting rarely used meanings and very peripheral taxons. The second way is to rebuild the structure of a dictionary. The automatic dictionary should be based not on the full descriptions of the lexical meanings but rather on semantic features that describe words valence. In other words, the definition may be incomplete from the lexical point and include some relational characteristics which lie beyond the lexical meaning. For example, in objective phrase like *споживач пшениці 'the consumer of wheat'* the adjunct must have such marker as 'to be an object of consumption' and its peculiar lexical description doesn't affect the relation (compare: *споживач хліба 'the consumer of bread', споживач енергії 'the consumer of energy'* etc.). So the machine-oriented dictionary must be more functional and relational than encyclopedical.

References

[1] Anisimov, A., Marchenko, O., Nikonenko, A., Porkhun, E., and Taranukha, V. (2013). Ukrainian WordNet: Creation and Filling. In Larsen, H. L., Martin-Bautista, M. J., Vila, M. A., Andreasen, T., and Christiansen, H., editors, *Flexible Query Answering Systems. FQAS 2013. Lecture Notes in Computer Science, vol 8132*, pages 649–660, Springer, Berlin – Heidelberg, Germany.

[2] Apresyan, Yu. (2014). *Izbrannye trudy, tom I. Leksicheskaya semantika. Sinonimicheskie sredstva yazyka*. Izdatelskaya firma «Vostochnaya literatura» RAN, Moscow.

[3] Baker Collin, F., Fillmore, Ch. J., and Lowe, J. B. (1998). The berkeley framenet project. In *COLING-ACL '98: Proceedings of the Conference*, pages 86–90, Association for Computational Linguistics, Montreal, Canada.

[4] Balkova, V., Sukhonogov, A., and Yablonsky S. (2004). Russian wordnet: From UML-notation to Internet/Intranet Database Implementation. In *Proceedings of the Second International WordNet Conference*, pages 31–38, GWC, Brno, Czech Republic.

[5] Broda, B., Piasecki M., and Maziarz, M. (2010). Evaluating LexCSD – a weakly-supervised method on improved semantically annotated corpus in a large scale experiment. In *Proceedings of Intelligent Information Systems (2010)*, pages 63–76, Publishing House of University of Podlasie, Siedlce, Poland.

[6] Darchuk, N., Zuban', O., Lanhenbakh, M., and Khodakivs'ka, Ya. (2016). AGAT semantyka: semantychne rozmichuvannya Korpusu ukrayins'koyi movy. *Ukrayins'ke movoznavstvo*, 46(1):92–103.

[7] Darchuk, N. (2013). *Komp"yuterne anotuvannia ukrayins'koho tekstu: rezul'taty i perspektyvy*. Osvita Ukrayiny, Kyiv.

[8] Derwojedowa, M., Piasecki, M., Szpakowicz, S., Zawislawska, M., and Broda, B. (2008). Words, concepts and relations in the construction of Polish WordNet. In *Proceedings of the Global WordNet Conference*, pages 162–177, Seged, Hungary.

[9] Erjavec, T. and Fišer D. (2006). Building the Slovene Wordnet: first steps, first problems. In *Proceedings of the Third International WordNet Conference, Vol. 2006*, GWC, Jeju Island, Korea.

[10] Flanigan, J., Thomson, S., Carbonell, J., Dyer, C., and Smith, N. A. (2014). A discriminative graph-based parser for the abstract meaning representation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 1426–1436, Association for Computational Linguistics, Baltimore, Maryland.

[11] Haselow, A. (2011). *Typological changes in the lexicon*. 1st ed. Mouton de Gruyter, New York, NY.

[12] Katsnelson, S. (1972). *Tipologiya yazyka i rechevoe myshlenie*. Nauka, Leningrad.

[13] Koeva, S., Leseva, S., and Todorova, M. (2006). Bulgarian sense tagged corpus. In *Proceedings of the 5th SALTMIL Workshop on Minority Languages: Strategies for Developing Machine Translation for Minority Languages*, pages 79–86, Genoa, Italy.

[14] Kul'chyts'kyi, I., Romanyuk, A., and Khariv, B. (2010). Rozroblennya Wordnet-podibnoho slovnyka ukrayins'koyi movy. *Visnyk Natsional'noho universytetu "L'vivs'ka politekhnika". Informatsiyni systemy ta merezhi,* 673:306–318.

[15] Kustova, G., Lyashevskaya, O., Paducheva Ye., and Rakhilina Ye. (2005). Semanticheskaya razmetka leksiki v Natsionalnom korpuse russkogo yazyka. In *Natsionalnyy korpus russkogo yazyka 2003–2005*, pages 155–174, Moscow.

[16] Lanhenbakh, M. (2012). *Spoluchuvanist' imennykiv ukrayins'koyi movy (hlybynnyy semantyko-syntaksychnyy analiz)*. Kyiv.

[17] Miller, G. A., Beckwith, R., Fellbaum, Ch., Gross, D., and Miller, K. J (1990). Introduction to WordNet: An on-line lexical database. *International journal of lexicography*, 3(4):235–244.

[18] Palmer, M., Gildea, D., and Kingsbury P. (2005). The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, 31(1):71–106.

[19] Raffaelli, I., Bekavac, B., Agić, Ž., and Tadić, M. (2008). Building croatian wordnet. In *Proceedings of the Fourth Global WordNet Conference,* pages 349–359, Szeged, Hungary.

[20] Schenkel, R., Suchanek, F. M., and Kasneci, G. (2007). YAWN: A Semantically Annotated Wikipedia XML Corpus. In Kemper, A., Schöning, H., Rose, T., Jarke, M., Seidl, T., Quix, C., and Brochhaus, C., editors, *BTW. LNI, vol. 103*, pages 277–291, GI.

[21] Sternin, I. (2015). *Leksicheskoe znachenie slova v rechi*. Direct-Media, Moscow-Berlin.

[22] Van Valin, R. D. (2001). *An Introduction to Syntax*. Cambridge University Press, Cambridge.

[23] Tesnière, L. (2015). *Elements of structural syntax*. John Benjamins Publishing Company, Amsterdam – Philadelphia.

TEXT COLLECTIONS FOR EVALUATION OF RUSSIAN MORPHOLOGICAL TAGGERS

OLGA LYASHEVSKAYA[4,5,6] – VICTOR BOCHAROV[3] – ALEXEY SOROKIN[1,2] – TATIANA SHAVRINA[4,7] – DMITRY GRANOVSKY[3] – SVETLANA ALEXEEVA[3]
[1]Lomonosov Moscow State University, Russia
[2]Moscow Institute of Physics and Technology, Russia
[3]OpenCorpora.org
[4]Higher School of Economics, National Research University, Moscow, Russia
[5]Vinogradov Institute of the Russian Language RAS, Moscow, Russia
[6]Russian National Corpus, Moscow, Russia
[7]General Internet-Corpus of Russian, Moscow, Russia

LYASHEVSKAYA, Olga – BOCHAROV, Victor – SOROKIN, Alexey – SHAVRINA, Tatiana – GRANOVSKY, Dmitry – ALEXEEVA, Svetlana: Text collections for evaluation of Russian morphological taggers. Journal of Linguistics, 2017, Vol. 68, No 2, pp. 258 – 267.

Abstract: The paper describes the preparation and development of the text collections within the framework of MorphoRuEval-2017 shared task, an evaluation campaign designed to stimulate development of the automatic morphological processing technologies for Russian. The main challenge for the organizers was to standardize all available Russian corpora with the manually verified high-quality tagging to a single format (Universal Dependencies CONLL-U). The sources of the data were the disambiguated subcorpus of the Russian National Corpus, SynTagRus, OpenCorpora.org data and GICR corpus with the resolved homonymy, all exhibiting different tagsets, rules for lemmatization, pipeline architecture, technical solutions and error systematicity. The collections includes both normative texts (the news and modern literature) and more informal discourse (social media and spoken data), the texts are available under CC BY-NC-SA 3.0 license.

Keywords: text collection, shared task, morphological tagging, universal dependencies, morphological parsing, Russian corpora

1 MOTIVATION

Comparison of existing methods for automatic text processing on every level is one of the pledges of systematic development of NLP technologies for each language. MorphoRuEval-2017 [1] is an initiative in the framework of Dialogue-Evaluation, aimed at both assessing and improving the evaluation metrics of morphological tagging and lemmatization for the Russian language, as applied to different text registers (news, social media, literary texts). As part of this shared task, the organizers faced the challenge of compiling a large training collection using different sources with annotation of good quality. It was decided to unify all the main corpus collections for Russian, coming from all the principal corpus projects – RNC [2], GICR [3], OpenCorpora.org [4], and SynTagRus [5] – all sources with different tagsets, obtained by different algorithms, and using different dictionaries. Our assumptions were that the morphological data standard for training collection should be 1) concise, 2) compatible with international shared task results, 3) suitable for

DOI 10.1515/jazcas-2017-0035

rapid and consistent annotation by a human annotator, 4) suitable for computer parsing with high accuracy, 5) easily comprehended and used by a non-linguist (the last three are taken from "Manning's Laws" [6]). As an essential solution of the problem we have chosen a new standard of multilingual morphological tagging, Universal Dependencies[1] (UD) [7].

2 SOURCE DATA

During the shared task, the following annotated data was provided:

1) RNC Open: a manually disambiguated subcorpus of the Russian National Corpus – 1.35 million words, ca. 10 thousand sentences (a balanced sample of fiction, news, nonfiction, spoken data, and blogs). RNC project is regarded as the main source for research in literary language.
2) GICR corpus with the resolved homonymy – 1 million words. General Internet-Corpus of Russian provides rich amount of blogs and social media texts, and is used as an instrument for modern and non-normative language studies.
3) OpenCorpora.org data – 400 thousand tokens (news, wikipedia, nonfiction, blogs). OpenCorpora provides mainly blogs and news texts, mostly normative and modern.
4) SynTagRus – 900 thousand tokens (fiction, news). SynTagRus is a part of RNC, openly distributed for syntactic research.

In each corpus, information about word form, lemma, part of speech (POS), and grammatical features were provided. To unify the representation of the data, the conll-u format was chosen, as the most common, convenient, and simple, and for the unification of morphological tags, the format of the Universal Dependencies (further UD) 2.0 was used (with some specifications, see below). The text collections are now available under CC BY-NC-SA 3.0 license[2].

We have also provided for the comparison the following plain text collections: 30 million words from LiveJournal, 30 million words from Facebook, Twitter and VKontakte, and 300 million words from Librusec.

3 MORPHOLOGICAL STANDARD

3.1 Background

Historically, the first morphological standards of the publicly available Russian corpora were, generally taken, based on Zalizniak's grammatical dictionary [7] and its spin-offs, and adopted the output of a few programs for Russian morphological analysis (Dialing/AOT, Mystem, ETAP, Starling). As a prominent example, the POS list of the RNC standard [2] included 13 classes of Zalizniak and three more specific subcategories for adverbs and predicates, while the inventory of grammatical features incorporated the so called "secondary forms" such as locative II (e.g. (*v*) *les-u* 'in

[1] `http://universaldependencies.org/`

[2] All materials accessible at `https://github.com/dialogue-evaluation/morpho-RuEval-2017`

the forest', as opposed to the locative (*o*) *les-e* 'about the forest') and comparative II (e.g. *po-skoreje* 'faster'). In the SynTagRus treebank, a number of additional distinctions were motivated by the needs of machine translation (e.g. grammatical gender of the personal pronoun *ja* 'I').

Later on, a successful attempt was made to compile a tagset compatible with the international multilingual specifications developed with emphasis on the statistical processing, Multext-East [9]. The manually disambiguated portion of the RNC was converted into this format and used for training, and a number of models for TreeTagger, TnT, SVMTagger were provided (see `http://corpus.leeds.ac.uk/mocky/). The variants of the Multext-East are currently exploited in the Russian Internet Corpus, HANKO, ruTenTen, Araneum, and GICR corpora.`

Yet another multilingual standard was adopted for the Russian morphology in UD-Russian and UD-Russian-SynTagRus annotation schema [10]. It is mostly compatible with the RNC standard and annotation practice, but the feature set is reduced by dropping distinctions between the "primary" and "secondary" forms, whereas the POS list is expanded to the new categories of proper nouns, auxiliaries, subordinate conjunctions, symbols, and punctuation marks to agree with unified Universal Dependencies standard [13].

Unlike the above-mentioned standards, the OpenCorpora tagset was developed specifically to be convenient for manual disambiguation of grammatical forms taken into account that annotation is made by crowdsourcing. Since some distinctions made in reference grammars and dictionaries were considered difficult to be explained to the crowd and to be applied to real data by the crowd, a number of adjustments were made. For example, the comparative forms of adjectives and adverbs were collapsed in a single POS category. Participles, gerunds, infinitives, and finite verb forms were treated as four separate parts of speech since this reflected a classification used in some secondary school programs of the Russian language.

As a result, the morphological annotation of existing Russian corpora differs in the following respects:

(a) If the annotation is token-based (simplex forms), or the periphrastic forms are tagged as well (cf. analytical future tense forms such as *budem schitat* '(we) will assume');
(b) If the multiword units are tokenized as one token or several tokens, particular multiword expressions are treated as single units, if any;
(c) The number and borders of the POS categories;
(d) The structure of the inflectional categories and their values;
(e) The structure of the lemma-classifying categories and values (e.g. transitivity, personal names, etc.);
(f) Presence/absence of additional tags which signals the disambiguation status; not-in-dictionary-ness; violation of grammatical norms, etc.;
(g) Lexical attachment: for example, the animacy tag may be obligatorily assigned to the pronoun *kto* 'who' in some corpora and be omitted in others;
(h) Lemmatization rules are affected by the structure of POS-tags and grammatical tags, on the one hand, and by some internal agreements within the standard, on

the other hand. For example, the superlative forms can get (i) the lemma of the base adjective and the superlative degree tag (cf. SynTagRus); (ii) the lemma with the superlative affix and no degree tag (cf. RNC, GICR); or (iii) the lemma with the superlative affix and the superlative degree tag (cf. OpenCorpora). In RNC, the perfective and imperfective verbs are assigned two different lemmas, whereas in SynTagRus, the perfective verb will usually get the lemma of the imperfective aspectual counterpart.

3.2 Unified Representation

There is no clear benchmark for morphological tagging for Russian. Apart from Universal Dependencies for Russian, those advantages were already mentioned in Section 1, there are already several competing standards, such as AOT tagset, NLC tagset, Dialog-2010 tagset, positional tagset for Russian, etc[3]. In this way, with one's desire to evaluate morphological tagging quality, one should inevitably face the problem of unification. With respect to the work of our colleagues at the MorphoEval-2010 [11], we carefully summarized all the inconsistencies and tag matches of our data set (described in Section 4). Within our standard, we unified the mismatches, concerning closed-class mismatches (predicatives, particles, determiners, conjunctions and adpositions), yet some of the cases of open-class lexemes left as is (see Section 5).

4 CONVERSION AND EVALUATION OF THE DATA

4.1 The Tagsets of Four Corpora

RNC Open

RNC Open is a subcorpus of the manually disambiguated corpus made available for the offline processing under a non-commercial license. The texts of social media (blogs) were prepared specifically for the MorphoRuEval. The "deficient" tagsets [12] (those lacking some non-determined categories such as gender in pluralia tantum nouns) were normalized. Besides, in the cases where more than one possible grammatical parsing was present in the annotation, we left only one, usually the most frequent and pragmatically neutral. All grammatical categories which were not included in the MorphoRuEval list (such as transitivity, voice, indeclinability, anomalous and distort forms, etc.) are provided with their values in a separate field (in the UD notation).

OpenCorpora

OpenCorpora project works on crowdsourcing morphological annotation. All Russian language native speakers are encouraged to participate and volunteers' knowledge or ability to do this work isn't assessed before they start. Works on annotation aren't paid directly, nor indirectly. Participants are motivated by the fact that they create a freely available resource. About 5 thousand people have participated so far.

In order to maintain annotation quality three- or fourfold overlap is provided and all disagreements are verified manually by moderators with linguistic education. The part of OpenCorpora dataset which was used in MorphoRuEval-2017 shared task consists of randomly selected sentences only partially verified. Decision in all not moderated cases is taken by majority voting.

[3] https://github.com/kmike/russian-tagsets

SynTagRus

SynTagRus was one of the first Russian treebanks automatically converted into UD standard [10]. For the MorphoRuEval shared task, the data were reannotated in UD v.1.4 standard and then the morphological tags were converted into the unified standard. Unlike UD-SynTagRus, all the limitations and solutions of the shared task are applied to the data.

GICR

GICR corpus with resolved homonymy is a first GICR open-source subcorpus, tagged with the aid of Abbyy Compreno technologies. Natural Language Compiler tagset was converted to MSD-Russian with the following specifications: GICR now contains a special category for parenthesis, predicatives and digits, that leaded to extension of the common MSD format. These specifications were reduced to UD standard according to the instructions, with an exception of parenthesis – they were left in the training data with the tag "H". The procedure of conversion is not straightforward since, for example, GICR contain several classes of pronouns, which must be arranged to different classes in SynTagRus. For example, adjective pronouns (*ego* 'his', *kotoryj* 'which') become determiners and adverbial pronouns (*kak-to* 'somehow' *vsegda* 'always') become adverbs, Several GICR adjectives *drugoj* 'other', *kaghdyj* 'every' were also considered as determiners.

4.2 POS

Table 1 demonstrates mapping of POS-tags in four corpora and MorphoRuEval unified list. For the reference, the column for the UD 2.1 POS-tags is also provided.

Part of speech	RNC	GICR	Open Corpora	SynTagRus (UD 1.4)	UD 2.1	Morpho-RuEval
(common) noun	S	N	NOUN	NOUN	NOUN	NOUN
proper noun	S	N	NOUN	--	PROPN	PROPN
initial letter	INIT	=	NOUN + Init	=	=	=
pronoun	SPRO	P	NPRO	PRON	PRON	PRON
numeral	NUM	M	NUMR	NUM	NUM	NUM
adjective	A	A	ADJF	ADJ	ADJ	ADJ
adjective (short form)	=	=	ADJS	=	=	=
adjectival numeral	ANUM	=	ADJF / ADJS + Anum	ADJ	ADJ	=
adjectival pronoun / determiner	APRO	P	ADJF / ADJS + Apro	DET	DET	DET
participle, full form	V	A	PRTF	VERB	VERB	ADJ
participle, short form	=	A	PRTS	=	=	=
verb	V	V	VERB	VERB	VERB	VERB
Infinitive verb	=	=	INFN	=	=	=
gerund	=	=	GRND	=	=	=
auxiliary	=	=	--	=	AUX	=
adverb	ADV	R	ADVB	ADV	ADV	ADV
adverbial pronoun	ADVPRO	P	ADVB + Ques / Dmns	=	=	=

parenthetically used discourse markers	PARENTH	H	ADVB	ADV	ADV	H
preposition / postposition	PR	S	PREP	ADP	ADP	ADP
conjunction	CONJ	C	CONJ	CONJ	CCONJ	CONJ
subordinate conjunction	=	=	CONJ	=	SCONJ	=
particle	PART	Q	PRCL	PART	PART	PART
interjection	INTJ	I	INTJ	INTJ	INTJ	INTJ
symbol	SYM	X	SYMB	SYM	SYM	X
foreign words, non-words	NONLEX	X	LATN	X	X	X
punctuation mark	--	-	PNCT	PUNCT	PUNCT	PUNCT
comparative	--	A, R	COMP	--	--	ADJ, ADV
predicative, predicative pronoun	PRAEDIC, PRAEDIC PRO	W	PRED	--	--	ADJ, ADV, VERB

Tab. 1. POS-tags

In RNC, the nouns were divided into NOUNs and PROPNs using the grammatical features of personal names, patronymics, toponyms, etc.; inanimate nouns were checked manually. The participles, which were tagged as VERB in the original standard, were assigned the tag ADJ, but their lemma remains the form of the infinitive, and an additional tag. The adjectival numerals were converted to ADJ except *odin* ‘one’, which semantically belongs to the class of cardinal numerals (marked as NUM following the UD standards). The classes of SPRO and APRO roughly correspond to PRON and DET, respectively. We compiled word lists to define these categories, and all words outside the lists were treated as nouns and adjectives. Conversion of predicatives is shown below.

In GICR, there is a special category for parenthetical constructions (H), which cannot be simply mapped onto adverbs or predicatives, as they are often a complex token combination. H is left in the training data, but not considered in evaluation. Proper nouns were also mapped onto simple NOUN during conversion to UD, that also led to testing procedure constraints discussed in Section 5.

OpenCorpora uses its own morphological tagset developed to be convenient for manual annotation purposes. In order to convert this tagset to Universal Dependencies an “OpenCorpora to UD” module has been added to Russian-tagsets project[4].

There is a number of deviations from MorphoRuEval-2017 guidelines in morphological annotation of OpenCorpora subset:

- the concept of auxiliary verb doesn‘t exist in OpenCorpora on morphological level and VERB / AUX disambiguation isn‘t performed. The verb *byt’* ‚be‘ is always annotated with VERB tag;
- OpenCorpora treats comparative as a separate part of speech. Universal dependencies guideline considers comparative as a form of an adjective or an adverb. In UD version of OpenCorpora subset all comparatives are annotated with ADJ tag.

[4] https://github.com/kmike/russian-tagsets

SynTagRus shows the closest match with regard to POS tags, except proper names, participles, and symbols. The proper names are tagged as NOUNs, the participle forms were converted to ADJ, and SYM was converted to X.

5 REMAINING DISCREPANCIES

Concerning the fact that the irreducible standard difference can affect the training results of the track participants, we refused to use the part-of-speech SYM (symbol) and AUX (auxiliary verb), and coordinate and subordinate conjunctions are both marked as CONJ. Here are the left ones in our collection: noun (NOUN), proper name (PROPN), adjective (ADJ), pronoun (PRON) numeral (NUM), verb (VERB), adverb (ADV), determinant (DET), conjunction (CONJ), preposition (ADP), particle (PART), interjection (INTJ). Also on the data are marked punctuation marks (PUNCT) and non-word tokens (X).

The following categories are marked and unified for different parts of speech:

1. Noun: gender, number, case, animacy
2. Proper name: gender, number, case
3. Adjective: gender, number, case, brevity of form, degree of comparison
4. Pronoun: gender, number, case, person
5. Numeral: gender, case, graphic form
6. Verb: inclination, person, tense, number, gender
7. Adverb: degree of comparison
8. Determinant: gender, number, case
9. Conjunction, preposition, particle, parenthesis, interjection, other: none

Accepted values:

Case: nominative – Nom, genitive – Gen, dative – Dat, accusative – Acc, locative – Loc, instrumental – Ins
Gender: masculine – Masc, feminine – Fem, neuter – Neut
Number: singular – Sing, plural – Plur
Animacy: animated – Anim, inanimated – Inan
Tense: past – Past, present or future – Notpast
Person: first – 1, second – 2, third – 3
VerbForm: infinitive – Inf, finite – Fin, gerund – Conv
Mood: indicative – Ind, imperative – Imp
Variant: short form – Brev (if the form is complete, no mark is placed)
Degree: positive or superlative – Pos, comparable – Cmp
NumForm: numeric token – Digit (if the token is written in alphabetic form, no mark is placed).

In order to increase the annotation agreement in the collections converted from different sources and simplify semiautomatic verification of annotation correctness, the following decisions were made:

1) DET is a closed class which includes 44 pronouns used primarily in the attributive position, exceeding official list of 30 determiners – such cases as *vsyak 'any' (vernacular), ihniy 'their' (vernacular)* were also included.

2) Predicative words. Modal words such as *mozhno* ‘can’, *nelzja* ‘cannot’ are considered as adverbs. The word *net* ‘no, not’ is considered as a third-person form of a verb. The predicative words homonymous to the short neuter forms of adjectives are coded as adjectives. Unlike adverbs, the short adjectives always form a part of the predicate.

   That condition was checked automatically by extracting the subject and predicate from each sentence and verified manually afterwards. Except for several words, our algorithm discriminates between adverbs and short adjectives in the same way as the one use in UD-SynTagRus does.
3) The lemma of the verb is its infinitive form in a particular aspect (perfective or imperfective). The gerund forms constitute a part of the verb paradigm. Verbs in passive voice keep their passive suffix *-sya* in their infinitive form as well.
4) The participles are treated as adjectives and their lemma is the full nominative masculine singular form. This form is reconstructed using dictionary lookup and suffix transformations.
5) The ordinal numerals are considered adjectives.
6) The tense forms of the verb are divided into Past and Notpast (present or future).
7) The analytic (multi-word) forms of verbs, adjectives, and adverbs are not coded. For example, the analytic future tense form is annotated as two separate tokens: the future form of the verb *byt* ‚be‘ and the infinitive.
8) For all prepositions including phonetic variants *с/со*, *в/во* its lemma coincides with the word itself.
9) NOUN and PROPN were evaluated as a single tag.
10) CONJ and SCONJ\CCONJ were also regarded to one tag.
11) Differences between UD 1.4 and UD 2.0 were not penalized.

Several of categories received the status of “not rated”: they may be present or not in the output of the system under evaluation:

* animacy (nouns, pronouns);
* aspect, voice, and transitivity (verbs);
* POS tags of prepositions, conjunctions, particles, interjections, and X (others).

6 CONCLUSION

The dataset collected shows one of the most challenging issue in the Russian NLP domain: there exist a lot of competing standards, associated with different existing pipelines and different theoretical views on Russian morphology. From the point of view of technological development and increasing interest among developers to the field of NLP, the mentioned data sources will inevitably be unified to one format. One can only hope that this format will be widely used and won’t become just one of N+1 competing standards, as in comparison with the previous shared tasks, this unification is more detailed. The main merits of the work described are:

- the original data set which was annotated in a single format consistent with UD guidelines was prepared and presented;
- techniques and principles which correspond to the UD standard, at the same time considering current situation with disparate standards for the Russian;
- the comprehensive guidelines for testing procedure and evaluation in this format.

All materials of MorphoRuEval-2017 including training and test set are now available at the competition's github. We welcome NLP-researchers and specialists in machine learning to use this collection and we hope that the collection will stay practical and relevant for a long time.

ACKNOWLEDGEMENTS

Authors are grateful to all the colleagues who were also involved in the MorphoRuEval organization committee and provided data: Kira Droganova, Alena Fenogenova, Ilia Karpov. We also thank GICR and OpenCorpora.org teams for preparing their segments. Dmitry Sichinava and Svetlana Savchuk has done a lot to make the RNC Open collection available. We also express our gratitude to the participants of the MorphoRuEval tracks who took part in the discussion list and provided their suggestions and comments.

References

[1] Sorokin, A., Shavrina, T., Lyashevskaya, O., Bocharov, V., Alexeeva, S., Droganova, K., and Fenogenova, A. (forthcoming). MorphoRuEval-2017: an evaluation track for the automatic morphological analysis methods for Russian. In *Computational linguistics and intellectual technologies. Proceedings of International Workshop Dialogue'2017*, Moscow.

[2] Lyashevskaya, O. N., Plungian, V. A., and Sichinava, D. V. (2005). O morfologicheskom standarte Korpusa sovremennogo russkogo jazyka [Morphological standard of the Corpus of contemporary Russian]. In *Nacional'nyj korpus russkogo jazyka: 2003–2005* [Russian National Corpus: 2003-2005], pages 111–135, Moscow. Accessible at: `http://ruscorpora.ru/sbornik2005/08lashevs.pdf.`

[3] Selegey, D., Shavrina, T., Selegey, V., and Sharoff, S. (2016). Automatic morphological tagging of Russian social media corpora: training and testing. In *Computational linguistics and intellectual technologies. Proceedings of International Workshop Dialogue'2016*, Moscow.

[4] Bocharov, V. V., Alexeeva, S. V., Granovsky, D. V., Protopopova, E. V., Stepanova, M. E., and Surikov, A. V. (2013). Crowdsourcing morphological annotation. In *Computational linguistics and intellectual technologies. Proceedings of International Workshop Dialogue'2013*, Vol. 12 (19), Moscow.

[5] Boguslavsky, I. (2014). SynTagRus–a Deeply Annotated Corpus of Russian. In Blumenthal, P., Novakova, I., and Siepmann, D., editors, *Les émotions dans le discours-Emotions in Discourse*, pages 367–380, Peter Lang, Frankfurt am Main, Germany.

[6] Nivre, J. (2016). *Reflections on Universal Dependencies*. Department of Linguistics and Philology, Uppsala University.

[7] Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, Ch. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., and Zeman, D. (2016). Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of LREC 2016*, pages 1659–1666, Portorož, Slovenia.

[8] Zalizniak, A. A. (1977/2003). *Grammaticheskij slovar' russkogo jazyka* [A Grammatical Dictionary of Russian.] Moscow.

[9] Sharoff, S., Kopotev, M., Erjavec, T., Feldman, A., and Divjak, D. (2008). Designing and evaluating Russian tagsets. In *Proceedings of LREC 2008*, Marrakech, Marocco.

[10] Lyashevskaya, O., Droganova, K., Zeman, D., Alexeeva, M., Gavrilova, T., Mustafina, N., and Shakurova, E. (2016). Universal Dependencies for Russian: a New Syntactic Dependencies Tagset. In *Series: Linguistics, WP BRP 44/LNG/2016.*

[11] Toldova, S., Sokolova, E., Astafiyeva, I., Gareyshina, A., Koroleva, A., Privoznov, D., Sidorova, E., Tupikina, L., and Lyashevskaya, O. (2012). Ocenka metodov avtomaticheskogo analiza teksta

2011-2012: Sintaksicheskie parsery russkogo jazyka [NLP evaluation 2011-2012: Russian syntactic parsers.] In *Computational linguistics and intellectual technologies. Proceedings of International Workshop Dialogue 2012*. Vol. 11 (18), pages 797–809, RGGU, Moscow.

[12] Lyashevskaya, O. (2016). The grammatical tagset of Russian. In Lyashevskaya, O. *Korpusnye instrumenty v leksiko-grammaticheskikh issledovavijakh russkogo jazyka* [Corpus approach to Russian grammar and lexicon], pages 435–456, Languages of Slavic culture press, Moscow.

[13] McDonald, R., Nivre, J., Quirmbach-Brundage, Y., Goldberg, Y., Das, D., Ganchev, K., Hall, K., Petrov, S., Zhang, H., Täckström, O., Bedini, C., Bertomeu Castelló, N., and Lee, J. (2013). Universal Dependency Annotation for Multilingual Parsing. In *Proceedings of ACL*. Accessible at: `https://ryanmcd.github.io/papers/treebanksACL2013.pdf.`

SUBCATEGORIZATION OF ADVERBIAL MEANINGS BASED ON CORPUS DATA

MARIE MIKULOVÁ – EDUARD BEJČEK – VERONIKA KOLÁŘOVÁ – JARMILA PANEVOVÁ
Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic

MIKULOVÁ, Marie – BEJČEK, Eduard – KOLÁŘOVÁ, Veronika – PANEVOVÁ, Jarmila: Subcategorization of Adverbial Meanings Based on Corpus Data. Journal of Linguistics, 2017, Vol. 68, No 2, pp. 268 – 277.

Abstract: We introduce a corpus based description of selected adverbial meanings in Czech sentences. Its basic repertory is one of a long lasting tradition in both scientific and school grammars. However, before the corpus era, researchers had to rely on their own excerption; but nowadays, current syntax has a vast material basis in the form of electronic corpora available. On the case of spatial adverbials, we describe our methodology which we used to acquire a detailed, comprehensive, well-arranged description of meanings of adverbials including a list of formal realizations with examples. Theoretical knowledge stemming from this work will lead into an improval of the annotation of the meanings in the Prague Dependency Treebanks which serve as the corpus sources for our research. The Prague Dependency Treebanks include data manually annotated on the layer of deep syntax and thus provide a large amount of valuable examples on the basis of which the meanings of adverbials can be defined more accurately and subcategorized more precisely. Both theoretical and practical results will subsequently be used in NLP, such as machine translation.

Keywords: adverbial meanings, deep syntax, annotation, treebank

1 INTRODUCTION

The description of adverbial meanings (local, temporal, manner, etc.) has a long lasting tradition and has been covered so far in Czech grammars and syntactic monographs in a varying granularity, with more or less detailed specification of the meanings (e.g. [1], [5], [6], [7], [8], [13], [23], [26], [27]).

However, it is well known that the traditional subclassification of adverbials is not grained enough for NLP tasks. For a deep syntax based machine translation (e.g. [4], [16], [28]), it is assumed that deep syntactic annotation narrows "the distance" between the source and the target language. For a successful transfer of a sentence from one language to another, it is necessary to capture all substantial information about the sentence meaning within the deep syntactic representation. The most important part of this representation is an accurate specification of meanings of particular modifications. On the deep syntactic layer in the Prague Dependency Treebanks (which serve as the corpus sources for our research; see their description in Sect. 2), the units of the sentence, i.e. content words together with their auxiliary words, such as prepositions and conjunctions, are represented by nodes of a tree-shaped graph. The tree reflects the underlying dependency structure of the sentence.

DOI 10.1515/jazcas-2017-0036

The types of the (semantic) dependency relations are represented by the "functor" attribute attached to all nodes. The functors represent relatively general categories. However, from the point of view of machine translation, they are not differentiated enough. For example, all the following modifications *na stole* 'on the table'*, pod stolem* 'under the table'*, za stolem* 'behind the table'*, poblíž stolu* 'near the table', etc. are covered by a single functor with a static meaning "where" (marked as LOC). Each of these modifications expresses the general meaning "where"; however, the introduction of a set of "narrower" meanings ("on the given place", "under the given place", "behind the given place", etc.) makes it possible to reflect the semantic differences among them. Thus, it is obvious that such a differentiation among the partial meanings is needed for a complete meaning of the sentence (and for its translation to another language). The requirement of the splitting one functor into more subtle units (called subfunctors here) occurs not only with spatial or temporal adverbials, but it concerns the other functors, e.g. accompaniment (with/without), regard (with respect to/without respect to), comparison (similarity, difference), etc. Illustrative examples of subfunctors were given in [22]. However, a comprehensive list of fine grained categories has not yet been developed.

To carry out a comprehensive and detailed subcategorization of all adverbial meanings and use it as a basis for creating a complete proposal of subfunctors requires a complex view on the theoretical core of the problem together with constant comparisons of proposed solutions with real data. In this paper, we shortly introduce our corpus sources and on the case of spatial adverbials, we describe our methodology used to fulfil our aims.

2 DATA: PRAGUE DEPENDENCY TREEBANKS

Large corpus sources are inevitable for a comprehensive study of subcategorization of all adverbial meanings. While many Czech corpora has morphological annotation (done automatically), we have to take into account the syntax. Nowadays, several richly syntactically annotated corpora, collectively called Prague Dependency Treebanks (PDTs in the sequel; [9]), have been already developed. These corpora provide a large amount of valuable examples that are used as a basis for the determination of subcategorized meanings of adverbials.

The annotation scenario of PDTs is reflected in several detailed annotation manuals (see [17], [18], and [19]). The main features of the annotation style are:

- well-developed dependency syntax theory which is known as the Functional Generative Description (FGD in the sequel; see [22], [24], [25]),
- interlinked hierarchical layers of standoff annotation,
- deep syntactic layer.

In the years 1996 through 2005, the first **Prague Dependency Treebank**[1] (PDT in the sequel; [11], for the latest version 3.0 see [2]) was designed and built. The data in PDT are composed by articles from the Czech daily newspapers. The slightly modified scenario was then used for the annotation of the Prague Czech-English Dependency Treebank, the Prague Dependency Treebank of Spoken Czech, and PDT-Faust corpus.

[1] `http://ufal.mff.cuni.cz/prague-dependency-treebank`

In contrast to the anchoring original project of PDT, in these treebanks, the morphological and surface syntactic annotations were done automatically and the manually annotated deep syntactic layer does not contain annotation of information structure and some other special annotations. However, annotation of functors, which we are mainly interested in here, is done manually in all four treebanks.

The **Prague Czech-English Dependency Treebank**[2] version 2.0 (PCEDT in the sequel, see [10]) is a manually parsed Czech-English parallel corpus. The English part consists of the Wall Street Journal section of the Penn Treebank [15]. The Czech part, which is used in our research, was translated from the English source sentence by sentence.

The **Prague Dependency Treebank of Spoken Czech**[3] version 2.0 (PDTSC in the sequel, see [20]) contains slightly moderated testimonies of Holocaust survivors from the Malach project corpus[4] and dialogues (two participants chat over a collection of photographs) recorded for the Companions project.[5]

The **PDT-Faust** is a small treebank containing short segments (very often with vulgar content) translated by the various users on the webpage reverso.net.

	PDT	**PCEDT**	**PDTSC**	**Faust**	**Total**
Tokens	833195	1162072	742257	33772	**2771296**
Sentences	49431	49208	73835	3000	**175474**

Tab. 1. Volume of data in Prague Dependency Treebanks

It is obvious that the Prague Dependency Treebank family provides rich language data for our purpose. Altogether, the treebanks include around 180 000 sentences with their deep syntactic annotation (see Table 1 and 2). Moreover, the PCEDT, PDTSC, and PDT-Faust treebanks will be also extended and corrected by manual annotation on the morphological and surface syntactic layers, and together with PDT, they will become a part of the **Consolidated Prague Dependency Treebanks** release in 2018, which will thus contain four different treebanks of Czech, uniformly annotated using the same scenario, with data coming from text, speech and Internet sources.

3 ANALYSIS OF ADVERBIAL MEANINGS

Our approach following the principles of the FGD is based namely on classification given in Novočeská skladba by Vladimír Šmilauer and Mluvnice současné češtiny 2 by Jarmila Panevová et al. Šmilauer´s classification of modifications ([26], pp. č-334) was used as the basis for constituting the set of functors for FGD as well as for the annotation on the deep syntactic layer in PDTs. The description of modifications given in Mluvnice současné češtiny 2 ([23], pp. 39-100) corresponds to the list of functors used for the annotation scenario applied in the PDTs. Similar detailed analysis of adverbials for English is given in [12].

[2] https://ufal.mff.cuni.cz/pcedt2.0/; https://catalog.ldc.upenn.edu/ldc2012t08

[3] http://ufal.mff.cuni.cz/pdtsc2.0

[4] http://ufal.mff.cuni.cz/cvhm/vha-info.htm

[5] http://www.companions-project.org

The starting point for our research is a subdivision of adverbial meanings into related groups which roughly correspond to the categories described in traditional Czech grammars (spatial, temporal, manner, causal, etc.). Then we gradually analyze one group at a time and generalize individual partial meanings of these modifications. The proposed set of subcategorized meanings is based on the detailed analysis of real examples gained from the PDTs.

Firstly, we study all formal realizations for each functor in the PDTs, i.e. we determine which parts of speech, cases, prepositions, and subordinate conjunctions were used to express the meaning of that particular functor. It means that for each functor we create a list of its formal realizations with a sufficient number of examples.

	LOC	**DIR1**	**DIR2**	**DIR3**
Occurrences	79874	17394	1590	28165
The most frequent forms (in all PDTs)	31531 v+6 17215 adv 13122 na+6 3566 u+2 1396 mezi+7 539 za+7 539 pod+7 461 před+7 393 po+6 350 nad+7	11965 z+2 692 od+2 594 adv 18 ze strany+2 4 zpoza+2 2 zpod+2	496 Instr 393 po+6 327 přes+4 73 adv 33 kolem+2 17 mezi+7 14 okolo+2 13 v+6 13 skrz+4 10 podél+2	9415 do+2 4740 adv 3644 na+4 2254 k+3 233 mezi+4 177 pod+4 170 za+7 106 za+4 90 na+6 61 před+4
The most frequent forms in *written* corpus (PDT)	11894 v+6 3902 na+6 1676 adv 1073 u+2 619 mezi+7 198 před+7 144 za+7 135 pod+7 110 kolem+2 94 v oblasti+2	4362 z+2 202 od+2 75 adv 3 zpoza+2 3 ze strany+2 1 zpod+2	207 Instr 96 přes+4 78 po+6 14 adv 10 mezi+7 3 skrz+4 2 vedle+2 2 nad+7 2 na+6 2 mimo+4	2936 do+2 1101 na+4 765 k+3 439 adv 97 mezi+4 60 pod+4 57 za+7 42 za+4 28 před+4 14 proti+3
The most frequent forms in *spoken* corpus (PDTSC)	14151 adv 7494 v+6 5145 na+6 1628 u+2 284 za+7 209 po+6 196 vedle+2 196 mezi+7 186 pod+7 142 před+7	2812 z+2 470 adv 312 od+2 6 ze strany+2 1 zpod+2	266 po+6 163 přes+4 114 Instr 52 adv 23 kolem+2 13 okolo+2 7 podél+2 5 skrz+4 4 mezi+7 3 podle+2	4002 do+2 3579 adv 1670 na+4 908 k+3 98 za+7 63 na+6 59 pod+4 42 za+4 27 mezi+4 24 po+6

Forms which are *only in written* corpus (PDT)	blízko+3 kol+2 na čele+2 nad+4 na úrovni+2 po boku+2 uvnitř+2 v čele+2 v rámci+4	zpoza+2	před+7 skrze+4 vedle+2	do čela+2 na roveň+2 vůči+3

Tab. 2. Raw frequency of forms of spatial adverbials in PDTs. Usually a preposition plus a case or adverbial phrase (adv) or a direct case (Instr).

The values of functors on the deep syntactic layer of PDTs reflect the semantic distinctions roughly corresponding to the traditional classification of adverbials. However, in the PDT scenario, the repertory of functors is used not only for the modifications dependent on verbs, adjectives and adverbs (i.e. of traditional adverbials, e.g. *Kniha leží na stole.* 'The book is lying on the table.'), but also for the modifications dependent on nouns (e.g. *kniha na stole je černá,* 'book on the table is black'). All modifications (dependent on verbs, adjectives, adverbs and nouns) with particular meanings are objects of our studies.

	LOC (where)	**DIR1 (where from)**	**DIR2 (which way)**	**DIR3 (where to)**
in	*v domě*	*z domu*	*domem*	*do domu*
inside	*uvnitř domu*	*zevnitř domu*	*vnitřkem domu*	*dovnitř domu*
inmiddle	*uprostřed domu*	*zprostřed domu*	*prostředkem domu*	*doprostřed domu*
athead	*na čele domu*	*z čela domu*	-	*do čela domu*
indiff	*po domech*	-	-	*po domech*
intarget	-	-	-	*Střílí po lidech.*
on	*na domě*	*s domu*	*po domě*	*na dům*
above	*nad domem*	*znad domu*	*nad domem*	*nad dům*
below	*pod domem*	*zpod domu*	*pod domem*	*pod dům*
behind	*za domem*	*zpoza domu*	*za domem*	*za dům*
front	*před domem*	*zpřed domu*	*před domem*	*před dům*
frontopp	*naproti domu*	*odnaproti domu*	*naproti domu*	*naproti domu*
near	*u domu*	*od domu*	-	*k domu*
beside	*vedle domu*	-	*vedle domu*	*vedle domu*
alongside	*podél domu*	-	*podél domu*	*podél domu*
around	*kolem domu*	-	*kolem domu*	*kolem domu*
across	*přes dům*	-	*přes dům*	*přes dům*
between	*mezi domy*	-	*mezi domy*	*mezi domy*
among	*mezi domy*	-	*mezi domy*	*mezi domy*
outside	*vně domu*	*zvnějšku domu*	*vně domu*	*vně domu*

Tab. 3. Distribution of subfunctors of four spatial modifications

We are aware that a theoretical description based on the relation of form and function needs a transparent and a systematic treatment reflecting the hierarchy functor – subfunctor. Since the correspondence between forms and their semantic functions is not one-to-one within a single functor, and not even between the form and the meaning within subfunctors, the determination and systemization of these units is considered to be a part of scientific description of language. The necessity of the subcategorization of the functors is further demonstrated by splitting of spatial modifications into 20 subfunctors (cf. Table 3).

4 SUBCATEGORIZATION OF SPATIAL MEANINGS

The functors for spatial meanings are distinguished according to the question specifying the location as follows (cf. [19, p. 474]):

LOC: where? (static modification, simple localization),

DIR1: where from? (directional modification with the meaning of setting out the starting point),

DIR2: which way? (directional modification; the path rather than starting point or destination),

DIR3: where to? (directional modification with the meaning of approaching a destination).

Subfunctors	**Forms**	**Examples**
in	*v+6* *na+6* *u+2*	*V tom údolí byly obrovské plantáže čaje.* (PDTSC)
inside	*uvnitř+2*	*A hle, uvnitř paláce stojí nový palác a nové hradby.* (PDT)
inmiddle	*uprostřed+2* *veprostřed+2* *ve středu+2* *vprostřed+2*	*Táborovou kaplí se stal indiánský stan teepee uprostřed tábora.* (PDT)
athead	*v čele+2* *na čele+2*	*Tank v čele kolony obrněnců se řítí na studenta.* (PDT)
indiff	*po+6*	*Kantoroval po mnoha městech.* (PDT)
on	*na+6* *po+6*	*Poskakoval kolem dokola po jevišti.* (Faust)
above	*nad+7*	*Vyvolal jsem to, nad kamny usušil film, nadělal fotky a večer je přinesl.* (PDTSC)
below	*pod+7*	*My jsme bydleli nahoře a oni bydleli pod námi.* (PDTSC)
behind	*za+7*	*Za hranicemi na mě čekala teta.* (PDTSC)
beside	*vedle+2* *po boku+2*	*Když se manželka oběti vrátila domů, pes pokojně seděl vedle mrtvého těla.* (PDT)
alongside	*dle+2* *podél+2* *podle+2*	*Podle Labe jsou břehy osázené duby.* (PDTSC)
front	*před+7*	*Vejprava fauloval před jabloneckou brankou Krejčíka.* (PDT)
frontopp	*proti+3* *naproti+3* *tváří v tvář+3* *čelem k+3*	*Jak je tam ten dům na fotografii, tak ten byl proti domu, kde jsem tehdy bydlela já.* (PDTSC)

near	*u+2* *při+2* *blízko+2* *blízko+3* *v blízkosti+2* *poblíž+2* *nedaleko+2*	*Na nádvoří odcizil i zaparkovanou Škodu 120, ale vozidlo odstavil nedaleko objektu.* (PDT)
around	*kolem+2* *okolo+2*	*Otevřeně se pokračuje v prodeji drog okolo škol, parků a sídlišť.* (PCEDT)
across	*ob+2* *přes+4*	*Bydlely jsme blízko sebe, přes ulici*. (PDTSC)
between	*mezi+7*	*hodina mezi psem a vlkem* (PDT)
among	*mezi+7*	*Bylo to otevřené, ale já jsem byla mezi posledními.* (PDTSC)
outside	*vně+2* *stranou+2* *mimo+2*	*To musí být strašně těžké být o prázdninách mimo domov.* (Faust)
indomain	*v oblasti+2* *v oboru+2* *na poli+2* *v rámci+2*	*Náklady na zaměstnance stoupají mnohem rychlejším tempem v oblasti zdravotní péče než v jiných odvětvích.* (PCEDT)
inlevel	*na úrovni+2*	*Přesun důležitých pravomocí se nezastaví na úrovni republik.* (PDT)

Tab. 4. Subfunctors, forms and examples for LOC functor

Based on the comprised lists of formal realizations and real examples (acquired from all PDT treebanks; comp. the large amount of acquired material in Table 2), we have proposed subfunctors for each of the four spatial functors. An overall overview of 20 proposed subfunctors is shown in Table 3; a detailed list with forms and examples for one spatial functor (LOC) is given in the Table 4. For the labelling of the subfunctors the preposition prototypical for the given meaning is used instead of a metalanguage signs.

	LOC	**DIR1**	**DIR2**	**DIR3**
on	*Děti běhají po trávníku.*	-	*Pojedeme po náměstí.*	-
indiff	*Vysedávali po náměstích.*	-	-	*Putoval po hradech.*
intarget	-	-	-	*Stříleli po lidech.*

Tab. 5. Functors and subfunctors of *po*+6 form

It is obvious that the boundaries between individual semantic distinctions are not always clear; many ambiguities have to be solved. There is no form – meaning isomorphy, one form is used for expressing more meanings and one meaning can be expressed using various forms. For example, with the form *po*+6 (‘on/along/around’) three different meanings of spatial modifications were distinguished. These meanings are schematically represented in Table 5. Combining the LOC functor with the subfunctor *on* captures a move (an action) which has no target but merely happens on the surface (e.g. *Děti běhají po trávníku/na trávníku.* ‘The children are running on the lawn.’). Combining the DIR2 functor with the subfunctor *on* captures a move on a surface from somewhere to another place (neither the starting point nor the destination is expressed; e.g. *Pojedeme po náměstí (až ke kostelu*). ‘We shall go along the square (till to the church).’ The *indiff* subfunctor means that the specified location takes place on several places of the same kind at the same time. It applies to all locations where the action usually/often happens (static LOC; e.g. *Vysedávali po náměstích.* ‘They used to sit around squares.’), or places where all individual action heads to (dynamic DIR3; e.g. *Putoval po hradech.* ‘He travelled around castles.’). The *po*+6 applied for DIR3 modification conveys a specific meaning with semantically limited group of verbs. The direction is here connected with a live target, a victim, at whom the action (mostly negative) is aimed (e.g. *Stříleli po lidech.* ‘They shot at people.’). This meaning is captured by *intarget* subfunctor.

The conditions for the distribution of the forms expressing closely related meanings (such as *Děti běhají po trávníku*. ‘The children are running on/along the lawn.’ vs. *Děti běhají na trávníku*. ‘The children are running on the lawn.’; *Stříleli po lidech.* vs. *Stříleli na lidi.* vs. *Stříleli do lidí*. ‘They shot at people.’; *Vysedávali po náměstích.* ‘They used to sit on/around squares.’ vs. *Vysedávali na náměstích.* *‘*They used to sit on squares.’) as well as the cases of lexicalization where two different prepositions express the same meaning (e.g. *Bydlí v Praze/v Dejvicích* ‘He lives in Prague/in Dejvice’ vs. *Bydlí na Kladně / na Letné*. ‘He lives in (lit. ‘on’) Kladno/on Letná.’) are studied. A study of the overlapping of meanings can contribute to the introduction of the new subfunctors. Our goal is to describe and analyse the cases of overlapping meanings from the theoretical point of view as well as in the form of practical guidelines for annotation procedure. Reliable criteria ensuing from the language system itself will be formulated in order to specify the partial meanings and subtle semantic distinctions.

The secondary prepositions and their specific meanings are studied as well as a wide range of expressions which more or less correspond with expressions generally perceived as secondary prepositions. They are temporarily tagged as potential candidates for the word-class of prepositions. For the LOC functor, there are, e.g., the following secondary prepositions: *ve středu*+2 ‘in the centre of’, *v čele*+2 ‘at the head of’, *tváří v tvář*+3 ‘face to face to’, *v oblasti*+2 ‘in the domain of’, *na poli*+2 ‘in the field of’ (see Table 4). The study of criteria for determination of the class of secondary prepositions in Czech and for their semantic and/or stylistic contribution to the meaning of the sentence with regard to the examples from corpora as well as to the results proposed in the printed papers and monographs (e.g. [3], [14]) is needed and it will be presented elsewhere.

5 EXPRESSING OF ADVERBIAL MEANINGS IN WRITTEN AND SPOKEN CZECH

The fact that we currently have different types of annotated corpora of the Czech language, particularly written texts corpus PDT and spoken texts corpus PDTSC offers a unique opportunity to compare expressions of adverbial meanings in written and spoken Czech in a precise and reliable way. The repertory of adverbial meanings and their formal realizations in both types of data has to be compared in more detail.

We expect a refinement of forms for expressing adverbial meanings in written text on the one hand, and marked, peculiar forms in spontaneous speech on the other hand (cf. similar observations in PDT corpora for valency modifications in [21]). Likewise, in a general and simple overview in the Table 2, we can observe that secondary prepositions for abstract and refined meaning (cf. *v oblasti+2* 'in the domain of', *v rámci*+2 'within the frame of', *na úrovni*+2 'at the level of', *po boku*+2 'alongside with', *na čele*+2 'at the head of', *do čela*+2 'to the head of') are more typical for written text. The secondary prepositions occur among the forms which are present only in written corpus and do not occur in spoken one.

6 CONCLUSION

We introduced here our research focused on a description of selected adverbial meanings in Czech sentences. On the case of spatial adverbials, we described our methodology and demonstrated that the Prague Dependency Treebanks provide us with valuable and rich material allowing us to elaborate the issue in depth. We believe that a systematic and accurate description of adverbial meanings verified on the basis of corpus material is necessary for comparative studies and for an application in NLP tasks as well as for a comprehensive syntactic description.

ACKNOWLEDGEMENTS

The research reported in the paper has been supported by the Czech Science Foundation under the projects GA16-02196S and GA17-12624S. The research has also been supported by the LINDAT/CLARIN project of Ministry of Education, Youth and Sports of the Czech Republic (project LM2015071).

References

[1] Bauer, J. and Grepl, M. (1972). *Skladba spisovné češtiny*. SPN, Praha.

[2] Bejček, E., Hajičová, E., Hajič, J., Jínová, P., Kettnerová, V., Kolářová, V., Mikulová, M., Mírovský, J., Nedoluzhko, A., Panevová, J., Poláková, L., Ševčíková, M., Štěpánek, J., and Zikánová, Š. (2013). Prague Dependency Treebank 3.0. Data/software, MFF, ÚFAL, Prague.

[3] Blatná, R. (2006). *Víceslovné předložky v současné češtině*. Lidové noviny, Praha.

[4] Bojar, O., Callison-Burch, Ch., Hajič, J., and Koehn, P. (2009). Special Issue on Open Source Machine Translation Tools. *The Prague Bulletin of Mathematical Linguistics*, 91.

[5] Daneš, F. et al. (1987). *Mluvnice češtiny 3*. Academia, Praha.

[6] Grepl, M. et al. (1997). *Příruční mluvnice češtiny*. Druhé opravené vydání. Lidové noviny, Praha.

[7] Grepl, M. and Karlík, P. (1986). *Skladba spisovné češtiny*. SPN, Praha.

[8] Grepl, M. and Karlík, P. (1998). *Skladba češtiny*. Votobia, Olomouc.

[9] Hajič, J., Hajičová, E., Mikulová, M., and Mírovský, J. (2017). Prague Dependency Treebank. In *Handbook on Linguistic Annotation. Volume II*, pages 555–594, Springer Science+Business Media, Dordrecht, Netherlands.

[10] Hajič, J., Hajičová, E., Panevová, J., Sgall, P., Bojar, O., Cinková, S., Fučíková, E., Mikulová, M., Pajas, P., Popelka, J., Semecký, J., Šindlerová, J., Štěpánek, J., Toman, J., Urešová, Z., Žabokrtský, Z. (2012). Announcing Prague Czech-English Dependency Treebank 2.0. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, pages 3153–3160, European Language Resources Association, Istanbul, Turkey.

[11] Hajič, J., Panevová, J., Hajičová, E., Sgall, P., Pajas, P., Štěpánek, J., Havelka, J., Mikulová, M., Žabokrtský, Z., Ševčíková-Razímová, M., and Urešová, Z. (2006). Prague Dependency Treebank 2.0. Data/Software, Linguistic Data Consortium, Philadelphia.

[12] Hasselgård, H. (2010). *Adjunct adverbials in English*. Cambridge University, Cambridge.

[13] Havránek, B. and Jedlička, A. (1960). *Česká mluvnice*. SPN, Praha.

[14] Kroupová, L. (1985). *Sekundární předložky v současné češtině*. ÚJČ ČSAV, Praha.

[15] Marcus, M., Santorini, B., and Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: the Penn Treebank, *Computational Linguistics*, 19(2):313–330.

[16] Mareček, D., Popel, M., and Žabokrtský, Z. (2010). Maximum Entropy Translation Model in Dependency-Based MT Framework. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 201–202, Association for Computational Linguistics, Uppsala, Sweden.

[17] Mikulová, M. (2014). Annotation on the tectogrammatical level. Additions to annotation manual (with respect to PDTSC and PCEDT). Technical report no. 2014/ÚFAL TR-2013-52, ÚFAL MFF UK, Prague.

[18] Mikulová, M., Bejček, E., Mírovský, J., Nedoluzhko, A., Panevová, J., Poláková, L., Straňák, P., Ševčíková, M., and Žabokrtský, Z. (2013). From PDT 2.0 to PDT 3.0 (Modifications and Complements). Technical report no. 2013/ÚFAL TR-2013-54, ÚFAL MFF UK, Prague.

[19] Mikulová, M., Bémová, A., Hajič, J., Hajičová, E., Havelka, J., Kolářová, V., Kučová, L., Lopatková, M., Pajas, P., Panevová, J., Razímová, M., Sgall, P., Štěpánek, J., Urešová, Z., Veselá, K., and Žabokrtský, Z. (2006). Annotation on the tectogrammatical level in the Prague Dependency Treebank. Annotation manual. Technical report no. 2006/30, ÚFAL MFF UK, Prague.

[20] Mikulová, M., Mírovský, J., Nedoluzhko, A., Pajas, P., Štěpánek, J., and Hajič, J. (2017, in press). PDTSC 2.0 – Spoken Corpus with Rich Multi-layer Structural Annotation. In *Lecture Notes in Computer Science*, Springer, Dordrecht, Netherlands.

[21] Mikulová, M., Štěpánek, J., and Urešová, Z. (2013). Liší se mluvené a psané texty ve valenci? *Korpus – gramatika – axiologie*, 8:36–46.

[22] Panevová, J. (1980). *Formy a funkce ve stavbě české věty*. Academia, Praha.

[23] Panevová, J., Hajičová, E., Kettnerová, V., Lopatková, M., Mikulová, M., and Ševčíková, M. (2014). *Mluvnice současné češtiny 2. Syntax na základě anotovaného korpusu*. Karolinum, Praha.

[24] Sgall, P. (1967). *Generativní popis jazyka a česká deklinace*. Academia, Praha.

[25] Sgall, P. et al. (1986). *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. D. Reidel Publishing Company, Dordrecht.

[26] Šmilauer, V. (1969). *Novočeská skladba*. 2. vydání. SPN, Praha.

[27] Štícha, F. et al. (2013). *Akademická gramatika spisovné češtiny*. Academia, Praha.

[28] Tamchyna, A., Popel, M., Rosa, R., and Bojar, O. (2014). CUNI in WMT14: Chimera Still Awaits Bellerophon. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 195–200, Association for Computational Linguistics, Baltimore.

MEASURING AND IMPROVING CHILDREN'S READING ALOUD ATTRIBUTES BY COMPUTERS

MAREK NAGY
Faculty of Mathematics, Physics and Informatics,
Comenius University in Bratislava, Slovakia

NAGY, Marek: Measuring and Improving Children's Reading Aloud Attributes by Computers. Journal of Linguistics, 2017, Vol. 68, No 2, pp. 278 – 286.

Abstract: In this paper, method of an automated measuring reading aloud attributes is presented. The forced alignment as a part of speech recognition technique is used. The recorded reading aloud is forced aligned to the known text and the attributes are computed from it. The tempo and fluency of children are monitored and used for an individual motivation. The length of the read text is chosen according to readers' skills so that children end up reading at about the same time and poor readers are not frustrated. This approach has been tested and improved at the elementary school for five years and brought positive results.

Keywords: speech recognition, teaching reading, reading aloud

1 INTRODUCTION

The reading is the main skill that children acquire in elementary schools. Step by step, children learn to recognize letters and combine them into words. When teachers want an effective feedback the children must read aloud. The teacher listens to the reading and corrects errors. Computer applications that use speech recognition can also do it automatically [1], [2], [3]. The children improve their reading aloud and read faster and smoother. In the second grade, children read longer texts that develop their reading skills. Reading aloud is only a temporary level to silent reading. However, it will be better to remain at this level some time because children have to learn to read words as a whole with its meaning at first. Whether it happens it is possible to find out by watching the tempo and fluency of reading. If children read fast it is clear that they are not syllabifying words. This means that they know and understand the words. It can be supposed that if the tempo and fluency of reading aloud is higher then children understand more words of the text. Monitoring of the reading tempo and fluency is an important thing for teachers. To help teachers these two attributes will be automatically measured by computer.

2 READING ALOUD ATTRIBUTES

The progress in the reading aloud can be measured by the tempo of reading. It is a common attribute that has been being used for long time. To measure this attribute a stopwatch is only needed. Teachers prepare text, which has already countered and signed words. Children start reading when the teacher starts the stopwatch. After one minute the teacher stops the stopwatch, counts the correctly read words and computes

DOI 10.1515/jazcas-2017-0037

words correct per minute (WCPM) ratio [4]. This procedure can be simply used in a computer. Children start reading into a microphone and when they finish the computer stops recording. The whole reading time is computed from the length of the recorded speech. A problem is at the beginning and end of the recording where silence can occur. It can be reduced by a silence detector. But a main problem is with automatic checking correctly read words. A computer speech recognition is used to identify words in recordings [1], [2], [5]. The methods are based on Hidden Markov Models (HMM) recognition approach. HMMs are trained from speech data as subunits e.g. phones, triphones, … [6] A substantial part of the recognizer is a recognition grammar or network, which determines possible sequences of subunits that can be recognized. To design such grammar is quite a problem. If the grammar is too complex (i.e. many possible sequences) an accuracy of the recognizer is lower and a processor time consumption is higher. In case of children's reading it must also be taken into account a fact that children can read aloud meaningless words e.g. when they syllabify words or they omit a phone.

Another requirement is a real-time response. Children will see their reading attributes immediately after they have read the text. Our experiments showed that if the feedback is late it is not a motivation factor.

We were used our Slovak speech recognizer, that was trained on children's speech data from Multimedia Reader [7]. The chosen HMM subunits are triphones. The isolated word accuracy is 98% with small dictionary (ca. hundred words) and phone accuracy is 88%. The recognizer is running on a server and must manage ca 20 reading children at once. If children are reading the same text then the recognition network can be shared what reduces a memory consumption.

We start to design grammar from the forced alignment approach. It means that if we know a potential transcription (the text), then we can compute time boundaries of individual words (of the text). Ideally, the transcription is a plain text as can children see and read from the computer monitor (see Figure 1). The syntax is taken from [6]. #sil represents a special unit that models silence. It can absorb a short or long pause between words.

#sil word$_1$ #sil word$_2$ #sil word$_3$ #sil …

Fig. 1. The optimistic transcription-grammar used for the forced alignment

However, when children make mistakes the transcription has to contain an appropriate correction. The typical mistake is repeating words (a double/triple reading) or omitting words (due to inattention). A situation of swapping words is infrequent and can be modeled as an incorrect reading and omitting. Therefore, the transcription is made as a recognition network that is build from a recognition grammar (see Figure 2). The curly brackets mean that their content can be repeating zero or more times.

#sil {word$_1$ #sil} {word$_2$ #sil} {word$_3$ #sil} #sil …

Fig. 2. The repeating and omitting transcription-grammar used for the forced alignment. The curly brackets denote zero or more repetitions.

This approach consumes too much processor time to compute the best sequence of words because the number of subsequences (with repetitions and omissions) is very high for long texts (more than 200 words). In this case, unfortunately, the recognizer will accept the reading with many omitted words, too. E.g. the tempo will be computed from one word only, what is undesirable. Therefore, the grammar was changed (see Figure 3). The angle brackets mean one or more times repeating. Now, words can’t be omitted. If children omit too many words then the recognizer will not find out a result and the whole reading is rejected. That is the reason why we introduce possibility for the recognizer to reject the reading. Our goal is that children will read the whole text correctly and if it needed they can read it again. It is an advantage against approaches where the recordings are static and processed off-line. Our experiments showed that the omitting is not a significant problem and therefore the rejecting can appears only infrequently.

#sil <word$_1$ #sil> <word$_2$ #sil> <word$_3$ #sil> …

Fig. 3. The repeating and omitting transcription-grammar used for the forced alignment. The angle brackets denote one or more repetitions.

According to [6] the recognition network is constructed using a dictionary that contains phonetic sequences for words. One word can have more phonetic sequences. See Figure 4.

word$_1$ = $w_1p_{1,1}$ $w_1p_{1,2}$ $w_1p_{1,3}$ … | $w_1p_{2,1}$ $w_1p_{2,2}$ $w_1p_{2,3}$ … | …

Fig. 4. The different phonetic sequences of the word$_1$. The vertical bars denote alternatives.

If we want to omit dictionary then the phonetic sequences have to be written into the grammar directly. See Figure 5. Of course, the dictionary must exist but it will contain direct phone-phone mapping only.

#w0
<($w_1p_{1,1}$ $w_1p_{1,2}$ $w_1p_{1,3}$ … | $w_1p_{2,1}$ $w_1p_{2,2}$ $w_1p_{2,3}$ …) #w1>
<($w_2p_{1,1}$ $w_2p_{1,2}$ $w_2p_{1,3}$ … | $w_2p_{2,1}$ $w_2p_{2,2}$ $w_2p_{2,3}$ …) #w2>

<($w_Np_{1,1}$ $w_Np_{1,2}$ $w_Np_{1,3}$ … | $w_Np_{2,1}$ $w_Np_{2,2}$ $w_Np_{2,3}$ …) #wN>

Fig. 5. The grammar used for the forced alignment. Every word is substituted by its phonetic transcriptions. Special non-terminals #w0- #wN (equal to #sil) are introduced.

The terminals word$_i$, which represent composed HMMs, can’t be used because phones are the terminals (HMMs) now. Therefore, we are using non-terminals #w1-#wN that represents silence (equal to #sil) between two neighboring words. The #w0 is a special case. The recognition algorithm (Vitterbi search) will generate sequence of phones, which is intermittented by silences #w0-#wN. The time boundaries of words can be deduced from the boundaries of silences #w0-#wN.

```
#w0
<($w_1p_{1,1}$ [#sil] $w_1p_{1,2}$ [#sil] $w_1p_{1,3}$ … | $w_1p_{2,1}$ [#sil] $w_1p_{2,2}$
[#sil] $w_1p_{2,3}$ …) #w1>
<($w_2p_{1,1}$ [#sil] $w_2p_{1,2}$ [#sil] $w_2p_{1,3}$ … | $w_2p_{2,1}$ [#sil] $w_2p_{2,2}$
[#sil] $w_2p_{2,3}$ …) #w2>

<($w_Np_{1,1}$ [#sil] $w_Np_{1,2}$ [#sil] $w_Np_{1,3}$ … | $w_Np_{2,1}$ [#sil] $w_Np_{2,2}$
[#sil] $w_Np_{2,3}$ …) #wN>
```

Fig. 6. The grammar used for the forced alignment. Every word is substituted by its phonetic transcriptions. Special non-terminals #w0-#wN (equal to #sil) are introduced and #sil is inserted between phones.

When children are reading they can syllabify words (or can use an unwanted double/triple reading). It introduces inaccuracy in a time alignment. So silences are inserted between phones, too (see Figure 6). Now, the algorithm is able to compute the time boundaries of words relatively good. The side effect is determination of silence between phones. It helps to express the fluency.

The grammar from Figure 6 is based on the assumption that children have read the whole text. If they leave out (or swap) more words (grater then ca. 10) then Vitterbi alignment algorithm [6] can't find the appropriate path – the word sequence. In our case it is not a problem because children are motivated by this way to read the whole text without too many mistakes. However, if the problem occurs children will read the text again.

Now we can compute time of the reading (in seconds), summary time of the words and summary time of the silences . It must be met . If we have a number of words the tempo can be computed as

$$tempo_{WO} = \frac{N_{WO}}{T} 60 \quad [words/min] \tag{1}$$

The tempo based on a number of words is not very reliable. Our experiments showed that the tempo based on a number of syllables is better, because our texts are not normalized and it happens that the text contains too many – 3 and more – syllabic words. (The tempo in the prosody is also based on syllables.) The syllables can be easily counted as a number of vowels. Diphthongs are a special case. And syllabic consonants r, l surrounded by consonants are understood as syllables.

$$tempo_{SY} = \frac{N_{SY}}{T} 60 \quad [syllables/min] \tag{2}$$

We proposed computing the fluency as:

$$fluency = \frac{T_W}{T_S} \tag{3}$$

The tempo and fluency are computed for the text as a whole and have global meaning. Our experiments showed that the attributes change during reading. In the future, the graph of it can help teacher identify the difficult sequences of words.

3 MEASURING THE READING ATTRIBUTES ON ELEMENTARY SCHOOL

Children read texts at school at lessons. A special subject has been introduced for this reason called Multimediálne čítanie, '*Multimedia Reading*'. It is done once in a week. Children use computers with the web application Multimediálna čítanka, 'Multimedia Reader' [7] that records their reading aloud by a microphone. When they finish the computer shows a simple one line diagram of the tempo and the fluency where children see their progress. Children are rewarded when they go up. It is an individual approach because a child only compares his/her actual personal performance with his/her previous ones. Children immediately see whether their attributes are rising or not. If they are not satisfied they can read again and improve their rating. After the corrective reading children can mostly see an instant progress, which is a motivating factor. The second and the other attempts make sense only for instant evaluation on the actual lesson. On the next lesson, the first attempts of previous lessons are only used in all statistics and graphs.

In this manner, children proceeded from the 1st to the 3rd grade. It happens that in 3rd grade the proficient readers tend to read aloud too quickly, so that they are not understood by others. E.g. they swallow word endings. To suppress inarticulate readings we set the limit on the tempo as **260** syllables/min. It is ca. 130 words/min. Kids who exceeded the limit were rewarded regardless whether they go up or down. Similarly, the fluency has limit set to **3**. We determined these limits experimentally. We started at 300 syllables/min but many children were frustrated by impossibility to reach it. Teachers also expressed objections that this limit is too high. We think that it relates to physical parameters of a vocal tract.

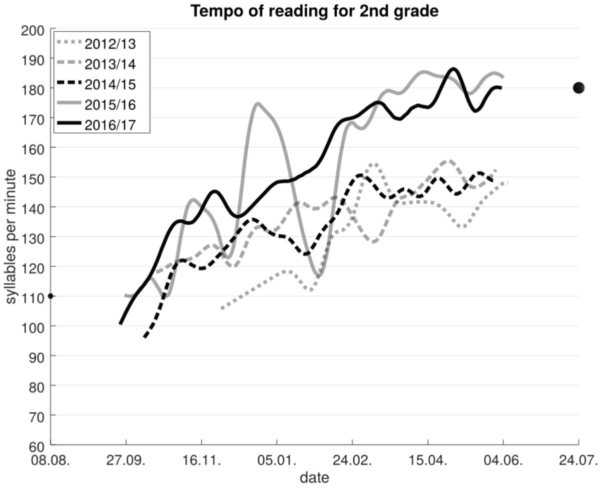


Fig. 7a. The graph of the average tempo of reading-aloud for 2nd graders. The actual 2016-17 and previous school years are showed.

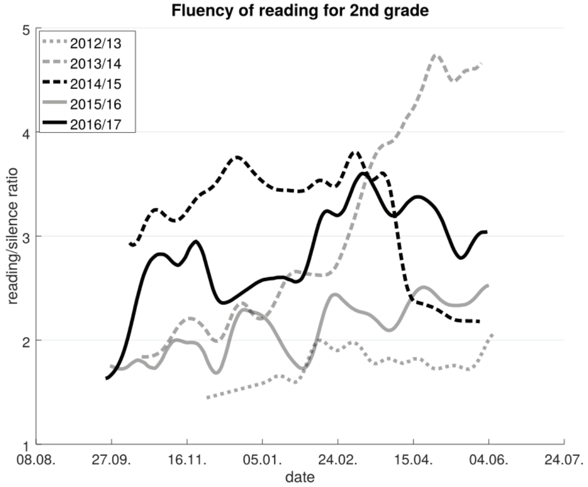


Fig. 7b. The graph of the average fluency of reading-aloud for 2nd graders. The actual 2016-2017 and previous school years are showed.

On the Figures 7a and 7b, improvements of the tempo and fluency are shown. Five consecutive school years are included. It were 2nd grade children in age 7 – 8 yo. The graph of averages is constructed on week base and every child performance is included only once a week. The yellow circles on boundaries of the graph (Figure 7a) correspond to average oral reading fluency norms, which are taken from [4]. As can be seen, within five years, we managed to increase the tempo of reading at an adequate level. The school years 2015/16 and 2016/17 are already within the norm. It should be noted that the number of syllables has been extrapolated from the number of words in the norm. It is ca. twice the number. We make average ratio for all 2 069 usable texts included in the Multimedia reader [7] and it gives syllables/words ratio equal to 1.9. The number of children, which are taken, is presented in Table 1.

school year	**girls**	**boys**	**total**
2012/13	21	16	37
2013/14	24	15	39
2014/15	21	19	40
2015/16	12	22	34
2016/17	29	19	48

Tab. 1. The number of 2nd grade children in school years

Statistical graphs for teachers and parents are more detailed as it can be seen on Figure 8. There are two examples. On the left side is a worse (below average) reader and on the right a better (above average) who is attacking the limit 260 syllables/min. The graphs show all improvement of reading from the 1st to 4th grade in general. The teacher can compare the child with the mean of all class or with the mean of all grade. It gives teachers a better chance to synchronize performances of kids.

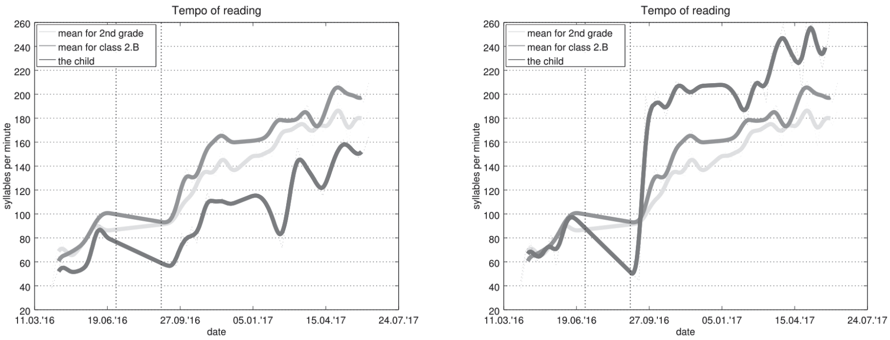


Fig. 8. The graph of the reading tempo of two 2nd graders. The graph shows the improving of reading skills from the 1st grade to the 2nd grade. The middle strip bounds summer holidays.

4 CHOOSING APPROPRIATE TEXT LENGTH FOR READING

At the beginning the length of the text that children are reading-aloud was determined roughly. On the 1st grade shorter and on the 2nd grade longer. But a problem appears. The children did not end at same time in reading. Better readers read the text sooner and started to have a fun. From the beginning, some extra activities were provided, but then kids were dispersed and could not concentrate on continuing the work with the text. How fast the text is read, depends on the tempo.

The development of the mean of tempo can be analyzed from experimental data. Typical standard deviation of the tempo was experimentally measured as from our collected data. If we imagine Gaussian distribution then the vast majority of readers will have the reading tempo in interval . Concretely, statistically, 68% of children. Now we can compute what time difference will be between a worse and a better reader. We must choose the appropriate tempo in syllables per minute. If we denote number of syllables of text by we can write:

$$\Delta t = \frac{N_{SY}}{\mu - \sigma} - \frac{N_{SY}}{\mu + \sigma} \quad [min] \qquad (4)$$

$$\Delta t = N_{SY}\frac{2\sigma}{\mu^2 - \sigma^2} \qquad (5)$$

It is good if children finish reading aloud at about the same time. If we choose a time difference , then we can compute how many syllables the text must have. See formula 6.

$$N = \Delta t \frac{\mu^2 - \sigma^2}{2\sigma} \quad [syllables] \qquad (6)$$

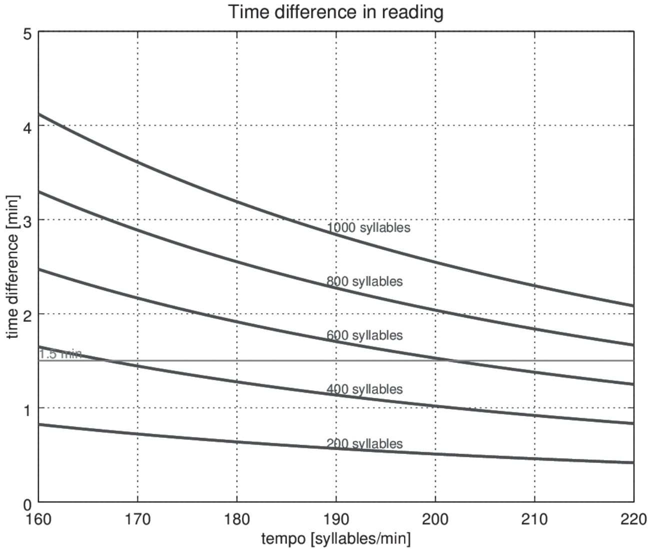


Fig. 9a. Graphs of dependency among the tempo and the finish time difference

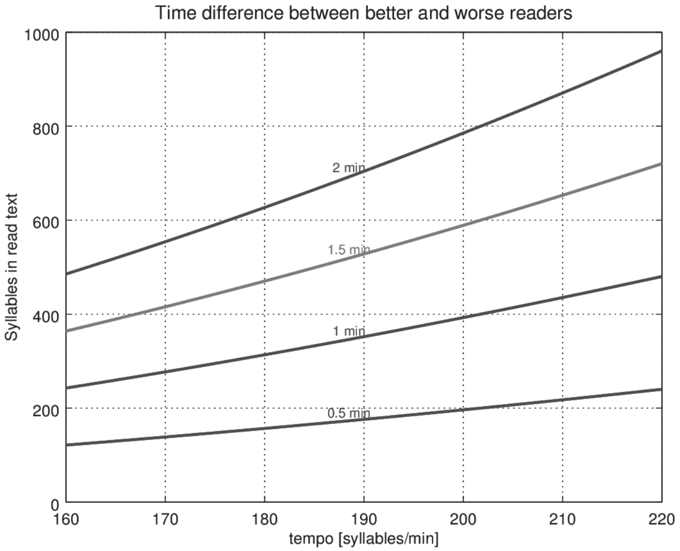


Fig. 9b. Graphs of dependency among the tempo and the text length

We have experimentally found out that the time difference must be less than 1.5 minutes. On the Figures 9a and 9b are graphs for the tempo 160 – 220 syllables/min. According to the Figure 7a the 2nd graders read ca. 180 syllables/min at the end of school year. The suitable text should have length ca. 500 syllables (ca. 250 words).

5 CONCLUSION

The measuring reading attributes, which is presented above, has brought automation in monitoring reading skills of children. Among other things, automation also brings a certain degree of objectivity. Teachers can compare children's performances and identify problematic ones. On the other hand, kids are proceeding in individual manner and are motivated to overcome themselves. Children are rewarded relative to their own performance and not on established values. This is particularly important for poor readers, who would be everytime at the end. The mentioned approach has improved the overall readership levels as is shown in Figure 7. Another motivating factor is the time aspect. It is important that children end up reading a text at about the same time. Otherwise, they are frustrated because they see that they are objectively slower. To avoid this problem it is necessary to choose an appropriate range of texts. If children end up reading at about the same time, it allows us to better manage the teaching, so children have no time for distraction.

References

[1] Mostow, J. and Roth, S. F. (1995). Demonstration of a reading coach that listens. In Hauptmann, A. G., editor, *Proceedings of the 8th Annual ACM Symposium on User Interface and Software Technology. UIST '95*, pages 77–78, ACM, New York, NY.

[2] Bolaños, D., Cole, R. A., Ward, W., Borts, E., and Svirsky, E. (2011). Flora: Fluent oral reading assessment of children's speech. *ACM Trans. Speech Lang. Process.*, 7(4):16:1–16:19.

[3] Patel, R. and Furr, W. (2011). Readn'karaoke: Visualizing prosody in children's books for expressive oral reading. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI '11*, pages 3203–3206, ACM, New York, NY.

[4] Hasbrouck, J. and Tindal, G. A. (2006) Oral Reading Fluency Norms: A Valuable Assessment Tool for Reading Teachers. *Journal of the Reading Teacher*, 59:636–644.

[5] Zechner, K., Sabatini, J., and Chen, L. (2009). Automatic Scoring of Children's Read-Aloud Text Passages and Word Lists. In *Proceedings of the NAACL HLT Workshop on Innovative Use of NLP for Building Educational Applications*, pages 10–18, Boulder, Colorado.

[6] Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., and Woodland, P. (2006). The HTK Book Version 3.4, Cambridge University, Cambridge.

[7] Nagy, M. (2005). Multimedia Reader (Multimediálna čítanka), a web application, `https://www.mmcitanka.sk`, (in Slovak).

THREE ASPECTS OF PROCESSING OPHTHALMOLOGICAL TERMINOLOGY IN A "SMALL LANGUAGE": A CASE OF CROATIAN TERM BANK STRUNA

BRUNO NAHOD[1] – PERINA VUKŠA NAHOD[1] – MIRJANA BJELOŠ[2]
[1] Institute of Croatian Language and Linguistics
[2] Clinical Hospital "Sveti Duh"

NAHOD, Bruno – NAHOD, Perina Vukša – BJELOŠ, Mirjana: Three Aspects of Processing Ophthalmological Terminology in a "Small Language": A Case of Croatian Term Bank Struna. Journal of Linguistics, 2017, Vol. 68, No 2, pp. 287 – 295.

Abstract: In this paper, we will present the problems we have observed while editing terminological units as a part of the specialized language of ophthalmology that is currently being processed as part of the program Struna. Struna is the Croatian National term bank (`http://struna.ihjj.hr/).` Its aim is to gradually standardize Croatian terminology, for all professional domains, by coordinating the work of domain experts, terminologists and language experts [1], [2]. *The Croatian Ophthalmological Terminology*[1] is the first Struna project that encompasses a subfield of an already existing field in the database. Namely, in 2013 the general medical terminology was processed as a part of the project *Croatian Anatomy and Physiology.* This situation has revealed a new set of problems that previously were not taken into account and has forced us to re-evaluate methodology and adapt accordingly.

Keywords: Struna, terminology, specialized language of ophthalmology, terminology management

1 INTRODUCTION

The Croatian national terminological database – Struna was first inaugurated in early 2012, introducing terminology from 10, mostly technical, fields. Over the years, new specialized languages were included with the terminology from 18 fields open to the public with additional six in various stages of processing. The currently ongoing project of *Croatian Ophthalmological Terminology* is unique in a sense that it is the first project where a highly specialized domain is being processed following the language of the wider domain. Namely, in 2013 the general medical terminology was processed and open to the public, presenting 2 575 terminological units from human anatomy and physiology.

As expected, a number of those terminological units did cover the domain of ophthalmology. The fact that terminological units processed in the past were processed from the general medical point of view where current one are being processed from a highly specialized one, lead us to the point where we were forced to rethink the whole methodology in Struna.

[1] This work has been fully supported by Croatian Science Foundation under the project Croatian Ophthalmological Terminology (Struna-2016-05-01).

DE GRUYTER OPEN

DOI 10.1515/jazcas-2017-0038

The problems of harmonization of multiple entries in Struna were discussed before by various authors, incorporating different aspects of terminology management [1], [2], [3]. The case discussed in this paper does transcend the scope of problems we have encountered earlier. Aside from the usual aspects of harmonization and standardization on the language and terminographic level, for the first time, the conceptual aspect did not include just a "simple" harmonization between different specialized fields but also a level of harmonization in-between two levels of expertise within the same domain.

In the following chapters, we will present the examples of newly encountered problems as well as solutions to them (in this stage of processing) from three inseparable aspects of terminology management: domain expertise, Croatian language standardization and conceptual one.

2 THREE ASPECTS OF TERMINOLOGY MANAGEMENT

2.1 The Specialized Language of Ophthalmology

The field of medicine comprises specific terminology that is estimated to include around 20 000 terms alone, apart from the nomenclature of the diseases, drugs and human anatomy. In the last two decades, there have been excessive developments in technology and revolution of Internet communication which have imposed challenges to generating new terminology as new diagnostic tools and diseases were accredited. When formulating terminology, we have to acknowledge the importance of the national and global effects of the strong points and shortcomings of these new terms' transcription. This being the case, a medical professional is likely to be accustomed to limited native vocabularies introducing local language expressiveness in opposition to providing discriminative, right and indexed national term. There are several international tools providing standardized medical terminology that can facilitate assistance to manual extracting terms to national medical corpora. Unfortunately, to our knowledge, the automatic term extraction tool optimized for Croatian language has yet to be developed. One such controlled thesaurus providing a hierarchically organized terminology for indexing articles and designating biomedical information is the NLM's (U.S. National Library of Medicine) MeSH (Medical Subject Headings). In addition, the World Health Organization's International classification of diseases (ICD) defines diseases, disorders, injuries and other related health conditions and stands as the international coding tool for reporting health status for all clinical and research evidence-based objectives. International Nonproprietary Names (INN) as determined by WHO alleviate the identification of globally recognized pharmaceutical substances. Terminologia Anatomica released in 1998 and developed by the Federative Committee on Anatomical Terminology (FCAT) and the International Federation of Associations of Anatomists (IFAA) is the international standard when it comes to human anatomic terminology. It comprises about 7 500 terms.

In ophthalmology, a wide variety of medical concepts: diseases, symptoms, diagnostic tests and results, therapeutic strategies is related to terms with unique context on the one hand and exceedingly inconclusive medical context on the other hand, the latter being associated with inaccurate interpretation of information and development of local community terminology, abbreviations and acronyms. The task of retrieving Croatian medical speech, phrases, definitions and words in the field of ophthalmology has always represented a challenge as no standardized specific bibliographical tool existed. In addressing those problems, the healthcare professional mostly relies on its own motivation, skills and language knowledge to bring high-quality morphosyntactic features of information through professional dictionaries, lexicons, encyclopedia and published evidence-based literature search. Possessing traits from Latin, British-American and German, many words have just been adapted to Croatian specific language forms with only minor differences in transliteration, retaining the originality of the item (examples like: ekscimer laser – excimer laser, hipotalamus – hypothalamus, keratoplastika – keratoplasty). The designation of visual acuity represents an everyday problem, as Latin, British-American English and Croatian spelling, acronyms and abbreviations are used simultaneously. In addition, multiword terms represent a special issue.

The example presented below is selected from the STRUNA dental medicine and physics catalogue tailored to their specific demands. Some of those lexical items overlap with the field of ophthalmology where they are associated with the different level of granularity.

Dental medicine catalogue:

Retina – the innermost coat of the eyeball containing photoreceptors sensitive to light. Remark: Retina is divided into a blind part (ciliary body and iris) and perceptive part (choroid). A better definition would be – the light receptive, innermost nervous coat of the posterior part of the eyeball consisted of ten layers, lying between the choroid and vitreous body, extending from the optic disk to the ciliary body. (Uvea is the vascular coat of the eye comprising iris, ciliary body and choroid.)

Astigmatism – refractive anomaly of the eye in which parallel rays of light refract in the dioptric system and are focused at more than one focal point. It is our strong opinion that the definition of astigmatism should be – refractive anomaly of the eye in which parallel rays of light from an external single point luminous source are not focused as a single point of an optical system, but instead are focused as two line images at different distances from the retina, generally at right angles to each other.

Stereopsis – binocular ability to perceive the relative distance between two near objects in order to perceive the depth of field. Remark: interchangeable with: binocular vision. The definition we are suggesting is – perception of relative distance, or the depth separation, between objects that occur as a result of neural processing of the relative horizontal binocular disparities between the monocular retinal images. Related term: Stereoacuity – acuity for the smallest relative binocular disparity stimulus (smallest relative binocular difference in distance of two objects) for depth that can be detected, specified by arcsec of disparity at the threshold.

Tonometry – indirect method of intraocular pressure measurement by measuring the tension of the eyeball. A more reasonable definition would be – measurement of ocular tension with a tonometer.

Intraocular pressure – pressure of aqueous humor that distends the eyeball. We, on the other hand, are proposing – the pressure of the intraocular fluid, measurable by a manometer.

Physics catalogue:

Myopia – insufficiency of the human eye by which the image produced by the lens is focused in front of the retina, corrected with the diverging lens. The definition, our team has agreed upon, is – the refractive condition of the eye represented as one in which parallel rays of light entering the eye with relaxed accommodation, focus in front of the retina.

Hypermetropia – insufficiency of the human eye by which the image produced by the lens is focused behind the retina, corrected with a converging lens. Farsightedness or hypermetropia should be defined as – the refractive condition of the eye represented as one in which parallel rays of light entering the eye with relaxed accommodation, focus behind the retina.

Astigmatism – error of the lens due to its different horizontal and vertical curvature. Based on our research the definition of astigmatism should be – refractive anomaly of the eye in which parallel rays of light from an external single point luminous source are not focused as a single point by an optical system, but instead are focused as two line images at different distances from the retina, generally at right angles to each other. □ corneal a. Astigmatism caused by the toroidal surface of the cornea. □ lenticular a. Astigmatism of the crystalline lens due to variations of curvature or to inequalities of refractive index.

2.2 Croatian Language Standardization

Medical terminology has from the start been closely connected to Greek and Latin languages which have served as a basis for medical communication on the national and international level. Most of the European languages have at some point used these ancient languages as a linguistic pool for their national medical terms. The idea behind using Latin for official medical documentation was a valid one when it was introduced. It was the main mean of allowing patients to have medical documentation written in *lingua franca* and enabling them a comprehensible medical history no matter the language barrier between a patient and doctor. Unfortunately (or luckily), the technological and medical advances in the 21st century have led to a widespread acceptance of English as the unofficial *lingua franca* in most scientific domains, and, consequently, medicine also.

The problem that has emerged in the late 20th and early 21st century is the rapid decline of usage of the native languages in medicine. This phenomenon is especially notable in the so-called small languages such as Croatian. Most of the medicine research done today is published in English, and we have been witnessing current textbooks published in English as well as lectures offered at universities. As a result, there is an evident shortage of Croatian terms, especially for new technologies and procedures.

A certain kind of renaissance of the awareness of the importance of systematized terminology in the Croatian language did happen in the last decade, with Struna being just one of its products.

Consequently, aside from producing terminological collections and making them available to the general public, one of the fundamental goals of Struna is popularization of existing Croatian terms, and when applicable, introducing new ones. Each terminological unit in Struna contains a preferred term with the associated part of speech information. A preferred term is the one in which both field expert and language experts have agreed upon as the best Croatian term. Considering how each terminological unit can be presented in various ways in a textual discourse, a synonymy section was introduced as a crucial language aspect of the terminological unit processing. The synonyms are categorized in six separated fields in the database, according to their level of acceptances within the Croatian standard. Each field is capable of keeping multiple terms, is related to the main table of the terminological unit, and makes a crucial part of the terminological unit both in editing stage and in public presentation. The categories are: admitted, deprecated, obsolete, colloquial and proposed. The admitted category contains the terms that are actively used by the field experts but not marked as preferred term due to a linguistic reason or overlapping with similar terms in other fields. The deprecated term is the one that is used in specialized texts but has been found as not appropriate according to Croatian standardization principles or as semantically inadequate to transfer the proper concept and its properties in the discourse – usually marked as such by the field expert. The obsolete term is the one that is no longer used in specialized texts and the colloquial term is the one used by the domain experts in informal communication.

As stated before, all of the terms are recorded in appropriate category by the experts-linguists consensus. The sole exception to this practice is the proposed term category which is activated in rare cases where domain experts and Croatian standard experts can't agree on the preferred term. In such cases, the domain experts' candidate is categorized as preferred and the linguists' one is categorized as proposed. The idea is that over the time the experts would possibly accept the proposed term and it will become the preferred one.

The main idea behind recording all the existing synonyms is to offer the end user a possibility to find the preferred term no matter what synonymous term is used in searching the database.

The Croatian terms can come into existence in a few different ways: by the 'pure' Croatian compounding (implantat – *usadak*), by the acceptance of internationalisms from ancient Greek and Latin or using elements from those languages in compounding new terms (mortality – *mortalitet*), by introducing foreign terms from modern languages (shock therapy – *šok-terapija*), by terminologization of general language lexemes (neck – *vrat*, root – *korijen* – of the teeth), reterminologization of existing terms in other domains (concrete – *cement*) and by compounding multiple words (*farmakotolerancija* – ability to take drugs) [5].

By analyzing the corpus of the Croatian ophthalmological terms we have identified three main problems with the existing terms extracted from medical dictionaries, textbooks and scientific papers.

1. The usage of internationalisms of Greek and Latin origin even though valid Croatian terms exist. The subcase of this problem should be noted where we have observed terms that were compounded using Latin or Greek elements.

2. Croatian equivalents for English terms don’t exist, therefore, a phonetized version of English term is used.
3. Wrong word formation – i.e. using English adjective instead of the noun when forming Croatian adjective.

retina		**mrežnica**
adjective	retinalni	mrežnični
multiword term	periferna retina	periferna mrežnica
multiword term	retinalna vena	mrežnična vena
multiword term	ablacija retine	odignuće mrežnice
multiword term	anomalna retinalna korespondencija	anomalna mrežnična korespondencija
pupila		**zjenica**
adjective	pupilarni	zjenični
multiword term	pupilarna membrana	zjenična membrana
multiword term	pupilarni refleks na svjetlost	zjenični refleks na svjetlost
sklera		**bjeloočnica**
adjective	skleralni	bjeloočni
multiword term	skleralni prsten	bjeloočni prsten
multiword term	skleralna leća	bjeloočna leća

Tab. 1. An example of preferred Croatian terms for Latin and Greek synonyms

Table 1 shows the examples of Latin and Greek synonyms being replaced with purely Croatian terms. One of the main conditions for this kind of procedure is that the Croatian term is productive in a sense of related terms formation, in most cases this being the ability to make a valid adjective from a noun which is used in multiword terms.

The same principles are applied when English terms are translated into Croatian (Table 2). The corpus analysis has shown that most of those terms in Croatian texts are used in their English version, and when Croatian terms do appear they are usually noted in braces.

crowding	zbijanje
overlap masking	prekrivanje
crosslinking	umreživanje
cover test	test pokrivanja
uncover test	test otkrivanja

Tab. 2. An example of preferred Croatian terms for English synonyms

A special case involving English terms has been observed while analyzing medical corpus. A certain number of terms were found that were multilingual. Namely, a part of the multiword term was left in English and the other part was translated into Croatian.

Typically, these terms deprecate new concepts for which there is no traditional Croatian synonym or related term. Therefore, the experts, when using them in a text, simply leave the part of the term that can't be easily translated into English, and only translate the part of the term. Examples of such terms are *frequency-doubling perimetrija* – from eng. frequency-doubling perimetry and *double-void tehnika* from eng. double-void technique. The problem of standardizing these terms comes from the fact that by the time they are 'marked' as problematic by language expert, they are widely used in scientific discourse, and it is hard to change them.

The third case of problematic terms refers to the ones that are simply phonetized from foreign language. Such as *skrinig* – screening, *distraktor* – distractor etc. In these cases, it is preferable to find a proper Croatian term i.e. *probir* for screening and *ometač* for distractor.

There are several problems observed that occur in the creation of Croatian terms incurred by eponym or from the English adjective or from the Latin prefix. Eponyms are commonly used in medicine terminology, and ophthalmology is no exception. The English language has several ways of forming eponyms. Until recently, the most numerous were the eponyms containing synthetic genitive: e. g. Purtscher's retinopathy, Horner's syndrome. Today, they are being replaced by another way of forming eponyms: substantively adjunct + principal noun: e. g. Edinger-Westphal nucleus, but in Croatian these eponyms must be changed into a construction with a possessive adjective: *Purtscherova retinopatija*, *Hornerov sindrom*, *Edinger-Westphalova jezgra* [4].

In spite of the clear and explicit term-forming principles, terms that are simply left in original English form or treated as an abbreviation, such as *Hess-Lancaster test/Hess-Lancaster-ov test* and *Hirschberg test/Hirschberg-ov test,* are often found in texts.

2.3 Conceptual Aspects

As we have mentioned before, along with the more common problems of forming terms for new concepts using terms from languages such as English, Greek or Latin, a problem we had not encouraged before is the one of the different semantic extent of the same term, based on the more narrow specialization of the domain.

concept	definition in anatomy
eye	visual organ located in the orbit
retina	part of the inner layer eyeball which contains a light-sensitive cells

Tab. 3. Ophthalmological concepts that were defined as part of anatomy terminology.

Table 3 shows the most basic examples of the ophthalmological concepts that were defined previously as a part of anatomical terminology. Both of these definitions are good when we consider them from a discourse of general medical anatomy. As soon as an ophthalmologist observes them it is clear they are not acceptable as a part of the specialized language of ophthalmology.

During our work on various specialized languages in Struna, we have observed numerous examples of this kind of conceptual variance between two or more different domains [5], [6].

Considering that our end users find multiple occurrences of the same term with different definitions distracting and confusing even when they appear as search results in two or more different domains, we can assume that two different definitions for the same concept inside the field of medicine would be even more unwanted advent.

It has been argued by many researchers that the classical approach to terminology (based on the so-called Vienna School [7], [8]) is not flexible enough to deal with this kind of conceptual variations [8], [9], [10], [11], [12]. Unfortunately, terminology management in Struna is currently based on the Vienna School and no elegant solution can be offered for this kind of problems in the present. The only 'solution' is to enter new terminological units in the domain of ophthalmology, which will coexist independently of all the terms that were edited in the past.

Starting in the year 2014, the researchers working on Struna have started to develop a new model for terminology management; Domain Cognitive Models (DCM) [5], [12] [13], [14]. The DCM is a sociocognitive based paradigm for processing and presenting specialized languages that is trying to solve exactly this kind of problems. It is currently in a testing stage (`http://skm.ihjj.hr/`), showing promising results. Hopefully, it will soon be implemented in Struna as an additional method for processing terminological units, not as an alternative but as an integral module for dealing with conceptual substructures that are impossible to process using the traditional terminological principles.

3 CONCLUSION

Struna is the Croatian national term bank the aim of which is to (eventually) include processed specialized languages from most specialized domains that are being researched in Croatia. *Croatian Ophthalmological Terminology* is the first project under the Croatian Special Field Terminology program (Struna) that is covering terminological units from the domain that can be considered to be a highly specialized subfield of the domain that was previously processed in Struna. Furthermore, besides that ophthalmology is a subfield of medicine it is also a profession that has experienced an incredible progress in theory, praxis and technology in the last few decades. This has led to numerous new problems that we have not encountered before.

We have categorized the observed problems and presented them according to three unique, yet obviously mutually dependent aspects: the one of the domain specialist, linguistic one, with emphasis on the Croatian standard, and the conceptual or terminographic aspect.

Even though the problems we have observed and identified during our work on processing ophthalmological terminology can be categorized in three seemingly independent categories, it is evident that none of them can be solved by the aspect's expert respectfully. The domain expert, in our case the ophthalmology practitioner, the Croatian standard expert and the terminologist have to work together on each individual case and solve the problems by coming as close as possible to a consensus, bringing all three aspects of terminology processing into a unified model of terminology management.

We have shown that most of the problems can, and will be solved using the well-established principles and praxis that are employed in Struna. On the other hand, some of the problems that have arisen during our work on the specialized language of ophthalmology will not be able to solve within the classical terminology principles, and will eventually lead to further research of both terminological theory and terminographic praxis.

References

[1] Mihaljević, M. and Nahod, B. (2009). Croatian Terminology in a Time of Globalization. In *Terminologija in sodobna terminografija*, pages 17–26.

[2] Bergovec, M. and Runjaić, S. (2012). Harmonization of Multiple Entries in the Terminology Database Struna (Croatian Special Field Terminology). In *Proceedings of the 10th Terminology and Knowledge Engineering Conference (TKE 2012)*, pages 231–241.

[3] Nahod, B. (2016). Can Big National Term Banks Maintain Complex Cross- Domain Conceptual Relations ? *In Term Bases and Linguistic Linked Open Data / 12th International conference on Terminology and Knowledge Engineering*, pages 1–13.

[4] Ostroški Anić, A. and Lončar, M. (2013). Eponymous medical terms as a source of terminological variation. In *Languages for Special Purposes in a Multilingual, Transcultural World, Proceedings of the 19th European Symposium on Languages for Special Purposes*, pages 36–44.

[5] Nahod, B. and Vukša Nahod, P. (2014). On Problems in Defining Abstract and Metaphysical Concepts – Emergence of a New Model. *Coll. Antropol.*, 38(2):181–190.

[6] Bergovec, M. and Runjaić, S. (2015). Teorijske dvojbe i mogućnosti usklađivanja višestrukih terminoloških zapisa u Struni. In Bratanić, M., Brač, I., and Pritchard, B., editors, *Od Šuleka do Schengena*, pages 237–248, Institut za hrvatski jezik i jezikoslovlje, Zagreb.

[7] Wüster, E. (1979). *Einführung in die allgemeine Terminologielehre und terminologische Lexikographie*. Springer, Wien – New York.

[8] Felber, H. (1984). *Terminology Manual.* Infoterm, Vienna.

[9] Cabré Castellví, M. T. (2000). Elements for a Theory of Terminology: towards an alternative paradigm. *Technology*, 6(1):35–57.

[10] Temmerman, R. and Kerremans, K. (2003). Termontography : Ontology Building and the Sociocognitive Approach to Terminology Description. *Appl. Linguist.*, Cvc:1–10.

[11] Faber, P. and Martín, A. S. (2010). Conceptual Modeling in Specialized Knowledge Resources. *International Journal Information Technologies & Knowledge*, 4(2):110–121.

[12] Nahod, B. (2015). Domain – specific Cognitive Models in a Multi – Domain Term Base. *Suvremena lingvistika*, 41(80):105–128.

[13] Nahod, B. (2015). Brak čestice i prostora: sociokognitivna proedbena analiza pojmovnih struktura strukovnih jezika fizike i antropologije. In Bratanić, M., Brač, I., and Pritchard, B., editors, *Od Šuleka do Schengena*, pages 169–196, Institut za hrvatski jezik i jezikoslovlje, Zagreb.

[14] Nahod, B. (2016). *O umu stručnjaka*. Institut za hrvatski jezik i jezikoslovlje, Zagreb.

TERMINOLOGY AND LABELLING WORDS BY SUBJECT IN MONOLINGUAL DICTIONARIES – WHAT DO DOMAIN LABELS SAY TO DICTIONARY USERS?

JANA NOVÁ – HANA MŽOURKOVÁ
The Institute of the Czech Language, Academy of Sciences
of the Czech Republic, Prague, Czech Republic

NOVÁ, Jana – MŽOURKOVÁ, Hana: Terminology and Labelling Words by Subject in Monolingual Dictionaries – What Do Domain Labels Say to Dictionary Users? Journal of Linguistics, 2017, Vol. 68, No 2, pp. 296 – 304.

Abstract: The paper focuses on labelling words by subject in a non-specialized dictionary. We compare the existing monolingual dictionaries of Czech and their ways of labelling terms of medicine and related fields; besides apparent differences between dictionaries, there are also inconsistencies within one dictionary. We consider pros and cons of domain labels as such and their usability in the light of needs and limits of dictionary users, with the aim to motivate further discussion on related issues.

Keywords: terminology, terms, lexicography, monolingual dictionary, e-dictionary, domain labels, Czech

1 INTRODUCTION

The accelerated development of science during the past century and the development of mass communication are followed by growing interest of the public in terminology, especially in several past decades. Specialized vocabularies of all fields of interest become part of laypersons' lives as a part of popular culture because of increasing use of scientific terms in mass media [1], [2]. Naturally, in monolingual dictionaries of Czech the number of terms steadily increases.[1] However, as it will be argued in this article, concepts of terminology processing in monolingual dictionaries of Czech vary and so does the level of their usability and user-friendliness.

Despite the aforesaid quick development of terminology, in Czech especially since the 1990', linguistic attention to this field is not sufficient. Except for major works of Ivana Bozděchová [4], [5], most papers devoted to (selected aspects of) terminology of specific fields were published with medicine being the most frequent topic [6, 7, 8]. There is a lack of studies dealing with the treatment of terminology in (non-specialized) dictionaries,[2] a few authors comment on domain labels [15], [12], [16], [17].

In our paper we will focus on labelling process and domain labels in monolingual dictionaries of Czech. Our examples are mostly taken from medicine and rela-

[1] We have proved it for the field of medicine in our paper [3].

[2] Not considering prefaces of monolingual dictionaries and similar conceptual materials [9], [10], [11], we have only one specialized study by Jaroslav Machač [12] in Czech linguistics; more recent works come from Slovak authors [13], [14].

DOI 10.1515/jazcas-2017-0039

ted fields.[3] Traditionally, special attention is paid to medical terminology, both by Czech (see above) and foreign linguists [18], [19], [20], [21], reflecting the importance of medicine and growing interest of the public in the matters of health and illness. Even more significantly for our purposes, in the field of medicine there exist apparent distinction between terminology and substandard, slang expressions[4] (comparing for example with computer science) and between terminology and general vocabulary (comparing with some humanities).

The paper is organized as follows. Firstly, we conceptualize the subject of terminological research and define what the term is from the lexicographic point of view, mentioning also the process of de-terminologization and its consequences for dictionaries. We present how the term is "signalized" with its label in a monolingual dictionary, and give a comparison with the following Czech academic dictionaries: Slovník spisovného jazyka českého, SSJČ, publ. 1960–1971 (The Dictionary of Standard Czech Language), Slovník spisovné češtiny, SSČ, publ. 1978, 3rd ed. 2003 (The Dictionary of Standard Czech), Nový akademický slovník cizích slov, NASCS, publ. 2005 (The New Academic Dictionary of Loanwords), and Akademický slovník současné češtiny, ASSČ, publ. on-line since 2017 (The Academic Dictionary of Contemporary Czech).[5]

Secondly, we assess usability of domain labels for dictionary users, relating to the problems of terminology and labelling presented previously. We are aware of the fact that we present more questions than definite answers and we are ready to discuss the topics at the Slovko conference.

2 TERM AS A RESULT OF CONCEPT FORMATION IN A SPECIFIC FIELD OF SUBJECT[6]

What is the term? From the lexicographic point of view, a scientific term is "a lexical item [...] used in a particular domain of expertise" where it is identified with "a rigidly fixed obligatory range of meaning" [2]. In ASSČ, a scientific term is "name of a concept in the concept system of a particular scientific, technical, economic or other field"[7]. Nomenclature is sometimes treated separately, but for dictionary purposes we regard it as an integral part of terminology. Most scientific terms are multiword units [4] which brings a question how to incorporate them into a dictionary structure.

Despite aforementioned definitions of terms, it is often difficult to determine whether a particular word is or is not a term. Regarding parts-of-speech, only nouns,

[3] Borderlines between medicine and related sciences (pharmacy, biology, biochemistry, psychology) are often unclear, also medical disciplines such as anatomy or physiology can be treated as separate sciences. Dictionaries choose various ways how to label headwords belonging to these fields, as we will discuss further.

[4] For the definition and concept of slang and professional vernacular in Czech linguistics see [22].

[5] Authors of this paper are members of the team of ASSČ.

[6] [23]

[7] "Termín se v ASSČ chápe jako pojmenování pojmu v systému pojmů některého vědního nebo technického oboru, hospodářského odvětví a dalších oborů lidské činnosti" [11]; following [24].

(underived) adjectives and a specific group of adverbs[8] are usually considered as terms [25], [15], [26]. Adjectives derived from nouns considered to be terms (such as *arteriální,* ‘arterial’) are sometimes treated as terms on their own, sometimes only as parts of terminological collocations (*arteriální hypertenze,* ‘arterial hypertension’). This uncertainty is reflected by variation of labelling in dictionaries, see below. Verbs are mostly not accepted as scientific terms [27], [6] and thus not labelled as terminological units in general dictionaries, not even when a terminological noun is actually derived from the verb (*aspirovat* ,‘to aspirate’ → *aspirac*e, ‘aspiration’).

Monolingual dictionaries include terms of the common-use while items belonging to the supernorm [28] too specialized for a general dictionary, with a low frequency or only occurring in specialized scientific literature,[9] are commonly absent.

Sager [31] makes a difference between primary and secondary term formation. The former is a process of designating a new concept, the formed terms come from the general vocabulary. The latter is a process starting with an already existing term. Terms can be borrowed from another language, too, English being their primary source these days.

When terms move from the specialized to general language[10] and they are no longer used exclusively in expert discourse, they become a part of laypersons’s communication and their “fixed” nature changes; there is “a variety of semantic, grammatical and pragmatic changes that may occur during de-terminologization” [2]. For the treatment of de-terminologized units in a dictionary, it appears to be the most important whether a clearly distinct new meaning of the original term has developed in the general language (*adrenaline* 1. a hormone, 2. strong excitement and emotions, 3. a thrilling activity causing such excitement[11]), or whether the new usage of the word means “only” blurring or shifting the original meaning, making it less definite (*angína* ‘tonsilitis’ → any disease with sore throat). The second case brings difficulties how to deal with the particular entry in a dictionary, also whether to keep labelling it as a term.

3 LABELLING OF TERMS

In monolingual dictionaries domain labels are used to signalize the subject matter of the headword (or one of its meanings) in a specific field of interest.

[8] In the field of music (terms of Italian origin like *allegro*, *andante*) and sport (*snožmo* ‘with legs together’, *obouruč* ‘both-handed’).

[9] For SSJČ, SSČ and NASCS headwords-terms were mostly picked from a huge lexical archive built by manual excerption and from older general encyclopedias [12]. The wordlist of ASSČ is built on using Czech language corpus SYN [29]; a minimal frequency is given for all words to be included into the wordlist, and there is another condition for scientific terms, they must occur in non-scientific literature too [30].

[10] “A determinologized lexeme is the result of the transition of a term from a specific terminology to a general lexical inventory, or to put it another way, from scientific texts to texts aimed at the general public” [32].

[11] Paraphrased version of the entry *adrenalin* in ASSČ: “1. chem., biol. hormon dřeně nadledvin regulující krevní tlak a ovlivňující činnost centrálního nervového systému; 2. kolokv. △ silné vzrušení, napětí, silné emoce; 3. kolokv. △ sportovní aktivita nebo jiná činnost spojená s rizikem a nebezpečím, vyvolávající silné vzrušení, emoce účastníků”.

Comparing the existing monolingual dictionaries of Czech, slight inconsistency in labelling principles is evident, especially:

a) The same domain is represented by various labels in various dictionaries, see pharmacy and veterinary medicine in Tab. 1.

b) The same headword is sometimes labelled for a specific domain and sometimes stands without any label, see *afázie* (‘aphasia’), *angína* (‘tonsilitis’), *astma* (‘asthma’), *astenik* (‘asthenic’), *aspirin*, *antikoncepce* (‘contraception’) in Tab. 1. In the small dictionary SSČ domain labels are used very rarely, while in NASCS derived adjectives etc. are often labelled, too.

c) The same headword is sometimes labelled for different domains in different dictionaries, see *akrální* (‘acral’), *apofýza* (‘apophyses’), *kyselina askorbová* (‘ascorbic acid’) in Tab. 1.-

d) Combinations of labels are sometimes used for closely related domains (medicine + psychology: *adolescence, apatie* ‘apathy’; medicine + pharmacy: *analgetikum* ‘analgetic’*, anestetikum* ‘anaesthetic’*, antipyretikum* ‘antipyretic’*, antiseptikum* ‘antiseptic’), but also for major discipline and its subdiscipline (medicine + anatomy: *abdominální* ‘abdominal’, *apendix, autonomní* ‘autonomous’); see Tab. 1. We noticed different labelling of the same word category in SSJČ and different order of labels in NASCS (*analgetikum* × *antipyretikum* etc.). Preparing the first part of ASSČ for publication, we did our best to make the label system clear and consistent.

Lastly, we should mention that domain labels are traditionally used in the Czech and Slovak dictionaries and ASSČ follows this tradition, while most e-dictionaries of English do not use domain labels and it seems their users do not miss this kind of information.

4 DOMAIN LABELS AND USERS

Now we come to the crucial part of our paper: what do domain labels say to dictionary users? In Germany, a survey among users of the e-dictionary *elexico* (mostly linguists and professional translators) was performed [33]; the participants considered domain labels as generally useful. A direct and detailed research about stylistic and domain labels in dictionaries of Czech would be welcomed, too.

A domain label makes a direct linkage between a headword and a field where the word is used. However, this kind of information is obvious from the headword definition and exemplification, as well: a dictionary entry beginning with “zánětlivé onemocnění” (‘inflammatory disease’) or “lékařský přístroj” (‘medical apparatus’) indicates the field of medicine.

Let us consider other assets of domain labels. Within a polysemic word, a domain label could make searching faster, telling the users immediately whether they are/are not in the field of their interest without reading the full entry. But we disagree that there should be no difference in the way how polysemic and monosemic entries are presented in a dictionary.

Domain labels are also used to mark multi-word items containing words of general vocabulary within the entries of their one-word components: e. g. large entry *bílý* (‘white’) in Czech dictionaries contains terminological collocations from medicine *bílé krvinky* (‘white blood cells’), *bílá hmota* (‘white matter’) etc. Nevertheless, condensation of multi-word units within the one-word entry used to be motivated by limited space in printed dictionaries while in electronic dictionaries there are other ways to present or highlight multi-word terms (separate paragraphs, separate entries, colours – cf. [34]) and a domain label is not necessary just to say “notice me”.

The fact that dictionaries use different labels for one domain (lékár. – farm. – farmac.) or combine more labels for one headword should not cause troubles for dictionary users as long as the system is consistent within the particular dictionary and as long as there are not too many labels which would complicate user’s orientation in the entry. However, we see as a problem the situation presented in Tab. 1, case d, when headwords of the same category are labelled differently or the order of labels varies. Then users can be confused whether it is an intention (and what does the different label/order signalize?) or a mistake of editors (more likely). Using an electronic dictionary writing systems (DWS Alexis is used for making ASSČ, cf. [35]) allows lexicographers to check and unify labels before publishing far better than in the pre-computer era.

A domain label in Czech dictionaries actually gives two kinds of information at the same time: 1) the word belongs to a specific field of interest; 2) the word is used in specialized communication [12]. In SSJČ and ASSČ a general label odb. (“odborný”, ‘professional’) is used where 3 or more labels of different domains would have to be used; this label system is explained in prefaces of these dictionaries. However, common users are not used to reading dictionary prefaces and then they can be confused, considering the label odb. to be a stylistic one (like slang., expr., hanl. = ‘pejorative‘, etc.), thus headwords without odb. may cause misperception and be understood as non-professional, especially when there is no domain label – e. g. derived adjectives like *arteriální,* ‘arterial’, not labelled even in the “maximalist” NASCS, but certainly belonging to specialized communication and not to the field of general vocabulary. One might argue that SSJČ and NASCS treat derived adjectives within the entry of their base (the nesting principle), so for *arteriální*, the base noun *arterie* is labelled as anat. and users still get this information. In contrast, ASSČ does not use nesting and the web interface of this e-dictionary always presents a single entry, without the context of related words as it was on a book page;[12] when users see the entry *arteriální*, they are not very likely to check the entry *arterie* for the information about the domain. To be as precise as possible, several types of labels (the domain; specialized/general communication; whether the word is/is not an exactly defined term) would have to be combined for each headword; but such system would be very complicated and would break the lexicographic rule that entries must be clearly arranged and must not overload users with too much information (cf. [36]). Homoláč and Mrázková [16] presented an elaborated system of stylistic labels where the field of scientific communication is represented by an abbreviation vkc (“vyšší komunikační cíle” – ‘higher communication intentions’) combined with a domain label such

[12] This applies to alphabetically close headwords only. Base nouns of words derived by the prefix anti-, for example, as *antidekubitní* (‘anti-decubitus’) or *antiretrovirální* (‘antiretroviral’), would be distant in a printed dictionary, too.

as ‘lékařský‘ (medical); a noun *arterie* and a derived adjective *arteriální* would have the same label vkc, lékařský then. However, the system as a whole was assessed as too complicated for common dictionary users and, therefore, not adopted for ASSČ.

Now let’s take a look at what information dictionary users expect to get from dictionary labels. There are different groups of users with different needs [37, 38], it is assumed that producers of texts need more information and different kind of information than readers [39]. Considering terminology, an author or a translator of scientific texts may need to know whether a word (or which one of several synonyms) is an exact term, while a reader focuses more on the meaning of the word. Then labelling only strictly defined terms could be useful, but producers of scientific texts are more likely to use specialized dictionaries[13] than general ones. For ASSČ, it was originally intended to label only defined terms; however, after discussions with other Czech linguists the decision was changed to follow the tradition of Czech and Slovak dictionaries and label all words belonging mostly to specialized communication, including derived adjectives, adverbs and verbs.

A user-friendly function of an e-dictionary, searching all words from the particular field of interest, might be appreciated by linguists who use a dictionary for additional lexicological studies. Then the approach of NASCS, labelling nearly all words from a domain and combining close fields, would be more useful; however, some headwords remain unlabelled anyway: general expressions like *nemocnice* (‘hospital’) or *pacient* (‘patient’). For this purpose, thematic labels might be more appropriate: there can be as many as necessary for each headword or meaning (for instance, *atropin* can be classified for medicine, pharmacy, biology, chemistry, biochemistry, botany, toxicology… – such a combination of domain labels would be excessive), they could be invisible so as not to glut the entry, and be used just for searching. The DWS used for compiling ASSČ includes thematic labelling and we hope we will be able to offer this function to ASSČ users in the future.

Domain or thematic labels could be optional in a web dictionary interface and users themselves would choose whether they want or do not want to see referred labels, use them for searching, sorting etc. While web versions of SSJČ, SSČ and NASCS only re-publish original printed versions of those dictionaries, ASSČ is the first Czech academic dictionary compiled as an electronic database and intended for web publication from the very beginning. We can use this advantage to offer customization of a dictionary entry layout; a survey among dictionary users could say whether they would find it useful.

5 CONCLUSION

To sum up, labelling words by subject appears to be a complicated issue. The approach of various dictionaries differs remarkably and it is difficult to follow the set rules of labelling within one dictionary, too. There is little literature on this subject and besides that, lexicographers lack information what dictionary users would actually

[13] Terminological dictionaries often use other means than labelling to mark terms: the preferred, standardised headword is followed by full definition while less desirable synonyms only refer to that headword [40].

want and find useful. We hope that further discussions on this topic will bring new ideas and improve lexicographers approach to label usage in dictionaries.

		SSJČ	SSČ	NASCS	ASSČ
a	domain “pharmacy”	lékár.	lékár.	farm.	farmac.
	domain “veterinary medicine”	zvěr.	vet.	vet.	vet.
b	*afázie, angína, astma*	med.	-	med.	med.
	astenik	-	-	med.	-
	aspirin	lékár.	-	farm.	-
	antikoncepce	N/A	-	med.	-
c	*akrální*, *apofýza*	anat.	N/A	med.	N/A
	kyselina askorbová	chem., med.	N/A	med.	chem., biol.[14]
d	*adolescence*	ped. [= pedagogy]	N/A	med., psych.	-
	apatie	-	N/A	psych., med.	med., psych.
	analgetikum	med.	N/A	farm., med.	farmac.
	anestetikum	med.	N/A	med., farm.	farmac.
	antipyretikum, antiseptikum	lékár.	N/A	med., farm.	farmac.
	abdominální	anat., med.	N/A	med.	N/A
	apendix	anat.	anat.	anat., med.	anat.
	autonomní	med.	N/A	med., anat.	N/A

Tab. 1. Differences in labelling in monolingual dictionaries of Czech. Symbols stand for: - = the headword is not labelled in the dictionary; N/A = the headword is not included in the dictionary. Cases a, b, c, d – see in the text; compared labels are sorted and placed in order following the text.

References

[1] Cabré, M. T. (1996). *Terminology: Theory, methods and applications.* John Benjamins B. V., Amsterdam – Philadelphia.

[2] Meyer, I. and Mackintosh, K. (2000). When Terms Move into Our Everyday Lives: An Overview of De-terminologization. *Terminology*, 6(1):111–138.

[3] Mžourková, H., Nová, J., and Pernicová, H. (2017, submitted). Proměny lékařské terminologie v jednojazyčných výkladových slovnících. *Naše řeč*.

[4] Bozděchová, I. (2009). *Současná terminologie (se zaměřením na kolokační termíny z lékařství).* Karolinum, Praha.

[5] Bozděchová, I. (2010). Teorie terminologie a kognitivní lingvistika: k pojetí kategorizace, definice a nominace. *Slovo a slovesnost*, 71(3):163–175.

[6] Bozděchová, I. (2006). Morbus professionalis. (K motivovanosti českých názvů nemocí). *Naše řeč*, 89(3):113–122.

[7] Kolenčíková, E. and Šír, A. (2004). Vývoj lékařské terminologie za posledních deset let. In Žemlička, M., editor, *Termina 2003*, pages 56–63, Technická univerzita v Liberci, Liberec, Czech Republic.

[8] Nečas, P. and Hejna, P. (2010). K užití české anatomické nomenklatury v současné soudnělékařské praxi. *Naše řeč*, 93(1):25–36.

[9] Daneš, F., Filipec, J., and Machač, V., editors (1978). *Slovník spisovné češtiny pro školu a veřejnost.* Academia, Praha.

[10] Filipec, J. (1995). Teorie a praxe jednojazyčného slovníku výkladového. In Čermák, F. and Blatná, R., editors, *Manuál lexikografie*, pages 14–49, H&H, Jinočany.

[14] In ASSČ we treat all chemical substances, including drugs, hormones etc., primarily from the chemical point of view. Combination of labels chem., biol. stands for biochemistry. See [41] for details.

[11] Lišková, M., Nová, J., and Pernicová, H. (2016b). Terminologie. In Kochová, P. and Opavská, Z., editors, *Kapitoly z koncepce Akademického slovníku současné češtiny*, pages 176–185, Ústav pro jazyk český AV ČR, v. v. i., Praha.

[12] Machač, J. (1964). Odborná terminologie ve výkladovém slovníku. *Československý terminologický časopis*, 3:65–76.

[13] Masár, I. (1984). Definícia termínu v Krátkom slovníku slovenského jazyka. In *Obsah a forma v slovnej zásobe: materiály z vedeckej konferencie o výskume a opise slovnej zásoby slovenčiny (Smolenice 1.–4. marca 1983),* pages 68–71, Jazykovedný ústav Ľudovíta Štúra Slovenskej akadémie vied, Bratislava.

[14] Buzássyová, K. (2000). Odborná lexika vo všeobecnom výkladovom slovníku. In K. Buzássyová, editor, *Človek a jeho jazyk. 1. Jazyk ako fenomén kultúry. Na počesť prof. Jána Horeckého*, pages 513–523, Veda, Bratislava.

[15] Blatná, L. (1964). Ke zpracování lékařské terminologie ve Velkém rusko-českém slovníku. *Československý terminologický časopis*, 3:143–150.

[16] Homoláč, J. and Mrázková, K. (2014). K stylistickému hodnocení jazykových prostředků, zvláště lexikálních. *Slovo a slovesnost*, 75(1): 3–38.

[17] Atkins, B. T. S. and Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford University Press, Oxford.

[18] Tsuji, K. and Kageura, K. (1998). An Analysis of Medical Synonyms: the Word-structure of Preferred Terms. *Terminology*, 5(2):229–249.

[19] Wang, J. L. S. and Ge, G. (2008). Establishment of a Medical Academic Word List. *English for Specific Purposes*, 27:442–458.

[20] León-Araúz, P. and Reimerink, A. (2014). From Term Dynamics to Concept Dynamics: Term Variation and Multidimensionality in the Psychiatric Domain. In Abel, A., Vettori, Ch., and Ralli, N., editors, *Proceedings of the XVI EURALEX International Congress: The User in Focus.* pages 657–668, Institute for Specialised Communication and Multilingualism, Bolzano, Italy. Accessible at: `http://euralex.org/publications/from-term-dynamics-to-concept-dynamics-term-variation-and-multidimensionality-in-the-psychiatric-domain/`, retrieved 2017-03-21.

[21] Panocová, R. (2016). A Descriptive Approach to Medical English Vocabulary. In Margalitadze, T. and Meladze, G., editors, *Proceedings of the XVII EURALEX International Congress: Lexicography and Linguistic Diversity*, pages 529–540, Ivane Javakhisvili Tbilisi State University, Tbilisi. Accessible at: `http://euralex.org/publications/a-descriptive-approach-to-medical-english-vocabulary/, retrived 2017-03-21.`

[22] Hubáček, J. and Krčmová, M. (2016). Sociolekt (slang). In Karlík, P., Nekula, M., and Pleskalová, J., editors, *Nový encyklopedický slovník češtiny online*. Accessible at: `https://www.czechency.org/slovnik/SOCIOLEKT`, retrieved 2017-03-03.

[23] Durkin, P., editor (2016). *The Oxford Handbook of Lexicography.* Oxford University Press, Oxford.

[24] Martincová, O. and Bozděchová, I. (2016). Termín (odborný název). In Karlík, P., Nekula, M., and Pleskalová, J., editors, *Nový encyklopedický slovník češtiny online*. Accessible at: `https://www.czechency.org/slovnik/TERMIN`, retrieved 2017-03-03.

[25] Hausenblas, K. (1962). K specifickým rysům odborné terminologie. In *Problémy marxistické jazykovědy*, pages 248–262, Nakladatelství Československé akademie věd, Praha.

[26] Dvonč, L. (1965). K problému slovies ako termínov. *Československý terminologický časopis*, 4:59–64.

[27] Man, O. (1964). Postavení slovesa v systému terminologie. *Slavica Pragensia*, VI:129–138.

[28] Abecassis, M. (2008). The Ideology of the Perfect Dictionary: How Efficient Can a Dictionary Be? *Lexikos 18* (AFRILEX-reeks/series 18), pages 1–14.

[29] Hnátková, M., Křen, M., Procházka, P., and Skoumalová, H. (2014). The SYN-series corpora of written Czech. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 160–164, ELRA, Reykjavík, Iceland.

[30] Lišková, M., Michalec, V., and Pernicová, H. (2016a). Hlavní heslář. In Kochová, P. and Opavská, Z., editors, *Kapitoly z koncepce Akademického slovníku současné češtiny*, pages 20–24, Ústav pro jazyk český AV ČR, v. v. i., Praha.

[31] Sager, J. C. (1990). *A Practical Course in Terminology Processing*. John Benjamins, Amsterdam.

[32] Žagar Karer, M. (2007). Determinologizacija v splošnih in terminoloških slovarjih. In Orel, I., editor, *Obdobja 24: Metode in zvrsti. Razvoj slovenskega strokovnega jezika*, pages 599–609, Center za slovenščino kot drugi/tuji jezik, Univerza v Ljubljani Filozofska fakulteta Oddelek za slovenistiko, Ljubljana.

[33] Müller-Spitzer, C., editor (2014). *Using Online Dictionaries*. Lexicographica Series Maior 145. Walter de Gruyter, Berlin.

[34] Dziemianko, A. (2015). Colours in Online Dictionaries: A Case of Functional Labels. *International Journal of Lexicography*, 28(1):27–61.

[35] Barbierik, K., Bodlák, M., Děngeová, Z., Jarý, V., Liška, T., Lišková, M., Nový, J., and Virius, M. (2015). The Current Status of the Development of the ALEXIS Dictionary Writing System. In Gajdošová, K. and Žáková, A., editors, *Natural Language Processing, Corpus Linguistics, Lexicography*, pages 17–25, RAM-Verlag, Lüdenscheid, Germany.

[36] L´Homme, M.-C. and Cormier, M. C. (2014). Dictionaries and the Digital Revolution: A Focus on Users and Lexical Databases. *International Journal of Lexicography*, 27(4):331–340.

[37] Ledinek, N. (2015). Obravnava izhodiščno terminološke leksike v novem Slovarju slovenskega knjižnega jezika. In Smolej, M., editor, *Obdobja 34: Slovnica in slovar – aktualni jezikovni opis (2. del)*, pages 441–448, Univerza v Ljubljani Filozofska fakulteta Oddelek za slovenistiko, Center za slovenščino kot drugi/tuji jezik, Ljubljana.

[38] Müller-Spitzer, C., Wolfer, S., and Koplenig, A. (2015). Observing Online Dictionary Users: Studies Using Wiktionary Log Files. *International Journal of Lexicography*, 28(1):1–26.

[39] Lew, R. and de Schryver, G.-H. (2014). Dictionary Users in the Digital Revolution. *International Journal of Lexicograph*y, 27(4):341–254.

[40] Machová, S. (1995). Terminografie – speciální případ lexikografie. In Žemlička, M., editor, *Termina 94*, pages 133–137, Katedra českého jazyka a literatury Pedagogické fakulty Technické univerzity v Liberci a Ústav pro jazyk český, Praha, Czech Republic.

[41] Nová, J. (2016). Zpracování názvů chemických sloučenin v ASSČ. In Kochová, P. and Opavská, Z., editors, *Kapitoly z koncepce Akademického slovníku současné češtiny*, pages 184–185, Ústav pro jazyk český AV ČR, v. v. i., Praha, Czech Republic.

CORRELATIVE CONJUNCTIONS IN SPOKEN TEXTS

PETRA POUKAROVÁ
The Institute of the Czech Language of the Academy of Sciences of the Czech Republic, Prague, Czech Republic

POUKAROVÁ, Petra: Correlative Conjunctions in Spoken Texts. Journal of Linguistics, 2017, Vol. 68, No 2, pp. 305 – 315.

Abstract: Correlative conjunctions (such as *buď – anebo* (either – or), *jednak – jednak* (firstly – secondly) etc.) represent one means of textual cohesion. The occurrence of one component of the pair implies the use of the other, which contributes to the cohesiveness of a text. Using data provided by the corpus of informal spoken Czech ORAL2013, I will try to demonstrate their use in a prototypical spoken language, that is commonly considered less coherent and more fragmentary compared to written language.

Keywords: correlative conjunctions, spoken Czech, ORAL2013, corpus

1 INTRODUCTION

There are several ways of expressing textual cohesion in Czech. Lexical means are most frequently represented by deictic expressions, synonyms, hyperonyms, word repetition, etc. In spoken language, paralinguistic means such as gestures, facial expressions, or direct pointing can also be used. Grammatical means of textual cohesion include various connectors, that in the broadest sense fulfil a conjoining function, express mutual relations between words or signal continuity [12, p. 912]. A large group of connectors is represented by conjunctions (as the name of the word class already suggests), that can be both simple or multi-word. Conjunctions connect parts of a text with what immediately precedes or follows so that the resulting text reads as cohesive. Apart from the connective function, they also express semantic relations, that can be either objective (for instance signalling a temporal sequence) or subjective.

In this article, we are going to deal with the so-called correlative conjunctions expressing textual cohesion, i.e. sets of two expressions that most frequently occur at the beginning of clauses or before clause elements that are being conjoined [1, p. 343]. The examples are *buď – nebo* (either – or)*, jednak – jednak* (firstly – secondly) etc. The occurrence of one of the conjunctions implies the use of the other; the speaker often opts for these pairs of words to explicitly express mutual relations at both clausal and textual level and at the same time creates a compact, cohesive text. The listener expects the use of the other component of the correlative pair of conjunctions as it contributes towards his or her understanding and relating of the information. This understanding enables the speaker to fulfil his or her communicative goal.

DE GRUYTER OPEN
DOI 10.1515/jazcas-2017-0040

In the case of a prototypical spoken language[1], cohesion or complexity of a text can be hindered because it is being produced “here and now”. Spoken texts are presented in contrast with written texts and are being characterised as less cohesive [2, p. 121]. The speaker does not have any time in advance to prepare his or her speech, which consequently excludes the use of complex syntactic constructions etc. The production of a spoken language is also influenced by short-term memory – the speaker might not be able to remember his or her previous words and might continue differently from what has been originally intended. In the case of correlative conjunctions, this might lead to the situation when the other item of a pair is not used although its occurrence is expected due to the use of the first conjunction. In the next few pages, the Czech correlative conjunctions *buď – nebo* (either – or)*, sice – ale* (although) and *jednak – jednak* (firstly – secondly) and the instances of a main clause introduced by the word *tak* (so) are going to be analysed in greater detail.

Using examples from the corpus of informal spoken Czech ORAL2013, I will try to demonstrate the real use of correlative conjunctions; whether speakers do express them both and therefore make the text more coherent and cohesive, or not.

2 CONJUNCTIONS

2.1 Conjunctions in General

Conjunctions are generally defined as an uninflectional synsemantic word class (with the exception of *kdyby* (if) and *aby* (so as)) that is closed (there are no more than 2 000 simple conjunctions in a language) but thanks to their function, they are used very frequently [1, p. 342]. This is confirmed by the lemma frequency statistics in SYN2015 and ORAL2013[2] corpora. The basic functions of conjunctions are also generally agreed on: they conjoin syntactic clauses and/or parts of clauses, both at the same syntactic level (the relation of coordination) and at different levels (the relation of subordination) [6, p. 36], and semantically specify the nature of syntactic relations they express [12, p. 524]. They also provide the semantic motivation of various conjoined elements.

2.2 Conjunctions as Text Organisers

Conjunctions are usually introduced together with the concepts of a main and a subordinate clause, that are used to define complex and compound sentences. Due to the fact that coordinative conjunctions join (and signal) clauses at the same syntactic level and are not a part of their syntactic structure, their distribution in a text is freer. “Their function is to express semantic relation and the same syntactic level of the conjoined elements and thus join them into a higher syntactic and

[1] The term *prototypical language* here represents commonly spoken, spontaneous unprepared Czech [4, p. 118].

[2] The conjunction *a* (and) is the third most frequent conjunction in the SYN2015 corpus. It can also be found on the third position in the spoken corpus ORAL2013. Among the first ten words, there is also the conjunction *že* (that). This conjunction can, however, also fulfil different functions – it can be found in collocations such as *že jo/že ano* (right), in which case it is a particle. Its other function can be that of a question tag.

semantic unit" [7, p. 139]. Unlike coordinative conjunctions, subordinative conjunctions form a part of the subordinate clause they introduce and are related to the verb in the main clause. To demonstrate this, the conjunction *A* (and) at the beginning of a clause suggests connection with the previous context (and semantically signals, for instance, a summary of information that precedes) but the conjunction *Jestli* (whether) at the beginning of a clause always refers to the previous clause (it expresses parcellation).

Correlative conjunctions in a way follow both the principles at the same time. The first item of the pair connects the text with its context and expresses their mutual semantic relations. Its word class categorisation is relativised, and because it occurs at the beginning of a clause, it fucntions as a connective particle [9, p. 693]. The second constituent of the pair does not, unlike subordinative conjunctions, express the relation of dependency between the conjoined units, but it refers to the preceding clause because the use of the first element of the pair requires its presence (as same as a verb requires complementation in the form of a subordinate clause).

2.3 Conjunctions from the Phraseological Point of View

A phraseme is defined as a "unique expression consisting of minimally two elements in which at least one of them cannot occur in a different word combination in the same way, i.e. it occurs only in one such combination, or possibly very few of them" [10, p. 140]. Such definition relate to correlative conjunctions too. They can be described as fixed and therefore systemic combinations of expressions, that can be combined with other elements in a limited way [5, p. 449]. We can support this claim by looking at the description of *buď* (either) in [11] where it is defined as *a coordinative conjunction used in the correlative pair buď(to) – (a)nebo* (either – or). Speakers´ own experience of the use of the language also confirms the definition – the use of the first expression of the pair implies the use of the other; or, choosing the other perspective, the occurrence of the second item of the pair "justifies" the use of the first.

3 METHOD

The correlative conjunctions were looked up in the corpus of spontaneous spoken Czech ORAL2013, that includes recordings of dialogues between speakers older than 18 that know each other well. The corpus maps all regions of the Czech Republic. It is a part of the fifth version of a series of united corpora ORAL2006, 2008, and 2013 that were experimentally lemmatized and tagged.

Thanks to that, I could look up expressions using categories such as word class, lemma etc. In the case of correlative conjunctions *buď – (a)nebo* (either – or)*, sice – ale* (although) and *jednak – jednak* (firstly – secondly), I used the interface KonText and entered the first component of the pair with its possible pronunciation variants, as the recordings had been transcribed in accordance with the rule "write what you hear". That is why there is the option *buďto* (the informal variant of the conjunction either) occurring along the more standard *buď* (either) in the corpus.

Searching for instances of a main clause introduced by the expression *tak* (so) proved more difficult. I needed to look up examples in which the subordinate clause precedes the main clause. First I tried to find such sequences using the query <sp> [tag="J,.*"], in which <sp> stands for the "search for the expressions at the beginning of the speaker´s utterance", tag J, marks subordinative conjunctions. I expected that by looking up subordinative conjunctions at the beginning of the speaker´s turn, I would get relevant results, but there was a problem with the segmentation of the transcripts. The transcribers were instructed to arbitrarily divide long utterances of one speaker into parts no longer than 15 words, creating the boundary where there was a pause. The beginning of a communication unit thus does not have to correspond to the beginning of a new syntactic unit (the main clause could have occurred in the previous segment) or to the beginning of a new semantic unit.

That is why I modified the query and searched only for the expressions *a když* (and when) and *že když* (that when). The conjunctions *a* (and) and *že* (that) ensure that the main clause follows the subordinate one[3] (with the exception of the cases where the unit *a když* (and when) expresses a multiple relation; there were, however, only few such examples). I was interested in how many times the main clause would be introduced by expression *tak* (so). In the case of *buď – (a)nebo* (either – or) and *jednak – jednak* (firstly – secondly), I worked with all the occurrences, and in the case of all other expressions, I manually analysed a random sample of 250 occurrences (see Results).

4 RESULTS

BUĎ – NEBO (either – or)

There were 703 occurrences of the word *buď* (or *buďto*) (either) in the corpus ORAL2013. I had to look at all the instances because some cases of the word *buď* (either) could represent the imperative of the verb *být* (to be) and thus be homonymous with the conjunction, and I wanted to get as many instances of *buď* (either) as a conjunction as possible. There were 601 cases of *buď* (either) used as a conjunction (the other cases were either homonymous, or they represented repetition of the expression *buď* (either)); in 445 (74%) out of these, the speakers also provided the other part of the correlative pair, i.e. *nebo* (or). In this respect, we could see that they are supporting the cohesion of the text, although their compactness can be influenced by many other factors. *Nebo* (or) was not used in 156 of the cases.

Informal spoken language enables the two components of a correlative pair to occur at a various distance from each other. The distance is, on the other hand, limited by short-term memory (see above), i.e. the bigger the distance between the two items, the lower the probability of occurrence of the second element of the pair. In the data I analysed, there were both extremes; the two components were used one immediately after the other (example 1), and also in the distance that was greater

[3] I do realise that using the terms *main* and *subordinate clause* appears problematic. When we use the conjunction *že když* (that when), both of the clauses are actually subordinate. Compare: *řekl mi, že když přijde, uleví se mu* (he told me that if he came, he would feel better) a *řekl mi* (he told me) *->že se mu uleví* (that he would feel better) -> *když přijde* (when he came), in which *kdy přijde* (when he came) is subordinate of the clause *že se mu uleví* (that he would feel better)*,* which is also a subordinate clause.

than 54 words (plus there is the second speaker´s utterance in between) (example 2). The average distance between the two is four words.

1. S1: *za to může .. můžou ptáci no .* (we should blame… the birds are to blame for it)
   S2: ***buď*** . ***anebo*** *počasí* (either . or weather)

   In this case, S2 reacts to S1´s utterance and by using a multi-word unit specifies an alternative causer of the discussed condition (birds or weather).

2. S1: *a teď* ***buď*** *. eee .. pro . jakože . eee pěší turistiku . to by se šlo . to by se určila trasa* (and now either . eee .. for . like . eee hiking . we should follow . a trail would be fixed)
   S2: *no* (yeah)
   S1: *a . šlo by se z kempu do kempu . a dycky v tom kempu by čekala stage .. a ňákej program . jo a ty lidi by tam museli dojít . ňák .. se dopravit . jo ? . a to by byla takle celá akce .* ***anebo*** *že by to bylo podél ňáký řeky pro vodáky .. jo ?* (and . we could go from a camp to another one . and there a stage would wait in the camp .. and some programme . yeah and the people would have to get there . somehow .. to get . yeah ? . and it would be the whole event like this . or it would be along some river for watermen .. yeah ?)

Example 2 is interesting because even the second item of the pair could be seen as a particle structuring the text. Among conjunctions, Hrbáček distinguishes connectors (where he places structuring conjunctions) and junctives. "Conjoining devices (junctives) and connective devices (connectors) can be distinguished roughly in the sense that conjoining devices express semantic relations of the utterances inside of one unit, whereas connective devices express semantic relations between semantic units across utterance/sentence boundaries" [8, p. 56]. That is why they often occur at the beginning of clauses. In the case of spoken language, we cannot talk about a sentence or its beginning and end. The instances that were found correspond to connectors due to the placement of the first item of the pair at the beginning of the turn, after a pause etc. *Anebo* (or) in example 2 can also be characterised as a connector rather than a junctive (a conjoining device) as it, first, occurs at the beginning of a part of an utterance after a pause, and second, as it appears at quite of a distance from the first component of the unit (in that way it looks more like a connective device).

Among the correlative conjunctions with both components of the pair used, there are also phrasemes *buď tak nebo tak* (either in this way or that way); *buď ta pravá strana je těžší nebo co já vím* (either the right side is heavier or what do I know); *buď byla sjetá nebo co* (either she was high or something like that).

There are two reasons why the second element of a pair of correlative conjunctions is not expressed. First, it is the speaker him- or herself who decides not to continue with the original utterance (example 3), and second, it could be his or her communication partner who interrupts and choses a different topic (example 4).

3. *S1: a eee říkám no tak . tak* ***buď*** *sou v Kamenickym Šenově . protože von zas ňákou horolezeckou knihu ňáký . z Děčína něco prže byl . deset let předseda horolezeckýho oddílu ve Varnsdorfu . ve Slovanu a to takže chtěj ňákou knihu a ňáký materiály . a že musí jet za ňákym klukem ve Chřibský a . a s tim zas kamarádí Franta NP tak sem říkala třeba jeli tam a to* (and eee I say yeah so . so either they are in Kamenický Šenov . because he again some mountaineering book some . from Děčín something because he has been . a chair of the moutaineering group in Varnsdorf for ten years . in Slovan and so so they want some book and some materials . and that they have to go to see some boy to Chřibská and . and Franta NP is friends with him so I said they had gone there maybe and so)

The speaker linearly develops and specifies her utterance and follows the topics that emerge. It seems that the speaker moves forwards with her narrative rather than coming back to her original idea and providing the other element of the correlative pair. Nevertheless, based on the occurrence of the first item, we are able to determine in which context the second part, *nebo* (or), should appear. In this case, we can consider the goal of providing the other part as fulfilled – the people the speaker is talking about are either in Kamenický Šenov, or in Chřibská.

4. S1: *chcu udělat posezení a támhle to nějak zastinit . mmm buďto nějakou roletu* (I want to make seating and there to shade it somehow . mmm either with some blind)
S2: *dyť *s míval roletu nebo co *s tam míval ?* (you used to have a blind, didn´t you or what did you use to have there ?
S1: *míval* (I used to have)
S2: *ale víš co chcu udělat ?* (but you know what I want to do ?)
S1: *hadr* (a tatter)

The use of both of the elements is therefore considerably influenced by the communicative situation. The communication partner can always interrupt the first speaker in a dialogue, as in example 4 where the communication partner reacts to the information about the roller blind (*roleta* “blind”) and thematises it in the next utterance to confirm his understanding. The speaker reacts to it and the conversation moves on to another topic.

SICE – ALE (although)
After searching for the expression *sice* (although) in the corpus, I obtained 506 instances and I analysed a random sample of 250 of them. Out of these, both of the elements *sice – ale* (although) were formulated 169 times (68%), in remaining 81 cases, the second element did not appear. This situation is therefore similar to the use of the conjunctions *buď – nebo* (either – or). Again, it was either the speaker who

linearly developed his or her topic, not referring back to the first element (example 5), or the communication partner interrupting the speaker´s utterance (example 6). Other possible interpretation is the use of the conjunction *a* (and) instead of *ale* (but).

5. S1: *je tam* ***sice*** *asi . zima že jako . že Rosťa mi říká všichni chodí v mikinách že jako . Šimona oblíkne . a Šimon byl hlavně spokojenej že má balónek sem nesla balónek a už odešel . a mi říká balónek . a vrátil se za mnou* (it might be . cold there . that Rosťa tells me everybody wears a sweatshirt . he will give Šimon some warm clothes . and Šimon was happy primarilly because he had a baloon I took the baloon and he left . and he tells me the baloon . and he came back to me)
   S2: *no jo už je z něho velký prďola* (yeah, he is a big guy)

6. S1: *takže ty tam můžeš přijet autem* ***sice*** *vybíraj [padesát]* (so you can go t here by car although they take [fifty])
   S2: *[nasadíš] lyže* (you will put the skis on)
   S1: *padesát* (fifty)
   S2: *hmm* (hmm)
   S1: *korun @ za tři hodiny anebo sedumdesát korun za celej den* (crowns @ for three hours or seventy for all day)
   S3: *ježišmarja* (Jesus!!!)
   S1: *necháš tam* ***a*** *celej den můžeš jezdit po celejch . po celý Šumavě* (you will leave it there and you can ski through the whole Šumava)

In the case of *sice – ale* (although), there were examples where the second component of the pair was replaced by a different expression with the same or similar meaning (expressing adversative relation between the connected parts). The question therefore arises as to how obligatory the second element is and how fixed these expressions are. The potential of conjunctions to combine is influenced by their logico-semantic functions [5, p. 450]. That is why we can hardly find incompatible cases of conjunctions such as *a ale* (and but), *aby že* (so as that) etc. occurring together.

The same is true for correlative conjunctions, as can be illustrated in the following examples: *ale já sem tam jako elév jo* ***sice*** *mě jako raděj mezi vědecký pracovníky což si mysim že je úžasně nadsazený . a* ***na druhou stranu*** *se tam bavíš s těma ženskejma* (but I am there as a beginner yeah although they think of me as of a scientific worker which I think is amazingly exaggerated . and on the other hand you speak there with the women); *bylo to* ***sice*** *. jako zaručeně nejlevnější ve srovnání se všema ostatníma chtěli tam jen vosum set na den zálohu jen deset tisíc asi . což teda bylo jakoby stejný jako všude .* ***akorát*** *ňák sem s tim brousil asi dvě hodiny a najednou z toho začly lítat jiskry* (it was . like definitely the cheapest in comparison with all others they wanted only eight hundred per day like an advance only ten thousand maybe . which was really the same as everywhere . but I used it for sharpening for about two hours and suddenly sparks started to fly from it); *to* ***sice*** *jo* ***jenomže*** *jako ty kulatiny podle mě stojí za to ňák hezky oslavit* (yes I think so but it is worth to celebrate this birthday well). The formulations *na druhou stranu* (on the other hand)*, akorát, jenomže* (but) express a contradiction between the content of the clause they introduce and the preceding context. The semantic relation is therefore fulfilled although it is not realised by the "prototypical" conjunction *ale* (but).

Example 7 does not come from the corpus ORAL2013 but from the newspaper Lidové noviny. I am using it to further illustrate a phenomenon that also occurred with the conjunctions *buď – nebo* (either – or). It shows that the second element of the pair does not have to fulfil the connective function and thus express semantic relations, but it can also structure the text.

7. *Během příštích šesti měsíců, které ještě zbývají do německých celostátních voleb, se* ***sice*** *může ještě leccos přihodit. // V každém případě* ***ale*** *začíná s trana, která byla založena před 153 lety, přicházet zjevně opět do módy.* (So much can happen during the next six months, which are left until the German statewide elections. // But anyway the party, which was established 153 years ago, is obviously trendy again.)

Here, not only do the two parts of the pair occur in two independent sentences, but the sentences are also parts of two different paragraphs (the boundary is marked by //). The word *ale* (but) is a part of a structuring chain, in which the expression *v každém případě* (but anyway) summarises the information from the preceding context and the "particle – conjunction" *ale* (but) provides the adversative connotation.

JEDNAK – JEDNAK (firstly – secondly)

I found 159 occurrences of the word *jednak* (firstly) in the ORAL2013 corpus. Unlike the previous correlative conjunctions, the instances where the second item was not expressed were dominant here (in 58 cases, i.e. 36.5%). The second *jednak* (secondly) was used in 31 cases (19.5%). The third, and more interesting, variant were the cases where the semantic relation was expressed but by the use of a different expression. There were 36 such cases, i.e. 22.6% (the other instances are of the kind where *jednak* (firstly) is immediately repeated and the resulting concordance was counted twice).

The question arises as to why the realisation of the last pair of conjunctions differs from the previously discussed ones. One possibility is that while the correlative conjunctions *buď – nebo* (either – or) and *sice – ale* (although) express alternative relation, or possibly an adversative relation between two situations in which the first element requires the presence of the second so that the utterance is complete, the conjunctions *jednak – jednak* (firstly – secondly) signal an interrelating relationship and the utterances that are being connected are at the same level, as is often the case with various listings. (A comparison could be provided by syntactic hypo- and paratactic relations.)

In the cases where a different conjunction was used instead of *jednak* (secondly), it is again possible to determine the combinatorial potential of the expression. The second position in the pair could be taken by: **jako** (*jednak vypadá dobře a jako je hodnej* "he looks good and he´s also kind"); **pak**/**potom** (*máme to napsaný jednak* *na seznamu literatury a potom sme to no dělali v dějepise* “firstly, it is written on the list of the books and we also did it in the history classes); the pair *jednak – a pak* (firstly – also) even occurs on the list of phrasemes in [3]; **druhak** / **za druhé** (*jednak výjezdní z Maroka . a druhak příjezdní do Čech* “firstly exit from Marocco . and secondly arrival to Czechia); **zároveň** (at the same time), **a taky** (and also).

Some other conjunctions, actually creating an illogical pair, were also used: ***jednak*** *na všechny dohromady .* ***ale*** *je to i na každej zvlášť* (it is applicable for all together but also for each of them separetely) or a connection with a subordinate clause: *že tam je to hodně .. nahnutý že* ***jednak*** *toho maj moc a* ***že*** *asi nebudou až tak čistý* (that it is very .. tilted and so they are really busy and also „the things“ might not be that clean). In this case, repetition of the conjunction *že* (that/so) expresses a multiple relation and signalizes that these two sentences are at the same syntactic level (it corresponds to the description of *jednak* (firstly), see above).

SUBORDINATE CLAUSE – ***TAK*** *(SO) MAIN CLAUSE*

In this case it is not a typically multi-word conjunction but the frequency of occurrence is very high, which is why I decided to include it. We can say that the occurrence of a conjunction introducing a subordinate clause calls for the need to use the expression *tak* in the main clause in postposition. From all the 500 instances of the pairs *a když* (and when) + *main clause* and *že když* (that when) + *main clause* that were analysed, there were 365 (71% of the cases) in which the main clause was introduced by the word *tak* (so). This phenomenon is connected with spoken texts and is being avoided in, for instance, journalistic texts (mainly from dialogues) and it is considered a sign of inarticulacy, something that should not be used in written texts.

In spoken discourse that is being created linearly at the present moment, the speaker uses the expression *tak* (so) to structure his or her utterance and create a more cohesive text. The word *tak* (so) basically refers to the previous utterance and connects with it, and at the same time signals that the speaker is still talking about the same topic. The listener can thus more easily recognise pieces of information as connected and will understand better. *Tak* (so) is interpreted more as a particle as it does not function as an obligatory element of the clause and the utterance will not change when it is not used. Neither does it express any syntactic relation or carry meaning, its function is limited to a text-structuring device only: *a dyž už to bylo skoro na konci Teplic a voni začli houkat . tak sem zastavila* (and so it was at the end of Teplice and they started to hoot . so I stopped).

The connecting expression *že* (that) was used in 23 instances instead of *tak* (so): *možná čekal* ***že*** *když ti je pochválí* ***že*** *ty se ohneš aby sis je narovnal . víš ?* (maybe he expected that when he praised them, you would bend down to straighten them . you know ?). In this construction, the complexity and cohesiveness of the clauses is strengthened by the word *že* (that). We could also imagine the utterance being formulated as: *možná čekal, že ty se ohneš, když ti je pochválí* (maybe he expected, that you would bend down when he had praised them). The speaker thus structures his or her utterance and therefore expresses the relations between its parts and their mutual connection more explicitly.

5 CONCLUSION

The aim of this article was to analyse correlative conjunctions *buď – (a)nebo* (either – or)*, sice – ale* (although)*, jednak – jednak* (firstly – secondly) and a special combination *subordinate clause + tak (so) + main clause* in the corpus of informal Czech and to determine whether and how they contribute to cohesion of spoken texts. Specialised literature often describes spoken language as less coherent based on the fact that it is being produced "online" without preparation. Using both elements of the pair of correlative conjunctions, the speaker creates a text that is more compact, as he or she relates its parts one to another, and by using the second item of the pair, the speaker points back to the first one, which results in a structure that is connected. If we consider the use of the two conjunctions as fixed, we could describe correlative conjunctions as phrasemes.

In the case of the pairs *buď – (a)nebo* (either – or) and *sice – ale* (although), the speakers mostly opted for the formulation of the second element of the pair (in 74% and in 68% respectively). The first component of the pair, when used at the beginning of the utterance or the speaker´s turn, functions as a particle structuring the text and at the same time connecting the preceding context with what follows. This element has to be always realised by a coordinative conjunction. According to the data analysed, the second constituent of the pair could also be classified as a particle – it does not immediately follow the utterance that is introduced by the first item but there could be more words inserted in between (in the case of *buď – (a)nebo* (either – or), the number of words in between the two expressions can be even 54). By using both parts of the pair, the speaker stabilises the theme and signals that he or she is continuing talking about the same topic. If the second constituent did not appear, it either meant that the speaker digressed from the main topic, or it was his or her communication partner who interrupted and changed the perspective of the following utterance.

The situation with the pair *jednak – jednak* (firstly – secondly) was opposite – both elements were expressed only in 19.5% of the cases, in 22.6% of the instances was the second item replaced by another with the same meaning (*pak/potom* "and then"*, druhak* "secondly"*, a taky* "and also"). We could explain this by analysing the meaning of the expressions – while the pairs *buď – (a)nebo* (either – or) and *sice – ale* (although) signal adversative or alternative relation and both elements must be used (so that it is clear what alternates and what contradicts what), the additive relation, or listing, expressed by *jednak – jednak* (firstly – secondly) can be signalled by mere juxtaposition, the conjunction *a* (and) etc. As for word combinations, *jednak* (firstly) seems to have the highest potential to be combined with other expressions, as the examples clearly demonstrated. Nevertheless, we cannot say that the resulting text was less compact because its cohesiveness could have been realised by various other devices.

The last group consisting of a *subordinate clause + tak (so) + main clause* occurs both in spoken language and in journalistic texts (most frequently in dialogues where it represents features of oral communication and is usually avoided). *Tak* (so) introducing a main clause appeared in 71% of the cases of the 500 instances analysed. The alternative to it proved to be the conjunction *že* (that). By using these elements, the speaker highlights connection with the previous utterance, stabilises the theme, and signals that he or she is still talking about the same topic. We can see a clear attempt to create a cohesive text that consequently facilitates the listener´s understanding.

ACKNOWLEDGEMENTS

The present article was written under the auspices of grant project No. 15-011165 (Syntax mluvené češtiny) [Syntax of spoken Czech]), supported by the Grant Agency of the Czech Republic.

References

[1] Cvrček, V. a kol. (2015). *Mluvnice současné češtiny.* Karolinum, Praha.

[2] Čechová, M., Krčmová, M., and Minářová, E. (2008). *Současná stylistika*. Lidové noviny, Praha.

[3] Čermák, F. (2009). *Slovník české frazeologie a idiomatiky – výrazy neslovesné.* Leda, Praha.

[4] Čermák, F. (2009). Spoken Corpora Design: Their Constitutive Parameters. *International Journal of Corpus Linguistics*, 14(1):113–123.

[5] Čermák, F. (2014). Syntagmatika, kombinace a kumulace konjunkcí. In *Jazyk a slovník*, pages 447–453, Karolinum, Praha.

[6] Hošnová, E. (2005). *Studie z vývoje novočeské syntaxe.* Karolinum, Praha.

[7] Hrbáček, J. (1967). K poměru mezi spojovacími prostředky členskými a větnými. *Naše řeč,* 50(3):138–144.

[8] Hrbáček, J. (1994). *Nárys textové syntaxe*. Trizonia, Praha.

[9] Karlík, P., Nekula, M., and Rusínová, Z. (1995). *Příruční mluvnice češtiny*. Lidové noviny, Praha.

[10] Karlík, P. a kol. (2002). *Encyklopedický slovník češtiny*. Lidové noviny, Praha.

[11] Štícha, F. (2013). *Akademická gramatika spisovné češtiny*. Academia, Praha.

Corpus ORAL2013:
Benešová, L., Křen, M., and Waclawičová, M. (2013): *ORAL2013: reprezentativní korpus neformální mluvené češtiny*. Ústav Českého národního korpusu FF UK, Praha. Accessible at: `http://www.korpus.cz.`

ISSUES OF POS TAGGING OF THE (DIACHRONIC) CORPUS OF CZECH: PREPARING A MORPHOLOGICAL DICTIONARY

ANNA ŘEHOŘKOVÁ
Institute of the Czech National Corpus, Charles University, Prague, Czech Republic

ŘEHOŘKOVÁ, Anna: Issues of POS Tagging of the (Diachronic) Corpus of Czech: Preparing a Morphological Dictionary. Journal of Linguistics, 2017, Vol. 68, No 2, pp. 316 – 325.

Abstract: Many important decisions concerning the part-of-speech categorization remain unexplained in the current practice, only reported in corpus manuals. The aim of this paper is to offer a different perspective on the problems of morphological annotation of corpora – the perspective of mapping and analyzing conceptual problems in the annotation. Focused mainly on function words in Czech, we discuss the possibilities of the POS tagging of the inherently ambiguous category of particles and we introduce criteria for distinguishing particles from interjections.

Keywords: corpus, function words, morphological annotation, Czech

1 INTRODUCTION

The motivation of this paper is to share the experience from the preparation of a new diachronic corpus of Czech, covering the 19th century. Dealing with shifts and changes in the older language, where the lack of native speaker knowledge is perceptible, led us to rethink the principles of morphological annotation, concerning function words in particular, and to seek for inspiration in other corpora (cf. [2]).

Words considered as secondary prepositions, conjunctions, adverbs, particles and interjections, namely all those that have undergone a grammaticalization and conventionalization process, are often difficult to classify. Clues provided by grammars and dictionaries turned out to be insufficient for corpus annotation where every token needs to be tagged. For example, in the Oxford English Dictionary and elsewhere, prepositional, adverbial and conjunctional use of *notwithstanding* is distinguished, the adverbial one according to the meaning ‚nevertheless, all the same‘ (*he must be told, notwithstanding*). On the contrary, the annotation of the BNC2 corpus is based on contextual features which are recognizable to the automatic tagger, and therefore it is the instances that come after an NP and precede punctuation that are mostly tagged as adverbs:

(1) The author *notwithstanding*, many conclusions can be drawn from this steel-trap of a book [...]

According to the OED, though, (1) is an example of a preposition (used postpositively) meaning ‚in spite of‘. Thus, it seems that the adverbial category might have been redefined in the corpus with respect to the formal recognizability of

DOI 10.1515/jazcas-2017-0041

the word in context.[1] Nevertheless, cases like (2), (3) and (4) still can be found where the sentence has the same structure but the word is tagged in three different ways (AV0 - general adverb, PRP – preposition, PRP-CJS: the ambiguity tag for preposition/conjunction):

(2) AV0: *Notwithstanding* all these problems, the bank has kept faith with us [...]

(3) PRP: *Notwithstanding* this promise, the use of road pricing to change travel habits still seems some way off.

(4) PRP-CJS: *Notwithstanding* the re-election of Mrs Thatcher in 1983 and 1987, a clear majority of voters have favoured increased taxes [...]

These examples indicate the complexity of interfaces between various function words. In this article we will focus on the case of particles in Czech.

2 PARTICLES VERSUS OTHER PARTS OF SPEECH

In Czech grammatical theory, particles were not fully recognized as a part of speech until the 1980s [16]. The oldest contemporary grammar [7] introduced a wide and heterogeneous category of adverbs, consisting of content words as well as function words, including idiosyncratic cases like *ne* 'no'. This grammar became a widely used school book and a base for part-of-speech classification in dictionaries of Czech ([8], [14], [19]). Later [13] the definition of adverbs was refined and only clause constituents were considered adverbs, the others being classified as particles (e. g. *snad* 'perhaps' which does not bring any information about the circumstances of the action expressed by a verb and, therefore, unlike other adverbials, can not be used as an answer to any question about the action – how? when? etc.). Interestingly, this criterion was not accepted by Quirk et al. ([20]) who argue that all adverbials (unlike objects, complements etc.) are optional elements to the structure of a clause. Furthermore, in the Czech tradition not only adverbs but also conjunctions, pronouns, nouns, verbs or even phrases have been viewed as particles in cases where they displayed signs of semantic bleaching and/or a shift in their function towards pragmatics of interaction (cf. [6]). Thus, particles, instead of adverbs, became a new heterogeneous category and, in addition, the identification of many of its instances became context-dependent.

2.1 Identification of Particles

To our knowledge, there is no universal criterion for defining particles, except the negative one (a non-declined word which is not a conjunction, an adverb, a preposition nor an interjection). In an attempt to define this category on a functional basis, several sets of subcategories have already been proposed and the research remains ongoing (see [16] for an extensive enumeration). Bearing in mind a practical goal of the delimitation of particles, we chose a bottom-up approach: the first step was to compile a list of particle candidates based on example words obtained from grammars and related works ([22], [5], [3], [10], [13], [9]) and on lists of words tagged as particles (CNC – SYN2015, SNK – prim-7.0) or as similar classes (ATT,

[1] The Collins COBUILD Dictionary, based also on a corpus, probably introduced such an interpretation for the first time.

CM and MOD functor in the Prague Dependency Treebank 3.0) in corpora. In the next step, the items were sorted approximately according to prominent features they had in common, in relation to their function. Inspired by previously suggested subclasses (namely by [13], [10]), we built a generalized system which integrates commonly used perspectives. With many overlaps between the groups, we identified particles:

1. structuring discourse and/or information in an utterance (sentence adverbials, restrictive particles):
   mimochodem 'by the way', *obzvlášť* 'particularly, especially', *ostatně* 'anyway', *také* 'also'
2. indicating sentence mood/type or its illocutionary function (questions, wishes, appeals, threats etc.), often adding expressivity:
   ***Kéž** bych měla dítě* '**If only** I had a child' (CNC – InterCorp v9)
   ***Běda, jestli** za to můžeš ty* ‚This had **better not** be your fault' (CNC – InterCorp v9)
3. implying a presupposition:
   ***ještě** větší* '**even** bigger' (assuming smaller)
   *to je **teprve** začátek* 'that's **just** the start' (despite the assumption that nothing more is to come, CNC – InterCorp v9)
4. commenting on a proposition and its wording, in terms of modality, emotions or attitude (hedges, amplifiers, emphatics):
   asi 'perhaps', *jaksi* 'somehow', *naštěstí* 'fortunately', *naprosto* 'absolutely', *opravdu* 'really'
5. expressing affirmation and negation:
   ***Pravda**, ale nemáme na vybranou*. '**True**, but we have no choice.' (CNC – InterCorp v9)
   *žádné plachty, **kdepak*** 'no oars, **nay**' (CNC – InterCorp v9)
6. serving as fillers:
   tentononc 'whatsit', *jako* 'like (colloquial)'

The list of particle candidates was further refined. Firstly, since our goal is to tag texts from the 19th century, we checked the items against the first modern dictionary of Czech ([19], 1935–1957), which captures the language of classic writers of the period in question, and removed words that started to be used as pragmatic devices only later (e. g. *prakticky* 'practically, basically') and also foreign words (e. g. *apropos*) due to the unknown degree of their integration into Czech vocabulary. Secondly, we extracted older derived forms, variants and synonyms from the dictionary using the categories obtained from the list.

The main decisions made throughout the whole procedure concerned the extent to which we should adhere to the criterion of function. This criterion goes across established boundaries of parts of speech, and when there is no additional feature distinguishing particles from the other classes, as mentioned above, the decision about the inclusion or exclusion of particular words can be made only on the basis of convention and with respect to a practical purpose. For example, we did not include many of the words which specify the intensity of particular action or quality (degree adverbs in Czech school tradition, e. g. *velmi* 'very', ***moc*** *hezký* '**pretty** good', ***strašně*** *dobře* '**awful** good') into the fourth subclass because such intensifiers are largely metonymy- and metaphor-based and therefore still productive. The subclass would thus be unpredictably extensive (cf. **hodinářsky** přesná práce, lit. 'watchmaker.ADV accurate work', 'very accurate'). We chose only the words explicitly expressing the highest/lowest grade of intensity, which also function as rheme indicators in an utterance (e. g. *maximálně* 'maximally, a maximum of'). Similarly, we distinguished between two types of "commenting words" (*hlavně* 'mainly' vs *většinou* 'mostly') according to the difference between "limit" and "degree". When a borderline case occurred (e. g. *nadmíru* 'above the line', 'extraordinarily'), we tended to make a decision according to the semantics of the word (*nadmíru* refers to the usualness rather than to the highest extent, and therefore we classified it as an adverb). The overall aim thus was not to come up with the one and only right set of principles to identify particles but to keep them as a category „for the remaining cases" while understanding what makes them different (and which cases can be still counted as less typical representatives of other parts of speech).

2.2 The Estimate of Particle Ambiguity

Having adjusted the compiled list to 19th century language, we arrived at a final list (further referred to as P-list and P-words) consisting of more than 500 items (available at `https://trnka.korpus.cz/~zitova/`). This number was quite surprising given that the list obtained from the CNC – SYN2015 contains 214 items (excluding words with hyphens that were incorrectly tagged as particles) and even the more extensive list from the SNK – prim-7.0 comprises 374 items.[2] We would also expect more particles identified in newer texts than in the older ones given a general shift towards oral discourse during the time (cf. [21]: 254). Our assumption is that the class of particles is intentionally maintained rather small to leave out words with multiple morphological interpretation. Therefore, to estimate the ambiguity rate of the items in the P-list and to map the approach to tagging particles in the corpus of present-day Czech, we tested the P-list against the CNC – SYN2015 corpus.

We used a multi-level frequency distribution function of the KonText interface to get a list of matched words and their tags. Despite the adjustments of the P-list to the older language, the vast majority of words was found in the corpus (442 items of the original 512). Words with more than one part-of-speech tag were counted as ambiguous.

	particles	**%**	**non-particles**	**%**
ambiguous	67	48.55	35	11.55
unambiguous	71	51.45	268	88.45
total	138	100	303	100

Tab. 1. Part of speech assigned to the words from the P-list in the CNC – SYN2015 corpus

[2] We found also 395 particles in the CNC - Prague Spoken Corpus but the list largely consists of phonological variants of a limited set of words, preserved in the transcription.

As can be seen from Table 1, roughly a half of the P-words tagged as particles in the CNC – SYN2015 is, according to the tagging scheme, homonymous with representatives of another part of speech. On the contrary, almost 90% of the P-words not tagged as particles do not need to be disambiguated in the context. The tendency to somewhat avoid particles in the POS tagging is thus understandable given its ambiguity rate. In most cases, particles are homonymous with adverbs: 66% of ambiguous particles (44 out of 67) also have an adverbial interpretation and adverbs represent 68% of non-particles in this analysis (183 out of 268, the rest is accounted for by 8 other POS).

It is precisely the difference between particles and adverbs that is most difficult to recognize. Examples 5 and 6 show one of the cases that are fairly impossible to distinguish for an automatic tagger (stochastic or rule-based), example 7 poses a problem even for a human:

(5) *"Uzavřeme sázku," řekl Lukáš. [...] "**Dobře**," řekl nakonec [Richard].* "Let's make a bet," said Lucas. - "**Alright** then," said Richard. (CNC – SYN2015, affirmative particle in Czech)

(6) *"Jak se ti vede?" - "**Dobře**."* "How are you?" "I'm **fine**." (CNC – SYN2015, adverb in Czech)

(7) *hebrejština se normálně píše zprava doleva, ale átbaš můžeme **jednoduše** použít i takto* (CNC – InterCorp v9)

‚Hebrew is normally written in the opposite direction, but we can just as **easily** use Atbash this way' (adverb in Czech)

‚Hebrew is normally written in the opposite direction, but **in short**, we can use Atbash this way' (alternative interpretation; discourse-structuring particle in Czech)

Thus it seems recommendable not to integrate particles into the morphological tagging scheme unless there is a possibility of their manual disambiguation (and even in that case only with certain restrictions, see section 2.4). Standard dictionaries of Czech, containing example sentences or phrases, continue the tradition of treating such words as adverbs probably for similar reasons. Another option is to introduce ambiguity tags with information about the probable accuracy in large corpora which, however, presupposes at least the identification of the typical cases in their contexts.

2.3 The Current State of the Tagging of Particles in the CNC – SYN2015

Concerning the original set (214 items after refinement), 42% of particles have more than one tag which is less than in the case of the P-list. Nevertheless, we have found certain inconsistency in the tagging scheme. The original set contains also salutations and swear words (e. g. *ahoj* 'Hi!', *kčertu* 'Damn!', *ježíši* 'Jesus!') which are traditionally regarded as interjections (cf. [1]). It is to be said, though, that the difference between interjections and particles is not always clear (cf. category names like "particles of contact" and "particles of emotions" [13]). We will focus on this issue in section 3.

Overall, there does not seem to be any function-based conception of particles behind the CNC - SYN2015. Candidate words have thus been probably assessed independently, as can be seen from the different tagging of close variants and synonyms, e. g. *nejspíš* and *nejspíše*, both meaning 'probably' (1. adverb or particle, distinguished without any obvious contextual clue by the stochastic module of the tagger; 2. adverb only), *opravdu* and *doopravdy*, both 'really, truly' (the same case) or *bezesporu* and *nepochybně*, both 'undoubtely, certainly' (1. particle, 2. adverb). Although there certainly exist some different features of contexts of these words, they are rather subtle or their importance for the POS categorization is questionable (e. g. there are 60,22 i.p.m. of *opravdu* before an adjective, whereas only 2,58 i.p.m. of *doopravdy* in the same position in the corpus, however, this has not been recognized as an important feature yet).

Another consequence of the lack of conception is the uncertain boundary between particles, adverbs and conjunctions. For example, *vždyť* 'after all; because', *však* 'well; however, though' and *přece* 'surely, after all; though' are all able to express a syntactic relationship as well as a pragmatic meaning but they are tagged differently (1. conjunction only, 2. conjunction or particle, 3. adverb or particle). We deal with this issue in the next section.

2.4 Particle as a Functional Attribute

Trying to avoid loss of information about the pragmatics of texts (which comes with using adverbial tags only) on the one hand and unreliability of tagging on the other, we suggest to follow a morphological criterion first (almost every particle is morphologically an adverb, having similar affixes etc.), as the dictionaries usually do, and then to optionally add information about the function of such an adverb, which can be not only pragmatic but also syntactic (connective), as mentioned above. As examples 8 and 9 show, the same word can have different functions and none of them is typical for adverbs (primarily used to denote circumstances) to which it points with its formation (the suffix *-ak* occurs also in *tak*, *jinak* and a few other adverbs).

(8) ***Však*** *víte*. '**Well**, you know.' (CNC – InterCorp v9, pragmatic)

(9) *...první večer padla volba na ni. Nazítří ráno* ***však*** *došlo ke změně* ,...for the first evening she was his settled choice. The next morning, **however**, made an alteration...' (CNC – InterCorp v9, syntactic)

Tagging the first case as an adverb serving as a particle due to its pragmatic function (ADV + PART) and the second case as an adverb with a connective function (ADV + CONJ) allows us to avoid the difficult clear-cut decision whether the word *však* is still a particle when it connects two adjacent utterances (should we conceive it as a discourse-structuring particle, to keep the interpretation close to its other usage, as an adverbial connector or as a conjunction?). This manner of annotation also enables us to capture the connective function of traditional adverbs like *přesto* (lit. 'over it', 'yet, still, however'), *proto* (lit. 'for it', 'therefore') etc. which can not only modify a conjunction but also substitute it, so they are partially grammaticalized as connective devices.

The introduction of multiple tags, however, also presupposes clear rules for their application. For example, when there is a collision between pragmatic and syntactic function (e. g. *vždyť* indicating reproach and marking an explicative relationship at the same time in some cases), there are at least two possible solutions: 1. the pragmatic function (ADV + PART) will be given precedence for the relationship between the two utterances is implied by their propositions and does not need to be expressed overtly (explication is based on a partial reformulation of the previous proposition; more on the nature of such relationships in [18]); 2. a new tag (e. g. ADV + MIX1) will be introduced to denote this combination (to avoid a triplet of tags), which seems to capture the nature of the problem more accurately. Nevertheless, despite the difficulties with setting rules, this system allows more space to deal with problematic cases than a single-tag solution and well documented rules will be informative both for the users of the corpus and for an automatic tagger.

2.5 The Interface Between Particles and Interjections

Words, that can be found included either in the category of particles or interjections, are especially response words, *ano* 'yes' and *ne* 'no'. As opposed to our view in 2.1 (also e. g. [10]), which conforms to the school tradition, some Czech papers ([5], [23]) argue that *ano*, *ne* are interjections due to the criterion of forming independent non-elliptical utterances (cf. [1], [20]). Cvrček et al. ([5]) mention *ne* along with content words used in rejections (cf. example 10 and 11). As interjections are supposed to be closer to content words than particles, the analogy with content words of rejection would support the view that *ne* is an interjection.

(10) *Jseš na flámu, bejby? – **Hovno**, já jsem na flámu pořád.* ‚You been partyin‘, baby? **Shit**, I been partyin‘ all the time.‘ (CNC – InterCorp v9)

(11) *Mrzí mě to. – Ale, **houby** se stalo.* 'I'm sorry about that. – Hey, **shit** happens.' (lit. 'mushrooms', CNC – InterCorp v9)

On the other hand, interjections are also supposed to express a rudimentary proposition which should be paraphrasable (e. g. *Ouch, it hurts!*) and it is hard to imagine how to paraphrase *ne* otherwise than by repeating the previous utterance (usually a question), only with the negative polarity. The non-elliptical nature of *ne* is thus questionable.

Example 11 is further complicated by the fact that *houby* 'shit', originally a noun, is a clause constituent which is untypical both for particles and interjections. Although Komárek et al. ([13]) and Kleňhová ([11]) argue that interjections can perform a function of any other part of speech in the clause structure (with an implicit reference to their primarily independent use), the concept of secondary interjection in its secondary function, which would be the case here, seems too complex. As in section 2.4, we prefer to tag the word according to its morphology first (NOUN) and its function second (PART). It is obvious that the word was reanalysed as uninflected thanks to the homonymy of its inflectional suffix *-y* with an adverbial suffix *-y* (*hovn-o* is the same case).

Overall, it seems that the devices of negation and affirmation should be conceived as particles rather than as interjections. Besides the reanalysed cases mentioned above, there may certainly be a problem with disambiguation of sound-like words like *hm* (does it express a response, hesitation or something else?) and even with *ano*, *ne* 'yes, no' expressing emotional reaction to an event (success, loss etc., cf. below). The most appropriate solution seems to be to tag them as borderline cases between particles and interjections, though it indicates a need for another type of multiple tag: the OR tag (different from the AND tag suggested in section 2.4), denoting two competing interpretations.

Particles expressing emotional comment on the formulation of an utterance are closely related to interjections. In an attempt to distinguish between them, Vondráček ([23]) proposed to follow the criterion of syntactic independency (examples are taken from [23]):

(12) ***Bohužel*** *se ještě nevyjádřila* 'She **unfortunately** has not commented on it yet' (PART)

(13) ***Bohužel****, ještě se nevyjádřila.* '**Unfortunately**, she has not commented on it yet' (INTJ)

Unlike the English equivalent of *bohužel* 'unfortunately', the Czech expression can be either intervowen with the structure of a clause through a change in the word order of enclitics (e. g. *se* above) or separated by a comma as an independent element. When the word is separated, Vondráček draws a parallel with interjections and their paraphrases. However, it remains unclear what to do with clauses without such a change in the word order (in 14, the enclitic *tam* 'there' stays in Wackernagel's position):

(14) *Dámy tam,* ***bohužel****, přístup nemají* 'Ladies, **unfortunately**, are not allowed to enter there' (CNC – SYN2015)

Furthermore, graphically separated occurrences of *bohužel* are quite infrequent and may thus be the result of a stylistic rather than a functional variation. Examining the frequency of such occurrences in related particles of emotions (*naštěstí* 'fortunately'; *naneštěstí* 'unfortunately'; *díkybohu*, *bohudíky*, *bohudík*, *chválabohu*, *zaplaťpánbůh* 'thank God'), however, we found substantial differences between particular words indicating that relying purely on the analogy with one of them could be misleading.

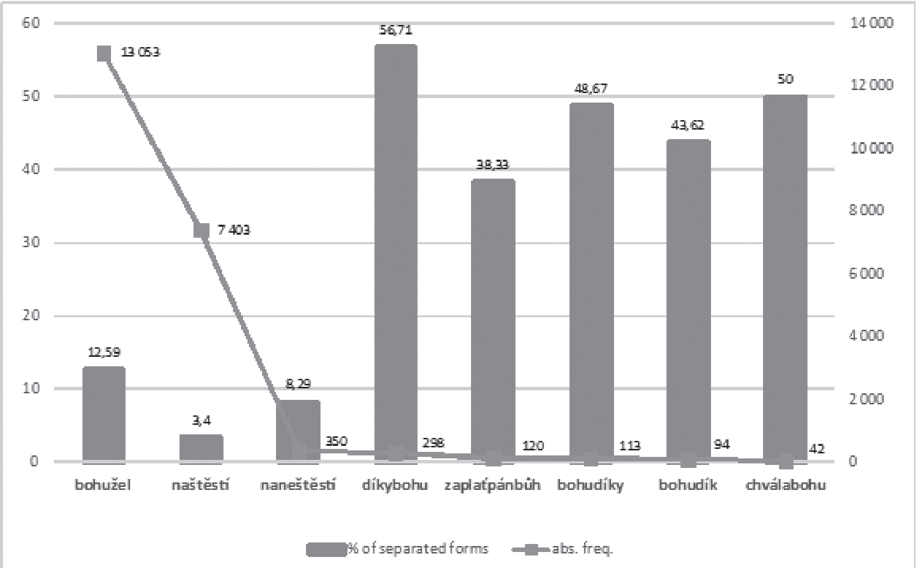


Fig. 1. The percentage of graphically separated particles and their absolute frequency in the CNC – SYN2015

As can be seen from Graph 1, a group of compound words with the element -*bůh*, -*bohu* ('God') besides another noun or verbal element tend to be separated more often than the others, unless they are too frequent (as is the case of *bohužel*). On the other hand, *naneštěstí* (lit. 'to unhapiness', with a prepositional element), though rather infrequent, is mostly accepted to a clause structure. Word formation and frequency thus have an impact on whether a word is perceived as an integral part of a clause (and therefore should not constitute a truly non-elliptical utterance) or still as a parenthesis. Given that various stages of conventionalization are visible even in contemporary language, let alone the older periods, when the word is graphically separated, we suggest to tag it 1. as adverb due to the compound form, 2. both interjection and particle (e. g. ADV + MIX2).

3 CONCLUSION

Showing problematic cases of function words, we aimed to draw attention to theoretical backgrounds of morphological annotation of texts in corpora. The analysis of the corpus of present-day Czech allowed us to considered the complexity of including the category of particles into a tagging scheme and we arrived at a recommendation not to apply this category to large and automatically tagged corpora because of a high rate of ambiguity of respective words. Inspired by the BNC2 and Czech dictionaries, we recommend rather the extensive use of the category of adverbs and the application of ambiguity tags. This seem to be reasonable also for the diachronic corpus of Czech in preparation because of the language change that affects this pragmatic means considerably. The basic interpretation of word forms should lean on formal morphology and word formation and then attributes of particular function should be added if the word is listed in a list of functionally-conventionally defined particles. When such a word has also a clause-linking function, it should be given also a tag for conjunction. Multiple tags and tags with attributes seems to be the right mean to tackle the problem of categorization of scalar phenomena like those of language.

References

[1] Ameka, F. (2006). Interjections. In Brown, K., editor, *Encyclopaedia of Language and Linguistics*, pages 743–746, Elsevier, Amsterdam.

[2] Atwell, E. S. (2008). Development of tag sets for part-of-speech tagging. In Ludeling, A. and Kyto, M., editors, *Corpus Linguistics: An International Handbook*, Volume 1, pages 501–526, Walter de Gruyter.

[3] Bedřichová, Z. (2008). Částice implikující presupozici jako podstatná složka větného významu. *Čeština doma a ve světě* 3–4:119–126.

[4] *Collins COBUILD Dictionary.* Accessible at: `http://collinsdictionary.com`, retrieved 2017-03-20.

[5] Cvrček, V. et al. (2010). *Mluvnice současné češtiny: Jak se píše a jak se mluví*, Karolinum, Praha.

[6] Grepl, M. (1989). Partikulizace v češtině. *Jazykovědné aktuality*, 26:95–100.

[7] Havránek, B. and Jedlička, A. (1960). *Česká mluvnice*. Státní pedagogické nakladatelství, Praha.

[8] Havránek, B. et al., editor (1989). *Slovník spisovného jazyka českého*. Academia, Praha.

[9] Hoffmannová, J. (1983). *Sémantické a pragmatické aspekty koherence textu.* Ústav pro jazyk český ČSAV, Praha.

[10] Karlík, P., Nekula, M., and Rusínová, Z. (1995). *Příruční mluvnice češtiny*. Nakladatelství Lidové noviny, Praha.

[11] Kleňhová, E. (2011). Pojetí interjekcí v některých českých mluvnicích. *Naše řeč*, 94(5):242–255.

[12] Kleňhová, E. (2012). Postavení a užívání interjekcí v současné češtině. *Naše řeč*, 95(5):238–254.

[13] Komárek, M. et al. (1986). *Mluvnice češtiny: vysokoškolská učebnice pro studenty filozofických a pedagogických fakult, aprobace český jazyk. [Díl] 2. Tvarosloví*. Academia, Praha.

[14] Kroupová, L. et al. (eds.) *Slovník spisovné češtiny pro školu a veřejnost: s Dodatkem Ministerstva školství, mládeže a tělovýchovy České republiky*. Academia, Praha.

[15] Milička, J. (2013). Bootstrapper [software]. Accessible at: `http://milicka.cz/en/bootstrapper.`

[16] Nekula, M. (2017). Částice. In Karlík, P., Nekula, M., and Pleskalová, J., editors, *Nový encyklopedický slovník češtiny*. Accessible at: `https://www.czechency.org/slovnik/ČÁSTICE`, retrieved 2017-03-25.

[17] *Oxford English Dictionary.* Accessible at: `http://www.oed.com`, retrieved 2017-03-20.

[18] Poláková L. et al. (2012). *Manual for Annotation of Discourse Relations in the Prague Dependency Treebank.* Technical Report No. 47, ÚFAL, Charles University, Prague. Accessible at: `http://ufal.mff.cuni.cz/discourse/publications.`

[19] *Příruční slovník jazyka českého* (1935-1957). Státní nakladatelství, Praha.

[20] Quirk, R., Greenbaum, S., Leech, G., and Svartvik, J. (1985). A Comprehensive Grammar of the English language. Longman, London and New York.

[21] Reppen, R., Fitzmaurice, S. M., and Biber, D., editors (2002). *Using corpora to explore linguistic variation* (Vol. 9). John Benjamins Publishing.

[22] Štícha, F. et al. (2013). *Akademická gramatika spisovné češtiny.* Academia, Praha.

[23] Vondráček, M. (1998). Citoslovce a částice – hranice slovního druhu. *Naše řeč*, 81(1):29–37.

DESIGNING THE DATABASE OF SPEECH UNDER STRESS

RÓBERT SABO[1] – JAKUB RAJČÁNI[2]
[1]Institute of Informatics, Slovak Academy of Sciences, Bratislava, Slovakia
[2]Faculty of Arts, Comenius University, Bratislava, Slovakia

SABO, Róbert – RAJČÁNI, Jakub: Designing the Database of Speech Under Stress. Journal of Linguistics, 2017, Vol. 68, No 2, pp. 326 – 335.

Abstract: This study describes the methodology used for designing a database of speech under real stress. Based on limits of existing stress databases, we used a communication task via a computer game to collect speech data. To validate the presence of stress, known psychophysiological indicators such as heart rate and electrodermal activity, as well as subjective self-assessment were used. This paper presents the data from first 5 speakers (3 men, 2 women) who participated in initial tests of the proposed design. In 4 out of 5 speakers increases in fundamental frequency and intensity of speech were registered. Similarly, in 4 out of 5 speakers heart rate was significantly increased during the task, when compared with reference measurement from before the task. These first results show that proposed design might be appropriate for building a speech under stress database. However, there are still considerations that need to be addressed.

Keywords: stress, arousal, stress detection, heart rate, speech under stress, speech database

1 INTRODUCTION

Research in the field of speech processing is increasingly drawn to specific manifestations of speech such as speech under stress. This area of speech research is closely linked with psychology and physiology, which should answer the question, what is stress and how to identify it in speech. Our goal in this study is to establish methodology for creating a database of speech under real stress, which may be used in other experiments investigating speech under stress in future.

2 RESEARCH OF STRESS IN SPEECH, EXISTING SPEECH DATABASES

One of the most widely used speech databases in speech under stress is the SUSAS database – Speech Under Simulated and Actual Stress [1], [2]. The database consists of four domains, encompassing a wide variety of stresses and emotions. It contains 32 speakers (13 female, 19 male), with ages ranging from 22 to 76 years who have made more than 16 000 utterances. SUSAS also contains several longer speech files from four Apache helicopter pilots and a common highly confusable vocabulary set of 35 aircraft communication words. Unfortunately, in carrying out acoustic analyses, researchers are limited by noisy channel and the 8 kHz sampling frequency.

Speech database containing speech under stress with high quality recordings is the CRISIS database [3]. This database contains acted expressive speech from 15

DOI 10.1515/jazcas-2017-0042

speakers. Each speaker records a set of 150 sentences, each in different arousal level. Once in a neutral manner (referred to as level 1 of tense arousal), then with higher imperativeness, like a serious command or directive (level 2), and finally like an extremely urgent command or statement being declared in a situation when human lives are directly in danger (level 3). Even though high-quality recordings allow to perform a number of acoustic analyses [4], [5], database is a missing part with speech under realistic stress.

In our approach, we propose a method to obtain high-quality recordings of speech under real stress. One of the important questions, that needs to be answered first, is: What is stress and how it can be measured?

2.1 Definition of Stress

Proposing a scientific definition of stress is a difficult problem, in a large part, due to the term being too general and hardly usable in different contexts [6]. In the general sense, stress is a state in which internal integrity (or homeostasis) of an individual is challenged via external or internal means – called stressors [7], [8].

Stress results in a complex physiological reaction, which can be marked by changes in bodily systems, such as autonomic nervous system (ANS), endocrine and immune system [8]. Sympathetic branch of ANS becomes predominant during stress reaction, which leads to acceleration of heart rate (HR), secretion of noradrenaline and adrenaline, as well as inhibition of gastrointestinal function, changes in electrodermal activity (EDA) and many other physiological changes. All these bodily reactions serve as preparatory measures for behavioral reaction to stress and successful adaptation. Increased preparatory physiological activation, may be labelled by term "arousal" which is also used in context of emotions as a level of overall physiological activation.

Investigating stress biomarkers, such as heart rate, electrodermal changes, stress hormones, etc., are a large part of current stress research. On the other hand, speech changes in stress are not so well examined. Though, there are studies investigating speech changes, they differ in proposed understanding of stress and used methods and therefore yield different results.

Current research shows that speech changes that are a result of both involuntary bodily changes and voluntary effort, are also dependent on a particular type of stressor. Hansen [9] proposed a taxonomy of stressors and their impact on speech, based on the mechanism in which they perturbate speech process. Stressors were sorted to several categories such as: "zero order" – stressors with direct physical impact on speech (e.g. acceleration), "first order" – biological or chemical stressors (e.g. dehydration), "second order" which involves perception (e.g. Lombard effect) and "third order" – psychological, emotional and social stressors.

Besides a lot of research findings on stress detection from studies using acted stress, studies of real-life stress also show a detectable difference in speech. Lu et al. [10] obtained stress identification accuracy of 71.3% when comparing job interview with indoor neutral speech, and accuracy 82.9% when personalized model was used. Increased skin conductance level as an electrodermal stress related phenomenon was used to validate stress during job interview. Similarly, Luig, et al. [11] proposed heart rate and heart rate variability as relevant physiological correlates to speech analysis.

Presence of stress may be detected via physiological, but also from speech parameters. This study follows findings of prior research on analysis of speech under stress. Our aim is to develop a database of speech under realistic stress and to further validate it by using physiological indicators, such as heart rate and electrodermal activity.

For obtaining relevant data, we have chosen laboratory setting aiming to implement these issues: 1. It must induce a strong enough stress reaction, 2. Person under stress needs to speak as much as possible, 3. The setting must be relevant to real-life applications.

One of the problems with speech databases of real stress is that speakers may speak very little, or that utterances included in the database are too short to yield any notable results. Our previous work with acted stress enabled us to detect stress with high accuracy, however, conducting a study of realistic stress is necessary [3]. For these reasons the following design was proposed.

3 METHOD

3.1 Research Setting

Based on the mentioned aims, we decided to use a communication-based task, in which the research subject must give instructions on solving the task to their partner via microphone. In a setup like this, subject is forced to speak as much as possible, however, inducing a strong enough stress reaction is also essential. This was realized considering following stress factors:

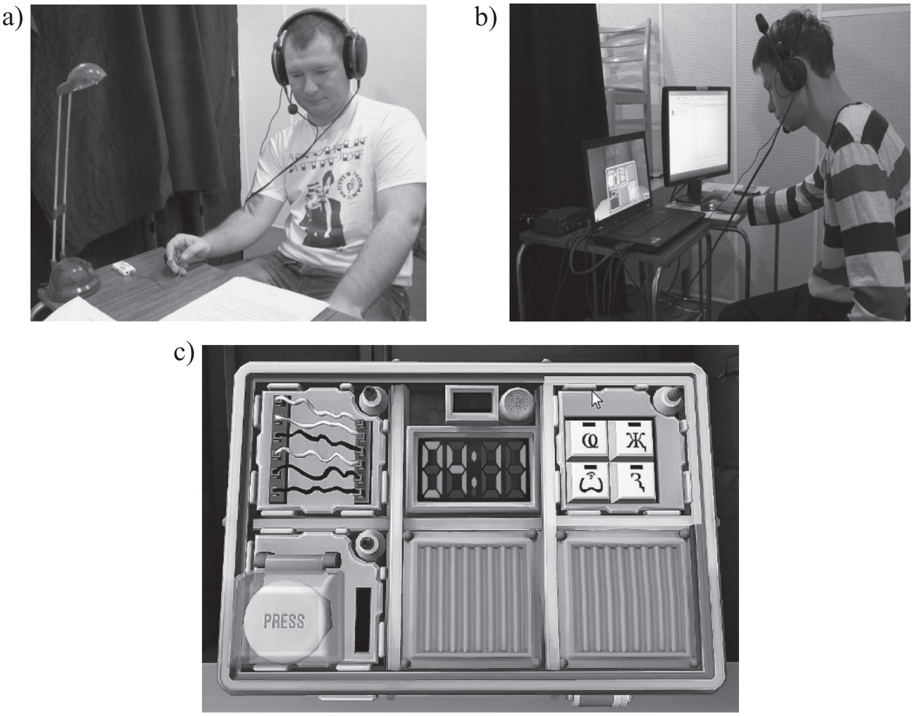


Fig. 1. Photo of a) research subject, b) researcher controlling the bomb on screen, c) bomb interface.

A) Task itself. For the purpose of data acquisition, we adapted a commercially available computer game "Keep talking and nobody explodes" [12]. The game itself is a moderate stress inducing task, in which two players dismantle a bomb composed of several modules, each representing a logical puzzle (Fig. 1c). While one player sees the bomb on screen (in our case a member of research team, to provide standard conditions for all subjects) (Fig. 1b), the other (a research subject) sees a manual with detailed instructions on solving individual puzzles (Fig. 1a). Two players do not see each other and they communicate only via microphone.

B) Time pressure. The game itself has a countdown timer, which can be adjusted to the task. Subject in our task sees the timer and hears beeping sounds in the headphones. Time pressure of 10 min for solving entire bomb composed of 6 modules provides a very hard, yet solvable task.

C) Environmental factors. During the task, subject is sitting in recording studio with lights off, only using a table lamp. At random times during dismantling the bomb, subject is disturbed by a siren in the headphones.

D) A reward and a set "best score." It is expected, that a research subject who solves the task has at least some degree of motivation to achieve a good result. For a task like this to become a stressor, it needs to be important and consequential to the subject. Therefore, we added incentives to enhance subjects' motivation. One incentive is financial reward. Subjects are instructed, that both they and their co-player (to increase their feeling of responsibility) will receive reward depending on their performance. They are told, they both receive 10€ for dismantling a bomb successfully, if they fail, they receive 1€ for each successfully solved module (Entire testing consists of three consecutive bombs so the reward can go up to 30€). Second incentive is information, that if players break the record, which is set to 8 min for solving the bomb, their reward doubles (20€ for solving a bomb).

To meet the conditions for quality of the recording testing was realized in an acoustically treated recording studio. Recordings were obtained via head-mounted close-talk microphone Sennheiser ME3 and Emu Tracker Pre USB audio interface with 48 KHz sampling frequency and 16 bit resolution. Participants used high-quality closed headphones Sennheiser HD 650. Each speaker was recorded in separate channel.

3.2 Participants and Procedure

The first recorded sample of the speech database contains speech from five speakers in the Slovak language. Subject A: female, 47 years; subject B: male, 28 years; subject C: male, 29 years; subject D: male, 45 years; subject E: female, 47 years.

Participants were contacted with a request to participate in a communication experiment, in which their voice and psychophysiology (heart rate – HR and electrodermal activity – EDA) will be recorded. All subjects were informed of the research procedure and signed informed consent. First, subjects were given bomb manual to study for 20 minutes to become acquainted with the game mechanics. Before studying the manual subjects were told they will communicate with a co-player, who is also a subject playing for the same reward. During debriefing after the test, subjects were explained, that the co-player was a member of research team, they could talk together and all subjects' questions about the research were answered.

The recording consisted of 10-minutes training game, which was realized using the same task with the experimenter in an easygoing manner. Training was used to collect reference values; stress factors were not present during training. Subsequently, three trials using the described procedure with the co-player were realized.

In this first test, the selection of speakers was not strictly limited of age and sex. Number of subjects in the research sample is only preliminary for initial tests of the research setting.

3.3 Analyzed Speech Features

The fundamental frequency (F0) and intensity values are specific for neutral speech of each speaker. Changes in frequency and intensity of speech can point to changes in speaker's emotional state. In the first data analysis, we evaluated mean fundamental frequency and mean intensity of speech for each task (training, trial 1, trial 2, trial 3), which represent approximately 8 minutes of speech for each task. When analyzing such long period of time, impact of various phonetic content and non-speech events such as hesitations should not be significant.

3.4 Physiological Measures

Beat to beat heart rate signal was obtained from all test subjects using FAROS 90° ECG device (Fig. 2a). Measurement of ECG was carried out using two electrodes, one positioned under right clavicle, the other on the left under ribs. Sampling rate for ECG was 250Hz, which is appropriate for high precision ECG and HR analysis.

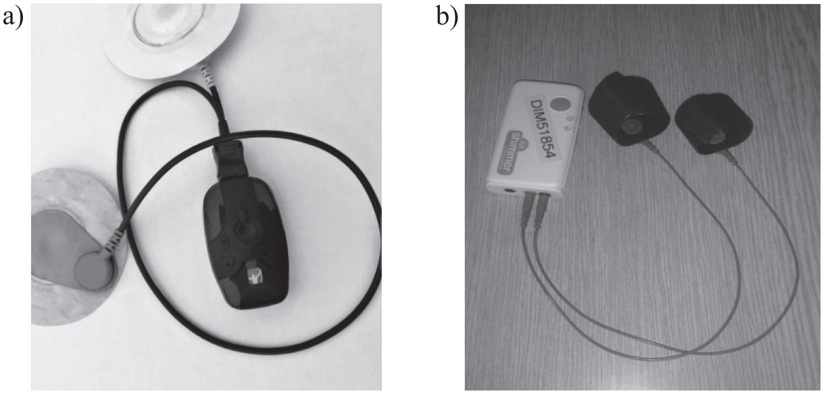


Fig. 2. a) FAROS 90° ECG device b) Consensys Shimmer GSR device

Though heart rate can be used as a reliable index of overall arousal and sympathetic activity, we may also calculate heart rate variability (HRV), which offers more information on autonomic nervous system influences of heart [11], [13].

Although Heart rate can be calculated from duration of single beat to beat interval, it is necessary to take HR changes during breathing cycle to consideration. Therefore, we analyzed 10s HR intervals, corresponding to events which occurred during data acquisition. On the other hand, heart rate variability measures can be reliably calculated only from longer segments of HR (at least 2–4 min) [13].

Electrodermal activity (EDA) is another useful indicator in stress research, which was previously used as a reference measure in study of speech [10]. From possible electrodermal phenomena, we measured skin resistance (in kΩ) via Consensys, Shimmer device (Fig. 2b). Both tonic, relatively stable skin resistance level and phasic, skin resistance responses can be further analyzed as stress indicators. This study will not include results from EDA analysis.

3.5 Subjective Stress Assessment

For assessment of subjective experience of stress and anxiety, we administered Slovak version of state anxiety inventory (STAI-X) [14]. STAI-X inventory is composed of 20 statements to which subjects answer on a 4-point scale. Test is used to describe an extent, to which a person feels anxiety at the given time. This inventory may be used for repeated measurements; we administered it before and after recording.

Moreover, after the recording, subjects also answered several standard questions regarding their motivation, feelings of stress and satisfaction with the achieved result.

4 RESULTS

Of all the subjects in the research sample, none could dismantle any of the given bombs in time, however two subjects were able to solve 5 of 6 modules before the bomb exploded.

Data analysis showed differences in both speech features and heart rate. Table 1 summarized increases in F0 and intensity of speech.

Speaker ID	Task ID	F0 [Hz]	Intensity[dB]
A	Training	173	60.9
	Trial 1	195	67.6
	Trial 2	198	68.3
	Trial 3	198	68.3
B	Training	144	64.8
	Trial 1	153	68.4
	Trial 2	153	69
	Trial 3	153	68.5
C	Training	126	52
	Trial 1	128	57
	Trial 2	126	57.8
D	Training	129	66
	Trial 1	153	79.9
	Trial 2	173	83.6
	Trial 3	170	82.8
E	Training	240	62.5
	Trial 1	246	52.5
	Trial 2	247	52.6
	Trial 3	242	52.6

Tab. 1. Average values of F0 and Intensity for training and each trial

4 of 5 speakers (except C) proved a significant increase of speaker's fundamental frequency in average of 14% (Fig. 3).

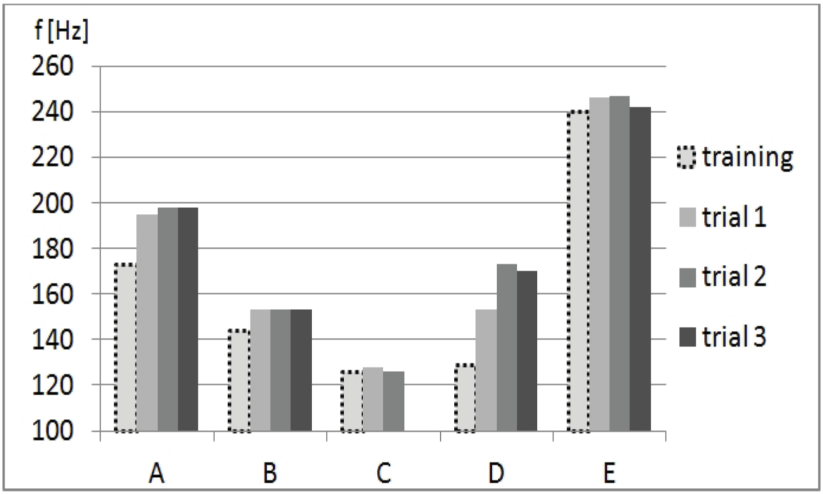


Fig. 3. Fundamental frequency for each speaker and each task

4 of 5 speakers (except E) proved an increase of the speech intensity in average of 16%.

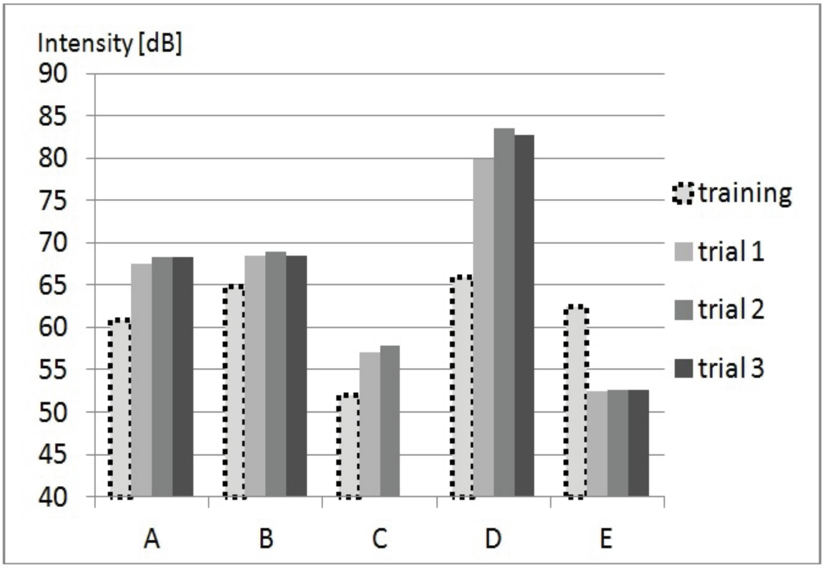


Fig. 4. Intensity of speech signal for each speaker and each task

Following Table 2 shows changes in heart rate (HR) expressed in beats per minute between training and three trials for each subject.

Speaker ID	Heart rate [bpm] Training	 Trial
A	82.21	87.35 82.79 83.18
B	87.04	93.96 96.64 90.30
C	67.87	69.57 74.85
D	80.92	84.92 85.94 85.72
E	108.20	110.98 109.36 106.51

Tab. 2. Average values heart rate (in beats per minute) for each subject and trial

Figure 5 illustrates changes in HR during entire recording. It contains detailed analysis of 10s HR windows from the recording (data from subject "D" were chosen for illustration). Increases in HR during individual trials (dismantling of bomb 1, 2 & 3) may be observed in Figure 5.

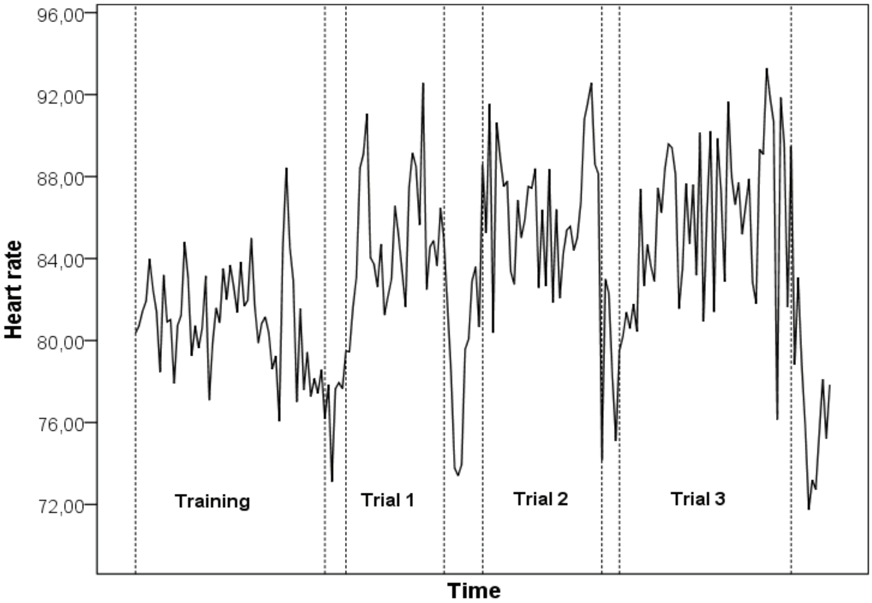


Fig. 5. Changes in HR during recording – differences between test and three trials (subject "D"). HR was sampled from 10s windows.

5 DISCUSSION

5.1 Speech Analysis and Physiological Findings

The initial analysis of five obtained recordings of speech under real stress show that proposed design should be appropriate for data acquisition. Even though a significant increases of F0 and intensity were observed only in 4 out of 5 speakers, balanced values between the trials point out, that speech obtained by the proposed method in trials is acoustically different from speech obtained in training. To identify whether this acoustic difference was induced by stress, analysis of psychophysiological correlates of stress was performed.

Findings from heart rate clearly indicate increase of physiological distress during trials. Moreover, as showed in Figure 5, increases of HR peaked in the last moments before the bomb exploded. However, due to low number of subjects so far, we cannot statistically evaluate these differences for the whole sample.

It is also important to note, that heart rate differs between individuals in a large extent. Factors such as age and sex must be taken into consideration when interpreting the HR data.

5.2 Further Methodological Issues and Considerations

Research Subjects. In the following data collection using this design, it is important to test at least 20 subjects, all of which fall into one age category. Because of using a computer game as a task interface, young subjects (age 18–30) would be optimal. Secondly, if we want to compare men and women, larger sample with balanced groups will be necessary. It is important that test subjects are naive to the task before participating in the study.

Design changes. In the first test, we needed to minimalize variables considered; therefore, we decided to use a member of the research team to stand in the role of co-player. The main advantage of this setup is that every subject had similar conditions during the game as their co-player responded in a standard manner (as somebody who sees this task for the first time). Moreover, if the subject spoke very little, or the utterances were very short, co-player encouraged them with asking more questions about the task at hand. However, there is a possibility of using two groups of subjects for both player positions in the game. This alternative may be useful to collect more speech data, however, subjects from different player positions will hardly be comparable.

Linguistic point of view. The speech in the database contained interesting linguistic, phonetic phenomena such as hesitations, repetitions, changes in speech rate, etc. High-quality stereo recordings allow us to perform precise analysis of overlapping speech patterns. If a design with real subjects on both player positions were used, the database might be a suitable for research on turn taking in speech.

In future, we plan to evaluate also other relevant acoustic features such as F0 maximum, intensity maximum, root mean square, spectral energy distribution etc., and also evaluate shorter speech segments, possibly related to annotated events during the game. A detailed phonetic annotation at the level of statements, words, phonemes will be carried out.

Other possible expansion of the database might be inclusion of another language, besides Slovak, if the prepared Slovak database yields good results in obtaining speech under stress.

ACKNOWLEDGEMENTS

The research leading to the results presented in this paper has received funding from the European Union FP7 under grant agreement n° 312382 (GAMMA – Global ATM Security Management project [15]. This study was also supported by grants APVV-0496-12, VEGA 1/0739/17.

References

[1] Hansen, J. H., Bou-Ghazale, S. E., Sarikaya, R., and Pellom, B. (1997). Getting started with SUSAS: a speech under simulated and actual stress database. *Eurospeech*, 97(4):1743–1746.

[2] Hansen, J. H. SUSAS LDC99S78. Web Download. Philadelphia: Linguistic Data Consortium, 1999. Accessible at: `https://catalog.ldc.upenn.edu/LDC99S78.`

[3] Sabo, R., Rusko, M., Ridzik, A., and Rajčáni, J. (2016). Stress, Arousal, and Stress Detector Trained on Acted Speech Database. In *International Conference on Speech and Computer*, pages 675–682.

[4] Rusko, M., Darjaa, S., Trnka, M., Sabo, R., and Ritomský, M. (2014). Expressive Speech Synthesis for Critical Situations. *Computing and Informatics*, 33(6):1312–1332.

[5] Rusko, M., Darjaa, S., Trnka, M., Ritomský, M., and Sabo, R. (2014). Alert!... Calm Down, There is Nothing to Worry About. Warning and Soothing Speech Synthesis. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, pages 1182–1187, Reykjavik, Iceland.

[6] Newport, D. J., and Nemeroff, C. B. (2002). Stress. In Ramachandran, V. et al., editors, *Encyclopedia of Human Brain*. vol. 4, pages 129–139, Academic Press.

[7] Mc Ewen, B., and Lupien, S. (2002). Stress: Hormonal and Neural Aspects. In Ramachandran, V. et al., editors, *Encyclopedia of Human Brain*. vol. 4, pages 129–139, Academic Press.

[8] Chrousos, G. P. (2009). Stress and disorders of the stress system. *Nature Reviews Endocrinology*, 5(7):374–381. Accessible at: `http://doi.org/10.1038/nrendo.2009.106.`

[9] Hansen, J. H. L. et al. (2000). The Impact of Speech Under 'Stress' on Military Speech Technology. NATO PROJECT 4 REPORT.

[10] Lu, H., Frauendorfer, D., Rabbi, M., Mast, M. S., Chittaranjan, G. T., Campbell, A. T., and Choudhury, T. (2012). Stresssense: Detecting stress in unconstrained acoustic environments using smartphones. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pages 351–360, ACM New York, NY, USA.

[11] Luig, J., Sontacchi, A., Goswami, N., Moser, M., and Shaw, C. (2010). Conception and Realization of Speech Recordings for Instantaneous Stress Level Assessment. In *9th EUROCONTROL Innovative Research Workshop and Exhibition*, Nice, France.

[12] Computer game *Keep talking and nobody explodes*. Accessible at: `http://www.keeptalking-game.com.`

[13] Berntson, G. G., Thomas Bigger Jr. J., Eckberg, D. L., Grossman, P., Kaufmann, P. et al. (1997). Heart rate variability: Origins methods, and interpretive caveats. *Psychophysiology*, 34(6):623–648.

[14] Müllner, J., Ruisel, I., and Farkaš, G. (1980). Príručka pre administráciu, interpretáciu a vyhodnocovanie dotazníka na meranie úzkosti a úzkostlivosti. Psychodiagnostické a didaktické testy. 93, Bratislava.

[15] GAMMA – Global ATM Security Management project. Accessible at: `http://www.gamma-project.eu.`

ANNOTATION OF THE EVALUATIVE LANGUAGE IN A DEPENDENCY TREEBANK

JANA ŠINDLEROVÁ
Faculty of Mathematics and Physics, Charles University, Prague,
Czech Republic

ŠINDLEROVÁ, Jana: Annotation of the Evaluative Language in a Dependency Treebank. Journal of Linguistics, 2017, Vol. 68, No 2, pp. 336 – 345.

Abstract: In the paper, we present our efforts to annotate evaluative language in the Prague Dependency Treebank 2.0. The project is a follow-up of the series of annotations of small plaintext corpora. It uses automatic identification of potentially evaluative nodes through mapping a Czech subjectivity lexicon to syntactically annotated data. These nodes are then manually checked by an annotator and either dismissed as standing in a non-evaluative context, or confirmed as evaluative. In the latter case, information about the polarity orientation, the source and target of evaluation is added by the annotator. The annotations unveiled several advantages and disadvantages of the chosen framework. The advantages involve more structured and easy-to-handle environment for the annotator, visibility of syntactic patterning of the evaluative state, effective solving of discontinuous structures or a new perspective on the influence of good/bad news. The disadvantages include little capability of treating cases with evaluation spread among more syntactically connected nodes at once, little capability of treating metaphorical expressions, or disregarding the effects of negation and intensification in the current scheme.

Keywords: dependency treebank, corpus, plaintext annotation

1 INTRODUCTION

The identification, description and analysis of evaluative language has been an important issue of computational linguistics since the rise of big data exploration. There are multiple ways to approach the issue, but basically, there are two main routes – one using the linguistically preprocessed training data to acquire reliable information about the structural properties of evaluative constructions, the other one believing in the power of unsupervised machine learning, extracting the information about evaluation from the textual data based on statistical co-occurrence of lemmata.

Within the linguistics-based approaches, a shift from plaintext annotations to the exploration of treebanks and employment of parsing mechanisms is noticeable, though both ways of data analysis have their advantages.

Plaintext annotation of evaluative states and roles is easy to learn for the annotator and in principle, does not require any specialized software. On the other hand, especially in case of large segments and less structured utterances, it may become confusing. Also, it can hardly be helped by automatic methods.

Using previously syntactically analyzed data requires availability of such data and specialized software, but it offers information helpful to the automatization of the annotation process. For example, a complex analysis of the targets as a unity of

DOI 10.1515/jazcas-2017-0043

the entity and its attributes is possible, even in case of discontinuous structures. Also, it is possible to trace sources and targets of evaluation easily via anaphora resolution. In the analysis, we can make use of explicit syntactic relations, such as dependency and valency. Considering the tectogrammatic (deep syntactic) layer as the layer of capturing evaluative relations, the problem of marking or not marking grammatical words as part of individual evaluative categories falls out of question, etc.

Our goal is to provide a sentiment annotation over an existing syntactically annotated treebank to be used in further sentiment classification and prediction tasks, and analyze its capacities to account for the persisting obstacles to the automatization of the sentiment identification and classification process. In this paper, we present an analysis of a small corpus of sentiment-annotated sentences that was created to verify the usability of the "evaluative state" annotation scheme on treebank data.

2 RELATED WORK

The current approaches to sentiment classification split basically into two branches, copying the two general approaches to machine learning: one branch promoting the use of syntactically parsed corpora as training data for the supervised learning of algorithms, and the other one favouring statistical methods (and, newly, also the neural networks) over the costly human annotation, i.e., the unsupervised learning. Both these approaches agree that using some kind of syntactic parsing yields better results than employing simple bag-of-words methods, because of the principle of compositionality of meaning, which says simply that the meaning of a compound expression is a function of the meaning of its parts and of the syntactic rules by which they are combined. Therefore, if we desire to interpret evaluation as a semantic issue in a complex and reliable way, we should use data capturing the mapping of syntactic and semantic functions.

A method to classify the sentiment polarity of a sentence based on compositional semantics was proposed, e. g., in [2]. A promising use of a treebank representation for predicting sentiment is described in [7]. The authors describe the creation of the Stanford Sentiment Treebank. The SST is an automatically parsed treebank of 11 855 movie review sentences, where each sentence was manually annotated for sentiment features by three (linguistically inexperienced) human annotators. The model trained on the SST computes sentiment using neural networks and deep learning based on the composition of meanings in the syntactic structure. The authors of [5] work with a dependency treebank and employ a probabilistic model counting polarities for each subtree. They also use a lexicon of polarity reversing words. In [6], the authors are concerned with solving metaphorical evaluations by a combination of a statistical and a rule-based system.

Though the newest studies suggest that unsupervised learning may yield optimal results at low costs in the task of automatic sentiment classification, the use of human annotated corpora lets us explore the linguistic dimension of evaluative constructions more reliably and describe properly the evaluative patterns in everyday language.

3 ANNOTATING EVALUATIVE LANGUAGE: THEORY AND DATA

3.1 Plaintext Annotation

The first phases of the project of capturing evaluative relations in Czech texts were carried out as series of plaintext annotations [9]. The individual parts of evaluative stance, the source, the target and the evaluative expression, were manually copied into the cells of a spreadsheet; each evaluative stance found in the text was treated separately. Thus, e.g., the Moilanen and Pulman [4] example *The senators supporting the leader failed to praise his hopeless preventive program*, which they use for computing the overall sentiment value for the sentence, would represent (at least) three separate evaluative states, see Table 1.

The plaintext data analysis suggested there are repeating patterns for expressing evaluative meaning in the language, but did not enable a clear extraction of such patterns, due to the lack of information considering the configuration of syntactic positions of the source, target and evaluative expression in the structure.

Evaluative state	Source	Evaluative expression	Target
1.	The senators	supporting	the leader
2.	The senators	failed to praise	preventive program
3.	AUTHOR	hopeless	preventive program

Tab. 1. Three evaluative states in the sentence *The senators supporting the leader failed to praise his hopeless preventive program*.

3.2 Treebank Annotation

In the second phase, we decided to use the data from the Prague Dependency Treebank 2.0 (PDT 2.0), a large and richly annotated treebank of Czech sentences [3], and apply the evaluative features to its tectogrammatic structures.

Since we need data analyzed for semantic and syntactic features, we make use of the tectogrammatical (deep syntactic) layer of PDT annotation. The choice of PDT data brought in several advantages, as well as disadvantages. It offers a complete, profound and reliable syntactic annotation with no extra annotator costs. On the other hand, a rather low amount of evaluative information is expected in the data, because the texts represent a rather objective journalist style.

The sentences of the Prague Dependency Treebank 2.0 were automatically searched for expressions matching the entries in the Czech SubLex 1.0. Czech Sublex 1.0 [8] is a Czech subjectivity lexicon, i.e., a list of subjectivity clues for sentiment analysis in Czech. It has been gained by automatic translation of a freely available English MPQA Subjectivity Lexicon [10] using a Czech-English parallel corpus CzEng 1.0 [1]. Additionally, some manual refinement of the lexicon followed in order to exclude controversial items. Finally, it contains 4626 domain-independent evaluative items (1672 positive and 2954 negative) together with their part of speech tags, polarity orientation and source English lemmas.

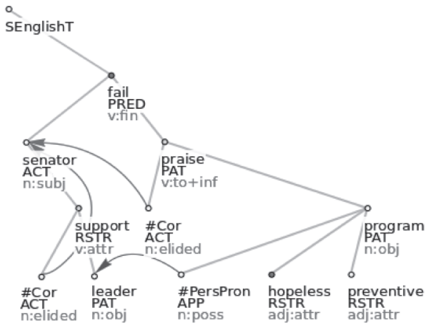


Fig. 1. Tectogrammatic representation of the sentence *The senators supporting the leader failed to praise his hopeless preventive program* with highlighted sentiments.

Fig. 1 shows a tectogrammatical representation of a typical sentence for annotation of evaluative states. There are four Sublex-suggested clues highlighted in green (for positive polarity, nodes *support* and *praise*) and red (for negative polarity, nodes *fail* and *hopeless*). The dependency links allow us to capture syntactic relations between the evaluative expressions and the sources and targets (if present overtly in the structure). The coreference arrows (blue for textual and brown for grammatical coreference) allow us to trace the lexical identity of sources and targets throughout the structure, and even beyond the sentence boundaries.

4 ANNOTATION ENVIRONMENT

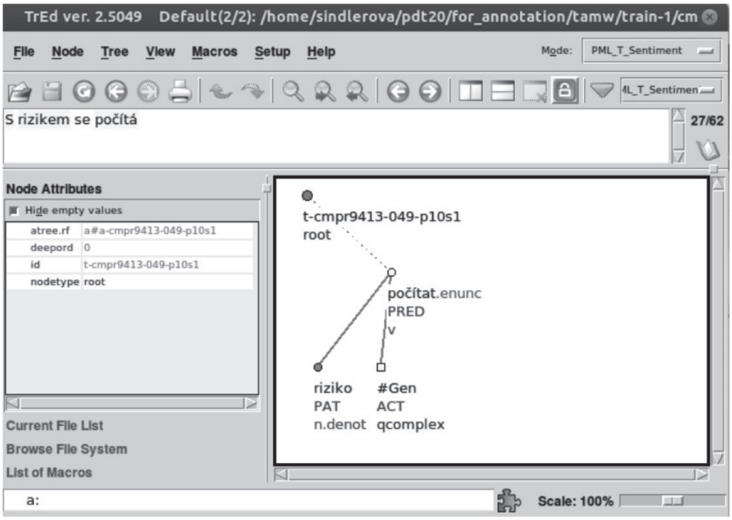
The annotation interface was designed as an extension of the tree editor (TrEd) environment, see. Fig. 2. TrEd is a fully customizable and programmable graphical editor and viewer for tree-like structures. Among other projects, it was used as the main annotation tool for the tectogrammatical annotation of the source treebanks (PDT). It allows displaying and annotating sentential tree structures on multiple linguistic layers with a variety of tags using either the Prague Markup Language (PML) or the Treex format.

The new extension, named PML_T_Sentiment, provides a GUI supporting the entry and modification of sentiment information. The information about the part of evaluative state the individual words stand for and their possible polarity value is stored in the attribute-value matrix. The sentiment information can be changed by the annotator via use of simple macros.

5 ANNOTATION PROCESS

The annotator is given a tectogrammatical tree for each sentence. Within the sentence, the potential candidates for evaluative nodes appear highlighted – nodes with potential positive orientation in green, nodes with potential negative orientation in red. The annotator is asked to annotate each separate highlighted node.

Annotating an evaluative node means making a decision and taking an action in each of the following issues:



TrEd ver. 2.5049 Default(2/2): /home/sindlerova/pdt20/for_annotation/tamw/train-1/cm

Mode: PML_T_Sentiment

S rizikem se počítá

27/62

Node Attributes

■ Hide empty values

atree.rf	a#a-cmpr9413-049-p10s1
deepord	0
id	t-cmpr9413-049-p10s1
nodetype	root

Fig. 2. Annotation environment

1) Is the node evaluative in the given context?
An annotator is obliged to decide whether the highlighted node is in fact evaluative in the given context. If so, the annotator selects the active evaluative node, decides on its sentiment value orientation, and selects the source and target in the context. If the node is not evaluative in the given context, the sentiment highlighting and the sentiment attributes for the given node can be removed.

2) What is the source and target of the evaluative expression?
Once the node is selected, the source and target of sentiment may be annotated. If the source or target is present in the immediate sentential context, the annotator is obliged to click on it (make it active) and set it as the source or target. This inscribes the node identifier into the value of the corresponding evaluative node attribute.
If the source or target is not to be found in relatively close context, the attribute "is_extern" of the corresponding role in the attribute list of the evaluative node must be set to value "1" manually.

3) What is the polarity orientation of the evaluative node?
For each highlighted node, a polarity value is originally ascribed from the Czech SubLex by an automatic procedure. This value can be confirmed or changed manually. Immediately after the value is manually set, the node is marked as "was annotated".

Apart from the nodes suggested by the automatic comparison with Czech SubLex items, any other node in the tree may be initiated as an evaluative expression by annotator. This is done using the function "Init Sentiment Value". By using this function, the attributes of sentiment are added into the list of attributes of the given node.

6 THE PILOT TREEBANK

The "pilot sentiment treebank" contains 1044 annotated sentences of the PDT 2.0 train data section. Since our previous work showed that the interannotator agreement on evaluative state and features identification is high [9], the sentences were annotated by a single annotator only.

184 of the annotated sentences contained at least one evaluative state, positive or negative. The overall number of evaluative states found in the data is 204. This means that only 17,6% of the sentences were evaluative.

The procedure using SubLex for identifying potential evaluative nodes highlighted 1091 candidate nodes, 754 positive and 337 negative. Strikingly, only 162 of the highlighted nodes, 79 positive and 83 negative, were confirmed as evaluative by the annotator. This means that eventually, the SubLex-based prediction does not give satisfying results. Also, the results suggest that the lexicon works far better for negative polarity clues (24,6% predicted successfully) than for positive clues (only 10,5% predicted successfully). We address the subjectivity lexicon limitations in the next section.

Apart from the SubLex-predicted nodes, the annotator assigned evaluation to 42 new nodes, i.e., 20,5% of all the confirmed evaluative nodes in the pilot treebank have not been recognized by the procedure.

7 ANNOTATION CHALLENGES

In this section, we address the common and widely known challenges to the sentiment classifying models and theories, and see how they manifest themselves in our treebank data annotation.

7.1 Subjectivity Lexicon Limitations

One of the underlying reasons for carrying out the annotation of sentiment in PDT was testing the feasibility of employing a subjectivity lexicon in automatic classification of structured data. While the Czech Sublex 1.0 has been originally created as a translation of the MPQA subjectivity lexicon [10], it includes lemmas falling within a much broader concept of subjectivity than the narrow concept of evaluativeness. Thus, words like *zdát se* ("to appear") or *skutečně* ("in fact"), which express (or just suggest) subjective attitude, but not specifically evaluation, appear superfluous to our purposes and are not annotated in the data.

The evaluation has been proved to be context sensitive in many cases, which makes the automatic identification of evaluative expressions even more difficult. Thus, the word *kladný* ("positive"), which comes from the subjectivity lexicon as inherently evaluative, loses its evaluative power in economic contexts (1), and other, non-evaluative words gain evaluative power in domain specific contexts (2), or when modified by an intensifier (3).

(1) I když letos a příští rok je nutné počítat se zpomalením růstu vývozu a zrychlením růstu dovozu, prognózujeme, že saldo přesto zůstane *kladné* ve výši 300 - 600 mil. USD ročně (1 - [1,6]1.6 % HDP).
'Even though it is necessary to expect a slowdown in the growth of export and a speed-up in the growth of import, we predict that the balance will remain positive*, $300-600 million a year.'*

(2) Má snad mobil nějaká negativa? Ano, má. *Nepodporuje* „české“ LTE.
‘Does the cell-phone have any negatives? Yes, it does. It does not support *“Czech” LTE.’*

(3) *Mimořádný výkon* podal Aleš Velc, který běžel druhý závod.
‘Aleš Velc, who ran the second race, exhibited an outstanding performance.’

Unfortunately, the effects of general context on the evaluative meaning of individual words is almost as hard to be solved in treebank data, as it is in plaintext data.

7.2 Negation and Other Polarity Reversing Items

Lexical negation is usually treated as a separate grammatical node in PDT 2.0. Thus, words like *nepříznivý* (“unfavourable”) are lemmatized as positives (*příznivý,* “favourable”) and a separate “Neg” node is added as a dependant. Since the subjectivity lexicon stores negated lemmata as separate entries, this complicates the automatized matching of lexicon entries to the data. The current system matching lexicon entries to the data nodes and assigning polarity to them does not take into account polarity reversing effects of certain dependent nodes yet, therefore the automatic polarity orientation prediction usually fails with negated nodes.[1]

Apart from negation, there are other words with polarity reversing (or neutralizing) effects in the data – verbs (*znemožnit,* ‘prevent*’*), prepositions (*bez,* ‘without*’*), adverbs (*nedostatečně,* ‘insufficiently*’, příliš,* ‘excessively*’*). Such expressions can be stored in the form of lists, or small lexicons of polarity reversing items and (together with a set of rules for negation effects) can be employed in the system.

7.3 Bad News/Good News (BGN)

Most sentiment lexicons and methodologies up to date do not discriminate evaluation from bad news/good new items properly. This is an important issue, because on one hand, BGN items in a way influence our subjective evaluative judgment of a text, on the other hand they often appear in informative, non-evaluative contexts.

The definiton of BGN was suggested in [9]. The main difference between evaluation proper and BGN lies in the fact that there is no target in case of BGN, or, more likely, the BGN items incorporate the evaluative expression and the target of evaluation both in a single word or phrase.

As the treebank data suggest, the most truth-like model will be the one showing the transition from evaluation to BGN as a scale, with unclear borderlines, since the evaluative power of BGN activates in domain specific contexts (4).

(4) *Inovovali* jsme také receptury pracích prášků, *zvýšili* podíl *účinných* látek a parfémů.

[1] The same problem was experienced the other way round in the SubLex creation process. Since Sublex was translated via bilingual treebank data, wrong polarity was often assigned in contexts where negation was employed on one side of the translation, but not on the other side. These cases were then manually corrected.

'We innovated *also the detergent formula, we* increased *the proportion of active ingredients and perfume.'*

BGN as a phenomenon tends to follow some basic tendencies noted already in cognitive linguistics studies – we praise what is big, high, nice and healthy and we defame the opposite. The most clear example are thus the words of rising and falling (5).

(5) Ekonomika jde do vzestupu už letos.
'The economics already rises this year.'

7.4 Comparisons, Graded Sentiments

So far, the annotation scheme is only able to capture absolute polarity values. It is not designed to work with relative evaluation, which is represented linguistically, e. g., by comparison sentences, see (6).

(6) Vláda kompetence celků považuje za důležitější než jejich množství a vymezení.
'The government considers the competences of the units more important than their number or delimitation.'

There are two important issues connected to the treatment of comparisons in PDT data. First, the comparative degree *důležitější* (*'*more important*'*) is lemmatized as *důležitý* (*'*important*'*) in the treebank, and second, the second part of the comparative structure, usually elided, is represented fully in the tectogrammatic structure. The comparative word, which is usually evaluative, is thus copied in the structure (7), and therefore identified also as bearing polarity.

(7) Vláda kompetence celků považuje za důležitější než [považuje za důležité] jejich množství a vymezení.
'The government considers the competences of the units more important than [it considers important] their number or delimitation.'

Nevertheless, the current scheme does not take into account any scale representation of polarity strength. Therefore, the treatment of comparisons is quite difficult and fully dependent on human annotator judgement.

7.5 Metaphors

One of the almost irresolvable issues in evaluative state identification tasks is the identification of sentiment in metaphors. (8,9)

(8) Ve srovnání s vládní bitvou o počet celků z konce června *byla* tato jednání *naprostou selankou*.
'Compared to the government battle over the number of units at the end of June, these negotiations were a piece of pie*'*

(9) Například naše zubní pasty obsadily dominantní podíl 55 procent, *čímž se nemůže pochlubit ani žádná světová firma*.
'For example, our toothpastes took a dominant share of 55%, which is something that no international company can boast of*'*

Since the meaning of metaphors is not derived compositionally, the treebank does not help with this task any way, nor it is easy to incorporate metaphorical expressions in the lexicon of subjective expressions due to their low frequency in language.

7.6 Complex Phrases

The annotation scheme, as it is designed, ties the evaluative state to the evaluative expression matching an entry in the lexicon. From this perspective, it becomes paradoxically difficult to treat syntactically complex expressions of evaluation, as in (10). Without further improvement of the scheme, the system is not able to capture the impact of "sentiment evoking" verbs, like *považovat* (*'*consider*'*).

(10) TTI Therm *považuje* tyto návštěvy *za nejlepší způsob* dalšího zvyšování odbytu.
'TTI Therm considers *these visits* the best way *to increase their sales.'*

8 CONCLUSION

We have described our efforts to annotate an existing dependency treebank with information about evaluative language. The annotations of structured data bring much light into the area of evaluative language patterns, but the currently used scheme must be further developed in order to be able to account for more complex phenomena.

1) To account for the effects of intensifiers, negation and other polarity reversing items we suggest creation of lists of polarity reversing and shifting items. Also, adding some kind of evaluation strength attribute would be beneficial.
2) The scheme should be enriched with additional attributes to account for the evaluative power of whole phrases and complex expressions, possibly also for some cases of BGN in the data.
3) It is probably not necessary to try to account for complex metaphorical expressions of evaluation.

ACKNOWLEDGEMENTS

This work has been supported by grant No. GA15-06894S of the Czech Science Foundation (GA ČR). This work has been using language resources developed and/or stored and/or distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth, and Sports of the Czech Republic (project LM2015071).

References

[1] Bojar, O. and Žabokrtský, Z. (2006). CzEng: Czech-English Parallel Corpus release version 0.5. *Prague Bulletin of Mathematical Linguistics*, 86:59–62.

[2] Choi, Y. and Cardie, C. (2008). Learning with compositional semantics as structural inference for subsentential sentiment analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 793–801, Association for Computational Linguistics, Honolulu, Hawaii.

[3] Hajič, J. (2005). Complex corpus annotation: The Prague dependency treebank. *Insight into Slovak and Czech Corpus Linguistics*, pages 54–73, Veda, Bratislava.

[4] Moilanen, K. and Pulman, S. (2007). Sentiment composition. In: *Proceedings of RANLP*, pages 378–382, Borovets, Bulgaria.

[5] Nakagawa, T., Kentaro, I., and Kurohashi, S. (2010). Dependency tree-based sentiment classification using CRFs with hidden variables. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Association for Computational Linguistics, Los Angeles, California, USA.

[6] Rentoumi, V., Petrakis, S., Klenner, M., Vouros, G. A., and Karkaletsis, V. (2010). United we Stand: Improving Sentiment Analysis by Joining Machine Learning and Rule Based Methods. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*, pages 1089–1094, Valletta, Malta.

[7] Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., and Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, Vol. 1631, pages 1642–1653.

[8] Šindlerová, J., Veselovská, K. and Hajič, J. jr. (2014). Tracing Sentiments: Syntactic and Semantic Features in a Subjectivity Lexicon. In *Proceedings of the 16th EURALEX International Congress*, pages 405–413, Bolzano/Bozen, Italy.

[9] Veselovská, K., Hajič, J. jr., and Šindlerová, J. (2012). Creating annotated resources for polarity classification in Czech. In *Proceedings of KONVENS 2012 (PATHOS 2012 Workshop)*, pages 296–304.

[10] Wilson, T., Wiebe, J., and Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing,* pages 347–354, Association for Computational Linguistics, Vancouver, British Columbia, Canada.

TEDXSK AND JUMPSK: A NEW SLOVAK SPEECH RECOGNITION DEDICATED CORPUS

JÁN STAŠ – DANIEL HLÁDEK – PETER VISZLAY – TOMÁŠ KOCTÚR
Faculty of Electrical Engineering and Informatics,
Technical University of Košice, Slovakia

STAŠ, Ján – HLÁDEK, Daniel – VISZLAY, Peter – KOCTÚR, Tomáš: TEDxSK and JumpSK: A New Slovak Speech Recognition Dedicated Corpus. Journal of Linguistics, 2017, Vol. 68, No 2, pp. 346 – 354.

Abstract: This paper describes a new Slovak speech recognition dedicated corpus built from TEDx talks and Jump Slovakia lectures. The proposed speech database consists of 220 talks and lectures in total duration of about 58 hours. Annotated speech database was generated automatically in an unsupervised manner by using acoustic speech segmentation based on principal component analysis and automatic speech transcription using two complementary speech recognition systems. The evaluation data consisting of 50 manually annotated talks and lectures in total duration of about 12 hours, has been created for evaluation of the quality of Slovak speech recognition. By unsupervised automatic annotation of TEDx talks and Jump Slovakia lectures we have obtained 21.26% of new speech segments with approximately 9.44% word error rate, suitable for retraining or adaptation of acoustic models trained beforehand.

Keywords: automatic annotation, speech recognition, speech corpus

1 INTRODUCTION

The development of more advanced and more precise large vocabulary continuous speech recognition (LVCSR) system requires huge amount of data for estimation of statistical parameters of acoustic and language models to cover the most possible real situations that usually occur in spontaneous speech. The complexity of speech recognition is mainly influenced by the speaker characteristics and speaking style. Robust acoustic models require phonetically rich and gender-balanced speech corpora that contain from hundreds to thousands of hours of annotated speech recordings.

Manual speech transcription and annotation of such amount of data requires much time and effort, as well as considerable amount of funds. Conventional transcription and annotation methodology requires training of the professional annotators on transcription guidelines, which lasts from a few hours to several weeks. Typical manual transcription speeds of spontaneous or conversational speech lasts around 7 to 12 times real-time, due to its complexity. The transcription and annotation of non-native speech is an even more difficult, slow and laborious process [1].

If even a few hours of manually annotated speech utterances is available, then it is possible to develop, using the latest approaches, principles and methods, a comprehensive speech transcription system for automatic annotation and creation of a new speech database that can be used for re-estimating selected parameters of an existing acoustic models or their adaptation to the characteristics of the given speaker.

DOI 10.1515/jazcas-2017-0044

Publicly available online spoken language resources are preferable because there are problems with obtaining licence agreement from source data providers. One such resource is the database of TED talks (Technology, Entertainment, Design) that promote "*ideas worth sharing*". They become a good online available resource for creation of speech databases for the number of languages, which are under-resourced because of their thematic, stylistic and natural richness.

One of the best known and widely used automatic speech recognition dedicated corpus is TED-LIUM [2] that consists of 1495 automatically annotated English talks. The initial speech recognition was performed using five-pass ASR system based on the open-source CMU Sphinx framework [3] with acoustic and language model adaptation, speaker adaptive training, re-computing the linguistic scores from updated word-graphs with 4-gram language model and algorithm for hypothesis re-scoring at different stages. The official speech recognition results discussed at the IWSLT 2011 evaluation campaign reached 17.40% word error rate (WER) in average.

Otherwise, one of recently proposed and automatically annotated spoken language resource built from TED talks is the SI TEDx-UM speech database [4]. It contains 242 talks in the Slovenian language in total duration of about 54 hours. The efficiency of unsupervised transcription was evaluated using the UMB Broadcast News speech recognition system and reached 50.70% WER in average.

Several previous works and research activities have been reported on enhancing the efficiency of automatic speech transcription and unsupervised annotation of speech corpora based on TED talks by using robust acoustic and language modeling [5-8].

A number of algorithms have been proposed for acoustic model adaptation in automatic transcription of TED talks based on discriminative training criteria, maximum a posteriori (MAP) estimation, maximum likelihood linear regression (MLLR), feature space adaptation (FSA), vocal tract length normalization (VTLN) [9], speaker adaptive training (SAT) [10], or statistical modeling with deep neural networks (DNN) [11].

Moreover, the problem of frequently appearing errors in automatic transcription of lecture speech was eliminated in [12] by correction of colloquial expressions, deletion of fillers and insertion of periods using statistical post-processing techniques. Authors in [13] and [14] explore output recognition hypotheses and effectiveness of supervised and unsupervised adaptation with varying amounts of user-provided transcripts to tune the language model parameters on a lecture transcription task in English.

This paper presents a new spoken language resource built from Slovak TEDx talks and Jump Slovakia lectures annotated automatically in an unsupervised manner using two complementary LVCSR systems with using filtration of output hypotheses with minimal amount of errors. The reason for building the corpus lies in the fact that modern and leading trends in building resources for LVCSR applications focus on fully-automatic annotation of speech, without any additional human effort [2][4][10][15]. The database of speech recordings and their transcripts will be publicly available from the end of june 2017 at the web page of our laboratory[1].

[1] `http://nlp.web.tuke.sk/pages/tedx`

2 CONTENT AND STRUCTURE OF THE SPEECH CORPUS

The goal of this research is to build a new automatically annotated speech database with the best possible quality and one or at least two speakers per talk. Source data consisting of Slovak TEDx talks[2] and Jump Slovakia[3] lectures were gathered from official YouTube channels. Foreign-language lectures and low-quality speech recordings were removed from the list of about 300 Slovak audiovisual recordings obtained from ten different events, publicly available between years 2010 and 2016.

event	number of lectures	number of speakers	males	females	duration	for male gender	for female gender
TEDx Bratislava	57	61	42	19	13:03:55	09:02:35	04:01:20
TEDx Kežmarok	9	10	6	4	02:48:06	01:59:18	00:48:48
TEDx Košice	30	30	24	6	08:50:03	07:24:35	01:25:28
TEDx Nitra	14	14	12	2	04:13:37	03:33:07	00:40:30
TEDx Prešov	17	17	11	6	05:57:31	04:07:32	01:49:59
TEDx Trenčín	24	25	14	11	05:50:43	03:36:40	02:14:03
TEDx Trnava	9	9	6	3	02:21:53	01:42:20	00:39:33
TEDxYouth Bratislava	20	20	15	5	05:36:39	04:06:05	01:30:24
TEDxYouth Žilina	6	6	4	2	01:41:34	01:06:59	00:34:35
Jump Slovensko	34	35	20	15	07:25:35	04:12:44	03:14:51
together	**220**	**227**	**154**	**73**	**57:51:36**	**40:51:55**	**16:59:41**

Tab. 1. Structure of the speech corpus of TEDx talks and Jump Slovakia lectures

2.1 Corpus Design

A total of 220 talks and lectures in the Slovak language were selected for automatic segmentation and unsupervised transcription using our proposed system architecture based on two complementary Slovak LVCSR systems. Audiovisual recordings downloaded from the official YouTube channels were encoded in H.264 video format. The captured audio stream was encoded in MPEG AAC format. Each recording has been converted into WAV audio and down-sampled to 16kHz/16bit PCM mono audio using SoX tool[4] to be compatible with acoustic models used in our LVCSR system.

2.2 Corpus Statistics

The presented speech corpus consists of about 58 hours, including silence and other malformed audio content. The useful part covers a total duration of about 55 hours. The speech corpus contains 227 unique speakers of both genders with 154 males and 73 females. Approximately 30% of the database is build of samples of female voices.

Manually annotated part of the speech corpus covers 50 randomly selected Slovak TEDx talks in total duration of about 12 hours. The speaking rate in this part

[2] `https://www.youtube.com/user/TEDxTalks`

[3] `https://www.youtube.com/user/jumpslovensko`

[4] `http://sox.sourceforge.net/`

varies from 115.53 up to 256.32 words per minute (*wpm*). The average rate of out-of-vocabulary (OOV) words is 3.23% and the average language model perplexity is 508.40. The detailed description about the number of lectures, number of speakers and total duration of speech in the presented corpus of 220 Slovak TEDx talks and Jump Slovakia lectures is summarized in the Table 1.

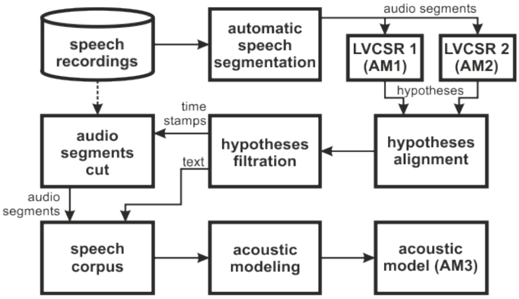


Fig. 1. Automatic segmentation and transcription of speech recordings using two complementary large vocabulary continuous speech recognition systems in the Slovak language

2.3 Characteristics of the Speech in the Corpus

During about 15-minute talk (lecture), speakers are often non-native, have a strong accent, and sometimes, are not fluent. Despite the fact that speaking style of a speaker being in general planned, spontaneous speech occurs more frequently. There are some differences among speakers in their grammar, articulation or speaking style with frequent errors (e.g. filled pauses, sentence restarts, phrase modifiers, repetitions, or mistakes). Although speech recordings are usually realized with close-talk, lapel or goose-neck microphones, the signal often contains some noise from the auditorium and from the speaker itself. Therefore, lecture speech transcription is a difficult task, both from the acoustic and linguistic point of view, due to the many hesitation fillers that occur in spontaneous speech, different and varying speaking rates, mixed topics and speaking style, or combining colloquial expressions with formal jargon [7]. Furthermore, a new speech corpus covers current events and hot topics in Slovakia, which is suitable ground for domain-based modeling and text summarization tasks.

3 AUTOMATIC SEGMENTATION AND ANNOTATION OF THE SPEECH CORPUS

The proposed approach for automatic segmentation and unsupervised speech transcription and annotation of the presented corpus of TEDx talks and Jump Slovakia lectures using two complementary LVCSR systems is depicted on Fig. 1. A brief review of the main building blocks of this system architecture will be described in the following sections.

3.1 Automatic Speech Segmentation

In general, it is possible to transcribe continuous audio stream without any segmentation, but the computation cost of the decoding may take a very long time.

Therefore, automatic speech segmentation is usually applied to speed up the speech recognition process and to improve the overall performance by identifying and handling the specific parts in the recognized speech (gender- and speaker-specific segments, speaker-change boundaries, different acoustic conditions, non-speech events, etc.).

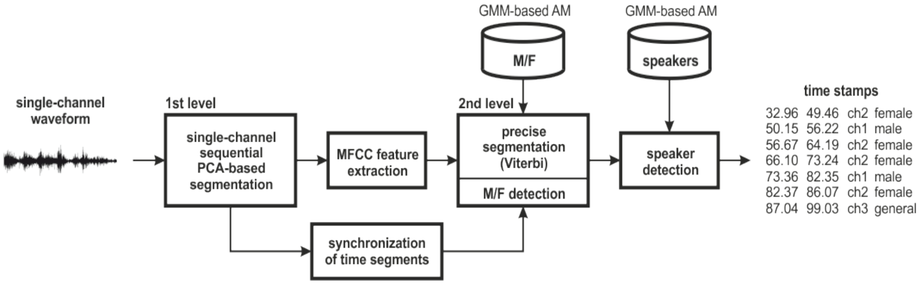


Fig. 2. Automatic speech segmentation

The proposed speech segmentation is able to process any kind of single-channel audio recordings (e.g. talks, lectures, discussions, broadcast news, etc.). The gender detection can be performed using the default gender-dependent acoustic models. The detection rate will be satisfactory, because the acoustic models were trained on a sufficient amount of acoustic representations for each gender.

The speaker-dependent segmentation is not supported implicitly, if the single-channel waveform contains voices of unknown speakers that were not included in the training data. On the other hand, there is a possibility to train a new speaker-dependent AM, if the recognized audio provides a sufficient amount of speaker examples.

The proposed system architecture employs two-level fully-automatic speech segmentation, depicted on Fig. 2.

At first level, the silence discrimination is performed by our proposed voice activity detection (VAD) algorithm [16]. In order to determine VAD labels, the waveform is processed in the time domain by overlapping blocks extracted by rectangular window with length of 25ms and 10ms frame step. After re-arranging samples into sample matrix, the time domain principal component analysis (PCA) is applied to each block. After that, *N* eigenvalues are computed for each block, where *N* is the dimension of the PCA space. The eigenvalues are used to determine the nature of the *i*-th segment (voice or silence). Finally, the VAD coefficients are smoothed by applying a sliding average window.

The second level uses only the speech active segments and it employs the Viterbi algorithm for precise gender- or speaker-dependent segmentation. In other words, gender- and speaker-dependent recognizers are run to detect and locate gender- and speaker-change points to enable these regions to be split into shorter segments. At this stage, time stamps are generated with gender labels and if needed, they can be extended with speaker labels. The overall segmentation requires a time synchronization between the first and second level due to eliminated silent parts at the first level.

3.2 Automatic Speech Transcription

For initial experiments with unsupervised transcription, annotation and acquisition of large speech databases we have created a new speech recognition system architecture based on the complementarity of two Slovak LVCSR systems (see Fig. 1, LVCSR 1 and LVCSR 2) [1], [17].

The Slovak LVCSR system uses an open-source recognition engine Julius [18] that was modified to support multi-threaded parallel speech recognition and sharing acoustic and language models among all instances for memory space saving purposes. For supporting the actual speech recognition with different configurations, the speech recognition server was created and is capable of parallel speech recognition supporting different configurations at the same time [19].

Complementarity of the LVCSR systems was achieved by using two acoustic models trained on the different training sets. The first acoustic model (AM 1) was trained on 320 hours of manually annotated speech recordings of judicial readings and parliament proceedings [20]. The second model (AM 2) was trained on a database of 330 hours of manually annotated speech recordings acquired from the main broadcast news [21] and Court TV shows with a high degree of spontaneity [1], [21]. Both acoustic models (AM 1 and AM 2) were generated from feature vectors with standard dimension of 39 mel-frequency cepstral coefficients, along with delta and acceleration coefficients and cepstral mean normalization enabled. The triphone context-dependent acoustic models are based on hidden Markov models (HMMs) with 32 Gaussian mixtures. The training sets also involve models of silence, short pause and additional noise events for filled pauses and prolongations. A typical tree-based state tying for HMMs has been replaced by the effective triphone mapping algorithm [22].

The proposed LVCSR system uses a trigram model of the Slovak language created by the SRILM Toolkit [23], restricted to the vocabulary size of about 500*k* unique words and smoothed with the Witten-Bell algorithm [24]. Language model has been trained on preprocessed web-based corpora of Slovak written texts of more than 2,150M tokens contained in 120M of sentences [25].

The process of speech recognition was enhanced using *N*-best output hypotheses rescoring with the ROVER algorithm [26], slightly modified to our needs to include confidence measure score context between words into consideration [27].

3.3 Filtration of Output Hypotheses

After transcription of segmented speech recordings, the output hypotheses from both complementary recognition systems (LVCSR 1 and LVCSR 2) are time aligned and compared. In the next step, overlapping transcribed speech segments obtained from aligned output hypotheses are filtered out. Proposed approach for filtration output hypotheses takes maximum time delay from the start and end of the speech segment, minimum number of equal words in aligned hypotheses and confidence measure score (CMS) value into account. Output of the process of filtration are short automatically annotated speech segments [19].

In this research, the parameters for filtration of output hypotheses were empirically set to 20ms maximum time delay from the start and end of each speech segment, minimum number of equal words in aligned hypotheses was set to 3 words, and the threshold value for confidence measure score varies from 0 to 0.75.

data set	actual duration	duration after segmentation	setting 1 ~ 13.57% WER	setting 2 ~ 9.44% WER	setting 3 ~ 4.94% WER
	amount of gathered data [hh:mm:ss]				
eval	12:26:07	11:50:37	05:39:30	02:47:35	00:39:43
dev	45:25:29	43:13:12	19:37:41	08:54:47	02:01:04
eval+dev	57:51:36	55:03:49	25:17:11	11:42:22	02:40:47
	amount of gathered data [%]				
eval		95.24	47.78	23.58	5.59
dev		95.15	45.41	20.62	4.67
eval + dev		95.17	45.92	21.26	4.87

Tab. 2. Amount of gathered data

4 EVALUATION

In the first step of creation of a new spoken language resource of Slovak TEDx talks and Jump Slovakia lectures we divided speech corpus into two parts – evaluation and development data set. The evaluation data, in total duration of 12 hours, was annotated manually on the word level by professional annotators. This set was used for evaluation of the transcription accuracy in various settings (see Table 2, setting 1 to 3). These settings of the systems were selected for the best quality of the transcription (setting 1) and for the biggest amount of annotated data (setting 3). Setting 2 is a trade-off between quality and quantity of automatic speech transcription.

Values obtained from the settings (1-3) were used for automatic transcription of the remaining (development) part of the corpus. The total amount of gained data after automatic transcription is summarized in Table 2.

The Table 2 shows that it is possible to get 45.92% of a new fully-automatic annotated speech segments from the total length and amount of speech recordings with approximately 13.57% WER. These automatically annotated speech segments can be directly used for re-estimation of the parameters of existing acoustic models or for their adaptation. Similarly, 21.26% of new fully-automatic annotated speech segments can be gained with approximately 9.44% WER and about 4.94% WER can bring only 4.87% of annotations from the total amount of 58 hours.

5 CONCLUSION

In this paper we introduced a new speech recognition dedicated corpus built from Slovak TEDx talks and Jump Slovakia lectures. We were motivated by the modern trends in corpora design based on fully-automatic annotation procedure to generate error-free transcripts. We expect that we will be able in the immediate future to move fully-automatic annotation of any kind of new data without the need for human annotation effort.

In the further research, we want to focus on eliminating common recognition errors by introducing unsupervised language model adaptation to the current topic and specific speaker speaking style and statistical modeling of most frequent hesitation fillers in spontaneous speech for improving system performance and automatic transcription and annotation of large acoustic corpora of the spoken Slovak language [28]. Also, we are planning to append the fully-annotated data from that corpus to the current training data in order to retrain the present acoustic and language models.

ACKNOWLEDGEMENTS

The research in this paper was supported by the Faculty of Electrical Engineering and Informatics at the Technical University of Košice under the project FEI-2015-30, by the Ministry of Education, Science, Research and Sport of the Slovak Republic under the research project VEGA 1/0511/17, and by the Slovak Research and Development Agency under the applied research project APVV-15-0517.

References

[1] Koctúr, T., Juhár, J., Viszlay, P., Staš, J., and Lojka, M. (2016). Unsupervised speech transcription and alignment based on two complementary ASR systems. In *Proceedings of RADIOELEKTRONIKA 2016*, pages 358–362, Košice, Slovakia.

[2] Rosseau, A., Deléglise, P., and Estève, Y. (2012). TED-LIUM: An automatic speech recognition dedicated corpus. In *Proceedings of LREC 2012*, pages 125–129, Istanbul, Turkey.

[3] Deléglise, P., Estève, Y., Meignier, S., and Merlin, T. (2009). Improvements to the LIUM French ASR system based on CMU Sphinx: What helps to significantly reduce the word error rate? In *Proceedings of INTERSPEECH 2009*, pages 2123–2126, Brighton, UK.

[4] Žgank, A., Maučec, M. S., Verdonik, D. (2016). The SI TEDx-UM speech database: A new Slovenian spoken language resource. In *Proceedings of LREC 2016*, pages 4670–4673, Portorož, Slovenia.

[5] Rosseau, A., Deléglise, P., and Estève, Y. (2014). Enhancing the TED-LIUM corpus with selected data for language modeling and more TED talks. In *Proceedings of LREC 2014*, pages 3935–3939, Reykjavik, Iceland.

[6] Leeuwis, E., Federico, M., and Cettolo, M. (2003). Language modeling and transcription of the TED corpus lectures. In *Proceedings of ICASSP 2003*, pages 232–235, Hong Kong, China.

[7] Cettolo, M., Brugnara, F. and Federico, M. (2004). Advances in the automatic transcription of lectures. In *Proceedings of ICASSP 2004*, pages 769–772, Montreal, Canada.

[8] Niesler, T. and Willet, D. (2002). Unsupervised language model adaptation for lecture speech transcription. In *Proceedings of ICSLP 2002*, pages 1413–1416, Denver, Colorado, USA.

[9] Wölfel, M. and Berger, S. (2005). *The ISL baseline lecture transcription system for the TED corpus*. Tech. Rep., Karlsruhe University, Germany.

[10] Naptali, W. and Kawahara, T. (2012). Automatic transcription of TED talks. In *Proceedings of the 6th Spoken Document Processing Workshop, SDPWS 2012*, Toyohashi, Japan.

[11] Bell, P., Yamamoto, H., Swietojanski, P., Wu, Y., McInnes, F., Hori, Ch., and Renals, S. (2013). A lecture transcription system combining neural network acoustic and language models. In *Proceedings of INTERSPEECH 2013*, pages 3081–3091, Lyon, France.

[12] Nanjo, H., Shitaoka, K., and Kawahara, T. (2003). Automatic transformation of lecture transcription into document style using statistical framework. In *Proceedings of ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition, SSPR 2003*, Tokyo, Japan.

[13] Hsu, B.-J. and Glass, J. (2009). Language model parameter estimation using user transcriptions. In *Proceedings of ICASSP 2009*, pages 4805–4808, Taipei, Taiwan.

[14] Akita, Y., Watanabe, M., and Kawahara, T. (2012). Automatic transcription of lecture speech using language model based on speaking-style transformation of proceedings texts. In *Proceedings of INTERSPEECH 2012*, pages 2326–2329, Portland, Oregon, USA.

[15] Viszlay, P., Staš, J., Koctúr, T., Lojka, M., and Juhár, J. (2016). An extension of the Slovak broadcast news corpus based on semi-automatic annotation. In *Proceedings of LREC 2016*, pages 4684–4687, Portorož, Slovenia.

[16] Vavrek, J., Viszlay, P., Kiktová, E., Lojka, M., Juhár, J., and Čižmár, A. (2014). Query-by-example retrieval via fast sequential dynamic time warping algorithm. In *Proceedings of the 37th International Conference on Telecommunications and Signal Processing, TSP 2014*, pages 453–457, Berlin, Germany.

[17] Staš, J., Viszlay, P., Lojka, M., Koctúr, T., Hládek, D., Kiktová, E., Pleva, M., and Juhár, J. (2015). Automatic subtitling system for transcription, archiving and indexing of Slovak audiovisual recordings. In *Proceedings of the 7th Language & Technology Conference, LTC 2015*, pages 186–191, Poznań, Poland.

[18] Lee, A., Kawahara, T., and Shikano, K. (2001). Julius – An open source real-time large vocabulary recognition engine. In *Proceedings of EUROSPEECH 2001*, pages 1691–1694, Aalborg, Denmark.

[19] Lojka, M., Ondáš, S., Pleva, M., and Juhár, J. (2014). Multi-threaded parallel speech recognition for mobile applications. *Journal of Electrical and Electronics Engineering*, 7(1):81–86.

[20] Rusko, M., Juhár, J., Trnka, M., Staš, J., Darjaa, S., Hládek, D., Sabo, R., Pleva, M., Ritomský, M., and Ondáš, S. (2016). Advances in the Slovak judicial domain dictation system. In Vertulani, Z., Uszkoreit, H., and Kubis, M., editors, *Human Language Technology: Challenges for Computer Science and Linguistics*, LNAI 9561, pages 55–67, Springer International Publishing Switzerland.

[21] Koctúr, T., Staš, J., and Juhár, J. (2016). Unsupervised acoustic corpora building based on variable confidence measure thresholding. In *Proceedings of the 58th International Symposium ELMAR 2016*, pages 31–34, Zadar, Croatia.

[22] Darjaa, S., Cerňak, M., Trnka, M., and Rusko, M. (2011). Effective triphone mapping for acoustic modeling in speech recognition. In *Proceedings of INTERSPEECH 2011*, pages 1717–1720, Florence, Italy.

[23] Stolcke, A. (2002). SRILM – An extensible language modeling toolkit. In *Proceedings of ICSLP 2002*, pages 901–904, Denver, Colorado, USA.

[24] Staš, J. and Juhár, J. (2015). Modeling of the Slovak language for broadcast news transcription. *Journal of Electrical and Electronics Engineering*, 8(2):43–46.

[25] Hládek, D., Ondáš, S., and Staš, J. (2014). Online natural language processing of the Slovak language. In *Proceedings of the 5th IEEE International Conference on Cognitive InfoCommunications, CogInfoCom 2014*, pages 315–316, Vietri sul Mare, Italy.

[26] Fiscus, J. G. (1997). A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER). In *Proceedings of ASRU 1997*, pages 347–352, Santa Barbara, CA, USA.

[27] Lojka, M. and Juhár, J. (2014). Hypothesis combination for Slovak dictation speech recognition. In *Proceedings of the 56th International Symposium ELMAR 2014*, pages 43–46, Zadar, Croatia.

[28] Staš, J., Hládek, D, and Juhár, J. (2016). Adding filled pauses and disfluent events into language models for speech recognition. In *Proceedings of the 7th IEEE International Conference on Cognitive InfoCommunications*, *CogInfoCom 2016*, Wroclaw, Poland.

HELPING THE TRANSLATOR CHOOSE: THE CONCEPT OF A DICTIONARY OF EQUIVALENTS

WERONIKA SZEMIŃSKA
Institute of Specialist and Intercultural Communication, University of Warsaw, Poland

SZEMIŃSKA, Weronika: Helping the Translator Choose: The Concept of a Dictionary of Equivalents. Journal of Linguistics, 2017, Vol. 68, No 2, pp. 355 – 363.

Abstract: The purpose of the article is to present the innovative concept of a dictionary of equivalents, a reference work designed specifically for translators of legal texts. The article describes the features of legal terminology which render legal translation particularly difficult, such as polysemy and synonymy as well as incongruence among legal systems. Then it proposes a classification and labelling system of equivalents which ought to be offered in a terminographic reference work for legal translators.

Keywords: dictionary of equivalents, legal language, legal translation, congruence, equivalence

1 INTRODUCTION

"As translators we all know deep in our hearts that legal translation is impossible. The very expression *legal translation* seems to be a contradiction in terms (…)." This statement by P. Chaffey [1, p. 69] may sound radical, but also rather insightful. What makes legal texts untranslatable? Unlike other disciplines, law is a national phenomenon. Each national legal system is independent and has its own terminological apparatus, conceptual structure, rules of classification, sources of law, methodological approaches and socioeconomic principles [2, p. 13]. Moreover, since law defines reality, we may venture to say that each society lives in its own legal universe, which develops throughout centuries and reflects a people's history and culture [3, pp. 23–24].

This disconcerting nature of law predictably finds its reflection in the legal language, or rather in legal languages. As a result of the culture-bound nature of law, there is no single universal legal language, with rare exceptions. Only few areas, such as public international law or European law, have an international character and therefore use international legal terminology. Apart from that, some common ground may be found in branches of law such as constitutional, administrative, civil or criminal law [4, p. 122].

P. Chaffey's statement quoted at the beginning ends almost resignedly: "(…) and yet we do have to translate legislation and legal documents." If so, how can this impossible task be completed? And how can terminography aid the translator in their job? The aim of the present article is to propose a tool that would greatly facilitate the translation process, and in particular one of its stages, namely the phase of transfer from the source into the target language, when equivalents are chosen. First, the legal lexicon will be characterised, followed by a description of how its features af-

DE GRUYTER OPEN
DOI 10.1515/jazcas-2017-0045

fect the process of translation. Finally, the concept of the dictionary of equivalents will be presented.

2 THE LEGAL LEXICON

The potentially most interesting feature of legal language is the legal lexicon. The vocabulary may be classified according to various criteria, for instance the type of text in which it is used, as proposed by G.-R. de Groot [5, pp. 18–19]. Thus we can divide legal vocabulary into (i) that used in statutes and other regulations (with the subcategory of expressions explicitly defined by the legislator), (ii) that used by lawyers or in commentaries, and (iii) that used in general publications concerning the legal system in question. Another classification, by Gizbert-Studnicki [6, pp. 44–45], refers to the layers of legal vocabulary: (i) vocabulary related to the object of regulation, i.e. the area of social practice which a given regulation concerns, (ii) vocabulary related to the method of regulation, i.e. a large set of characteristically legal expressions which recur frequently in a given branch of law, and (iii) vocabulary being a consequence of the normative character of legal texts, i.e. a set of frequently used expressions referring to the notions related to obligations, claims, rights, competences etc., which are common to the entire legal system in question. Similarly, M. Chromá [7, p. 15] lists: (i) 'pure' law terminology, which is a scarce group of expressions that are not used in other contexts, (ii) law terminology found in everyday speech, including expressions with related general and legal meanings, with differing general and legal meanings and with several meanings, and (iii) everyday words assigned a special connotation in a given legal context, i.e. expressions whose meaning is either expanded or narrowed by means of a legal definition.

The fact that everyday words may be assigned a precise and often differing meaning in the legal context evidences the autonomy of the legal language in relation to the general language. It stems from the legislator's right to define the expressions used by means of legal definitions [8, pp. 9–10]. This may lead to certain complications. It happens namely that the legislator is inconsistent in the use of a term or that they use the same expression derived from the general language now in the legal, now in the general sense. Furthermore, one term may carry differing meanings in various branches of a single legal system, as for instance the English term 'charge', which can refer to a formal accusation of a crime in criminal law, to a legal interest securing the payment of money in contract law, and to a financial punishment in administrative law. Another cause of terminological polysemy is the situation when various legal systems are expressed by means of one ethnolect, as in the case of i.a. English, German, or French.

H. E. S. Mattila [9, pp. 109–112] divides polysemy in legal language into two categories: consistent and inconsistent. Consistent polysemy characterises a term that has several closely related meanings, often hierarchical or overlapping. Inconsistent polysemy occurs when the meanings diverge to such an extent that they no longer have anything in common. Finally, legal language often involves synonymy and, in particular, partial synonymy. Often all partial synonyms are used in a text in order to guarantee that the entire semantic field of a concept is covered.

3 LEGAL TRANSLATION

The differences between law and other specialist disciplines mean that legal translation poses difficulties unknown to translators of other types of texts. The main problems have their source in the nature of law and legal language. Let us say it again: in contrast to many other areas of knowledge and professional activity, law is not an independent being described by humans, but a human construct which is entirely dependent on language. Moreover, it is not a single construct, but exists in as many variants as there are legal systems and ethnolects that express them. Since no single legal reality is given, translation of legal texts entails translating not only between two ethnolects, but also between differing realities – legal systems. It may be depicted in the form of the following figure:

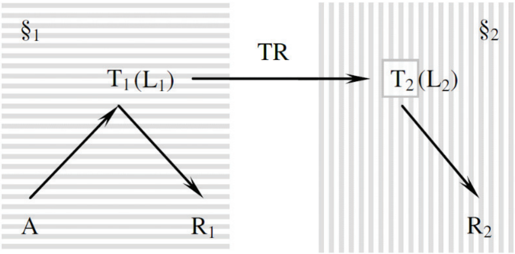


Fig. 1. Model of legal translation

The author (A) produces the source text (T_1) in the source language (L_1) directed at the primary recipient (R_1), all of whom function in the primary legal system ($§_1$). The translator (TR) produces the target text (T_2) in the target language (L_2), the text being directed at the target recipient (R_2). While the target language and the target recipient are immersed in the secondary legal system ($§_2$), the same cannot be said about the target text itself. A translation, albeit expressed in the target language, may have a different legal status in the secondary legal system than the original had in the primary legal system [3, p. 10], [10, pp. 198–199]. This phenomenon is related to the fact that, apart from rare exceptions mentioned previously, the object of the text does not exist in the secondary legal system. A more or less similar object may often be found, yet practically never an identical one. As a consequence, the target language often does not offer any expressions equivalent to the source ones.

This phenomenon is by some scholars [2], [3], [11], [12] referred to as incongruence. Šarčević [2, pp. 232–233] lists six major reasons for this phenomenon in legal terminology: (i) boundaries between the meanings of concepts in different legal systems are incongruent, (ii) the same term designates different concepts in different legal systems expressed by one language, (iii) legal concepts which were transplanted into another legal system have been gradually assimilated and altered their meaning, (iv) a number of terms in each legal system are strictly system-bound and have no comparable counterparts in other systems, (v) indefinite or vague terms are interpreted differently by courts in various jurisdictions, and (vi) some terms with ideological content have different connotations in various cultures. In fact, full congruence between legal terms occurs extremely rarely; de Groot [4, p. 124] even maintains that it is possible solely in the case of one legal system expressed by two languages. A large number of concepts may be described as convergent, i.e. partly equivalent.

In situations when no congruent or convergent equivalent exists or when it would be unacceptable for other reasons (e.g. as misleading to the recipient), compensation techniques may be applied, examples being: borrowing, possibly accompanied by a literal translation; periphrasis; neologism (based on legal language, other specialist language or general language); neutral (non-technical) term; literal equivalent; or Latin equivalent [4, p. 125], [2, p. 250]. In addition, the choice of translation techniques must be consistent throughout the text and depends not only on the individual translation problem, but predominantly on factors like the purpose of the translation, i.e. its prospective function in the target reality, as well as the knowledge of the recipient, the genre of the text and the relation of the two relevant languages and legal systems. Based on the analysis of these aspects, the translator has to pick the most appropriate translation strategy, which in turn determines the set of techniques which may be applied to handle individual translation problems.

Principally we may speak of two translation strategies: (i) source language and primary legal system-oriented strategy and (ii) target language and secondary legal system-oriented strategy [13, p. 145], which are parallel to the literary translation concepts of foreignisation and domestication. The L_1 and $§_1$-oriented strategy involves emphasising the differences between the legal systems by means of using expressions and structures strange to L_2 and $§_2$. The L_2 and $§_2$-oriented strategy, on the other hand, aims to blur said differences [14, p. 36]. A simple example illustrating both types of equivalence is the translation of the Polish expression ‘akt oskarżenia’ into American English. The American law knows a similar concept expressed by the term ‘indictment.’ Thus the original term could be translated as ‘indictment’ if we want to stress the similarities between the Polish and the American law, and for instance as ‘act of indictment’ if we want to stress the disparities. In the latter case, the American term is modified by introducing a foreign element which refers to the Polish legal system. The equivalent is still comprehensible to an American reader. A more literal translation of the term (‘act of accusation’) would go even further to foreignise the text.

As may be inferred from the above considerations, choosing a right equivalent in legal translation is a most arduous task. It requires analysing the original concept, studying the secondary legal system in search for similar concepts, and if any are found, comparing them to the source concept to determine the degree of equivalence/congruence, deciding whether they are acceptable translations, and if not, applying a fitting compensation technique. This in turn involves researching other legal systems expressed by the same language, if any exist, including Roman law, inspecting the potential connotations of a literal equivalent or an equivalent based on another specialist language or on the general language, and the list goes on. Plus, the final choice must be consistent with the translation strategy assumed. Completing this complex task requires much more than a perfect command of the source and the target language, namely deep legal expertise and ability to perform comparative legal analysis. Indeed few translators have all those competences and sufficient time while working on a translation. That is why they (rather logically) expect help from legal dictionaries. These are, however, often no more than word lists offering unsubstantiated translations without indications of differences in meaning between the source and the target language [15, p. 2], [16]. Labelling systems used in some dictionaries usually indicate the branch of law or the legal system from which a given equivalent stems, but no more than that. If any neologisms are offered, they are not marked as such. Similarly, there are no indications of translation strategies which a particular solution follows, i.e. whether it is a $§_1$ or $§_2$-oriented equivalent. Finally, the overall number of equivalents is often insufficient.

4 DICTIONARY OF EQUIVALENTS

The above considerations seem to suggest that the answer to the translator's problem is, firstly, to provide a maximum number of possible equivalents of various types, secondly, to indicate the differences between them, and thirdly, to explain any incongruencies between equivalents stemming from the secondary legal system or another legal system expressed by the same ethnolect and the original term. Only then will the translator be able to make a well-informed choice of the most fitting equivalent in the given context.

Designing a dictionary that would aid the translator in choosing the best equivalent should therefore commence with preparing a classification of equivalents. First and foremost, they may be divided into two categories: actual equivalents and neologisms. Actual equivalents are terms that refer to concepts existing in the secondary legal system. Neologisms, on the other hand, are not necessarily recent or isolated expressions outside mainstream language, but any expressions which do not refer to concepts in the secondary legal system. Thus, both an expression from another legal system expressed by the same ethnolect as $§_1$ and a literal translation of the original term may be seen as neologisms.

Actual equivalents may be further divided into congruent and convergent ones. The former are expressions representing $§_2$ concepts that share most vital characteristics with the $§_1$ concept in question (as has been said above, full congruence is virtually impossible). The latter in turn represent $§_2$ concepts which only partially correspond to the original concept. Each of these two types may be further subdivided into valid and outdated equivalents. While valid equivalents reflect concepts that are still in use under effective law, outdated equivalents represent notions that have become obsolete, but are still recognisable by target language recipients (an example could be the obsolete English term 'custody' referring to the rights and responsibilities of parents with respect to their child).

Neologisms in turn may be classified according to the translation technique applied or to the strategy they follow. From the point of view of translation techniques, we can discern neologisms stemming from other legal systems than $§_2$ (which again may be congruent or convergent, valid or outdated), including borrowings, neologisms from the language for general purposes, neologisms from another language for specialist purposes, literal translations, periphrases and institutional neologisms (i.e. expressions used by a particular institution, such as a state agency or a company).

The aspect of translation strategy allows us to differentiate between $\S_1$-oriented and $\S_2$-oriented neologisms.

The above classification may be presented in the form of the below figure, including the proposed labels for each class (the sign ‘x’ in the label for terms from other legal systems stands for an abbreviation of the name of the country whose legal system is involved, so in its final version the label could look like this: $N_{\S US\equiv}$, if the legal system in question were American law):

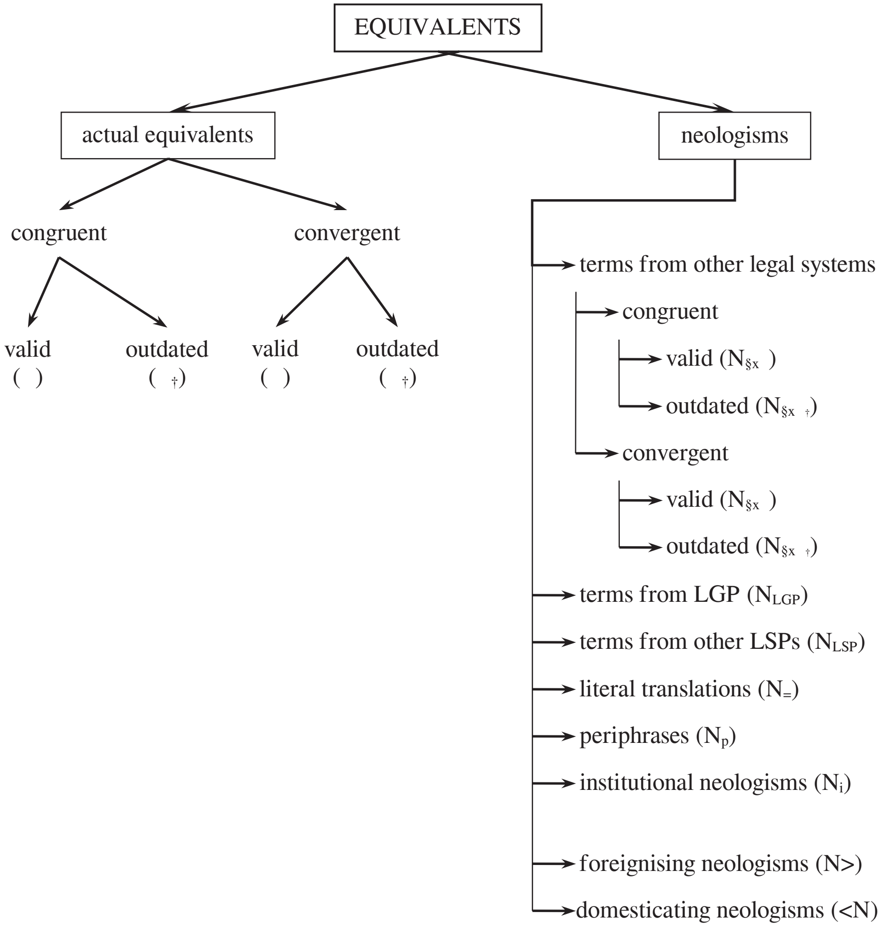
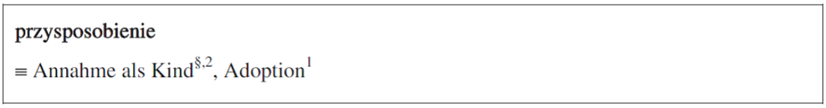


Fig. 2. Classification of equivalents and their labels

If an expression has several synonymous equivalents within the same class, two further labels may be introduced which will allow the translator to choose the right one: frequency and presence in normative acts. The fact that some expressions are used more often than others hardly requires an explanation. The second marker is based on the fact that some legal terms do appear in prescriptive texts, such as statutes, while others are used only in secondary legal texts, for instance commentaries or academic publications. An example could be the German terms for adoption: ‘Adoption’ and ‘Annahme als Kind,’ with the former being much more frequently used, but not present in the German Civil Code, which uses the latter expression. The proposed labels could be uppercase ‘§’ for the prescribed term and likewise uppercase numerals ‘1,’ ‘2’ etc. to indicate frequency.

Sample entries in a dictionary of equivalents could look as shown below:

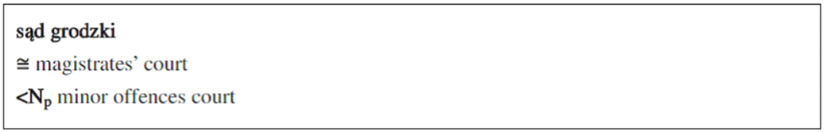


przysposobienie

≡ Annahme als Kind^{§,2}, Adoption^{1}

Fig. 3. Sample entry from a German–Polish dictionary of equivalents

Here the Polish term ‘przysposobienie’ (‘adoption’) has two actual congruent equivalents in the German legal system. They differ in terms of frequency and legal status: while ‘Annahme als Kind’ is the actual legal term used in the civil code, it is used less frequently than the common ‘Adoption’.



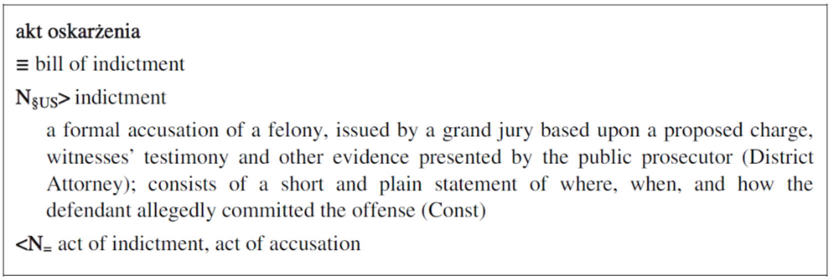
sąd grodzki

≅ magistrates’ court

<N$_{P}$ minor offences court

Fig. 4. Sample entry from an English-Polish dictionary of equivalents

In this case the dictionary offers two ways of translating the Polish term ‘sąd grodzki’: by means of either the convergent (that is only partially similar) term from the legal system of England and Wales, namely ‘magistrates’ court’, or the primary legal system-oriented neologism ‘minor offences court’. The latter expression does not refer to any institution in the secondary legal system, but merely succinctly explains the function of the Polish institution; hence it is marked as a periphrasis.



akt oskarżenia

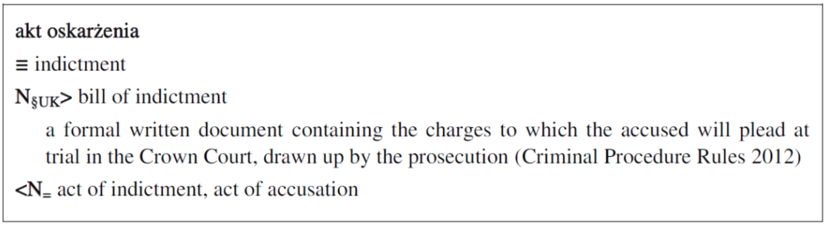
≡ bill of indictment

N$_{\S US}$> indictment

a formal accusation of a felony, issued by a grand jury based upon a proposed charge, witnesses’ testimony and other evidence presented by the public prosecutor (District Attorney); consists of a short and plain statement of where, when, and how the defendant allegedly committed the offense (Const)

<N$_{=}$ act of indictment, act of accusation

Fig. 5. Sample entry from a Polish-English dictionary of equivalents



akt oskarżenia

≡ indictment

N$_{\S UK}$> bill of indictment

a formal written document containing the charges to which the accused will plead at trial in the Crown Court, drawn up by the prosecution (Criminal Procedure Rules 2012)

<N$_{=}$ act of indictment, act of accusation

Fig. 6. Sample entry from a Polish-American dictionary of equivalents

The last two examples show why it is crucial to make separate dictionaries for each pair of particular legal systems and not languages. In a Polish-English dictionary (Fig. 5), that is one facilitating translation between the legal systems of Poland on the one hand and England and Wales on the other hand, the actual congruent equivalent of the term 'act oskarżenia' is 'bill of indictment'. However, the translator may want to emphasise that the two institutions are in fact not identical and rather use a term that will be less familiar to the final recipient. They may thus want to use a neologism in the form of a literal translation, namely 'act of indictment' or, even more exotically, 'act of accusation'. Another possibility is to use a neologism which stems from the American legal system (i.e. it is a neologism from the perspective of the law of England and Wales, not of the global English legal language in general), namely 'indictment'.

Figure 6 presents a reverse case: in a Polish-American dictionary the term 'indictment' will be the actual congruent equivalent, while the British term 'bill of indictment' will serve as a neologism. In each case the literal translation is marked as a source legal system-oriented neologism, while the terms from the respective foreign legal systems count as target legal system-oriented ones (being relatively familiar to the final recipients).

What might strike a careful reader is the lack of any definitions of actual equivalents (or the source terms, for that matter), while neologisms from other legal systems are followed by explanations and indications of the source of regulation. Worse still, the entries contain no grammatical or lexical information whatsoever. This stems from the fact that the dictionary of equivalents was designed not as an isolated reference work, but as an element of a system of dictionaries for translators of legal texts, comprising also an explicative dictionary, a contrastive dictionary, a combinatorial dictionary and a concise translation dictionary. The explicative dictionary handles issues encountered by translators in the first stage of their work, i.e. it aids them in understanding the source text by providing definitions of the source terms. The contrastive dictionary facilitates the comparison of the two legal systems involved by listing all actual equivalents and offering their definitions. The combinatorial dictionary is used in the final stage of translation, i.e. in the production of the target text, when the translator needs grammatical and lexical information concerning the use of the elected equivalent. The concise dictionary contains the most essential elements from all other volumes.

Full information the translator may need can be obtained through the study of all volumes. The idea to split information among four dictionaries is based on the finding, presented i.a. by S. Tarp [17, p. 37], that problems in translation do not necessarily occur at all the stages of the translation process: they can appear at one of them, at two, or at all. Depending on the problem complex, various dictionaries can be handy: a monolingual source language dictionary, a monolingual target language dictionary, or a bilingual dictionary. Sometimes the translator needs predominantly definitions, and at other times rather collocations. Thus, each element of the system of dictionaries can stand on its own, aiding the translator in solving problems related to a particular stage of the translation process. The dictionary of equivalents is assigned to the stage of transfer and therefore contains only information necessary to choose a proper equivalent.

5 CONCLUSION

The dictionary of equivalents is a concept of a new tool for legal translators. It is supposed to cater to their needs in respect of choosing a proper equivalent: it offers an extensive list of equivalents with relevant labels indicating their status, background and implications. The dictionary works best as an element of a system of dictionaries, which comprises reference works that are assigned to the remaining stages of the translation process and handle other aspects of legal terminology. However, the classification of equivalents and the system of their labelling may be applied in any dictionary designed for translators. Such a lexicon would go some of the way towards aiding the translators in achieving the impossible.

References

[1] Chaffey, P. N. (1997). Language, Law and Reality. In *On the Practice of Legal and Specialised Translation: Papers from the Third International Forum of Legal and Specialised Translation held in Cracow on 7th and 8th September, 1996*, pages 69–84, The Polish Society of Economic, Legal, and Court Translators TEPIS, Warsaw, Poland.

[2] Šarčević, S. (1997). *New Approach to Legal Translatio*n. Kluwer Law International, The Hague – London – Boston.

[3] Cao, D. (2007). *Translating Law*. Multilingual Matters Ltd., Clevendon – Buffalo – Toronto.

[4] De Groot, G.-R. (1990). Die relative Äquivalenz juristischer Begriffe und deren Folge für mehrsprachige juristische Wörterbücher. In *Translation and Meaning, Part 1. Proceedings of the Maastricht Session of the 1990 Maastricht-Łódz Duo Colloquium on 'Translation and Meaning', Held in Maastricht, The Netherlands, 4-6 January 1990*, pages 122–28, Euroterm, Maastricht, Netherlands.

[5] De Groot, G.-R. (1999). Guidelines for Choosing Neologisms. In *Aspects of Legal Language and Legal Translation*, pages 17–21, Łódź University Press, Łódź, Poland.

[6] Gizbert-Studnicki, T. (2004). Sytuacyjne uwarunkowanie językowych właściwości tekstów prawnych. In *Język – prawo – społeczeństwo*, pages 37–48, Uniwersytet Opolski, Opole, Poland.

[7] Chromá, M. (2004). *Legal Translation and the Dictionary*. Max Niemeyer Verlag, Tübingen.

[8] Roszkowski, S. (1999). The Language of the Law as Sublanguage. In *Aspects of Legal Language and Legal Translation*, pages 7–16, Łódź University Press, Łódź, Poland.

[9] Mattila, H. E. S. (2006). *Comparative Legal Linguistics*. Ashgate, Aldershot.

[10] Tognini-Bonelli, E. (1996). Towards Translation Equivalence from a Corpus Linguistics Perspective. *International Journal of Lexicography*, 9(3):197–217.

[11] Hausmann, F. J. (1977). *Einführung in die Benutzung der neufranzösischen Wörterbücher*. Max Niemeyer Verlag, Tübingen.

[12] Kubacki, A. D. (2002). Problemy konfrontacji polsko-niemieckiej terminologii podatkowej. In *Z problematyki języków specjalistycznych: materiały z konferencji*, pages 63–72, WSZMiJO, Katowice, Poland.

[13] Roelcke, T. (2005). *Fachsprachen*. Erich Schmidt Verlag, Berlin.

[14] Kielar, B. Z. (1977). *Language of the Law in the Aspect of Translation*. Wydawnictwa Uniwersytetu Warszawskiego, Warszawa.

[15] De Groot, G.-R. and van Laer, C. J. P. (2005). Bilingual and multilingual legal dictionaries in the European Union. A critical bibliography. Accessible at: `http://arno.unimaas.nl/show.cgi?fid=3130.`

[16] De Groot, G.-R. and van Laer, C. J. P. (2006). The Dubious Quality of Legal Dictionaries. *International Journal of Legal Information*, 34(1):65–86.

[17] Tarp, S. (2005). The concept of a bilingual dictionary. In *Schreiben, Verstehen, Übersetzen, Lernen. Zu ein- und zweisprachigen Wörterbüchern mit Deutsch*, pages 27–41, Peter Lang, Frankfurt am Main – Berlin – Bern – Bruxelles – New York – Oxford – Wien.

CZENGCLASS – TOWARDS A LEXICON OF VERB SYNONYMS WITH VALENCY LINKED TO SEMANTIC ROLES

ZDEŇKA UREŠOVÁ – EVA FUČÍKOVÁ – EVA HAJIČOVÁ
Institute of Formal and Applied Linguistics, Charles University, Prague,
Czech Republic

UREŠOVÁ, Zdeňka – FUČÍKOVÁ, Eva – HAJIČOVÁ, Eva: CzEngClass – Towards a Lexicon of Verb Synonyms with Valency Linked to Semantic Roles. Journal of Linguistics, 2017, Vol. 68, No 2, pp. 364 – 371.

Abstract: In this paper, we introduce our ongoing project about synonymy in bilingual context. This project aims at exploring semantic 'equivalence' of verb senses of generally different verbal lexemes in a bilingual (Czech-English) setting. Specifically, it focuses on their valency behavior within such equivalence groups. We believe that using bilingual context (translation) as an important factor in the delimitation of classes of synonymous lexical units (verbs, in our case) may help to specify the verb senses, also with regard to the (semantic) roles relation to other verb senses and roles of their arguments more precisely than when using monolingual corpora. In our project, we work "bottom-up", i.e., from an evidence as recorded in our corpora and not "top-down", from a predefined set of semantic classes.

Keywords: lexical resources, valency, synonymy, semantic roles, dependency corpus, multilingual

1 INTRODUCTION

It is widely accepted that verbs play a crucial role in a sentence structure – they form its core, relate other elements of the sentence to each other. Verbs can describe many events and states depending on the collocates they appear with, which in turn leads to the problem of ambiguity of verbs related to their meanings (senses). In addition, the same verb with no obvious meaning ambiguity can get translated into two or more different verbs in the target language, yet forming a perfect translation conveying the same meaning as in the source language. Take the verb "widen" in English, seen 32 times in the Penn Treebank [21] – in its Czech translation, 14 different verbs have been found: not only the most direct translation "rozšířit", but also "prohloubit" (lit. "deepen"), "rozrůst se" (lit. "grow [oneself]"), "stoupnout" (lit. "rise"), "zvětšit se" (lit. "enlarge"), "zvyšovat" (lit. "raise," "get higher") etc. Immediately, questions arise primarily about synonymy, but also about concrete vs. abstract distinction, relation to valency and argument structure, and more.

Different meanings of the same verb, or verb senses, are recorded and described – usually rather implicitly and informally – in both monolingual and bilingual dictionaries and we as humans can understand the sense distinctions well. However, our aim should be to describe verb senses precisely and explicitly. How do we know what is the explicit set of senses for any particular verb? Which senses (of different verb lexemes) are synonymous or near synonymous [7], [31] in the broader context

DOI 10.1515/jazcas-2017-0046

of use? It has been shown that if we let different people determine this, even on the same set of examples (i.e., using the same corpus), they inevitably come up with a different set. More precisely, the inter-annotator agreement [1] will be low, regardless of the level of linguistic expertise the annotators might have. Some researchers even go so far as to declare that they "do not believe in word senses" (legendary quote by the lexicographer Sue Atkins [2], explained by an article with the same title by Kilgarriff [16] that it should be interpreted as not believing in pre-determined, fixed set of word senses). Others try to find a sweet spot between a hard-to-agree-on, fine-grained set, represented e.g., by WordNet [6], and a coarse(r)-grained set, which does not provide enough detail–such as VerbNet lexicon [24], [14]. FrameNet [3], [8], [9] an English lexical resource which adds roles and uses semantic frames to group verbs and provide examples of use (based on attested corpus examples) is another well-known resource.

Regarding other languages, only some non-English WordNets link to the original English WordNet "synsets". FrameNet covers several languages, but it is not created systematically from parallel corpora. VerbNet is English-only. Moreover, these lexicons do not contain detailed morphosyntactic description of verb argument behavior (perhaps due to the selection of the original languages, which are in general not inflectional). There are no bilingual (or multilingual) resources describing verbs and their senses together with their semantic and morphosyntactic behavior in a bilingual setting. To fill this gap, our project will focus primarily on synonymy in bilingual context.

We believe that using the existing resources (mostly bilingual) based on the Functional Generative Description theory (FGD; [29]) will help us proceed in that direction. We are using two manually treebanked corpora: PDT (`http://ufal.mff.cuni.cz/pdt2.0`) and PCEDT [11], and the valency lexicons linked to these treebanks: PDT-Vallex [32], EngVallex [5], and a parallel valency lexicon CzEngVallex [33], [34]. We also take advantage of another FGD-based lexicon VALLEX [20], [19], [15] and other available resources, such as VerbNet [24], [14], FrameNet [3], English [6] and Czech WordNet [22], [23].

2 RESEARCH QUESTIONS

We will take advantage of the aforementioned lexical key resources as well as of large monolingual corpora, other parallel corpora such as Intercorp [28] and the NLP tools available for both languages, to work towards answering the following research questions:

– Do (verb) classes of synonyms based on monolingual and bilingual contexts differ, and if yes, in which respects? How are they related to structural representations (FGD, Abstract Meaning Representation (AMR, [4])?

– Crucially though, can the classes based on bilingual context be still kept disjoint (as the synsets in WordNet are)? Which consequences would overlapping classes have on the underlying theoretical approach(es)? Should any of the verb senses, as defined in the available dictionaries previously, be split or merged, based on the bilingual usage evidence?

– Which properties of a verb sense and the corresponding valency frame are relevant for grouping such verb senses into classes of synonyms, again in a bilingual vs. a monolingual context? Are they supported by corpus evidence?

– Conversely, what have the verb senses grouped in one class in common in terms of valency?

– Can a common set of verb "roles" (inspired, e.g., by FrameNet's Frame Elements [3] and by VerbNet's Thematic Roles [24]) be associated with one class, and how are these roles expressed in terms of valency (arguments, morphosyntactic expression)?

Our ultimate goal is in fact even a step deeper than to look at these questions in isolation: we hope to use the answers to these questions to confirm our hypothesis that using translation (i.e., bi-/multilingual context) as an important factor in determining the composition of such verb classes helps to define verb senses and their (primarily equivalent) relation to other verb senses and roles of their arguments more precisely than when using monolingual corpora, even if they follow Kilgarriff's postulate of giving substantial weight to individual occurrences in a corpus. We will create a lexicon of such synonymous verb pairs around representatively selected "seed" verbs; such similarity will be tested primarily against the translational equivalents in context, as found in the parallel corpora. We will compare the results with the approach of [18] as embodied, e.g., in the VerbNet [24], [14], as a representative of classes of semantically and syntactically similar verbs based on monolingual resources and research on one language (English). Last, but certainly not least, we will compare the resulting classes and their properties to the VALLEX lexicon [20], [19] and VerbaLex [13] on the Czech side. Results will be analyzed from the point of view of the Functional Generative Description theory [29] and its approach to valency [26], [27], [12] and the relation of form and meaning, and possibly generalized across the two languages we will work with.

3 PROJECT WORKFLOW

3.1 Preparatory Part

In the preparatory part of the project, we have been analyzing the existing Prague Czech-English Dependency Treebank (PCEDT), as well as the related valency lexicons: PDT-Vallex [32], EngVallex [5] and CzEngVallex [34]. We have been also studying the methodology of the VerbNet class-based verbal lexicon [17] and FrameNet [3]. We have performed a detailed analysis of the verbs contained in the CzEngVallex lexicon, in order to create classes of synonyms similar to those of VerbNet, but–importantly–in a bilingual setting, which needs the support of the PCEDT to see the use of such verbs in the parallel corpus, i.e., in the context of real usage. Next, we have selected a representative sample of verbs (about 50 classes centered around "seeds" from the sample selected), along the dimensions of frequency and richness of sense inventory and translation equivalents. Simultaneously, we have been preparing technical tools (software) allowing to manually (re-)group and refine verb senses, build classes of synonyms and assign them an appropriate semantic frame and roles.

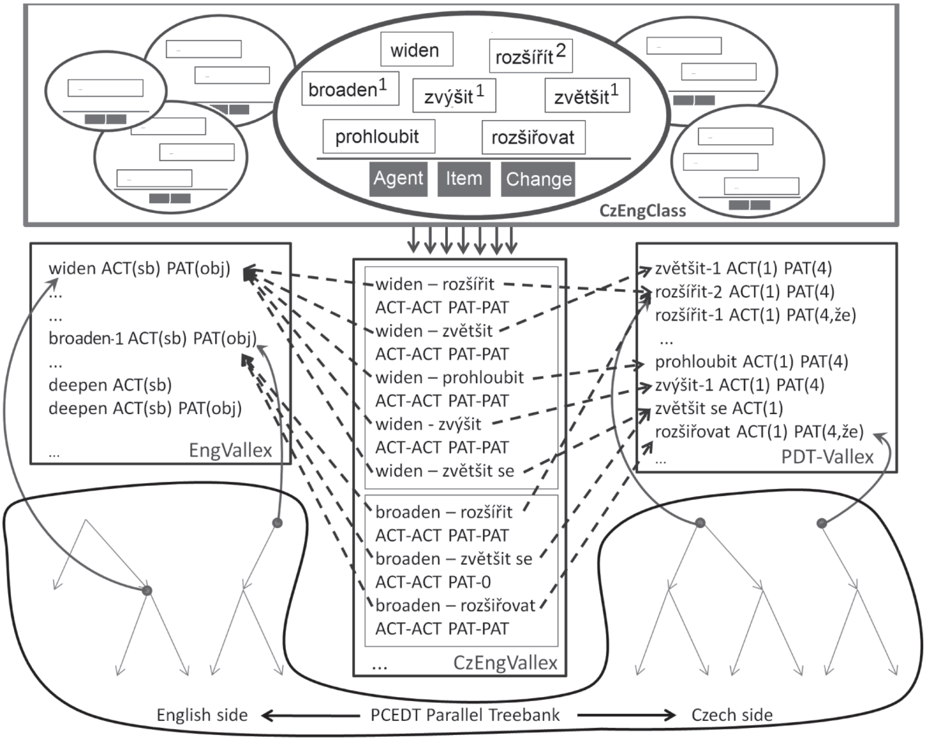


Fig. 1. Overall scheme of CzEngClass

3.2 Data Extraction

In the data extraction part, verb sense (~ verb valency frame) pairs selected from CzEngVallex have been grouped into classes corresponding to their semantic similarity (i.e., synonyms or near synonyms). While this is more complex in the bilingual setting, the translation context from PCEDT provided very strong, empirical evidence for their equivalence in context, as opposed to mere similarity in argument types or in their surface realization (and one's often unreliable intuition). The resulting "database" (Fig. 1, working name "CzEngClass") has a relatively simple form – it groups together frame pairs captured in CzEngVallex into classes, which represent synonym or near-synonym pairs of verbs (verb senses). However, every pair in every class is also linked to CzEngVallex, PDT-Vallex and EngVallex, and the PCEDT, allowing for relation-based search by computational tools in the analysis part.

Part ot the data extraction process will keep links to external resources as well. The following resources will be used: FrameNet, VerbNet, PropBank [25] and WordNet for English, and Czech WordNet for Czech. Most of these resources are accessible through the Unified Verb Index (`https://verbs.colorado.edu/verb-index/`). However, it will be necessary to find the right correspondences; for example, the senses as recorded in VerbNet "Groupings" have to be linked to EngVallex senses (frames), and of course the verb arguments, e.g, from PropBank are structured differently than in EngVallex.

3.3 Data Analysis

This part is the core part of the project. Here we plan to analyze the complex set of relations between meaning and form for the synonym classes of verb senses (as represented by their valency frames) created in the data extraction part. We will study the classes as a whole as well as its members individually in terms of arguments, their types, their surface morphosyntactic realization, and also all anomalies and deviations which we encounter either in the valency lexicons PDT-Vallex and EngVallex or in the PCEDT parallel data. We believe that such findings will lead to the description of bilingual-corpus-based semantically defined classes of synonyms or near synonyms. In this analysis, the external resources will also be consulted to get more insight into semantic role labeling, semantic classes etc.

4 PROJECT OUTPUT

The output of the project will be CzEngVallex, a lexicon of synonym classes, where each verb (verb sense), Czech and English alike, will be assigned to one class, and it will be linked to the other available resources for reference and to support other follow-up studies. In addition, each class will be also characterized by a set of semantic roles which will be shared about the class members, and verb arguments will be mapped to these roles. The data will be openly and freely available.

Verb lexemes	Closest FrameNet frame	Roles:			
		Cognizer	Means/ Instrument	Phenomenon	Source
dozvědět se[1]	Becoming_aware	ACT		PAT	ORIG
get[1]	Becoming_aware	ACT		PAT	ORIG
hear[1,2]	Becoming_aware	ACT		PAT	ORIG
know[1,3]	Becoming_aware	ACT		PAT/EFF	
learn[1]	Becoming_aware	ACT		PAT	ORIG
tell[3]	Becoming_aware	ADDR	ACT?	PAT/EFF	ACT?

Tab. 1. Example set for “learn” (“dozvědět se“) with (initial) argument mappings

An example of preliminary synonym set with equally preliminary mappings from verb argument labels to a set of roles initially identified for each class are in Table 1. It is clear that there are immediate problems to solve:

- what (FGD-)based roles should be used to map the candidate verb arguments to the Means/Instrument and the Source semantic roles?
- why the translation uses the word “know” as an translation equivalent of “dozvědět se”, given that “know” is more of a state-type of verb, while “dozvědět se” is describing the process of “getting to know“, albeit it is in perfective voice?
- is “tell” really a good synonym (even in the loose, contextually-based sense), given that the ACTor could well be assigned to both the Source and the Means semantic roles?

We also expect that the existing valency lexicons will be amended, since inconsistencies in the previous annotation may be found. The corrected lexicons PDT-Vallex and EngVallex will thus be also published openly.

The overall structure of the lexicon with the basic referencing (from CzEngClass to the two valency lexicons and the parallel corpus, but not to the external resources) are depicted schematically on Fig. 1. So far, an XML scheme for the lexicon has already been designed and a work on an editor is in progress (cf. also Sect. 3.1.) and it will be described in the final version of the paper.

5 SUMMARY

We have described (based on the grant No. GA17-07313S proposal, of which the authors of this article are participants) a project which is just starting and which is supposed to lead to an interconnected synonym bilingual lexicon based on parallel corpus and existing lexicons. Entries in this lexicon will share, for each class, a set of semantic roles mapped to arguments in the valency lexicons. The lexemes will also be linked to the Universal Verb Index to keep relations to the widely used verb lexicon resources, such as FrameNet, VerbNet or PropBank, whenever possible. An indispensable resource, which is directing the research, is the Czech-English parallel richly annotated corpus which brings a new view on cross-lingual (and multilingual) contextual synonymy.

All the new resources and the linking will be made public as open data.

ACKNOWLEDGMENTS

This work has been supported by the grant No. GA17-07313S of the Grant Agency of the Czech Republic, and it uses resources hosted by the LINDAT/CLARIN Research Infrastructure, project No. LM2015071, supported by the Ministry of Education, Youth and Sports of the Czech Republic.

References

[1] Artstein, R. and Poesio, M. (2008). Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555–596.

[2] Atkins, S. (1993). Tools for computer-aided corpus lexicography: the Hector project. In *Papers in Computational Lexicography: Complex'93*, pages 1–60, Budapest, Hungary.

[3] Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The Berkeley FrameNet project. In *Proceedings of the COLING-ACL*, pages 86–90, ACL, Montreal, Canada.

[4] Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M., and Schneider, N. (2013). Abstract Meaning Representation for Sembanking. In *Proceedings of the 7th LAW Workshop*, pages 178–186, ACL, Sophia, Bulgaria.

[5] Cinková, S. (2006). From PropBank to EngValLex: Adapting the PropBank-Lexicon to the Valency Theory of the Functional Generative Description. In *Proceedings LREC 2006*, pages 2170–2175, Genova, Italy.

[6] Fellbaum, Ch. (1998). WordNet: An Electronic Lexical Database. MIT Press, Cambridge, MA.

[7] Filipec, J. (1961). *Česká synonyma z hlediska stylistiky a lexikologie.* Nakladatelství Československé akademie věd, Praha.

[8] Fillmore, Ch. J. (1976). Frame semantics and the nature of language. *Annals of the New York Academy of Sciences*, 280:20–32. Accessible at: doi: 10.1111/j.1749-6632.1976.tb25467.x.

[9] Fillmore, Ch. J., Johnson, Ch., and Petruck, M. R. L. (2003). Background to FrameNet. *International Journal of Lexicography*, 16(3):235–250.

[10] Fučíková, E., Hajič, J., Šindlerová, J., and Urešová, Z. (2015). Czech-English Bilingual Valency Lexicon Online. In *Proceedings of the 14th TLT 2015*, pages 61–71, IPIPAN, Warszawa, Poland.

[11] Hajič, J., Hajičová, E., Panevová, J., Sgall, P., Cinková, S., Fučíková, E., Mikulová, M., Pajas, P., Popelka, J., Semecký, J., Šindlerová, J., Štěpánek, J., Toman, J., Urešová, Z., and Žabokrtský, Z. (2011). Prague Czech-English Dependency Treebank 2.0. Data/software, UFAL MFF UK, Prague, Czech Republic. Accessible at: `http://ufal.mff.cuni.cz/pcedt2.0` (23. 3. 2015).

[12] Hajičová, E. (1983). Remarks on the meanings of cases. *Prague Studies in Mathematical Linguistics*, 8:149–157.

[13] Hlaváčková, D., Horák, A., and Kadlec, V. (2006). Exploitation of the VerbaLex Verb Valency Lexicon in the Syntactic Analysis of Czech. In *Proceedings of 9th International Conference on Text, Speech, and Dialogue (TSD 2006)*, pages 85–92, Springer, Berlin – Heidelberg, Germany.

[14] Kawahara, D., Peterson, D., Popescu, O., and Palmer, M. (2014). Inducing Example-based Semantic Frames from a Massive Amount of Verb Uses. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, EACL-2014*, pages 58–67, Gothenburg, Sweden.

[15] Kettnerová, V. (2014). *Lexikálně-sémantické konverze ve valenčním slovníku*. Karolinum, Prague.

[16] Kilgarriff, A. (1997). I don't believe in word senses. *Computers and the Humanities*, 31(2):91–113.

[17] Kipper, K., Korhonen, A., Ryant, N., and Palmer, M. (2008). A large-scale classification of English verbs. *Language Resources and Evaluation Journal*, 42(1):21–40.

[18] Levin, B. (1993). *English Verb Classes and Alternations: A Preliminary Investigation*. The University of Chicago Press, Chicago.

[19] Lopatková, M., Kettnerová, V., Bejček, E., Vernerová, A., and Žabokrtský, Z. (2016). *Valenční slovník českých sloves VALLEX*. Karolinum, Praha.

[20] Lopatková, M., Žabokrtský, Z., Kettnerová, V., Skwarska, K., Bejček, E., Hrstková, K., Nová, M., and Tichý, M. (2008). *Valenční slovník českých sloves*. Karolinum, Praha.

[21] Marcus, M., Santorini, B., and Marcinkiewicz, M. A. (1993). Building A Large Annotated Corpus of English: The Penn Treebank. *Computational Linguisics*, 19(2):313–330.

[22] Pala, K. and Smrž, P. (2004). Building Czech Wordnet. *Romanian Journal of Information Science and Technology*, 7(2–3):79–88.

[23] Pala, K. and Všianský, J. (1994). *Slovník českých synonym*. 1. vyd. Nakladatelství Lidové Noviny, Praha.

[24] Palmer, M., Hwang, J. D., Brown, S. W., Kipper, S. K., and Lanfranchi, A. (2009). Leveraging lexical resources for the detection of event relations. In *Proceedings of the AAAI 2009 Spring Symposium on Learning by Reading*, pages 81–87, Stanford, CA.

[25] Palmer, M., Gildea, D., and Kingsbury, P. (2005). The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1):71–106.

[26] Panevová, J. (1974). On verbal frames in Functional Generative Description. *The Prague Bulletin of Mathematical Linguistics*, 22:3–40.

[27] Panevová, J. (1975). On verbal frames in Functional Generative Description. *The Prague Bulletin of Mathematical Linguistics*, 23:17–52.

[28] Rosen, A. and Vavřín, M. (2015). Korpus InterCorp, verze 8 z 4. 6. 2015. Ústav Českého národního korpusu FF UK, Praha. Accessible at: `http://www.korpus.cz.`

[29] Sgall, P., Hajičová, E., and Panevová, J. (1986). *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. Reidel, Dordrecht – Academia, Prague.

[30] Skoumalová, H. (2001). *Czech Syntactic Lexicon*. Charles University in Prague, Prague.

[31] Sparck, J. K. (1986). Synonymy and semantic classification. Edinburgh University Press (Edinburgh information technology series 1).

[32] Urešová, Z. (2011). *Valenční slovník Pražského závislostního korpusu (PDT-Vallex). Studies in Computational and Theoretical Linguistics 9*, UK Praha.

[33] Urešová, Z., Dušek, O., Fučíková, E., Hajič, J., and Šindlerová, J. (2015). Bilingual English-Czech Valency Lexicon Linked to a Parallel Corpus. In *Proceedings of the the 9th Linguistic Annotation Workshop (LAW IX 2015)*, pages 124–128, ACL, Stroudsburg, PA, USA.

[34] Urešová, Z., Fučíková, E., and Šindlerová, J. (2016). CzEngVallex: a bilingual Czech-English valency lexicon. *The Prague Bulletin of Mathematical Linguistics*, 105:17–50.

SLAVIC PHRASEOLOGY: A VIEW THROUGH CORPORA

VICTOR ZAKHAROV
Saint-Petersburg State University, Russia

ZAKHAROV, Victor: Slavic Phraseology: A View Through Corpora. Journal of Linguistics, 2017, Vol. 68, No 2, pp. 372 – 384.

Abstract: The study of word collocability is one of the main tasks of linguistics. The combinatory ability of language units, collocability, is one of the linguistic syntagmatic laws. This phenomenon is the main object of the phraseology and lexicography. The article deals with set phrases of different types in Russian, Czech and Slovak from the point of view of their quantitative evaluation. Corpus linguistics understand set phrases as statistically determined unities. This approach is the basic point of different automatic ways to extract idioms and collocations. The paper describes experiments which show how text corpora and corpus methods and tools can be used to expand the entries in existing dictionaries and how set phrases could be evaluated quantitatively. It is shown and maintained that corpus linguistics methods and tools allow to create dictionaries of new type which have to include a larger amount of set phrases and collocations than before.

Keywords: Slavic phraseology, phraseological units, set phrases, idioms, collocations, corpus, lexicography

1 INTRODUCTION

One of the popular topic in the science of language are set phrases. They are studied under different sections of linguistics and from various standpoints. The classical name for a set phrase in the linguistic terminology is a phraseological unit. Various scientists interpret this notion and its properties in various ways, and there are many different classifications of set phrases. However, when comparing them, we will see that the list of these properties and the very classification are often alike and have a lot in common.

The commonly adopted interpretation of phraseological units defines them as set, reproducible, expressive word combinations (*беречь пуще глаза* (literally, protect more than eyes); *драть по три шкуры* (literally, to skin three skins); *sedí na uších (literally, sit on ears), jako by do moře padl* (as it fell into the sea)). However, we can often find in translation dictionaries such units as, *и ничего больше* (and nothing else), *to je to* (that's just it)*, a tečka* (and period)*,* where the main property is reproducibility. And this is right, since such combinations are often stable, and can be compared to words by their frequency. In this article, we will consider phraseological units as "stable word complexes" [4].

However, despite the close attention of linguists to the phraseology, we can state that the computer-based methods of research have hardly been used to study phraseology. The dictionary description of phraseological units shall be elaborated. The phraseological reserve of a language is scattered in various lexicographic

DOI 10.1515/jazcas-2017-0047

publications, and no dictionary can be regarded as covering the phraseological lexicon in full, taking into account also the fact that it always expands.

Today, we can improve the situation by means of corpora. Due to the availability of large text corpora, including those that cover a long period of time, and the software tools that allow to estimate the compatibility quantitatively, all pre-requisite conditions have appeared for the creation of a large combinatory dictionary, obviously, in electronic form, based on corpora, and having quantitative parametrization inside.

It should be noted that the methodology of a larger understanding of phraseology has formed in linguistics, and the boundaries of phraseology have been significantly enlarged (or blurred) due to new approaches that have the notion of "statistical stability" in common. I.A. Melčuk spoke about it as early as in 1960. "The stability of a combination relative to this element is measured by the probability of this element forecasting the combined occurrence of the other elements of the combination (in a certain order relative to the forecasting element)" [7, p. 73].

In corpus linguistics, frequency characteristics and structural and syntactic models form the basis of the methods of calculating the strength of syntagmatic association between the word combination elements. Based on them, the association measure score, or, in other words, the uniqueness of this combination, is calculated.

As it is known, language is a dynamic system, which shall be reflected in dictionaries and grammars. However, maybE this chronological aspect has been less studied in terms of phraseological units and other stable combinations. One of the reasons is the absence of large historical (diachronic) corpora till our times.

The objective of this research is to show how phraseological and combinatory dictionaries can be improved using corpus-based methods. The basis of our approach are the analytic tools for phraseological units and stock-taking of the language material. Furthermore, we show how similar phraseological units correlate with each other in related Slavic languages. We also consider the issue of translation of phraseological units from one language to another. The Slavic phraseology is represented by Russian, Czech, and Slovak.

2 RESEARCH MATERIAL AND TOOLS

A family of comparable corpora Aranea of the Comenius University in Bratislava (`http://unesco.uniba.sk/`) and the Russian corpus of Google books Ngram Viewer (`https://books.google.com/ngrams`) have been used as the material and the tool. The corpora of the Aranea family [2] operate under the NoSketch Engine corpus manager. They are represented by the Maximum type in the research, and they have the following volume: Russian corpus is made up of 13.7 billion tokens, the Czech one – 5.17 billion tokens, and Slovak – 2.68 billion tokens. Also, some data were retrieved in [6], [10] and [13]. Google books Ngram Viewer system [8, p. 14] was used for the diachronic research. The system allows to plot graphs of word occurrences and collocations for a certain period of time. It also allows to select the most frequent collocations with such word form, both from the right and from the left, using wild cards. There is also the possibility of setting the part of speech of another components of the collocation.

Our chosen phraseological units for the research are the set phrases of two types: classical phraseological units (idioms) and syntactic idioms (idiomatic constructions).

3 REPRODUCIBILITY AND VARIABILITY OF PHRASEOLOGICAL UNITS

3.1 Idioms

Idioms are not only strongly reproducible, but they are, at the same time, very variable, and it is an important task of the phraseology to present this variability in the dictionaries. Many idioms and set phrases have lexical-syntactic variants when either the lexical meaning within a certain structural formula or the formula changes. There are a lot of examples of this phenomenon: *беречь (хранить) как зеницу ока* (keep as the apple of one's eye)*; беречь пуще глаза* [3]; *dát si na někoho/před někým majzla* (pay attention to somebody); *dát/uložit něco k ledu/z ruky* (literally, let smth to ice/from hand) [9]. Let us consider some of them.

Example 1:
Rus. *на сердце (на душе) кошки скребут* (literally, cats scratch the heart/soul) [5];
Czech. *je mi těžko (úzko) u srdce (v duši)* (be sick at heart/soul) [11];
Slov. *je mi ťažko (úzko) okolo srdca (na duši)* (near the heart/at soul) [12].

	на … кошки скребут	кошки скребут на ….	Total
на сердце	27	5	32
на душе	383	85	**468**
Total	**410**	90	500

Tab. 1. The frequency of occurrences of combinations with the phrase «кошки скребут» in the Araneum Russicum Maximum corpus

	je mi těžko …	je mi úzko …	Total
u srdce	17	3	**20**
v duši	0	1	1
Total	**17**	4	21

Tab. 2. The frequency of the occurrences of the combinations with the phrase «je mi těžko (úzko)» in the Araneum Bohemicum Maximum corpus

	je mi ťažko …	je mi úzko …	Total
okolo srdce	0	0	0
na duši	21	10	**31**
Total	**21**	10	31

Tab. 3. The frequency of the occurrences of the combinations with the phrase «je mi ťažko (úzko)» in the Araneum Slovacum Maximum corpus

What can be derived from this example? In Russian, this idiom is more commonly used than in the Czech and Slovak languages, and «на душе» (soul) is used with «кошки скребут» more often than «на сердце» (heart), i.e. 468 occurrences against 32 (Table 1). It is an interesting fact that an object noun usually precedes the verb (410 against 90). As contrasted with Russian, Czech prefers the "heart" variant (Table 2). It is possible that this is the influence of the German language. In the pair *těžko/úzko,* the variant *těžko* is used most often. The "object" (noun) virtually always stands after the verb. The same goes for the Slovak language: *ťažko* – 21 occurrences, and *úzko* – 10 occurrences, while the dictionary variant *okolo srdce* does not occur in the corpus at all.

Meanwhile, the Czech corpus returns word combinations with preposition "na" (on), which is not present in the dictionary [9]: *je mi těžko (úzko) na srdci (na duši)* (Table 4). And the combination *na duši* is far more often.

	je mi těžko …	je mi úzko …	Total
na srdci	4	0	4
na duši	15	6	**21**
Total	**19**	6	25

Tab. 4. Frequency of the occurrence of combinations with the phrase «je mi těžko (úzko) na ...» in the Araneum Bohemicum Maximum corpus

Similarly, the Slovak corpus adds to the dictionary data the following combinations: *na srdci* and *v duši* (Table 5).

	je mi ťažko …	je mi úzko …	Total
na srdci	29	1	**30**
v duši	1	2	3
Total	**30**	3	33

Tab. 5. Frequency of the occurrence of additional combinations with the phrase «je mi ťažko (úzko) ...» in the Araneum Slovacum Maximum corpus

For Russian, the dictionary [4] returns also a synonymic expression *камень лежит на сердце (на душе)* (literally, a stone lies on the hear/soul). Once again, the corpus analysis makes another correction: this expression occurs more often in the form ***камнем*** *что-то лежит на сердце (на душе)* (65 occurrences versus 15).

Example 2. Let us consider the phraseological units that are rather literal:
Rus. *камень преткновения* (literally, a stone of obstacle) [5];
Czech. *kámen úrazu* (literally, a stone of injury) [11];
Slov. *kameň úrazu* Id. [12];
Rus. *нашла коса на камень* (literally, a meak stumbled on a stone) [5];
Czech. *přišla/trefila/padla kosa na kámen* (came/dropped …) [11];
Slov. *padla kosa na kameň* Id. [12].

Idiom	frequency	ipm	variants
Rus. *камень преткновения*	16710	1,20	
Czech. *kámen úrazu*	11003	**2,10**	
Slov. *kameň úrazu*	5425	**2,00**	

Rus. *коса на камень*	1171	0,10	***найти – 945*** *пойти – 13* *налететь – 3*
Czech. *kosa na kámen*	649	**0,13**	*přijít – 9* *trefit – 5* ***padnout – 335*** *narazit – 201*
Slov. *kosa na kameň*	406	**0,15**	*prísť – 5* *trafiť – 24* *padnúť – 149* ***naraziť – 180***

Tab. 6. Frequency of occurrence of the combinations «камень преткновения» and «нашла коса на камень» and their equivalents in Czech and Slovak in three corpora

From Table 6 it is seen that in Czech and Slovak these expressions are used more often. While in the Russian corpus this expression is almost always found with the verb «найти» (*нашла*) (find), in the Czech and Slovak corpora, alongside with the word “padnout” (“padnúť”), the verb “narazit” (“naraziť”) (injure oneself, stumble upon smth) is used frequently, too, that is not given in dictionaries.

Example 3:
Rus. *делать (сделать) из мухи слона* (literally, make an elephant out of a fly) [5];
Czech. *dělat (udělat) z komára velblouda* (make a camel out of a mosquito) [11];
Slov. *robiť (urobiť) z komára somára* (make a donkey out of a mosquito) [12].

Idiom	frequency	ipm	variants
Rus. *из мухи* ***слона***	2343	**0.20**	*толстяка 1* *барсука 1*
делать/сделать	1098		
раздувать/раздуть	791		
Czech. *z komára* ***velblouda***	634	0.10	*vola (bull) 14* *slona 5 (3 of them appeared in texts in Slovak!)*
dělat/udělat	602		
Slov. *z komára* ***somára***	265	0.10	***slona*** *56* *velblouda 8 (4 of them appeared in texts in Czech!)* *vola - 4*
robiť/urobiť	240		

Tab. 7. Frequency of occurrence of the word combination «делать из мухи слона» and its equivalents in Czech and Slovak in three corpora

What is the essence of Table 7? First of all, different frequency of occurrences: it is twice more often in Russia; secondly, slightly different variations and, what is more interesting, the Russian corpus returns a lot of combinations with the verb «раздувать» (blow), which is not given in dictionaries. In the Slovak language, it is often an elephant that is made out of a mosquito, and the dictionary has no evidence of that.

3.2 Idiomatic Constructions

Set syntactic constructions that have variable lexical elements are singled out as a separate type of phraseological units. They are, so to say, syntactic patterns that are filled depending on the context, communicative aim of the author, and – as we will show – on the language. They include such expression as «X как X» (X as X), «тоже мне X» (some X), «всем X-ам X» (X to all Xs), «X X-ов» (X of Xs), «X – он и в Африке X» (X is the same in Africa), etc. Idiomatic constructions are studied within the framework of the construction grammar developed by Ch. Fillmore, A. Goldberg, and others. It is considered that "blank spaces" can be filled with anything at all. However, there are semantic limitations for the filling, and there is the language usage that changes in different languages. Let us consider several cases.

Example 4: X как X (X jako X, X ako X) (X as X).

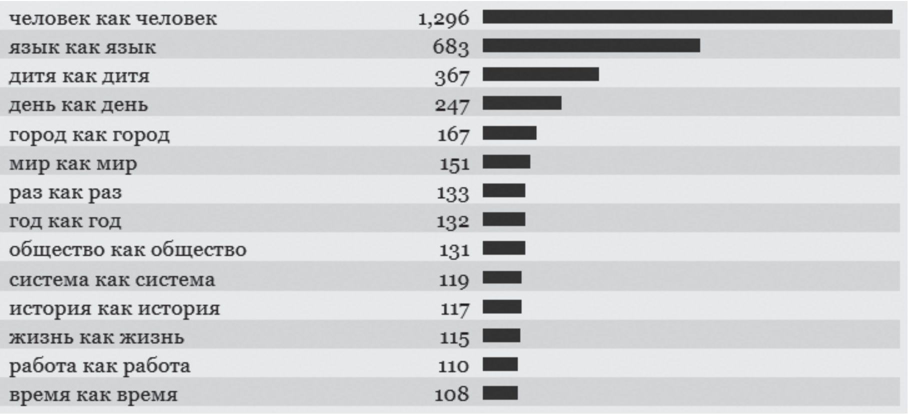


Fig. 1. Frequency of occurrence of expressions like «X как X» in the Araneum Russicum Maximum corpus

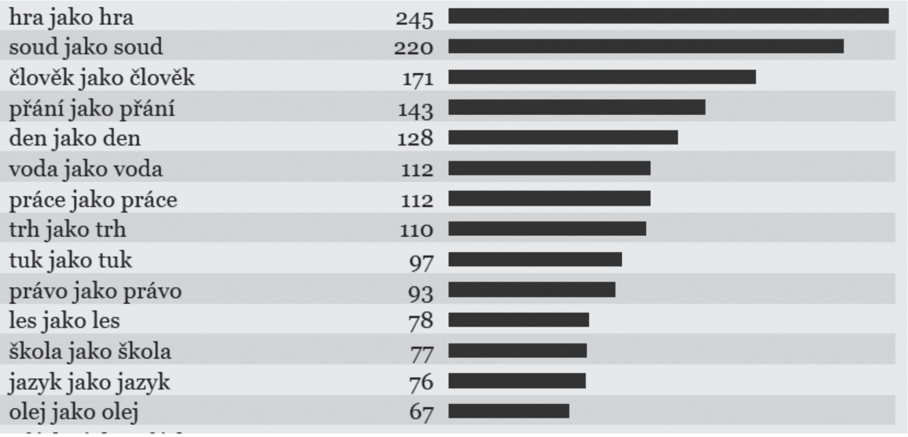


Fig. 2. Frequency of occurrence of expressions like «X jako X» in the Araneum Bohemicum Maximum corpus

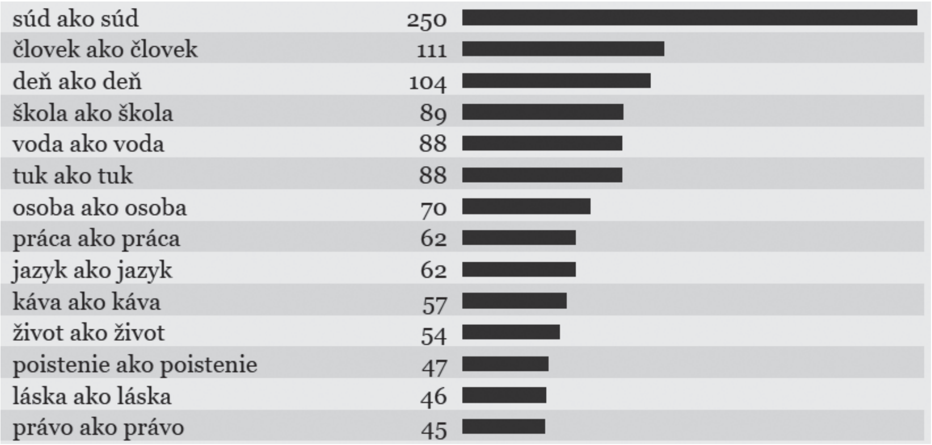


Fig. 3. Frequency of occurrence of expressions like «X ako X» in the Araneum Slovacum Maximum corpus

Based on the data in Figures 1-3, it is clear that there are frequent set combinations of such type (for Russian ipm=1.20, for Czech ipm=3.60, for Slovak ipm=4.40), and that they differ in different languages and that Czech and Slovak are more close between themselves.

Example 5: Все X суть Y (všichni X jsou Y, všetci X sú Y) (All X are Y).

Russian	freq.	Czech	freq.	Slovak	freq.
все люди братья	447	všichni muslimové jsou teroristé	33	všetci moslimovia sú teroristi	19
все мужики сволочи	167	všichni teroristé jsou muslimové	123	všetci ľudia sú bratia	15
все мужики козлы	163	všichni lidé jsou bratři	113	všetci politici sú zlodeji	9
все люди взрослые	90	všichni politici jsou zloději	69	všetci Rómovia sú zlodeji	6
все ребята молодцы	84	všichni muži jsou násilníci	61	všetci ľudia sú hriešnici	5

Tab. 7. Frequency of occurrence of the like-expressions in the three corpora

As in Figures 1-3, Table 7 shows different filling of variables X and Y for different Slavic languages. It is fair to say that these data to a certain extent reflect the linguistic consciousness of the native speakers. These data would have been hard to obtain without a large representative corpus.

4 COMPARABLE PHRASEOLOGY AND ISSUES OF TRANSLATION

4.1 Idiom Equivalency

The subject of comparable phraseology in the narrow sense of the word are phraseological units of different languages that have similar semantic or structural characteristics. One of the issues of a comparable analysis is the typology of interlingual equivalency. We have considered such idioms and constructions (without going into much detail) in Section 3.

The interlingual equivalency, including that of phraseological units, is also the equivalency in the system of a language (a dictionary problem) and the issue of translation of phraseological units in an individual text. While the translation of a set phrase of the "denotative" type is not much of a challenge (*silný čaj – крепкий чай* (strong tea), *nevlastní otec – nevlastný otec – неродной отец* (step-father)), the translation of other phraseological units is far from simple. This is explained both by their "non-denotativeness" and the absence of full, semantically organized collections of foreign idioms. Sometimes, such translation can be found quite easily (*делать из мухи слона – dělat z komára velblouda – robiť z komára somára*), and sometimes it requires a creative approach.

One should remember that the majority of idioms do not have direct equivalents in the compared language. Then, they can be translated descriptively or you can try and find one of the semantic equivalents. For example: *дуракам закон не писан* (literally, the law is not written to fools) – Czech. *hloupa kůže všechno může* (a stupid skin can do anything)*, на безрыбье и рак рыба* (literally, when there is no fish, a crawfish is fish) – Czech. *z nouze Franta dobrý* (even František is good when there is need) [11].

Besides, there is a partial equivalency that can be called the problem of the relation "many to many". For example, in Russian, one can say *молчит как ... рыба, партизан, истукан, вода, пень ... как язык проглотил* (silent as … fish, guerilla, statue, water, stub… as if has swallowed his tongue) [5]. In Czech, it will be *mlčí jako ... dub (*oak*), sfinga (*sphinx*), hrob (*grave*), pěna (*foam*), ryba (*fish*), jako zařezaný (*slaughtered*), jako by mu přimrzl jazyk (*as if his tongue froze down*)* [9]. In Slovak, it will be *mlčí ako ... hrob, kameň', peň (*stub*), ryba, sfinga, zarezaný, ako voš pod chrastou* (as a louse under blemish) [1]. This raises a question of what variant to choose in translation or to cite in a dictionary.

<u>Example 6</u>:

Rus. *Пьяный* (drunk) *как ... сапожник* (as a shoemaker)*, свинья* (pig)*, скотина* (cattle)*, в доску* (in board)*, в стельку* (in insole)*, в дым* (in smoke)*, вдрызг, вдребезги* (into smithereens) [5];
Czech. *opilý (vožralý, nalitý, zpitý)* (drunk, tight) *jako Dán* (as a Dutch)*, dráteník* (potter)*, zvíře* (animal)*, Holandr* (Dutchman)*, duha* (rainbow)*, námořník* (sailor)*, pod obraz (boží)* (unlike God's image) [9];
Slov. *opitý (spitý)* (drunk) *ako cepelín* (airship), *snop* (sheaf), *čík* (misgurnus (fish)), *prasa* (pig), *na mol* (in flat note), *do nemoty* (to muteness), *pod obraz boží* (unlike God's image) [1].

Which of the adjectives (or respective verbs) are used more often? Which of the comparisons are used more often? Which of the word combination have more association strength? A large corpus can give at least preliminary answers to these questions. However, naturally, those answers would not be exhaustive (see Table 8).

Russian	freq. (ipm)	Czech	freq. (ipm)	Slovak	freq. (ipm)
пьяный как ...	*1163* *(0,10)*	*opilý jako ...*	*165* *(0,03)*	*opitý ako ...*	*345* *(0,10)*
свинья	79	prase	21	čík	26
сапожник	53	Dán	10	doga	19
стелька	7	doga	6	delo	15
скотина	7	žok	4	prasa	14
собака	5	dělo	4	teľa	6
фортепьян, швед, ямщик, лорд, зюзя,извозчик...	…	slíva, konev, štoudev, kára, kráva, hovado, zeppelín...	…	sviňa, Dán, snop, činka, vôl, Rus, Poliak...	…

Tab. 8. Frequency of the combinations with «пьяный как …» and their equivalents in Czech and Slovak in three corpora

Table 8 gives only some examples whose aim is to show the differences in word combinations between the dictionaries and large corpora of a "live" language. Some of the expressions which are present in corpora and absent in dictionaries shall, without doubt, be included in them. For example, for the Czech language, those are *opilý na mol* (45 occurrences) and, possibly, *opilý na plech* (7) and *na šrot* (4). Besides, frequencies obtained from representative and balanced corpora allow to specify the sequence of phraseological units in dictionaries.

4.2 Parallel Corpora

One of the sources that can help in solving the tasks of comparable phraseology and translation are parallel corpora. Let us give several examples from the Intercorp which is a part of the Czech National Corpus [10].

Example 7: *pepka klepne* (apoplexy)

Otcenasek-Kulhavy_Orfe	Vo pár důstojnejch tatíků se pokoušela **pepka**.	Нескольких почтенных дядюшек чуть **кондрашка** не хватила.
Otcenasek-Kulhavy_Orfe	A nech toho chlastání, nebo mamulu klepne **pepk**a, až tě uvidí takhle zhulákanýho .	И перестань хлестать , а то мамулю **кондрашка** хватит, как увидит тебя такого развеселого.
Doncova-Manikura_pro	Říká se, že můj otec byl úplně stejný, do své smrti běhal za ženskými a pak ho klepla **pepka**.	Говорят , отец мой такой был, до смерти по бабам бегал и от **инфаркта** умер.

Tab. 9. Examples of translation of the «pepka klepne» word combination in the Intercorp corpus

Example 8: *obrátit vniveč, přijít vniveč* (render null, go down the drain)

Eco-Jmeno_ ruze	A je - li upálen člověk, shoří i jeho individuální substance a s ní je **vniveč** obráceno i konkrétní bytí, skutečné, a už jen proto dobré, aspoň v očích Boha, který je naživu držel.	Между тем, когда человек сгорает, раньше всего сгорает его индивидуальная субстанция, и при этом **аннулируется** то, что прежде составляло конкретный акт существования - очевидно, благой по своей идее, хотя бы на взгляд Господа Бога, который для чего-то потворствовал сему существованию.
Wells-Stroj_ casu	Podle mého by byla škoda, kdyby to jídlo přišlo **vniveč**, poznamenal redaktor chvalně známého deníku	Досадно, если обед будет **испорчен**, - сказал Редактор одной известной газеты.
Granin-Krasna_Uta	Kdepak, dobro nepřichází **vniveč**, spíš zlo může zmizet, ztratit se v něčí duši, zlo lze odpustit, zapomenout, ale dobro se podle všeho neodpouští.	Нет, нет, добро не пропадает, скорее зло может **пропасть**, сгинуть в чьей-то душе, зло можно простить, забыть, а добро, оказывается, не прощают.
Wells-Valka_ svetu	Všechna naše práce **vniveč**, všechna ta práce...	Все наши труды **пропали**, все труды...

Tab. 10. Examples of translation of the word combinations with the word «vniveč» in the Intercorp corpus

Use of parallel corpora, on the one hand, shows which of the multi-variant phraseological equivalents are most often used by translators, and, on the other hand, it can enrich dictionaries and text books.

5 ANALYSIS OF PHRASEOLOGICAL UNITS USAGE IN DIACHRONY

It is common knowledge that the language is a dynamic system, which shall be reflected in dictionaries and grammars. However, this chronological aspect is far less common for the study of phraseological units and other set phrases. One of the reasons is the absence (till the recent times) of large historical (diachronic) corpora.

When we tried to study the behaviour of some phraseological units in time based on the National Russian Corpus (`http://ruscorpora.ru`), the experiments showed that its volume (283 million tokens) was too small for such tasks.

Fortunately, there is a large diachronic corpus Google Books (books for the period from 1800 to 2008) that exists for the Russian language (and eight others). The volume of the Russian corpus is 67 billion tokens. Let us show the results which can be obtained using it. Unfortunately, neither Czech, nor Slovak are included.

Example 9: *кошки скребут*

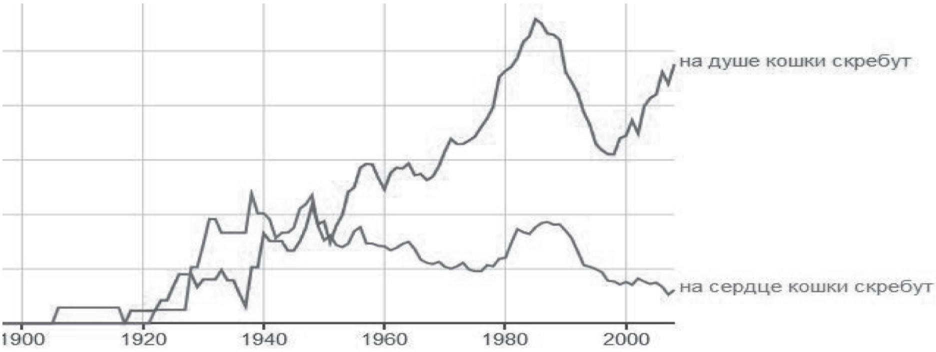


Fig. 4. The curves of occurrence of word combinations with the expression «кошки скребут» in the corpus Google books Ngram Viewer

The curves at Fig. 4 show that approximately up until the end of the 1940s the phraseological unit «на сердце кошки скребут» was used more often. Then, the situation changed drastically. And we see that the majority of the occurrences account for hard 1980–1990s.

Example 10: *ничтоже сумняшеся* (without a moment's hesitation)

The dictionaries present this phraseological unit in two forms: «ничтоже сумняшеся» and «ничтоже сумняся». Historians are well acquainted with this expression. But it is the corpus that will tell us how this expression "lived" in the language for centuries.

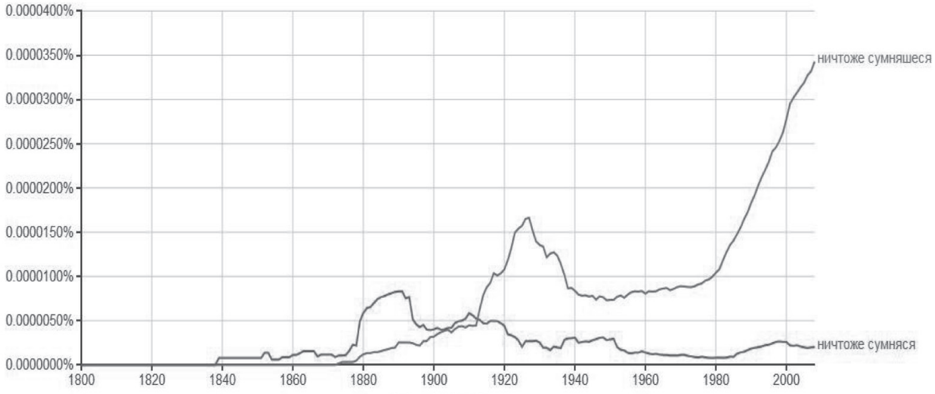


Fig. 5. The curves of occurrence of the bigram «ничтоже сумняшеся» in the corpus of Google books Ngram Viewer

We see in Fig. 5 that for quite a long time, the main form was «ничтоже сумняся». For example, in 1889 this form was used in literature 4 times more often.

It is no coincidence that we see this very form in Chekhov's books. «Для нее ясна была эта красивая смелость современного человека, с какою он, не задумываясь и ничтоже сумняся, решает большие вопросы и строит окончательные выводы» (A.P. Chekhov. «Несчастье» (Misfortune)). The variant «ничтоже сумняшеся» became preferred as late as in the second decade of the 20th century.

A search in the Google Ngram Viewer allows to identify other word combinations with the word «ничтоже», which can also be of interest for linguists: «ничтоже есть», «ничтоже суть», «ничтоже бысть».

Example 11: *перебиваться с ... на ...* (in the meaning "live in great poverty").

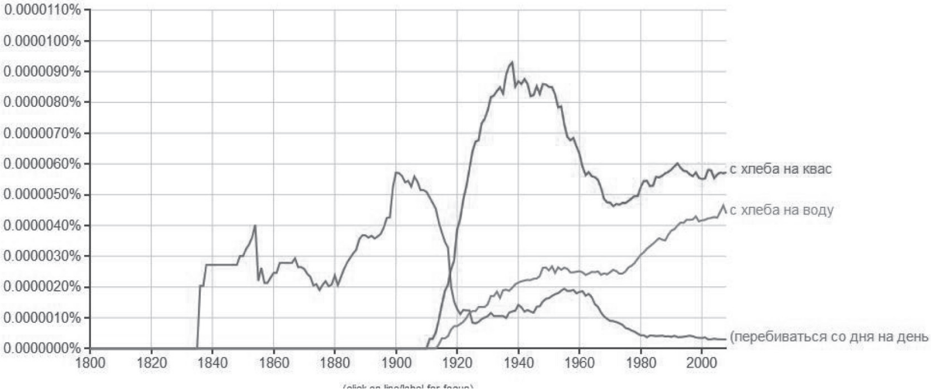


Fig. 6. The curves of occurrence of set phrases with the word «перебиваться» in the corpus of Google books Ngram Viewer

The curves of Fig. 6 show us that the most frequent set phrases with the verb «перебиваться» are those that are given in phraseological dictionaries (*с хлеба на квас, с хлеба на воду*) and that became used actively as late as in the 20th century.

6 CONCLUSION AND FURTHER WORK

The phraseology of any language is rich and variable. However, in order to see all this variability, we need large corpora, taking into account the relatively low frequency of usage of phraseological units in texts. Fortunately, for Russian, Czech, and Slovak, such corpora exist.

The research has shown that the corpus linguistics tools and corpora allow to identify and significantly enlarge the lexicon of set phrases of various types and peculiarities of their functioning. Based on corpora, linguists can create dictionaries and text books of a new generation, including phraseological dictionaries where the collocability will be represented far more widely than ever before. It is desirable that such dictionaries had such quantitative characteristics as the association strength in synchrony, and the history of usage in diachrony.

It is feasible to continue the research by choosing for experiments various types of phraseological units and, possibly, including other Slavic language. During the research, we also found repeatedly that in order to make credible conclusions based on corpus data one should be aware of the disadvantages and the limitations of the tools used.

ACKNOWLEDGEMENTS

This work was partially supported by the Russian Foundation of Basic Research (the Foundation for Humanities) as a part of the research project No. 17-04-00552 «Modern Russian Literary Language: Parametric Modelling of Lexical System».

References

[1] Avramovová, M et al. (2006–...). Slovník súčasného slovenského jazyka. Jarošová, A., editor, Veda, vydavateľstvo SAV, Bratislava.

[2] Benko, V. (2014). Aranea: Yet another family of (comparable) web corpora. In *Proceedings of the 17th International Conference Text, Speech and Dialogue,* pages 257–264, Springer International Publishing Switzerland (LNCS 8655).

[3] Birikh, A. K., Mokiyenko, V. M., and Stepanova, L. I. (1997). *Slovar' frazeologicheskikh sinonimov russkogo yazyka*. [Dictionary of phraseological synonyms of the Russian language.] Rostov-on-Don.

[4] Chernysheva, I. I. (1970). *Frazeologiya sovremennogo nemetskogo yazyka*. [Phraseology of the modern German.] Moscow.

[5] Denisov, P. N. and Morkovkin, V. V., editors (1983). *Slovar' sochetayemosti slov russkogo yazyka.* [Collocability dictionary of Russian language words.] Russkiy yazyk, Moscow.

[6] Čermák, F. and Hronek, J. (1994). *Slovník české frazeologie a idiomatiky. Výrazy slovesné*. Academia, Praha.

[7] Melčuk, I. A. (1960). *O terminakh 'ustoyvhivost'' i 'idiomatichnost''*. [About the terms steadiness and idiomaticity.] *Voprosy yazykoznaniya* [Questions of Linguistics], 4:73–80.

[8] Michel, J-B. et al. (2011). Quantitative analysis of culture using millions of digitized books. *Science* 331:176; DOI 1126/Science. 1199644. Accessible at: `http://www.sciencemag.org/content/331/6014/176.full.html`, retrieved 2017-01-30.

[9] Mokienko, V. and Wurm, A. (2002). *Česko-ruský frazeologický slovník*. Olomouc.

[10] Rajnochová, N., Runštuková, N., and Vavřín, M. (2016). Korpus InterCorp – ruština. Verze 9.9. Ústav Českého národního korpusu FF UK. Praha. Accessible at: `http://www.korpus.cz/`.

[11] *Russko-cheshskiy slovar*. [Russian-Czech dictionary.] (1978). Moskva – Praha.

[12] *Slovatsko-russkiy slovar*. [Slovak-Russian dictionary.] (1976). Bratislava – Moskva.

[13] Slovenský národný korpus – prim-7.0-public-all. Jazykovedný ústav Ľ. Štúra SAV, Bratislava. 2015. Accessible at: `http://korpus.juls.savba.sk/`.

[14] Zakharov, V. P. and Masevich, A. Ts. (2014). Diakhronicheskiye issledovaniya na osnove korpusa russkikh tekstov Google books Ngram Viewer [Diachronic researches on the base of the Russian Google books Ngram Viewer text corpus.] *Strtuctural and Applied Linguistics* [Strukturnaya i prikladnaya lingvistika], 10:303–327.

SLOVAK DEPENDENCY TREEBANK IN UNIVERSAL DEPENDENCIES

DANIEL ZEMAN
Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic

ZEMAN, Daniel: Slovak Dependency Treebank in Universal Dependencies. Journal of Linguistics, 2017, Vol. 68, No 2, pp. 385 – 395.

Abstract: We describe a conversion of the syntactically annotated part of the Slovak National Corpus into the annotation scheme known as Universal Dependencies. Only a small subset of the data has been converted so far; yet it is the first Slovak treebank that is publicly available for research. We list a number of research projects in which the dataset has been used so far, including the first parsing results.

Keywords: treebank, dependency, universal dependencies, syntax, morphology, tagging, parsing

1 INTRODUCTION

Syntactically annotated corpora (treebanks) are important language resources, indispensable for linguistic research and natural language processing alike. Modern treebanks are mostly built on the notion of dependency relations. With the increasing number of languages covered and amount of data available, there is a growing interest in finding one common, linguistically adequate and cross-linguistically applicable annotation style [5, 13]. Universal Dependencies (UD)[1] [6] is an international effort aimed at such an annotation standard; at the same time, UD also releases treebanks annotated according to the UD guidelines, and has arguably become the largest collection of freely available dependency treebanks worldwide.

UD treebanks are released twice a year and every release so far added several languages that had not been part of the previous releases. The group of Slavic languages is represented quite well. [12] gave an early account of Slavic languages in UD 1.1, as well as an overview of other Slavic treebanks outside UD (Table 1). At the time of this writing, UD 2.0 is the most recent release and it comprises 70 treebanks of 50 languages; among them, nearly all[2] Slavic languages are represented with at least a small dataset (Table 2).

In the present article we focus on one of the recent additions to UD, the Slovak Dependency Treebank.

[1] `http://universaldependencies.org/`

[2] Serbian and Upper Sorbian are ready to be released in UD 2.1. What remains missing is Lower Sorbian, Bosnian/Montenegrin and Macedonian; and one may also argue for some smaller languages with less clear status such as Kashubian or Rusyn.

DE GRUYTER OPEN
DOI 10.1515/jazcas-2017-0048

Language	Code	Treebank	Sent	Tok
Bulgarian	[bg]	BulTreeBank	13,221	196K
Church Slavonic	[cu]	PROIEL	7,818	72K
Croatian	[hr]	SETimes.HR	3,736	84K
Czech	[cs]	PDT	87,913	1504K
Polish	[pl]	IPI PAN	8,227	84K
Russian	[ru]	SynTagRus	63,000	900K
Slovak	[sl]	SNK	63,238	994K
Slovenian	[sl]	SSJ500K	27,829	500K

Tab. 1. Dependency treebanks of Slavic languages, as listed by [12] (only some of them were converted to UD at that time)

2 SLOVAK DEPENDENCY TREEBANK

The data in the Slovak treebank come from the Slovak National Corpus (*Slovenský národný korpus,* SNK)[3] [8]. Over 63,000 sentences (almost one million words) received manual morphological and syntactic annotation, making it one of the three largest treebanks of Slavic languages (after the Czech PDT [1],[4] and with similar size to the Russian SynTagRus [2]). [8] describe the composition of the treebank as 78% fiction, 13% scientific and 9% journalistic text. An important point is that it includes free sources like Wikipedia or folk tales, where intellectual property rights do not complicate access to and distribution of the annotated data. Most sentences were independently annotated by two annotators in order to identify difficult phenomena and reduce annotation errors. The positions where the two annotators disagree would eventually be decided by a third annotator. Unfortunately, this final step has not been completed for all the sentences, which also means that the treebank has yet to wait for its full official release.[5]

Language	Code	Treebank	Sent	Tok
Belarusian	[be]	UD	393	8K
Bulgarian	[bg]	BulTreeBank	11,138	156K
Church Slavonic	[cu]	PROIEL	6,337	58K
Croatian	[hr]	SETimes.HR	8,889	197K
Czech	[cs]	PDT	87,913	1506K
Czech	[cs]	CAC	24,709	494K
Czech	[cs]	CLTT	1,125	38K
Czech	[cs]	PUD	1,000	19K
Polish	[pl]	IPI PAN	8,227	84K
Russian	[ru]	Google	5,030	99K
Russian	[ru]	SynTagRus	61,889	1107K

[3] `http://korpus.juls.savba.sk/`

[4] `http://ufal.mff.cuni.cz/pdt`

[5] The Slovak Language Treebank has been listed in META-SHARE since 2011 (`http://metashare.korpus.sk/repository/browse/slovak-treebank/36e46d0a649311e292cd00163e00007874586ecb0acd48909e54babd7c5e7bc2/`) but it cannot be downloaded from there.

Russian	[ru]	PUD	1,000	19K
Serbian*	[sr]	SETimes.SR	3,891	87K
Slovak	[sl]	SNK	10,604	106K
Slovenian	[sl]	SSJ200K	8,000	141K
Slovenian	[sl]	SST	3,188	29K
Ukrainian	[uk]	UD	1,706	26K
Upper Sorbian	[hsb]	UD	646	11K

Tab. 2. Slavic treebanks in UD release 2.0 (plus Serbian, scheduled for UD 2.1). Note that there were two UD 2.0 releases and the counts in this table sum up both: First, training and development data were released in March 2017. The test sets were kept aside for the CoNLL 2017 Shared Task in dependency parsing, and they were released in May 2017 after the shared task.

Morphological annotation in SNK assigns to each word (token) its lemma and a morphological tag that encodes its part of speech and values of relevant morphological features: inflection type, gender, number, case, degree of comparison, agglutination (preposition + pronoun), verbal form, aspect, polarity, voice etc.[6]

The syntactic annotation follows the annotation guidelines of the "analytical layer" of the Prague Dependency Treebank.[7]

The following steps have been taken to ensure quality of the data. Note that these are filtering steps – on the first sign of a problem, the entire unit (sentence or file) is discarded. In most cases it should be possible to manually fix the problem and retain the sentence; however, the obvious short-term advantage of the filtering approach is that it requires fewer human resources and the problem-free part of the data can be made available sooner.

- Removed files where the morphological annotation was not manual.
- Removed files where the syntactic annotation was done only by one annotator.
- Removed files where the annotators disagree in sentence segmentation (different number of sentences).
- Removed sentences where the annotators disagree in tokenization (different number of tokens).
- Removed empty sentences and sentences with just one token.
- Removed sentences where one or more annotation items (lemmas, morphological tags, dependency relation labels) were empty.

The resulting corpus consisted of 40,350 sentences and 671,968 tokens. Every sentence in this data set had two complete dependency trees from two annotators. In general, the contrasted annotations were not expected to differ on the word level and in morphological annotation because the annotators were focusing on syntax (while morphology was inherited from pre-existing annotation of SNK). However, it seems that they occasionally modified the lower layers: there were 747 mismatches in word forms, 2 in lemmas and 3 in morphological tags. Again, all affected sentences were removed.

[6] See `http://korpus.juls.savba.sk/attachments/morpho/tagset-www.pdf` for documentation of the tagset (in Slovak).

[7] See `http://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/a-layer/html/index.html` for documentation.

As for the syntactic annotation, the two annotators agreed on 80.34% dependencies (both the parent node index and the dependency label). If we disregard the dependency labels and only look at the parent node assignment, the agreement rate rises to 87.31%. It means that in 6.97% of all tokens (35% of dependency errors) the sole disagreement is in the label (termed *analytical function* or *afun* in Prague-style treebanks).

Finally, let us consider “complete matches” – sentences whose dependency trees from the two annotators were identical in all aspects of annotation. These sentences constitute the most trustworthy core of the corpus, as it is unlikely that two annotators independently make the same error. Only completely matching sentences were selected for the first UD release of the treebank: reliability of the annotation got the top priority. Of course, there are also some downsides to this decision. The first of them is linked to filtering in general: the resulting corpus does not contain whole documents, making any discourse-level studies impossible. The second drawback is perhaps even more serious: the treebank contains a high proportion of short sentences because the more words in the sentence, the higher is the probability of an annotation error. Before removing sentences with annotation mismatches, the average sentence length in the treebank was 16.7 tokens. When only complete matches remained, the average length dropped to 10.0 tokens. (The longest completely matched sentence contained 54 tokens.) Such a corpus is unbalanced and some more complex grammatical structures may be seriously underrepresented in it. It is thus highly desirable to extend the corpus and add more sentences in the future. However, the filtered portion is arguably much better than nothing, and can be used to train statistical parsers for Slovak; with 10,604 sentences and 106,043 tokens it is still a medium-sized treebank, surpassing by an order of magnitude treebanks that are available for some other languages.

Given that there was no official download site for the Slovak treebank, the filtered part was first released, with the permission from the Ľudovít Štúr Institute of Linguistics, in the LINDAT/CLARIN digital library[8] [3]. This release retained the original Prague-style annotation before conversion to the UD standard.

3 CONVERSION TO UNIVERSAL DEPENDENCIES

The conversion of the annotation to the scheme defined in Universal Dependencies consists of two partially independent steps: 1. converting the morphological tags to universal POS tags and features, and 2. converting the dependency relations.

The Interset Perl library[9] [11] was used to convert the values of morphological categories to UD features; since the internal representation of Interset is defined as a kind of Interlingua for morphosyntactic tagsets, and because the features in UD are based directly on Interset, the conversion was rather straightforward. Most of the categories encoded in SNK tags were directly mappable to UD feature values, with the exception of *paradigma* (inflection class) for which there is no direct counterpart in UD.

[8] http://hdl.handle.net/11234/1-1822

[9] http://ufal.mff.cuni.cz/interset

The situation is less straightforward with part-of-speech categories (the first character of SNK tags). UD guidelines define 17 universal part-of-speech tags (UPOS) that are rather coarse-grained but assumed to be sufficient for any natural language. If more fine-grained distinctions are needed, they should be encoded by additional features (UPOS tags are separate from features in UD).

SNK	***Description***	***UPOS***	***Features***
S	noun	NOUN, PROPN	
A	adjective	ADJ	
P	pronoun	PRON, DET	
N	numeral	NUM	
V	verb	VERB, AUX	
G	participle	ADJ	VerbForm=Part
D	adverb	ADV	
E	preposition	ADP	
O	conjunction	CCONJ, SCONJ	
T	particle	PART	
J	interjection	INTJ	
R	reflexive pronoun	PRON	Reflex=Yes
Y	conditional morpheme	AUX	Mood=Cnd
W	abbreviation	X	Abbr=Yes
Z	punctuation	PUNCT	
Q	unidentifiable	X	Hyph=Yes
#	non-word element	X	
%	citation in foreign language	X	Foreign=Yes
0	digit	NUM	NumForm=Digit

Tab. 3. Correspondence between SNK POS tags and UPOS (universal part-of-speech tags)

Table 3 shows the correspondences between SNK POS tags and UPOS. Certain ambiguities are relatively easy to solve – for example, common and proper nouns are distinguished by subsequent characters in the SNK tag. Other ambiguities cannot be resolved by looking at tags alone, and they are addressed outside Interset, taking also the word and its lemma into account. Thus all *pro-adjectives* (pronouns inflecting and behaving like adjectives) are listed and re-tagged DET in UD (see [12] for a discussion of pronouns vs. determiners in Slavic languages). The feature PronType (pronominal type) is set for all pronouns, determiners and pronominal adverbs. Ordinal and multiplicative numerals are distinguished from cardinals by the feature NumType and by changing their tag to ADJ or ADV (the NUM tag is reserved for definite cardinal numbers). Similarly, a word list is used to distinguish coordinating and subordinating conjunctions.

Another change involves polarity of verbs. In SNK, the negative forms with the prefix *ne-* are treated as derivational morphology: they are not encoded in the morphological tags and negative verbs have different lemmas than their affirmative counterparts (e.g. *obviniť* “to accuse” – *neobviniť* “not to accuse”). The Slovak UD data, on the other hand, use the affirmative lemma for both forms and set the Polarity feature to either “Pos” or “Neg”. This is in line with the UD guidelines and improves parallelism to the Czech treebanks in UD. Note that negative verbs can be recognized using simple regular expressions, but one must watch for a few exceptions where an affirmative verb starts with *ne- (nechať, nechávať, nenávidieť)*.

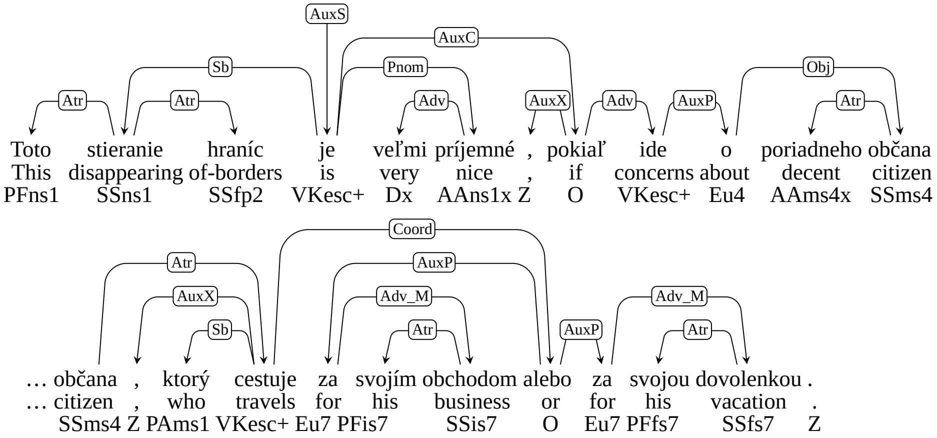

AuxS
AuxC
Sb
Pnom
Obj
Atr
Atr
Adv
AuxX
Adv
AuxP
Atr
Toto stieranie hraníc je veľmi príjemné , pokiaľ ide o poriadneho občana
This disappearing of-borders is very nice , if concerns about decent citizen
PFns1 SSns1 SSfp2 VKesc+ Dx AAns1x Z O VKesc+ Eu4 AAms4x SSms4
Coord
Atr
AuxP
AuxX
Adv_M
Sb
Atr
AuxP
Adv_M
Atr
… občana , ktorý cestuje za svojím obchodom alebo za svojou dovolenkou .
… citizen , who travels for his business or for his vacation .
SSms4 Z PAms1 VKesc+ Eu7 PFis7 SSis7 O Eu7 PFfs7 SSfs7 Z

Fig. 1. Example of an original Prague-style dependency tree

Further part-of-speech adjustments occur during the transformation of the syntactic structure. For instance, the verb *byť* “to be” usually functions as an auxiliary verb or a copula. If it is found in one of these functions, its tag is changed from VERB to AUX.

Conversion of syntactic annotation is illustrated in Figures 1 and 2. Besides simple relabeling of dependency relations (see also Table 4), it involves several structural transformations:

- Copula verb heads the non-verbal predicate in the Prague style while the non-verbal predicate is the head in UD: *je príjemné* “is nice”.
- In Prague, preposition is plugged as a connector between its noun and the parent of the prepositional phrase. In UD, prepositions are leaves attached to their nouns: *o občana* “about citizen”, *za obchodom* “for business”.
- In Prague, subordinating conjunction is plugged as a connector between the predicate of the subordinate clause and its parent. In UD, subordinating conjunctions are leaves attached to the predicates: *pokiaľ ide* “if it concerns”.
- In Prague, coordination is headed by a conjunction or punctuation symbol; the child nodes are marked as either members of coordination (“_M” attached to afun) or modifiers shared by the members (no suffix). In UD, coordination is headed by the first conjunct (member) and the subsequent conjuncts are attached to it. Shared modifiers cannot be distinguished from private modifiers of the first conjunct.

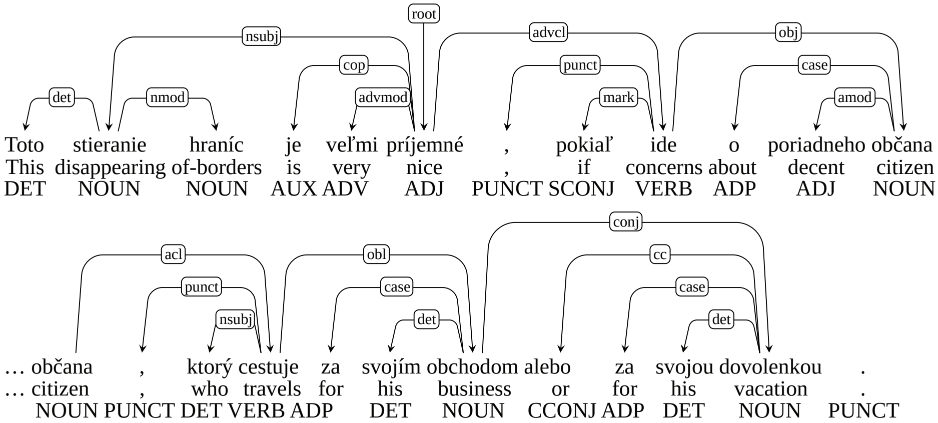

root
nsubj
advcl
obj
cop
punct
case
det
nmod
advmod
mark
amod
Toto stieranie hraníc je veľmi príjemné , pokiaľ ide o poriadneho občana
This disappearing of-borders is very nice , if concerns about decent citizen
DET NOUN NOUN AUX ADV ADJ PUNCT SCONJ VERB ADP ADJ NOUN
conj
acl
obl
cc
punct
case
case
nsubj
det
det
… občana , ktorý cestuje za svojím obchodom alebo za svojou dovolenkou .
… citizen , who travels for his business or for his vacation .
NOUN PUNCT DET VERB ADP DET NOUN CCONJ ADP DET NOUN PUNCT

Fig. 2. The tree from Figure 1 converted to Universal Dependencies

SNK	***Description***	***UD***
Adv	adverbial modifier	obl, advmod, advcl
Apos	apposition	appos*, punct*
Atr	attribute	amod, det, nummod, nmod, flat, acl
Atv	verbal attribute	acl
AtvV	verbal attribute	xcomp
AuxC	subordinating conjunction	mark*
AuxG	non-comma punctuation	punct
AuxK	sentence-final punctuation	punct*
AuxO	semantically redundant	discourse
AuxP	preposition	case*
AuxR	reflexive passive	expl:pass
AuxT	inherently reflexive verbs	expl:pv
AuxV	auxiliary verb	aux, aux:pass
AuxX	comma	punct
AuxY	extra conjunction	cc, mark
AuxZ	emphasizer	advmod:emph
Coord	coordination head	cc*, conj*, punct*
ExD	ex-dependent (ellipsis)	vocative, advcl, orphan*, dep
Obj	object	obj, iobj, ccomp, xcomp
Pnom	nominal predicate	cop*
Pred	main predicate	root, parataxis
Sb	subject	nsubj, nsubj:pass, csubj, csubj:pass

Tab. 4. Correspondence between SNK (Prague style) and UD dependency relations. The correspondences marked * are indirect: a structural transformation is necessary when the source relation occurs; the target relation may appear in the resulting structure but it will hold between a different pair of nodes.

The conversion procedure is not trivial because sometimes the rules outlined above interact. Notice how coordination is combined with prepositional phrases in our example—in the Prague style, the real type of the relation between *cestuje* and *za obchodom*, “Adv”, is revealed two levels lower than in the UD tree. Fortunately, there was already software for conversion between the Prague style and UD. The publicly available Treex package[10] [7] in the configuration described in [13] (with some extensions) was reused to convert the Slovak treebank.

The Slovak UD treebank first appeared in the Universal Dependencies release 1.4 in November 2016. In order to facilitate reproducibility of machine learning experiments, the dataset was split to training, development and test section, respectively (Table 5).

Section	***Sentences***	***Tokens***
Training	8,483	80,575
Development	1,060	12,440
Test	1,061	13,028

Tab. 5. The official split of the Slovak UD treebank into training, development and test data

The second edition that included Slovak, UD 2.0 in March 2017, followed the updated version of the UD guidelines, v2. (All examples in the present article also relate to the v2 guidelines.) The data split was the same as in UD 1.4 but the test sets were released separately after the CoNLL 2017 shared task in dependency parsing.

4 USAGE AND RELATED WORK

The mere fact that the treebank is available under a free license is very important for the Slovak language in the field of natural language processing. Being a part of a large collection like Universal Dependencies is a bonus that significantly increases visibility of the corpus. According to the statistics published by LINDAT/CLARIN, there have been 79 unique downloads of the Prague-style release of the Slovak Dependency Treebank, 2,592 unique downloads of UD 1.4 and 1985 unique downloads of UD 2.0 (as of July 19, 2017).

The treebank can be searched on-line in the PML-TQ search engine maintained by the Charles University in Prague[11] and in the SETS engine at the University of Turku.[12]

Soon after its first release, the treebank was picked (together with Czech, Slovenian, Croatian, Danish, Swedish and Norwegian) by the organizers of the VarDial 2017 shared task in parsing closely related languages [10]. A much larger shared task was organized as part of the CoNLL 2017 conference[13] [14]. The topic was end-to-end parsing from raw text, via automatic tokenization, sentence segmentation, lemmatization and morphological tagging to universal dependencies. The task set the new state of the art in dependency parsing for 45 languages, including Slovak. Baseline models were produced by the UDPipe system[14] [9]; this

[10] `http://ufal.mff.cuni.cz/treex`

[11] `https://lindat.mff.cuni.cz/services/pmltq/#!/treebank/ud20_sk/`

[12] `http://bionlp-www.utu.fi/dep_search/`

[13] `http://universaldependencies.org/conll17/`

[14] `http://ufal.mff.cuni.cz/udpipe`

parser is open-source and available together with the pre-trained language models. Twenty of the systems competing in the shared task managed to surpass the baseline result; some of them are freely available, too.

The best results for Slovak were achieved by the team from Stanford: 83.86% content-word labeled attachment score (CLAS), **86.04%** labeled attachment score (LAS) and 89.58% unlabeled attachment score (UAS). The parser was processing raw text (that is, it could not access gold-standard sentence segmentation, tokenization and morphology). All models were only trained on the training portion of the Slovak UD treebank. However, since much larger tagged data are available in the Slovak National Corpus, there is room for a significant boost of the tagging accuracy, which in turn may improve parsing results (but note that some parsers do not need morphology on input).

In connection to the shared task, large web corpora have been collected from CommonCrawl and Wikipedia for all the languages. The data have been automatically segmented, lemmatized, tagged and parsed by UDPipe, so there is now also a parsebank of Slovak comprising over 59 million sentences (811 million words) [4]. The first 2 million words have been indexed and made searchable through the SETS engine in Turku.

5 CONCLUSION AND OUTLOOK

We have presented the first public release of the Slovak Dependency Treebank and its automatic conversion to Universal Dependencies using rule-based heuristics and correspondence tables. We have shown that the release practically immediately put Slovak in several interesting NLP research projects where multilingual approaches are studied.

The current version contains only sentences with 100% inter-annotator agreement. This temporary measure ensures quality of the syntactic annotation but it also means that the released dataset is relatively small and unbalanced. It may not be easy to find human resources and funds to complete the disagreement resolution in the near future; however, we believe that there is room for checking additional sentences semi-automatically.

There are 7,564 sentences (95K tokens) where there was just one disagreement point between the annotators. Some of the mismatches mentioned in Section 2 may not be important for UD conversion or may be easily fixable. [8] notice that one of the most frequently confused pair of relations is `AuxT` (reflexive pronoun of an inherently reflexive verb) and `AuxR` (reflexive pronoun used to form reflexive passive). Both of them would be subtypes of `expl` in UD. Another frequent mismatch is `AuxX` vs. `Coord`. More research would be needed but if it signals inconsistent encoding of coordination, it could be normalized automatically.

Observations of this kind will hopefully help to speed up the completion of the remaining annotated data. Once all of them are added, the future releases of the Slovak Dependency Treebank will be four times bigger than the current one.

ACKNOWLEDGEMENTS

The work was supported by the grant 15-10472S of the Czech Science Foundation. We have been using language resources and tools developed, stored and distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2015071).

The author wishes to thank Mária Šimková, Katarína Gajdošová and the entire team of the Ľudovít Štúr Institute of Linguistics for their work on the treebank and for the permission to release it.

References

[1] Bejček, E., Hajičová, E., Hajič, J., Jínová, P., Kettnerová, V., Mikulová, M., Mírovský, J., Nedoluzhko, A., Panevová, J., Poláková, L., Ševčíková, M., Štěpánek, J., and Zikánová, Š. (2013). Prague dependency treebank 3.0. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University. Accessible at: `http://hdl.handle.net/11858/00-097C-0000-0023-1AAF-3.`

[2] Boguslavsky, I., Iomdin, L., Petrochenkov, V., Sizov, V., and Tsinman, L. (2013). A case of hybrid parsing: Rules refined by empirical and corpus statistics. In Gerdes, K., Hajičová, E., and Wanner, L., editors, *Computational Dependency Theory*, volume 258, pages 226–240, IOS Press, Amsterdam, Netherlands.

[3] Gajdošová, K., Šimková, M., et al. (2016). Slovak dependency treebank. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University. Accessible at: `http://hdl.handle.net/11234/1-1822.`

[4] Ginter, F., Hajič, J., Luotolahti, J., Straka, M., and Zeman, D. (2017). CoNLL 2017 shared task – automatically annotated raw texts and word embeddings. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University. Accessible at: `http://hdl.handle.net/11234/1-1989.`

[5] McDonald, R., Nivre, J., Quirmbach-Brundage, Y., Goldberg, Y., Das, D., Ganchev, K., Hall, K., Petrov, S., Zhang, H., Täckström, O., Bedini, C., Bertomeu Castelló, N., and Lee, J. (2013). Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 92–97, Sofija, Bulgaria.

[6] Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., and Zeman, D. (2016). Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1659–1666, Portorož, Slovenia.

[7] Popel, M. and Žabokrtský, Z. (2010). TectoMT: modular NLP framework. In *Advances in Natural Language Processing: 7th International Conference on NLP, IceTAL 2010*, pages 293–304, Springer, Berlin Heidelberg, Germany.

[8] Šimková, M. and Garabík, R. (2006). Синтаксическая разметка в Словацком национальном корпусе. In *Труды международной конференции Корпусная лингвистика – 2006*, pages 389–394, St. Petersburg University Press, Russia.

[9] Straka, M., Hajič, J., and Straková, J. (2016). UDPipe: trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, pages 4290–4297, ELRA, Portorož, Slovenia.

[10] Zampieri, M., Malmasi, S., Ljubešić, N., Nakov, P., Ali, A., Tiedemann, J., Scherrer, Y., and Aepli, N. (2017). Findings of the VarDial evaluation campaign 2017. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, Valencia, Spain.

[11] Zeman, D. (2008). Reusable tagset conversion using tagset drivers. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., and Tapias, D., editors, *Proceedings of the 6th*

International Language Resources and Evaluation Conference, LREC 2008, pages 28–30, Marrakech, Morocco.

[12] Zeman, D. (2015). Slavic languages in Universal Dependencies. In *Natural Language Processing, Corpus Linguistics, E-learning*, pages 151–163, RAM-Verlag, Lüdenscheid, Germany.

[13] Zeman, D., Dušek, O., Mareček, D., Popel, M., Ramasamy, L., Štěpánek, J., Žabokrtský, Z., and Hajič, J. (2014). HamleDT: Harmonized multi-language dependency treebank. *Language Resources and Evaluation*, 48(4):601–637.

[14] Zeman, D., Popel, M., Straka, M., Hajič, J., Nivre, J., Ginter, F., Luotolahti, J., Pyysalo, S., Petrov, S., Potthast, M., Tyers, F., Badmaeva, E., Gökırmak, M., Nedoluzhko, A., Cinková, S., Hajič jr., J., Hlaváčová, J., Kettnerová, V., Urešová, Z., Kanerva, J., Ojala, S., Missilä, A., Manning, C., Schuster, S., Reddy, S., Taji, D., Habash, N., Leung, H., de Marneffe, M.-C., Sanguinetti, M., Simi, M., Kanayama, H., de Paiva, V., Droganova, K., Martínez Alonso, H., Çöltekin, Ç., Sulubacak, U., Uszkoreit, H., Macketanz, V., Burchardt, A., Harris, K., Marheinecke, K., Rehm, G., Kayadelen, T., Attia, M., Elkahky, A., Yu, Z., Pitler, E., Lertpradit, S., Mandl, M., Kirchner, J., Fernandez Alcalde, H., Strnadová, J., Banerjee, E., Manurung, R., Stella, A., Shimada, A., Kwak, S., Mendonça, G., Lando, T., Nitisaroj, R., and Li, J. (2017). CoNLL 2017 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Association for Computational Linguistics. Vancouver, Canada.

COMPOUND ADVERBS AS AN ISSUE IN MACHINE ANALYSIS OF CZECH LANGUAGE

HANA ŽIŽKOVÁ
Faculty of Arts, Masaryk University, Brno, Czech Republic

ŽIŽKOVÁ, Hana: Compound Adverbs as an Issue in Machine Analysis of Czech language. Journal of Linguistics, 2017, Vol. 68, No 2, pp. 396 – 403.

Abstract: Compound adverbs represent an interesting issue in terms of Automatic Morphological Analysis (AMA). The reason is that compound adverbs in Czech are expressions formed by compounding existing words that are different parts of speech without any change in their form. An indicative sign of compound adverbs is that they can always be decomposed again. Compound adverbs may be written as one word but sometimes a multiword form coexists. A word that is originally a different part of speech gains an adverbial meaning and becomes an adverb. This article presents the results of a corpus probe aimed at mapping expressions that are demonstrably compound adverbs and were not recognized by AMA or were incorrectly tagged by AMA as another part of speech. Analysis of data obtained from the Czech National Corpus (ČNK) SYN v3 show that the unrecognized and incorrectly tagged units can be divided into several groups. Based on knowledge of these groups it is possible to refine part of speech tagging by AMA. The corpus probe examined units written in accordance with the current codification as well as substandard units.

Keywords: compound adverb, multiword expression, automatic morphological analysis, nominal form, corpus, tag

1 INTRODUCTION

Compound adverbs represent an interesting issue in terms of Automatic Morphological Analysis (AMA). Following the examination carried out by Osolsobě [1, p. 189n] of adverbs formed from prepositional cases, the questions of which of the units unrecognized by AMA are compound adverbs and which of the units may possibly be called “a nominal form that can be a part of a compound adverb” were raised.

Recognition of compound adverbs by AMA is difficult, as for instance Dokulil [2, p. 22] shows: “compound adverbs are formed by compounding frequently occurring words in a sentence, without any change in their form. It is characteristic for them that you can always divide the compound adverb again.” For the purpose of this paper it is essential that we write compound adverbs mostly together as one word, but often in parallel compound adverbs there exists a multiword expression and their meaning is the same (*na příklad – například*) [3]. Additionally, a member of the multiword expression can function independently of this expression as a separate word [4, p. 164]. Multiword expressions can be “defined as expressions which are made up of at least two words and which can be syntactically and/or semantically idiosyncratic in nature. Moreover, they act as a single unit at some level of linguistic analysis” [5].

DOI 10.1515/jazcas-2017-0049

There are contexts in which we hesitate whether to use a one-word adverb or a multiword expression [3]. Another important feature of the compound adverbs is that when written as two (or more) words, it is not possible to insert another expression between the two words that could develop the unit (*například – na příklad*, but not **na dobrý příklad*).

It is obvious that words which were previously other parts of speech are gaining adverbial meaning and becoming adverbs. Adverbialisation is an ongoing process during which flexible forms of parts of speech are changing into non-flexible forms, in the adverbs. Knapová claimed [6] several conditions must be valid in order for adverbialisation to occur: adverbialisation occurs in such cases where the units appear in sentences in the same syntactic position as adverbs, the head is a verb and autonomy of the preposition also plays a role. Spellings of compound adverbs thus depends largely on the extent of adverbialisation as shown by Trávníček [7, p. 1447]. Trávníček claims that: "The concept of compound adverbs is important also for spelling, since we write compound adverbs mostly together, because in our linguistic consciousness they act as a single word."

It is important for the compound adverb to be recognized by AMA in both cases (as a single word and also as a multiword expression) regardless of how the codification determines the correct spelling of the compound adverb. Automatic Morphological Analysis takes place in three steps: the first step is tokenization – a division of word forms, the second step is the assignment of one, but usually more interpretations from the morphological dictionary and the third step is disambiguation, which means assigning an interpretation.

The AMA recognizes and correctly identifies such compound adverbs that are written as one word and are listed in the morphological dictionary. The problem arises mainly in cases of multiword expressions. As indicated by Sag, Baldwin, Bond, Copestake and Flickinger [8, p. 1], multiword expressions are a real problem in natural language processing and the seriousness of this problem is much greater than has traditionally been considered in the context of linguistics.

Our aim was to map expressions that are clearly compound adverbs and were not recognized by the AMA or the AMA incorrectly tagged them as a different part of speech. Candidates of compound adverbs were extracted in August 2016, at that time the corpus ČNK SYN v3 [9] was most suitable for our analysis. We analyzed the obtained data and investigated whether the unrecognized and incorrectly tagged units can be somehow characterized and divided into groups based on their common characteristics.

2 APPROACH

We proceeded in several steps:

1. First, we typed queries for an unrecognized part of speech to ČNK SYN v3 corpus. The queries were as follows:

```
[tag=”X.*” & lemma=”po.*”]
[tag=”X.*” & lemma=”zpod.*”]
[tag=”X.*” & lemma=”po.*u”]
```

```
[tag=”X.*” & lemma=”do.*”]
[tag=”X.*” & lemma=”k.*”]
[tag=”X.*” & lemma=”ob.*”]
[tag=”X.*” & lemma=”mezi.*”]
[tag=”X.*” & lemma=”na.*”]
[tag=”X.*” & lemma=”od.*”]
[tag=”X.*” & lemma=”o[^db].*”]
[tag=”X.*” & lemma=”pro.*”]
[tag=”X.*” & lemma=”před.*”]
[tag=”X.*” & lemma=”při.*”]
[tag=”X.*” & lemma=”s.*”]
[tag=”X.*” & lemma=”u.*”]
[tag=”X.*” & lemma=”v.*”]
[tag=”X.*” & lemma=”za.*”]
[tag=”X.*” & lemma=”z[^a].*”]
```

2. The obtained data was sorted out manually and grouped by the prefixes: **do-, k-/ku-, mezi-, na-, nad-, o-, ob-, od-, po-, pro-, před-, při-, s-/sou-, u-, v-, z-, zpod-, za-** and, then further divided by their endings as follows:

-o: originally an accusative ending

-u: originally a genitive ending appended after prepositions *do/od,* a dative ending appended after preposition *k/ku*, an accusative ending appended after preposition *na*, a local ending appended after preposition *v*

-e/-ě: originally a genitive ending appended after prepositions *do/od*, *z*, an accusative ending appended after preposition *na*, a local ending appended after preposition *v*

-a: originally a genitive ending appended afer prepositions *do/od/s,*

-y: originally an accusative sg. ending appended after prepositions *do/od/s* or an accusative pl. ending appended after preposition *na*

-ou

- **consonant**: originally an accusative sg. ending or genitive pl.

-i: originally a genitive sg. ending or local sg.

-é

-[eě]m/-ím/_ám

3. We also investigated whether the expressions found are listed in available dictionaries. We used DEBDict [10] to investigate this.

4. Afterwards, we were interested in whether or not the AMA recognized expressions that we had found as a single word with the designation [tag=”X.*”] if we respread them into multiword expressions. Then, if the AMA recognized the multiword expressions, we wanted to know what tag the AMA would assign to them. So, we searched through the corpus for multiword expressions of one-word compound adverbs that we had previously determined while processing the first step.

5. The next step was finding out which of these multiword expressions are listed in available dictionaries. Again, we used DEBDict [10].

2.1 Example

We give an illustrative example of how we proceeded to process the data obtained from corpus ČNK SYN v3. We chose the query [tag=”X.*” & lemma=”k.*”] as an example. The process was identical for all queries.

Query [tag=”X.*” & lemma=”k.*”]

A. -u ending

For the above ending, we found only four words, namely: *kdobru, kpředu, kstáru, kuposledku*. One-word forms are not listed in dictionaries, except *kpředu*, which is listed in *Slovník spisovného jazyka českého* [10].

The list of found expressions

kdobru, kpředu, kstáru, kuposledku

Multiword expressions

Multiword expressions *k dobru, ku posledku* were tagged by the AMA as a preposition and a noun (N), *k stáru* was tagged as a preposition and an adverb (D), and *k předu* was mostly tagged as a preposition and an adverb (D), but there also appeared a tag for a preposition and for a verb (V).

N	*k dobru, ku posledku*
D	*k stáru*
D/V	*k předu*

Tab. 1. Tagged parts of speech and multiword expressions

Multiword expressions listed in Dictionaries

Only two multiword expressions were listed in the dictionary: *k dobru* and *k stáru*, both were listed in Slovník české frazeologie a idiomatiky [11].

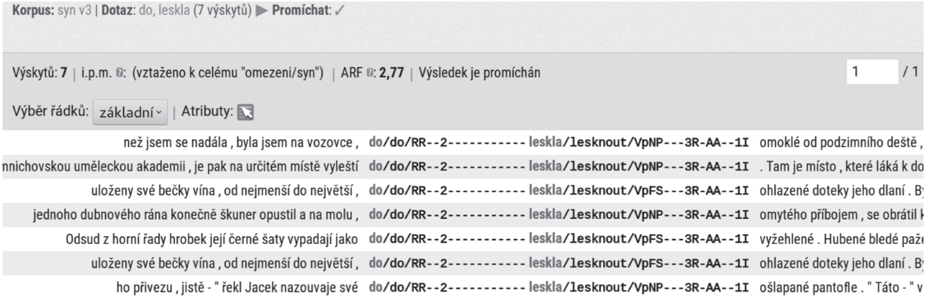
B. Other endings

In the remaining groups not a single unit was found which would correspond to the criteria of our investigation.

3 ANALYSIS

There is no specific tag for compound adverbs in ČNK SYN v3 corpus. Through detailed examination of the obtained data we found that when AMA recognizes the multiword expression, it is most often tagged as a preposition and such part of speech that the compound adverb was formed from. Mostly these are nouns, but there are also adjectives, adverbs or numerals and pronouns (eg. *nahromadě, kdobru, dogala, napřímo, poprvé, posvých, posvém, ...*).

We noticed several cases of incorrect tagging that did not follow the rule that in Czech a verb can not follow a preposition, eg. *do leskla* and *k předu* were both tagged as a preposition and a verb.



Korpus: syn v3 | **Dotaz:** do, leskla (7 výskytů) ▶ **Promíchat:** ✓

Výskytů: **7** | i.p.m.: (vztaženo k celému "omezeni/syn") | ARF: **2,77** | Výsledek je promíchán 1 / 1

Výběr řádků: základní | Atributy:

```
               než jsem se nadála , byla jsem na vozovce ,  do/do/RR--2---------- leskla/lesknout/VpNP---3R-AA--1I  omoklé od podzimního deště ,
nnichovskou uměleckou akademii , je pak na určitém místě vyleští  do/do/RR--2---------- leskla/lesknout/VpNP---3R-AA--1I  . Tam je místo , které láká k do
            uloženy své bečky vína , od nejmenší do největší ,  do/do/RR--2---------- leskla/lesknout/VpFS---3R-AA--1I  ohlazené doteky jeho dlaní . B
    jednoho dubnového rána konečně škuner opustil a na molu ,  do/do/RR--2---------- leskla/lesknout/VpNP---3R-AA--1I  omytého příbojem , se obrátil k
        Odsud z horní řady hrobek její černé šaty vypadají jako  do/do/RR--2---------- leskla/lesknout/VpFS---3R-AA--1I  vyžehlené . Hubené bledé paž
            uloženy své bečky vína , od nejmenší do největší ,  do/do/RR--2---------- leskla/lesknout/VpFS---3R-AA--1I  ohlazené doteky jeho dlaní . B
              ho přivezu , jistě - " řekl Jacek nazouvaje své  do/do/RR--2---------- leskla/lesknout/VpNP---3R-AA--1I  ošlapané pantofle . " Táto - " v
```

Fig. 1. Do leskla

Furthermore, we have found that part of the unrecognized units form semantic groups and belongs to commonly used vocabulary, for example:

- expressions meaning change of colour or quality (*dobíla, dorovna, dotvrda, natvrdo, namodro, ...*)
- expressions meaning to do something in a way / as somebody (*pochlapsku, poitalsku, poněmecku, podětsku, ...*)

We also noticed forms written in conflict with the current codification (*poitalsku, napamětnou, pocuď, ...*) and we included these in our study too, because it is important that the AMA recognizes the expression regardless of whether it is written in accordance with the current codification or not.

We discovered that an important part of the unrecognized one-word units is listed in one of the existing dictionaries (mostly in *Slovník spisovné češtiny* or in *Slovník spisovného jazyka českého*) and many of the multiword expressions are listed in *Slovník české frazeologie a idiomatiky* [11].

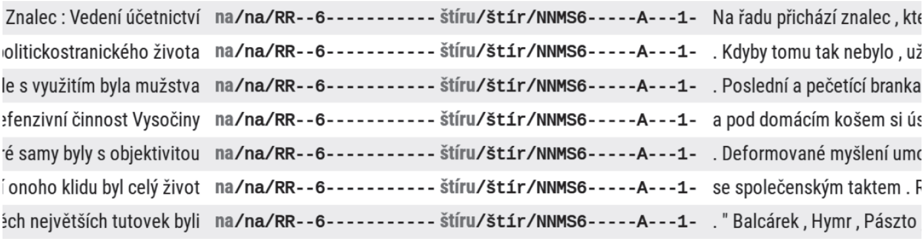
4 SUGGESTED SOLUTIONS

Thanks to the analysis, we identified five areas the implementation of which into AMA can refine compound adverbs recognition and reduce incorrect part of speech tagging.

The first area is the need for strict adherence to linguistic rules [12]. This will help to eliminate incorrect tagging of two consecutive parts of speech which cannot follow each other in Czech (e.g. a preposition cannot precede a verb).

The second area concerns the morphological tag set. We propose the introduction of a separate tag for a compound adverb or a nominal form that can be part of a compound adverb.

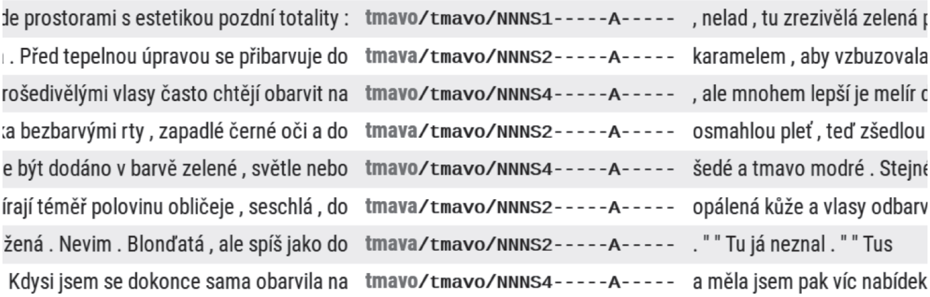
The third area deals with the utilization of the Automatic annotation of idioms and fixed collocations – FRANTA [13]. The pilot version of FRANTA, which primarily stems from *Slovník české frazeologie a idiomatiky* [11] and contains about 40,000 items, was released in December 2016 in ČNK SYN v4 corpus. The tagging could be improved if multiword compound adverbs which are listed in *Slovník české frazeologie a idiomatiky* [11] would also be tagged as compound adverbs. At the moment the part of speech tag does not differ, it only contains extra information about the idiom. FRANTA may also be of use in difficult cases where AMA recognizes and designates two-word units as a relocation of a preposition and p. ex. a noun. These two-word units can act as a preposition and a noun in certain contexts, but can have an adverbial meaning in other contexts (*do času, na koleně, na hromadě, z prázdna, ...*). We confirmed that some of these units are listed in *Slovník české frazeologie a idiomatiky* [11], which means utilization of the Automatic annotation of collocations and fixed collocations for these cases would also be possible. Also, most of the [tag=”X.*” & lemma=”on.*.u”] expressions are listed in *Slovník české frazeologie a idiomatiky* [11]. The expression *na štíru* caught our attention – we only registered this in the sense of *cannot do something well* and *do not like to deal with it* [11], but never in the sense of involving the animal Scorpion (*štír* in Czech).



Znalec : Vedení účetnictví	na/na/RR--6-----------	štíru/štír/NNMS6-----A---1-	Na řadu přichází znalec , kte
olitickostranického života	na/na/RR--6-----------	štíru/štír/NNMS6-----A---1-	. Kdyby tomu tak nebylo , už
le s využitím byla mužstva	na/na/RR--6-----------	štíru/štír/NNMS6-----A---1-	. Poslední a pečetící branka
efenzivní činnost Vysočiny	na/na/RR--6-----------	štíru/štír/NNMS6-----A---1-	a pod domácím košem si ús
é samy byly s objektivitou	na/na/RR--6-----------	štíru/štír/NNMS6-----A---1-	. Deformované myšlení umo
onoho klidu byl celý život	na/na/RR--6-----------	štíru/štír/NNMS6-----A---1-	se společenským taktem . F
ěch největších tutovek byli	na/na/RR--6-----------	štíru/štír/NNMS6-----A---1-	. " Balcárek , Hymr , Pászto

Fig. 2. Na štíru

The fourth area includes a very productive compounding by use of prepositions *do* and *na*. Most multiword expressions of the [tag=”X.*” & lemma=”do. *. a”] and [tag=”X.*” & lemma=”na.*.o”] groups were tagged by the AMA as a preposition and a noun. We found out that the corresponding lemma exists as a noun, but is never connected with the prepositions *do* or *na*. All of the expressions always have an adverbial meaning with these prepositions. The only exceptions are the lemmas *slovo* and *světlo*.

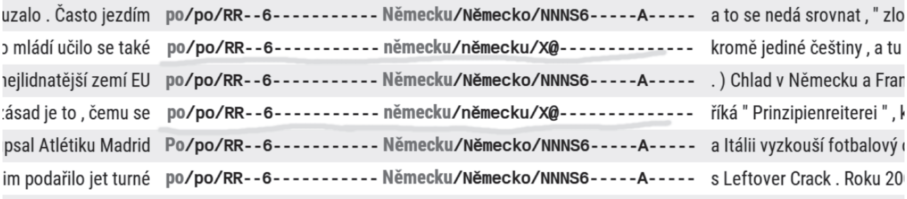


de prostorami s estetikou pozdní totality :	tmavo/tmavo/NNNS1-----A-----	, nelad , tu zrezivělá zelená p
. Před tepelnou úpravou se přibarvuje do	tmava/tmavo/NNNS2-----A-----	karamelem , aby vzbuzovala
rošedivělými vlasy často chtějí obarvit na	tmavo/tmavo/NNNS4-----A-----	, ale mnohem lepší je melír c
a bezbarvými rty , zapadlé černé oči a do	tmava/tmavo/NNNS2-----A-----	osmahlou pleť , teď zšedlou
e být dodáno v barvě zelené , světle nebo	tmavo/tmavo/NNNS4-----A-----	šedé a tmavo modré . Stejné
írají téměř polovinu obličeje , seschlá , do	tmava/tmavo/NNNS2-----A-----	opálená kůže a vlasy odbarv
žená . Nevim . Blonďatá , ale spíš jako do	tmava/tmavo/NNNS2-----A-----	. " " Tu já neznal . " " Tus
Kdysi jsem se dokonce sama obarvila na	tmavo/tmavo/NNNS4-----A-----	a měla jsem pak víc nabídek

Fig. 3. Do tmava, na tmavo

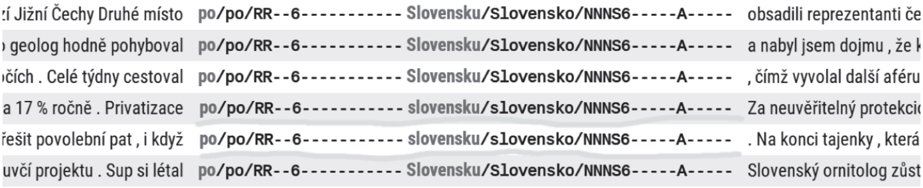
The same situation occurs when AMA tagged multiword expressions as a preposition and an adjective. Also these units always have an adverbial meaning and should be tagged as compound adverbs.

The fifth area contains fixed expressions that contain a geographic name. We discovered that, at the moment, there are cases where the uppercase of the initial letter of the lemma matters, but does not matter in other cases, although they are of the same type: the recognized noun exists in Czech and indicates a country *(po německu, po slovensku* X *po Německu, po Slovensku*).



uzalo . Často jezdím	po/po/RR--6-----------	Německu/Německo/NNNS6-----A-----	a to se nedá srovnat , " zlo
mládí učilo se také	po/po/RR--6-----------	německu/německu/X@--------------	kromě jediné češtiny , a tu
ejlidnatější zemí EU	po/po/RR--6-----------	Německu/Německo/NNNS6-----A-----	.) Chlad v Německu a Fran
ásad je to , čemu se	po/po/RR--6-----------	německu/německu/X@--------------	říká " Prinzipienreiterei " ,
psal Atlétiku Madrid	Po/po/RR--6-----------	Německu/Německo/NNNS6-----A-----	a Itálii vyzkouší fotbalový
im podařilo jet turné	po/po/RR--6-----------	Německu/Německo/NNNS6-----A-----	s Leftover Crack . Roku 20

Fig. 4. Po neměcku, po Německu

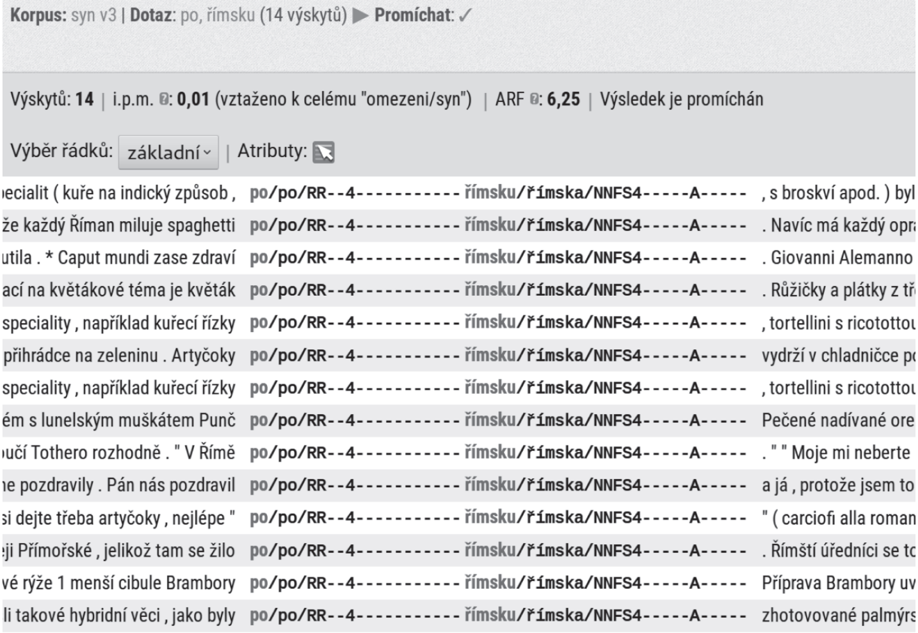


í Jižní Čechy Druhé místo	po/po/RR--6-----------	Slovensku/Slovensko/NNNS6-----A-----	obsadili reprezentanti če
geolog hodně pohyboval	po/po/RR--6-----------	Slovensku/Slovensko/NNNS6-----A-----	a nabyl jsem dojmu , že
čích . Celé týdny cestoval	po/po/RR--6-----------	Slovensku/Slovensko/NNNS6-----A-----	, čímž vyvolal další aféru
a 17 % ročně . Privatizace	po/po/RR--6-----------	slovensku/slovensko/NNNS6-----A-----	Za neuvěřitelný protekci
řešit povolební pat , i když	po/po/RR--6-----------	slovensku/slovensko/NNNS6-----A-----	. Na konci tajenky , která
uvčí projektu . Sup si létal	po/po/RR--6-----------	Slovensku/Slovensko/NNNS6-----A-----	Slovenský ornitolog zůst

Fig. 5. Po slovensku, po Slovensku

We assume that distinction of the case of the initial letter of the lemma would reduce incorrect tagging of the part of speech.

Related to this group we would like to note an interesting case where a multiword expression *po římsku* was assigned with the lemma *římska*, but there was no occurrence in the meaning *malá římsa = římska*.



Korpus: syn v3 | **Dotaz**: po, římsku (14 výskytů) ▶ **Promíchat**: ✓

Výskytů: **14** | i.p.m.: **0,01** (vztaženo k celému "omezeni/syn") | ARF: **6,25** | Výsledek je promíchán

Výběr řádků: základní | Atributy:

ecialit (kuře na indický způsob ,	po/po/RR--4-----------	římsku/římska/NNFS4-----A-----	, s broskví apod.) byl
že každý Říman miluje spaghetti	po/po/RR--4-----------	římsku/římska/NNFS4-----A-----	. Navíc má každý opr
utila . * Caput mundi zase zdraví	po/po/RR--4-----------	římsku/římska/NNFS4-----A-----	. Giovanni Alemanno
ací na květákové téma je květák	po/po/RR--4-----------	římsku/římska/NNFS4-----A-----	. Růžičky a plátky z tř
speciality , například kuřecí řízky	po/po/RR--4-----------	římsku/římska/NNFS4-----A-----	, tortellini s ricotottou
přihrádce na zeleninu . Artyčoky	po/po/RR--4-----------	římsku/římska/NNFS4-----A-----	vydrží v chladničce p
speciality , například kuřecí řízky	po/po/RR--4-----------	římsku/římska/NNFS4-----A-----	, tortellini s ricotottou
ém s lunelským muškátem Punč	po/po/RR--4-----------	římsku/římska/NNFS4-----A-----	Pečené nadívané ore
učí Tothero rozhodně . " V Římě	po/po/RR--4-----------	římsku/římska/NNFS4-----A-----	. " " Moje mi neberte
e pozdravily . Pán nás pozdravil	po/po/RR--4-----------	římsku/římska/NNFS4-----A-----	a já , protože jsem to
si dejte třeba artyčoky , nejlépe "	po/po/RR--4-----------	římsku/římska/NNFS4-----A-----	" (carciofi alla roman
ji Přímořské , jelikož tam se žilo	po/po/RR--4-----------	římsku/římska/NNFS4-----A-----	. Římští úředníci se to
vé rýže 1 menší cibule Brambory	po/po/RR--4-----------	římsku/římska/NNFS4-----A-----	Příprava Brambory uv
li takové hybridní věci , jako byly	po/po/RR--4-----------	římsku/římska/NNFS4-----A-----	zhotovované palmýrs

Fig. 6. Po římsku

5 SUMMARY

Compound adverb recognition by Automatic Morphological Analysis (AMA) is problematic for many reasons. We carried out a corpus probe aimed at mapping expressions that are demonstrably compound adverbs and were not recognized by the AMA as such or were incorrectly tagged by the AMA as a different part of speech. Thanks to precise analysis of the obtained data, we identified five areas the implementation of which into the AMA could refine compound adverb recognition and reduce incorrect part of speech tagging in such occurrences where the adverbial meaning is unambiguous even without context. We propose **a)** strict adherence to linguistic rules, **b)** a separate tag for compound adverbs or nominal forms that can be part of a compound adverb, **c)** use of Automatic annotation of idioms and fixed collocations (FRANTA), **d)** tagging the type [lemma="do.*.a"], [lemma="na.*.o"] as compound adverbs, the group being currently tagged incorrectly as a preposition and a noun, and finally **e)** distinction of the case of the initial letter of the lemma in fixed expressions including a geographic name.

We are aware that the proposed solutions do not cover the complete issue of compound adverb recognition, but we believe that the corpus probe and the proposed solutions can contribute to partial improvement of the AMA in this area at the least.

References

[1] Osolsobě, K. (2014). *Česká morfologie a korpusy*. Karolinum, Praha.

[2] Dokulil, M. (1962). *Tvoření slov v češtině, díl 1. Teorie odvozování*. Nakladatelství Československé akademie věd, Praha.

[3] Internetová jazyková příručka (2016). ÚJČ AV ČR, Praha.

[4] Cvrček, V. (2010). *Mluvnice současné češtiny*. Karolinum, Praha.

[5] Multiword Expressions, (2016). In *Wiki of the Association for Computational Linguistics*. Association for Computational Linguistics, Stroudsburg PA.

[6] Knapová, M. (1973). K otázkám adverbializace. Slovo a slovesnost, 34(2):150–157.

[7] Trávníček. F. (1951). Mluvnice spisovné češtiny. Slovanské nakladatelství, Praha.

[8] Sag, I. A., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. (2002). *Multiword Expressions: A Pain in the Neck for NLP*. Heidelberg, Berlin.

[9] Křen, M., Čermák, F., Hlaváčová, J., Hnátková, M., Jelínek, T., Kocek, J., Kopřivová, M., Novotná, R., Petkevič, V., Procházka, P., Schmiedtová V., Skoumalová, H., and Šulc, M. (2014). Korpus SYN, verze 3 z 27. 1. 2014. Ústav Českého národního korpusu FF UK, Praha.

[10] Horák, A., Pala, K., Rambousek, A., and Povolný, M. (2006). DEBVisDic – First Version of New Client-Server Wordnet Browsing and Editing Tool. In Proceedings of the Third International WordNet Conference – GWC 2006, pages 325–328, Masaryk University, Brno, Czech Republic.

[11] Čermák, F. (2009). *Slovník české frazeologie a idiomatiky*. Leda, Praha.

[12] Petkevič, V. (2014). Problémy automatické morfologické disambiguace češtiny. *Naše řeč*, 97(4-5):194–207.

[13] Křen, M. (2016). Automatická anotace frazémů a ustálených kolokací. In WIKI Český národní korpus. FF UK, Praha.

[14] Halfrunt, G. (1958). Psychosomatic aspects of bad poetry. *Prostetnik Research Methods*, 42(1):66–132.

THE USE OF AUTHORIAL CORPORA BEYOND LINGUISTICS

RICHARD ZMĚLÍK
Faculty of Arts, Palacký University, Olomouc

ZMĚLÍK, Richard: The Use of Authorial Corpora Beyond Linguistics. Journal of Linguistics, 2017, Vol. 68, No 2, pp. 404 – 414.

Abstract: The study concentrates on the issue of quantitative and qualitative methods within the context of literary theory. It intends namely to present the concept of the literary corpus of Czech prose and define main parameters of the corpus. Besides the project of a specialized corpus, primarily intended for the use in the field of literary theory, the study deals with current stochastic and corpus methods applied by foreign scholars in analysis of literary prosaic texts. The study tries to incorporate the original project of Czech prose literary corpus in this contemporary context that represents one form of a recently flourishing discipline called Digital Humanities (Digital Literary Studies).

Keywords: Literary Studies, Digital Humanities, Literary Corpus, Thematic Analysis

1 INTRODUCTION

Electronic corpora of national languages recently form a natural and truly essential part of linguistic studies. Specifically, the Czech National Corpus has recently not only been constantly extended, but also improved as for its functioning, the instruments used to utilize the corpus data, and even the concept of structuration of the corpus itself. The usability of language corpora has its source not only in the storage capacity of contemporary electronic platforms but also in the relevant way of stratification of linguistic data and in the method of their analysis. The notion of *representativeness* [10] thus becomes one of important features of a modern corpus. This feature can be understood in several different ways. In case of the Czech National Corpus, the concept of representativeness is based on relative gender balance of various types of texts part of which is naturally formed by artistic literary texts.[1]

Even though it is certainly possible to utilize such (sub)corpora in literary-theoretical research it is necessary to realize that the conception of the corpora is motivated by other than literary-theoretical purposes and it is primarily meant for linguistic exploitation. In spite of the fact that in SYN 2015 the genre taxonomization of texts[2] is more elaborate than in older corpora, the usability of the corpora for serious literary-theoretical study is problematic; the main reason for that has already been mentioned: the essential conception standpoint is linguistic – that is why it, quite naturally, ignores purely literary-theoretical criteria. Since the need to utilize modern

[1] In the pre-corpus era, linguists used namely literary texts as sources of quantitative analyses, serving predominantly to linguistic purposes, not to the purposes of literary theory [11], [12].

[2] As for fiction, namely the division (segmentation) between narrative (prosaic) and non-narrative (poetic) text has been implemented. Authors of the study presenting the new version of the corpus also emphasize that current tendencies prefer the use of smaller and specialized corpora.

DOI 10.1515/jazcas-2017-0050

corpora in literary-theoretical research rises we intend to define elementary requirements and preconditions for a special conception of fiction corpora, or rather prose corpora, that will be effectively used mainly by literary theoreticians.

2 Initial Conception of a Literary Corpus

If we now proceed to primary arrangement of the literary corpus we may take into consideration following fundamental intentions. First of all, the corpus should reflect a developmental viewpoint, specified by temporal relations between different groups of texts (texts written by the same author, texts written by different authors). Intra-textual and extra-textual markers for modelling and testing of particular phenomena should be available. Intra-textual markers will be related to the very analysis of the text matter (by analysis we mean primarily quantitative delimitation of thematic groups, entropy, lexical richness of texts, etc.). Extra-textual information, on the other hand, includes all specification data of the texts, such as data concerning authorship, gender phenomena, genres, location and time of origin, sequence of editions, etc.

A foundation from which a literary corpus can grow should be formed by corpora of smaller extent, primarily by authorial corpora that can be built independently but in accord with a predetermined strategy of building the corpus as a whole. (The final purpose of the corpus is to map the production of Czech prosaic literature from the 19th century to present times.) At the level of authorial corpora that will form elementary structure units of the intended corpus it is possible to stick to the following preliminary scheme:

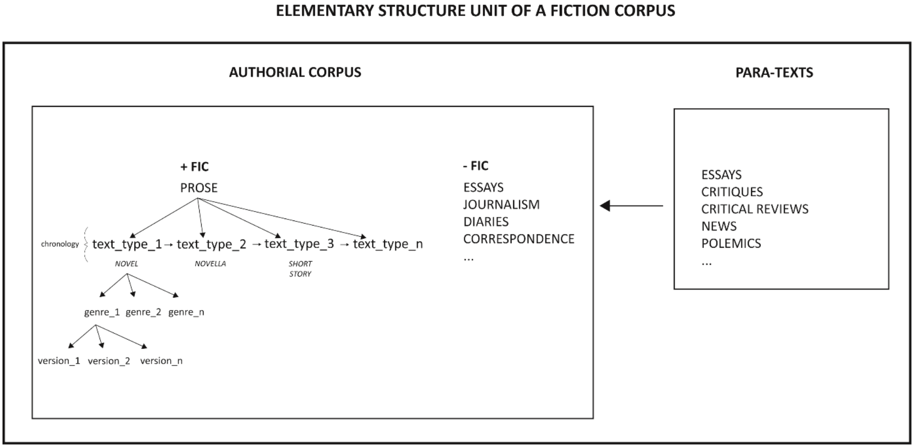


Fig. 1. A model of structuration of a partial authorial corpus serving as an elementary structure unit of the intended corpus of literary fiction

Each of authorial corpora will be divided in two main textual parts: fictional texts (prosaic, or possibly dramatic), and non-fictional texts. Special segmentation will pertain mainly prose that will be divided into individual genres and subgenres and subsequently to particular text variants. Para-texts, specified as to their type (review, critique…) and relation to a particular prosaic text, will form part of each authorial corpus. At this level of corpus building, a relatively sophisticated net will be formed within each segment of the corpus, connecting not only texts and para-texts but also individual literary texts. The net will enable corpus users to model particular literary texts in their relations to predefined chronological perspectives. We specifically mean the chronology defined according to the perspective of the first edition of each literary work and the chronology I accord with a subsequent, repeated edition (see Fig. 2).

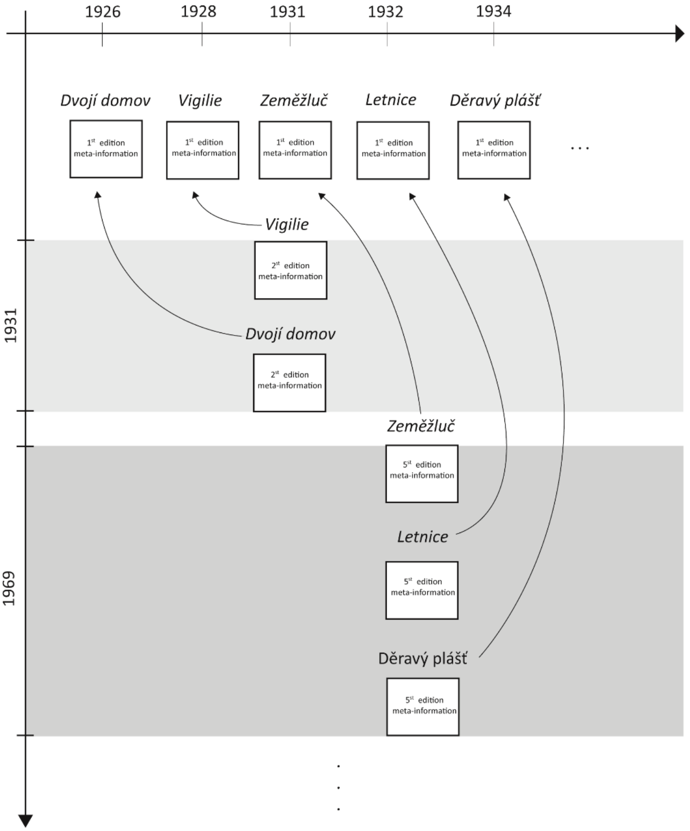
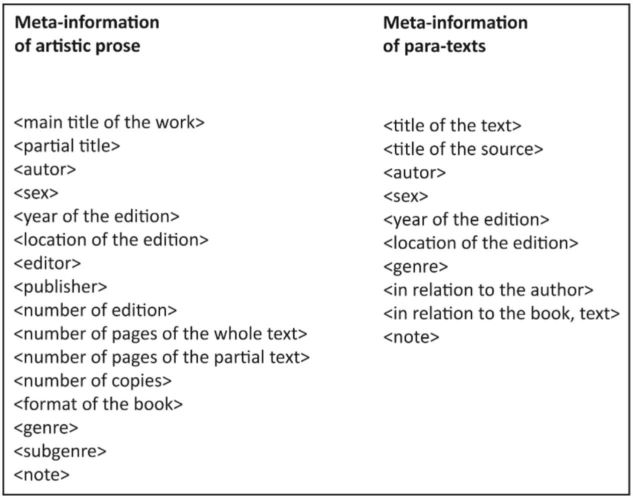

1926
1928
1931
1932
1934
Dvojí domov
Vigilie
Zeměžluč
Letnice
Děravý plášť
1^{st} edition meta-information
1^{st} edition meta-information
1^{st} edition meta-information
1^{st} edition meta-information
1^{st} edition meta-information
...
1931
Vigilie
2^{st} edition meta-information
Dvojí domov
2^{st} edition meta-information
Zeměžluč
5^{st} edition meta-information
1969
Letnice
5^{st} edition meta-information
Děravý plášť
5^{st} edition meta-information
...

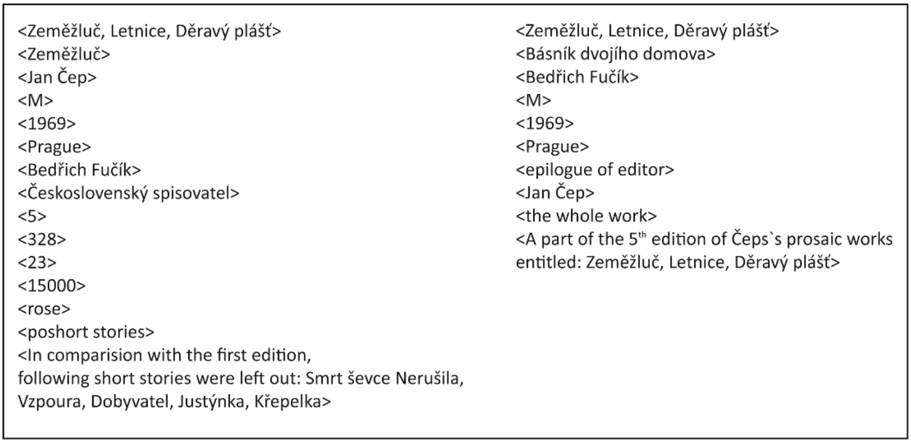
Fig. 2. The model of double chronology: first editions are marked on the horizontal axis, repeated editions on the vertical axis

Each text processed to become a part of the special literary corpus will be supplied with a set of metadata that can enable corpus users not only to find other ways of searching the corpus but also to analyse and model consistent with the criteria significant for literary history. Elementary data concerning the author and the text must be included in the meta-information, as well as other information pertaining location and date of a particular edition, number of copies, book format, etc.



Meta-information of artistic prose	**Meta-information of para-texts**
<main title of the work>	<title of the text>
<partial title>	<title of the source>
<autor>	<autor>
<sex>	<sex>
<year of the edition>	<year of the edition>
<location of the edition>	<location of the edition>
<editor>	<genre>
<publisher>	<in relation to the author>
<number of edition>	<in relation to the book, text>
<number of pages of the whole text>	<note>
<number of pages of the partial text>	
<number of copies>	
<format of the book>	
<genre>	
<subgenre>	
<note>	

Fig. 3a. Fundamental meta-information types for a different type of texts



<Zeměžluč, Letnice, Děravý plášť>	<Zeměžluč, Letnice, Děravý plášť>
<Zeměžluč>	<Básník dvojího domova>
<Jan Čep>	<Bedřich Fučík>
<M>	<M>
<1969>	<1969>
<Prague>	<Prague>
<Bedřich Fučík>	<epilogue of editor>
<Československý spisovatel>	<Jan Čep>
<5>	<the whole work>
<328>	<A part of the 5th edition of Čeps`s prosaic works entitled: Zeměžluč, Letnice, Děravý plášť>
<23>	
<15000>	
<rose>	
<poshort stories>	
<In comparision with the first edition, following short stories were left out: Smrt ševce Nerušila, Vzpoura, Dobyvatel, Justýnka, Křepelka>	

Fig. 3b. As an example we may state a complete edition of three short-story anthologies by Jan Čep, *Zeměžluč, Letnice* a *Děravý plášť*, in its edition from 1969, and a study written by Bedřich Fučík that supplements the edition.

The resulting model of the corpus will subsequently be open to numerous types of comparing and types of modelling. Corpus users will be able to compare partial texts of different authors as well as complete works or groups of texts defining literary schools, trends and tendencies. It will be possible to analyse even individual groups thanks to their inner structuration and differentiation. Importantly, comparisons will not have to be based exclusively on external criteria, i.e. on information supplied from literary-theoretical sources; corpus users will be able to test texts and works on the grounds of stochastic analyses and simultaneously observe and evaluate relations and bonds realized between individual sets of data. The above stated advancements will serve to systematic monitoring of developmental tendencies and more recent publishers' and editors' modifications. At the same time, on the quantitative level, it will be possible to model the course of reception of individual texts in newspapers and journals of the period (see Fig. 1), to monitor quantitative changes in the reception, and to compare thematic correlations between the texts and their critical reflections in the genre field of journalism.

2.1 Thematic Analysis

Thematic analysis is one of the significant functions of the intended corpus. Within the purview of quantitative-corpus approach, its methods can be divided in two main categories. The first category is based immediately on the stochastic principles that represent a central criterion of the analysis. Such methods fully rely on statistic algorithm applied on lexicon and as such they are fully quantitative.

In the Czech context, the method of so-called thematic concentration of text has attracted wider attention. The method uses division of lexicon in two main areas according to symptomatic occurrence of certain types of lexemes. In relation to frequency of occurrence, in each lexicon two areas can be separated: one area with prevailing synsemantic words, words that are found in higher frequency zones, and another area in which autosemantic words gradually dominate. For the method of measuring thematic concentration text, defining of so-called *h-point*, a point that statistically defines the border dividing the two areas, is crucial.[3] Filtering of autosemantic words in the area for which occurrence of synsemantic words is typical (that is in the area above the *h-point*), symptomatic lexemes that reflect thematic orientation of the text are detected. In subsequent steps of thematic analysis it is possible to taxonomize these lexemes in semantic classes and create more general models of semantic groups the paradigms of which can be mutually compared [15].[4]

Outside the Czech Republic, the research focused on thematic analysis of text currently also uses other methods. The study of Mathew L. Jokers [6] can serve as a typical example of such analytic methods, used in the context of literary-historical research. Jockers realizes that thematic analysis must be based on stochastic parameters different from those applied with so-called associative measures.[5] If, as Jockers suggests, collocations or key words are not sufficient for a relevant determination of

[3] "The method of TC (thematic concentration) measuring is based on the character of frequency structure of the text, particularly on so-called *h-point* [...] and on the sequence and frequency of words signalizing the topic of the text above this point" [2, pp. 15–16]. On the other hand, certain misunderstanding can arise in the situation when no autosemantic word appears above the border line. In such occasion, it would certainly be misleading and wrong to suppose that the analysed text has no theme.

[4] In the Czech context, the method of thematic concentration of text is developer namely by Radek Čech [3].

[5] "If our goal is to understand the narrative subjects and the recurrent themes and motifs that operate in the literary ecosystem then we must go beyond the study of individual n-grams, beyond the words, beyond the KWIC lists, and beyond even the collocates in order to capture what is at once more general and also more specific. Cultural memes and literary themes are not expressed in single words or even in single bigrams or trigrams. Themes are formed of bigger units and operate on a higher plane. If we may know the sense of a word by the company of words that surround it in a sentence, we may know a theme by the sentences, paragraphs, chapters, and even full books that express it. In short, simple word-to-word collocation and KWIC lists do not provide enough information to rise to the level of theme. What is needed in order to capture theme are collocations of collocations on a much larger scale" [6, p. 122].

thematic orientation of a text or a group of texts – namely because these qualities and their values are still overly determined by authorial style – it is necessary to develop an alternative quantitative procedure. As an alternative, Jockers suggests so-called *latent Dirichlet allocation* (LDA) which is able to cumulate thematically close lexemes in one paradigm. The LDA method of lexicon analysis surpasses collocations and KWIC since it can register even words that are very distant in the text (and undetectable by any available measurements of association) and include them in a common thematic set.[6] Currently many researchers who apply quantitative methods namely in the field of literary history use thematic and stylometric models based on quantification of immanent structural features of texts and text groups. Examples of the use of such methods can be found for example in a collective monograph entitled *Distant Readings: Topologies of German Culture in the Long Nineteenth Century* (2014).[7]

From the methodological viewpoint, the approach based on *a priori* thematic tax-

[6] In his work, Jockers mentions two lexemes, *stream(s)* and *Indians* that appear in mutually distant positions in an analysed text (Thomas Mayne-Reid: *The Scalp Hunters*, 1851). The LDA analysis can include them in one thematic cluster [6, p. 127]. Jockers worked with a corpus of 3346 prosaic texts written in English. With the use of MALLET software (`http://mallet.cs.umass.edu/`) he defined (named) 500 thematic groups [19]. This is how Jockers describes the method: "MALLET output includes two derivative files: a file containing topic 'keys' and file containing the proportions of each topic found in each text, or each text segment in this case. The keys file is a simple matrix in which the first column of data contains a unique topic identifier, and a second column contains the top words assigned to the topic. […] The second derivative file that MALLET produces provides data regarding the amount (proportion) of each topic found in each text segment. The modelling process assumes that every document is a 'mixture' of all the 500 possible topics in the corpus. Thus, each document is composed of some proportion of each of the 500 topics. Motivated by the work of the Veselovsky brothers [author means Alexander Nikolayevich Veselovsky, a Russian literary theoretician (1838–1906) – R. Z.] and their interest in studying literary evolution in terms of recurring motifs and national literatures, I began my analysis by plotting the mean proportions of every topic, in every year, separated first by nation, then by gender, and finally by nation and gender combined. Linking all of the thematic and topical data to the metadata facilitated the identification of thematic and topical patterns at the level of the corpus, at the level of the individual book, and across facets of time, gender, and nationality" [6, pp. 135–136].

[7] Title of the publication deliberately adverts to the book *Distant Reading* (2013) by Franco Moretti who significantly predefined current possibilities of application of quantitative methods in literary history. See also other Moretti's books in which the author deals not only with quantification but also with literary cartography [9]. The mentioned collective is dedicated to different aspects of quantitative study of literary texts of the 19th century. While presenting the elementary characteristic of the aspects, editors of the publication emphasize the possibility of systematic analysis of a large amount of textual data as one of the main benefits of applying the quantitative-corpus approach in literary theory: "For the literary historian concerned with incorporating larger numbers of texts and viewing works in a broader social, industrial, even transnational context toward discovery of the past, however, close reading of texts simply does not suffice. It does not adequately answer questions about the production and circulation of books, taste information, or even necessarily about the relative position of texts in the literary field" [5, p. 9]. They stress out the opposition of the *distant reading* and *close reading* methods [5, pp. 8–9]. What makes the approach they apply really significant is the fact that no matter how well quantitative methods and algorithms can process vast quantities of data, no matter how well they can generate meaningful models according to predefined criteria they cannot fully substitute literary-theoretical interpretation. Nevertheless, theoretical interpretation has, owing to the mentioned methods, a functioning instrument for analysis at its disposal that shifts literary theory closer to exact scientific procedures.

onomy, independent of analysed texts (external taxonomy), typologically represents another way how to recognize and analyse thematic level. In other words, in contrast to previous examples, we now have an initial paradigm at our disposal that represents a particular way of thematic taxonomization. It is a product of a corresponding individual lexical-semantic and lexicographic method and it has the same disadvantages as the method. To get a concrete idea, we suggest following advancement: *Tezaurus jazyka českého: Slovník českých slov a frází souznačných, blízkých a příbuzných* (2007)[8], a thematically organized dictionary, can represent the external paradigm for texts written in the Czech language. This extensive lexicographic work of Aleš Klégr, inspired directly by Roget's[9] *Thesaurus of English words and Phrases, Classified and Arranged so as to Facilitate the Expression of Ideas and to Assist in Literary Composition* (1852), can boast with its functional modification in relation to the current state of Czech lexicon and with division of the lexicon into individual thematic-semantic classes, sections, and sub-sections.[10] A thesaurus conceived in this way has the advantage of thematic (not alphabetical) taxonomization of the lexicon; the dictionary can thus be used as an initial paradigmatic filter during thematic stratification of texts.[11] The process of such segmentation will be based on the functional interconnection of the thesaurus and an authorial dictionary and its components that will be lemmatized; also, each lemma will be endowed with a frequency indication for a particular part of the authorial (sub)corpus. Due to algorithmic assigning of lemmas and their frequencies in corresponding thematic groups and classes declared by the thesaurus[12], each authorial (sub)corpus will acquire a frequency based, stratified model of thematic blocs, structured in accord with the parameters defined by the *Thesaurus*. Each of these partial models will be subsumed under a common authorial corpus, and, subsequently, under a pre-defined set, such as period of time, genre, gender, etc. (see chapter 2.2).[13]

[8] Aleš Klégr briefly defines the thesaurus as follows: "Meaning (notion) is the main principle of entry ordering; subsequently, the words that signify a particular notion are stated in a dictionary entry. Thus, the thesaurus functions as an onomasiologic handbook that translates the vocabulary of a language, its specific varieties (such as dialects), or a particular discipline in a way that signifies relations between words that belong to the same thematic or semantic areas (so-called semantic fields). In other words, the dictionary links of the same or similar (synonymic) meaning, may they be hyperonyms, hyponyms, or words on the same level" [8, p. 7].

[9] Peter Mark Roget (1779–1869); for elementary biographical data related to the thesaurus [7, pp. 65–66].

[10] Particularly, the Czech thesaurus is divided into six elementary thematic classes: abstract relations, space, matter, intellect, will, and emotions (with religion and morality). In total, the Czech thesaurus contains 885 entries that are subsequently assigned to particular sections and classes [8, pp. 16–21]. Naturally, it is not necessary to respect such taxonomization. Klégr himself added new entries (*Science, Sports, Gods*) to the original Roget's classification.

[11] Klégr mentions this potential use of the thematic thesaurus [8, pp. 9–10].

[12] Luckily for potential programmers, an electronic database, built by the team lead by prof. Klégr within the programme *A computerized thesaurus of the Czech language*, was finished before the publication of the thesaurus.

[13] Klégr mentions the decoding function of the thesaurus in relation to Julius Laffal dictionary *A Concept Dictionary of English* (1973), designed by the author "for purposes of automatized analysis of notional contents of texts" [8, p 74], particularly of spoken texts produced people suffering from mental diseases: "The dictionary also contains a demonstration analysis of a text entitled *Declaration of Independent*. The analysis revealed that notions belonging to four classes,

On the other hand, such advancement also inevitably involves certain distortion since it is dependent on a particular conception of external paradigm, i.e. on a particular chosen form of the thesaurus. Even though Klégr notes that the original Roget's thesaurus was meant as a descriptive, not normative, dictionary [8, p. 9] he also admits that the English thesaurus has become a model for numerous lexicographic works of this type *de facto* "reflects its author's world view" [8, p. 8]. As for the methodological orientation of thematic analysis, we rather prefer a combination of both types of orientation: *a priori* stochastic and *a priori* paradigmatic. The research can finally be focused on multiple methodological aspects. First of all, it will aim at the methodological issues pertaining thematic study of lexicon, at the ways of defining thematic classes and groups, at realization of partial thesauri [7, pp. 69–77],[8, p. 10], such as a literary thesaurus of the 19th century, and at realization of more extensive versions of the already existing thesaurus, etc. [8, p. 11]

Naturally, the initial version of the *Thesaurus* database will not include all lexemes that will appear in literary texts.[14] Therefore, the analysis must necessarily be accompanied by systematic adding of new lexeme in the original database or by modification of the present thematic paradigm.[15] Two main operations should thus become the basis of the whole process: selection and classification on one side, and incorporation of residual lexicon in the main database on the other hand; both operations must complement each other.

LEAD, LAW, GROUP, and MOTV, appear in the text most frequently [8, p 74]. František Čermák points out that electronic thesauri are called *ontologies*: "Besides systemization of gathered terms in a formalized system, production of ontologies is motivated by possibilities of easy computerized searching of the terms, of ontology-based classification, of the use of ontologies in the semantic web, etc." [4, p. 328].

[14] Specifically, autosemantic words will be excerpted, i.e. nouns, adjectives, verbs, and adverbs [4, p. 164].

[15] In relation to adopting the original Roget's conception of the thesaurus, Klégr mentions numerous modifications that influenced not only the extent of the excerpted lexicon, but also the way of taxonomization (many original categories and subcategories proposed by Roget had either to be reduced or extended [7, p. 70]. Analogically, even the thesaurus formed on the basis of excerption of literary texts could be modified functionally. After all, as František Čermák stresses, "there are many **systems** proposed for the thesaurus and there is no way how to asses them objectively. The questions, answered differently in individual systems, include for example questions pertaining (a) possible existence of one overall hierarchy, (b) the issue of a top notion, (c) the number of hierarchic levels, (d) the number of the very classes, or (e) the correlation of the proposed hierarchic taxonomy with intuitive perception on one side, and a scientific taxonomy on the other side (if such taxonomy, similar to the ones incorporated in natural sciences, actually exists) etc. […] If we summarize what has been said so far it turns out that each author conceives his/her hierarchy differently in all its aspects (see above a–e) and that none of the authors applies a scientific taxonomy" [4, p. 323; 324].

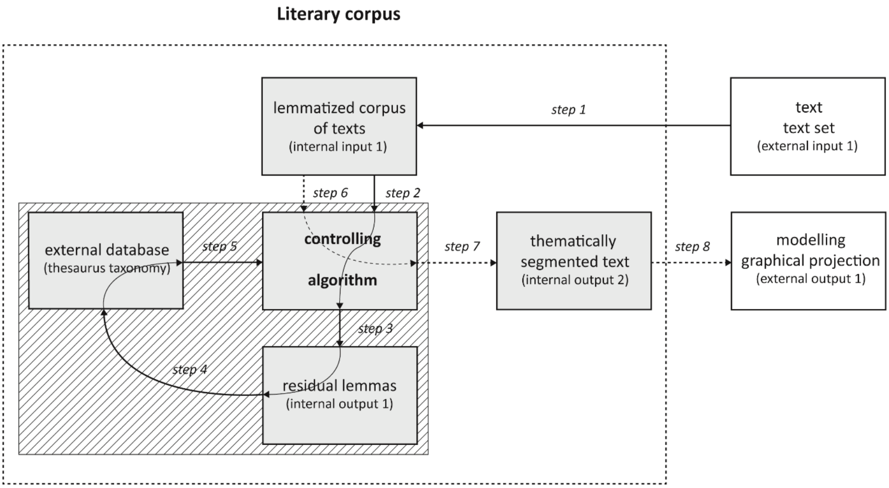


Fig. 4. The scheme of thematic analysis within the proposed literary corpus. At the level of *external input 1*, texts will be supplied in the electronic corpus database in the form intended for lemmatization. *Internal input 2* includes lemmatized texts, recognized as particular text types or text sets, that will pass through the controlling algorithm; the *thesaurus taxonomy* will form an external component of the algorithm. This procedure will separate those lemmas that have not been recognized, and subsequently (step 4) the lemmas will be incorporated in the external database. The controlling algorithm will carry out another analysis with the output of a thematically segmented text in which to each thematic unit (group, class, etc.) the value of relative frequency will be allocated, specified on the basis of accumulation of corresponding lexemes. *External output 3* is a graphical projection of this distribution, supplemented by optional functions in the extent of thematic layers (entry, section, class) and text groups, and by the function of graphical projection of frequency development of individual thematic classes etc.

Naturally, the results of the mentioned process can be used both by literary theoreticians – who will be able to register changes in distribution and stratification of thematic classes appearing in the course of a particular period of literary history and model them – and by linguists (see above). Linguistics will be provided not only with a constantly expanding thesaurus but also a large amount of material that can be used to study and analyse the issue of semantics and paradigmatics of the lexicon and for other purposes.[16]

2.2 Meta-information as a Source of Literary-Theoretical Modelling

Demands placed on a literary corpus must also include para-text information that should form an integral part of each lemmatized literary text (see Fig. 3.a). If in the context of current conception of building a synchronic corpus “explicit genre subclassification of fiction has been cancelled [1, p. 96] then in a literary corpus genre taxonomization must be preserved, namely with respect to the current literary-historical classification of texts. The corpus thus should be built in a way that serves the purposes of such research that would help to objectivize genre analysis and map or model transitional zones. Corpus-based research could incite more systematic and more variable study of genre categories and their specific features, initiate attempts

[16] It is not clear yet to what extent the still non-public database of Czech WordNet could be used in this process.

at empirical delimitation of temporal genre centres, peripheries and transition zones, and namely register the dynamics of these changes.

A sufficiently representative part of the corpus will enable elementary statistic operations, based on pre-defined criteria stated at the level of meta-information, to be carried out. It will thus be possible to register genre and gender production of text in individual periods of time. Reception, i.e. critiques and critical reviews, matched to corresponding texts in the database, will become a significant indicator. We can realistically expect the literary corpus to enable composition of so-called reception waves, a phenomenon that is currently often studied by literary historians [13]. Observation not only of changes and developments in genre categories, but also of gender specifications or the degree to which individual publishing houses, locations of publication, etc. participated in the development, is very useful for both literary history and theory of literature. The mentioned findings will have to be combined with other indicators, namely with values brought by thematic analysis and analyses of lexical richness and entropy.

3 CONCLUSION

The conception of building a special literary corpus of Czech prose, as it was briefly presented above, may make modern electronic corpora accessible to the literary theorists who currently do not work with corpora ore use them scarcely. First of all, the proposed corpus would enable systematic research based on exact data and stochastic methods. It was Pavel Vašák[17] who declared the use of such methods in literary theoretical research as reasonable and purposeful in the past. Currently, these tendencies are realized abroad; project such as Stanford Literary Lab [16] or .txtLAB [17] are partly focused on quantification of narrative segments in prosaic literary texts, etc.

As the project of Corpus of Czech verse [18] has suggested in the Czech context, specialized corpora using their own strategies and specific tools of statistic analysis appear as purposeful for literary theory. The literary corpus would finally be useful for other disciplines; namely with respect to the new conception of representativeness of the Czech language corpus it may also become a relevant instrument used by linguists.[18]

[17] "General methodological use of mathematic – and, in a wider context, even of cybernetic and system – methods has currently become common. […] It is therefore very advisable to use automatic computers that be of benefit to literary theory (and similarly to other humanist disciplines) at least in two respects:
1. they can gather data necessary for analysis and interpretation of literary works, may they be carried out either with mathematic or traditional methods. Formation of different concordance- or frequency-based dictionaries of significant authors, literary schools, movements, periods, generations, etc. would surely help literary theorists greatly.
2. Computers can also be used to look up and to gather bibliographic data" [14, pp. 52, 53].

[18] "Corpus-based research currently changes its focus; from the effort to describe the language as a whole it shifts to the description of individual varieties or genres of the language (finding obtained through analysis of the data are not related to all texts but only to a specific group of texts that is actually represented by the data). This development partly results from the

It has already been mentioned that specially processed authorial corpora should form the basis of the intended corpus; the building of these is exactly where the gradual systematic formation of the overall corpus should begin. Actually, testing of the proposed method of *thematic analysis* and its comparison with alternative methods applied to the lemmatized literary texts contained in the Czech National Corpus should be the very first step made towards the building of the Czech prose corpus in the future. Subsequently, two initial text centres, forming the basis for building of the corpus, will be founded – the first one situated to the 19th century, the other one to the 20th century.[19] To build such a corpus, much time and close cooperation with literary theorists, linguists and programmers will be required.

References

[1] Cvrček, V., Čermáková, A., and Křen, M. (2016). A new design of synchronic corpora of writen Czech. *Slovo a slovesnost*, 77(2):83–101.

[2] Čech, R., Popescu, I., and Altmann, G. (2014). *Metody kvantitativní analýzy (nejen) básnických textů*. Palacky University, Olomouc.

[3] Čech, R. (2016). *Tematická koncentrace textu v češtině*. Institute of Formal and Applied Linguistics, Prague.

[4] Čermák, F. (2010). *Lexikon a sémantika*. Nakladatelství Lidové noviny, Prague.

[5] Erlin, M. and Tatlock L. (2014). *Distant Readings: Topologies of German Culture in the Long Nineteen Century*. Camden House, Rochester – New York.

[6] Jockers, M. L. (2013). *Macroanalysis. Digital Methods and Literary History*. University of Illinois Press, Illinois.

[7] Klégr, A. (2000). Rogetův *Thesaurus* a onomaziologická lexikografie. *Časopis pro moderní filologii*, 82(2):65–84.

[8] Klégr, A. (2007). *Tezaurus jazyka českého: Slovník českých slov a frází souznačných, blízkých a příbuzných*. Nakladatelství Lidové noviny, Prague.

[9] Moreti, F. (2014). *Grafy, mapy, stromy: Abstraktní modely literární historie*. Karolinum, Prague.

[10] Šulc, M. (2001). Tematická reprezentativnost korpusů. *Slovo a slovesnost*, 62(1):53–61.

[11] Těšitelová, M. (1948). Frekvence slov a tvarů ve spise ‚Život a dílo skladatele Foltýna' od Karla Čapka. *Naše řeč*, 32(9):297–307.

[12] Těšitelová, M. (1955). Poznámky ke slovní zásobě v románě Karla Čapka ‚Život a dílo skladatele Foltýna'. *Naše řeč*, 38(9):297–307.

[13] Tureček, D. (2012). Synopticko-pulzační model českého literárního romantična. In *České literární romantično: synopticko-pulzační model kulturního jevu*, pages 92–142, Host, Brno, Czech Republic.

[14] Vašák, P. (1980). *Metody určování autorství*. Academia, Prague.

[15] Změlík, R. (2015). *Kvantitativně-korpusová analýzy a literární věda. Model a realizace autorského korpusu a slovníku Jana Čepa v kontextu zahraniční a české autorské lexikografie*. Palacký University, Olomouc.

[16] Stanford Literary Lab. Accessible at: `https://litlab.stanford.edu/`.

[17] .txtLAB. Accessible at: `https://txtlab.org/`.

[18] Korpus českého verše. Accessible at: `http://versologie.cz/kcv.html`.

[19] Jockers, M. L. 500 Themes from a corpus of 19th-Century Fiction. Accessible at: `http://www.matthewjockers.net/macroanalysisbook/macro-themes/`.

realization of the fact that a description at the level of the whole language would necessarily equalize the mutual dissimilarity of individual texts" [1, p. 92].

[19] As for the 19th century, the work will begin with building of authorial corporal of members of individual movements and schools (authors gathered around the almanacs and journals *Máj*, *Ruch, Lumír*, etc.).

AUTOMATIC MORPHEMIC ANALYSIS IN THE CORPUS OF THE UKRAINIAN LANGUAGE: RESULTS AND PROSPECTS

OKSANA ZUBAN
Taras Shevchenko National University of Kyiv, Ukraine

ZUBAN, Oksana: Automatic Morphemic Analysis in the Corpus of the Ukrainian Language: Results and Prospects. Journal of Linguistics, 2017, Vol. 68, No 2, pp. 415 – 425.

Abstract: The article describes theoretical issues, principles of constructing and functioning of the Automated System of Morphemic and Derivational Analysis (ASMDA). The ASMDA system performs the following functions: 1) information system; 2) automatic morphemic annotation of text; 3) automatic linguistic constructor for frequency dictionaries. Description of the use of ASMDA as an automatic morphemic analyser of Ukrainian texts' lexicon is in the centre of attention; this article also describes structure as well as search and classification options of electronic morphemic dictionaries presented in linguistic research system of the Corpus of the Ukrainian language.

Keywords: Morphemic-Derivational database, Corpus of the Ukrainian language, the morphic segmentator of the Ukrainian text, Electronic dictionary of frequency, automatic morphemic analysis.

1 INTRODUCTION

Most of the corpora of Slavic languages do not deal with text annotation at morphemic level, that is why the methodology of computer modelling at morphemic level of text structure is not developed enough in modern corpus linguistics. Corpus of the Ukrainian language [9] is a research tool directed to solve a wide class of linguistic problems, particularly in the field of morphemics and word formation.

The practice and theory of creating automated databases are being successfully developed in the field of morphemics and word formation in modern Ukrainian computational linguistics: Morphemic-Derivational database of the Ukrainian Language [8], created in the Department of Structural and Mathematical Linguistics at the Academy of Sciences of Ukraine; Automated System of Morphemic and Derivational Analysis (ASMDA) [1], [2], [5] created by the staff of Computational Linguistics Laboratory at the Taras Shevchenko National University of Kyiv.

These two morphemic-derivational databases have different tasks and, accordingly, are based on different techniques. Morphemic-Derivational database of the Ukrainian language is designed to function as a kind of guide for a researcher in the field of linguistics, and it is undoubtedly extremely important for organizing a full-scale study of language, but it is static, it can not be used in a mode of automated text analysis. Automated System of Morphemic and Derivational Analysis (ASMDA) is an electronic linguistic product that performs automatic morph segmentation of initial word forms (lemmatized tokens) from the text, and has the status of dynamic search engine; it can find information about morphemic language units from any parametrized text presented in the Corpus of the Ukrainian language in an automatic or an automated mode.

DE GRUYTER OPEN
DOI 10.1515/jazcas-2017-0051

The aim of this article is to describe methodological issues, structure and functions of the ASMDA.

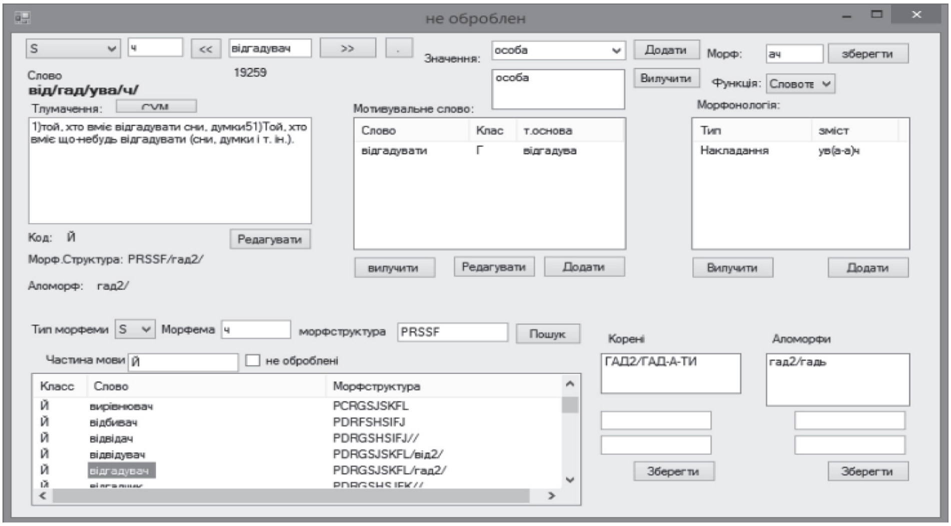
2 STRUCTURE AND STAGES OF CREATION OF MORPHEMIC DATABASE (MDB) OF ASMDA

The MDB of ASMDA system was based on I. T. Yatsenko's dictionary of morphemes [7]. One can define several stages of constructing ASMDA that aims at solving different tasks.

1st stage of MDB construction: 1) compilation of the database of morphemic word structure, where ≈ 170,000 of words were segmented into morphemes in automated mode; 2) compilation of the database of allomorphic roots (≈ 2,500 roots); 3) compilation of the database of homonymic roots (≈ 3,100 roots). The description of these tasks' automation and the obtained results are presented in the publications of the project authors [1], [2], [5], [13], [14].

2nd stage of MDB construction: 1) replenishment of MDB with new vocabulary, which has been automatically chosen from the texts of the Corpus of the Ukrainian language; at present the list of MDB contains ≈ 200,000 words; 2) ascribing the meaning and functions in morphemic word structure to each affixal morpheme of a word; 3) identification of a minimal derivational pair, where the analysed affix fulfills a word-formation function; 4) identification of morphonological phenomena in morphemic word structure of each de-rivative.

Today the MDB consists of six interrelated databases: 1) a database of morphemic word structure containing ≈ 200,000 words; 2) a database of allomorphic roots (≈ 2,500 roots); 3) a database of homonymic roots (≈ 3,100 roots); 4) a database of allomorphic affixes (in progress); 5) a database of homonymic affixes (in progress); 6) a database of morphonological alternations (in progress). The compilation of MDB is fulfilled online in automated mode with the help of a useful computer tool, which is provided by morfem.exe system (see Fig. 1).



не оброблен

S ч відгадувач

Слово 19259

від/гад/ува/ч/

Тлумачення:

1)той, хто вміє відгадувати сни, думки51)Той, хто вміє що-небудь відгадувати (сни, думки і т. ін.).

Код: Й

Морф.Структура: PRSSF/гад2/

Аломорф: гад2/

Значення: особа

особа

Морф: ач

Функція: Словоте

Мотивувальне слово:

Слово	Клас	т.основа
відгадувати	Г	відгадува

Морфонологія:

Тип	зміст
Накладання	ув(а-а)ч

Тип морфеми S Морфема ч морфструктура PRSSF

Частина мови Й □ не оброблені

Класс	Слово	Морфструктура
Й	вирівнювач	PCRGSJSKFL
Й	відбивач	PDRFSHSIFJ
Й	відвідач	PDRGSHSIFJ//
Й	відвідувач	PDRGSJSKFL/від2/
Й	відгадувач	PDRGSJSKFL/гад2/

Корені

ГАД2/ГАД-А-ТИ

Аломорфи

гад2/гадь

Fig. 1. Interface of the electronic card of morfem.exe system

3 METHODOLOGICAL ISSUES OF MDB COMPILATION OF ASMDA

3.1 Linguistic Issues of Modelling Morphemic Word Structure

The main linguistic task in the process of the MDB compilation was to define morphemic structure of words in Ukrainian language – a procedure of morphemic analysis.

Active development of theory and practice of derivational analysis in comparison to the morphemic one caused the definition of morphic word structure in terms of derivational analysis. Such methodology uses a notion of morphic word structure as a result of sequential derivational acts – a defined morph has to represent one derivational tact. It is contrary to the traditional definition of morpheme as an indivisible minimal linguistic sign, because it leads to legitimation of complex morphemes (polymorphemes [10]), and it also causes ambiguous morphic segmentation in different dictionaries and grammars.

Morphic word structure cannot always be explained by derivational processes, and according to its quantitative and morphic components does not necessarily represent morphic boundaries that emerge in derivational structures: 1) derivational formants may consist of several morphs; 2) non-derived word stems may also be divided into morphs. The morphs are determined not according to the derivational relations, but on the basis of distributional and paradigmatic relations of these morphs in the system of morphemics as a whole; it allows to use analogy method in morphemic segmentation of words. To sum up, the morphic structure of the word form may be considered as a stable structure which is not constructed, but is inherited by derivative from its analogy example [3, p. 58].

Typical nature of morphemic structures, regular repetition and reproducibility in words gives an opportunity to determine them as a special type of language units – ontological units of structural nature at morphemic level of language system. Morphemic structures of words are usually described with the help of models, in which Latin letters denote functional type of a morph, e.g., *ви/чит/к/а* PRSF. The basis of such modelling is a fundamental linguistic idea: a composition formula, which is regularly repeated and reproduced in language, is defined as a special unit of language structure along with phoneme and morpheme [11]. This methodological principle was taken as a basis for computer modelling of structural-functional connections of morphs in a word during compilation of MDB.

3.2 Computer Modelling of Morphemic Structures of Words

MDB is a list of Ukrainian words segmented into morphs. A basic list unit of MDB is a word with software procedure of determination of morphemic word structure (*задньоязиковий* RDSFIGRLSNFP), which models the morphemic structure of the given word (*зад-нь-о-язик-ов-ий* RSIRSF).

In the first stage of processing of the MDB lexical list, the automatic retransmission of literal word record into a simplified phonemic record takes place. The algorithm of retransmission of literal record into a simplified phonemic one takes into account only the positions of sounds that are pronounced with iotacism: **я**, **ю**, **є**, **ї**: makes the conversion **є à je**, **ї à ji**. This procedure is obligatory, since it makes it possible to define the boundary of morphs, when the one-letter spelling for two sounds appears at the junction of morphs: *клеїти à клеj-i-ти*. All the other peculiarities of a phonemic record are not taken into account.

Morphemic analysis of lexicon (initial word forms) takes place on the second stage. In MDB, each word has an ascribed model according to the methodology of computer modelling of structural-functional connections of morphs in a word. This model determines boundaries and functional type of morphs, and at the same time it is a formula of software procedure of morphemic word segmentation. E.g., *зад/нь/о/язик/ов/ий* R3S5I6R10S12F14: Latin letters denote a functional type of a morph: P – prefix, R – root, S – suffix, F – ending, I – interfix, X – postfix, and the numbers denote boundaries of a morph with a sequence number (from the beginning of a word) of a terminal letter of each morph. The formalized description of morphemic word structure in terms of software procedure represents functional pattern of word structure at morphemic level – RSIRSF; the substantial (literal) representation of this structure is presented in the form of quantitative-literal model (зад – 3; нь – 5; о – 6; язик – 10; ов – 12; ий – 14). Literal-number boundaries of morphs in the database are automatically converted into Latin letters: R3S5I6R10S12F14 = **R**D**S**F**I**G**R**L**S**N**F**P. The morphic structure, which is automatically formed with the help of a software procedure, gives full linguistic information about a morph, its structural and distributional relations with other morphs, and it is defined as a working unit of a dictionary of morphs.

The MDB list looks like an automatic dictionary:

безсонниця,К,PDRGSHSJFK/сон2/
безсонячний,А,PDRGSISJFL/сон1/
дисонанс,Й,PCRFSIFJ/сон3/,

where each word has an ascribed linguistic information about: a grammatical code of a word (К – feminine noun; Н – adverb, etc.); a model of software procedure of word segmentation into morphemes (PDRGSHSJFK; PDRGSISJFL); a root index in case of its homonymy or allomorphy (/сон1/; /сон2/; /сон3/).

4 THE USE OF ASMDA IN COMPUTER LEXICOGRAPHICAL TASKS

4.1 ASMDA as an Electronic Information System of Morphemics of Ukrainian Language

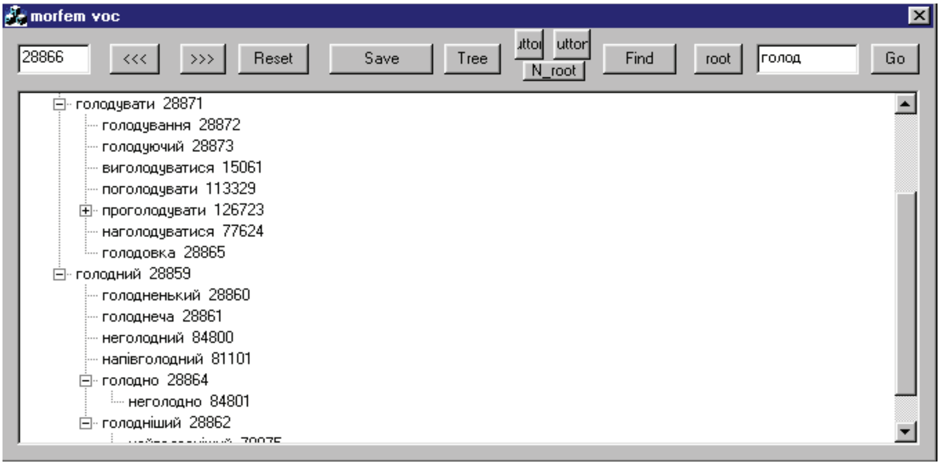
The software of ASMDA performs operations of automatic classification of Ukrainian language lexicon within the MDB lexical list (≈ 200 000 words) according to different parameters of morphemic word structure organisation: quantity of morphemes in a word; models of morphemic structures; specified affix or root morph. The ASMDA software also automatically compiles the lists of prefixes, roots, suffixes, interfixes, postfixes and endings (the system of endings is compiled with restriction, only on the basis of initial word forms).

ASMDA also performs the functions of the linguistic classifier in the process of compiling of an electronic derivational dictionary. The system of automatic derivational analysis is designed on the basis of this linguistic classifier. The formation of a derivational word family as an item of the electronic derivational dictionary is carried out on the basis of selection of all words from the MDB. The formation of selections of words having the same root is a difficult and laborious task, thus it is necessary to formalise the material on all the stages of its description, it enables to create software tools of the linguistic analysis.

Taking into account principles of derivation formalising principles of description of derivational relations between motivating and motivated words were designed. It enables to build the hypothesis-model in operation of a derivational word family. In this hypothesis-model every following word building act represents words with more complex affixal structures of stems in terms of quantity; i.e., a group of words of an each quantitative and affixal model of a word is a hypothetic stage of a derivational word family.

A linguist carries out further construction of a derivational word family automatically by means of toolset of the electronic card (see Fig. 2). This example demonstrates the part of the derivational word family of the words with the root – голод-, which is constructed by a linguist in automated mode on the basis of the selection of the words having one root of the morphemic database.

Each branch of the derivational tree reflects the relations of derivational motivation between the main word, which is marked by a square with either + or -, and the words that finish the branches of this main word. Marker "+" denotes that the word is the main one, i.e. that it is a starting point of the branch, while the marker "-"denotes that this branch is already extended. The modelling structural and motivational relations between the words of neighbouring derivational stages is carried out by means of establishing correspondences between numeric codes of words which belong to the morphemic database: *голодувати à голодування, голодувати à поголодувати.*



28866

голод

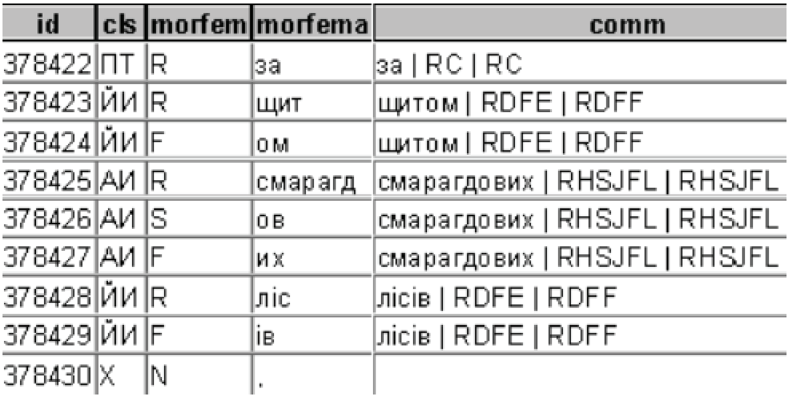
```
⊟ голодувати 28871
   голодування 28872
   голодуючий 28873
   виголодуватися 15061
   поголодувати 113329
 ⊞ проголодувати 126723
   наголодуватися 77624
   голодовка 28865
⊟ голодний 28859
   голодненький 28860
   голоднеча 28861
   неголодний 84800
   напівголодний 81101
 ⊟ голодно 28864
     неголодно 84801
 ⊟ голодніший 28862
```

Fig. 2. Interface of the derivational word family

Philologists of Taras Shevchenko National University of Kyiv actively use ASMDA as an information system of morphemics and word formation of Ukrainian language in different scientific research projects, although this database is not a lexicographical product for a wide range of users. That is why the staff of Computational Linguistics Laboratory started to work on the creation of lexicographical system “Morphograph”, which will be presented online as a public information resource for the users-philologists.

4.2 ASMDA as an Automatic Morphemic Segmentator of Ukrainian Text Tokens

Since the beginning of its creation, the system of ASMDA has been focused on automatic morphemic segmentation of text tokens. The morphemic segmentator of the Ukrainian text is a system, which on the input has the word forms of the analysed text with defined grammatical codes, which have been ascribed to the forms as a result of morphological annotation of text. On the output there are the same word forms split into morphs with proper indexes ascribed to each defined morph (P – prefix, R – root, S – suffix, F – ending, I – interfix, X – postfix).



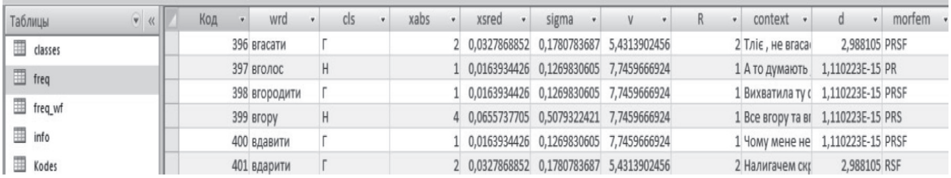
id	cls	morfem	morfema	comm
378422	ПТ	R	за	за | RC | RC
378423	ЙИ	R	щит	щитом | RDFE | RDFF
378424	ЙИ	F	ом	щитом | RDFE | RDFF
378425	АИ	R	смарагд	смарагдових | RHSJFL | RHSJFL
378426	АИ	S	ов	смарагдових | RHSJFL | RHSJFL
378427	АИ	F	их	смарагдових | RHSJFL | RHSJFL
378428	ЙИ	R	ліс	лісів | RDFE | RDFF
378429	ЙИ	F	ів	лісів | RDFE | RDFF
378430	X	N	.	

Fig. 3. A fragment of the morphemic annotation database of Ukrainian text

Fig. 3 shows the morphemic segmentation of the text fragment’s tokens *...за щитом смарагдових лісів*. Each token is automatically segmented into morphs with automatic ascription of two software procedures (5th column): the first one is a morphemic model of the initial word form of the lexeme; the second one is a morphemic model of the text fragment’s word form. The morphemic segmentation is performed according to the procedure of comparison of initial forms and text’s word forms on the basis of two databases: a morphemic database and a morphological database of the system for automated grammatical text analysis.

4.3 ASMDA as an Automatic Morphemic Analyser in the Corpus of the Ukrainian Language

While creating the automatic morphemic analyser in the Corpus of the Ukrainian language, we have rejected the method of morphemic text annotation for the purpose of search optimization in large text databases. There is no morphemic annotation of text tokens, the corpus texts are only a source for compilation of different alphabetic-frequency dictionaries of morphemic based on vocabulary samples of the corpus texts; the ASMDA performs the function of morphemic analyser-module.

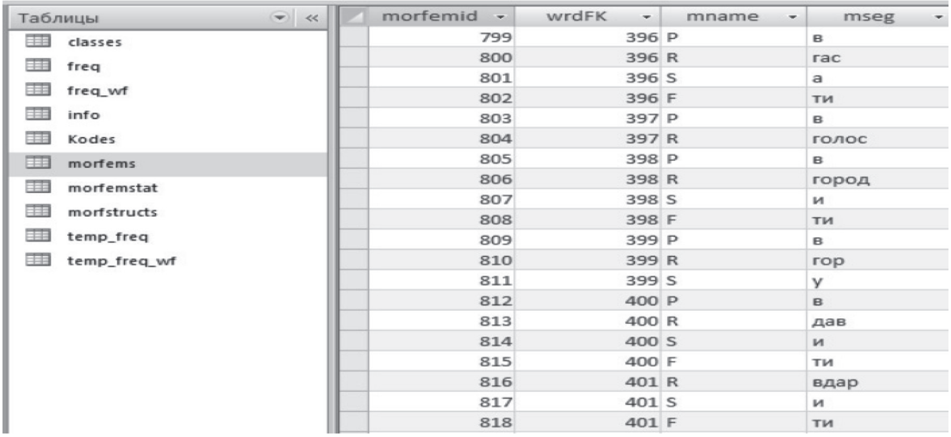


Код	wrd	cls	xabs	xsred	sigma	v	R	context	d	morfem
396	вгасати	Г	2	0,0327868852	0,1780783687	5,4313902456	2	Тліє , не вгаса	2,988105	PRSF
397	вголос	Н	1	0,0163934426	0,1269830605	7,7459666924	1	А то думають	1,110223E-15	PR
398	вгородити	Г	1	0,0163934426	0,1269830605	7,7459666924	1	Вихватила ту	1,110223E-15	PRSF
399	вгору	Н	4	0,0655737705	0,5079322421	7,7459666924	1	Все вгору та в	1,110223E-15	PRS
400	вдавити	Г	1	0,0163934426	0,1269830605	7,7459666924	1	Чому мене не	1,110223E-15	PRSF
401	вдарити	Г	2	0,0327868852	0,1780783687	5,4313902456	2	Налигачем ск	2,988105	RSF

Fig. 4. Fragment of morphemic database of initial word forms

On the stage of morphological module performance all the text tokens with morphological annotation are lemmatized into the initial forms, and then they enter the morphemic module in the form of the alphabetic-frequency dictionary of initial forms including information about a part of speech, frequency characteristics and contextual usage. A morphemic model of the word is ascribed to the initial forms in the morphemic module on the basis of comparison with this word in MDB (see Fig. 4).

A new database is compiled on the output of morphemic module; this database includes the same systematized initial word forms, which have been segmented into morphs, containing information about a functional type of morpheme (Fig. 5). The segmentation into morphs is carried out according to the rules of morphemic word structure organisation and the specified models of software procedures of word segmentation in the morphemic database of ASMDA.

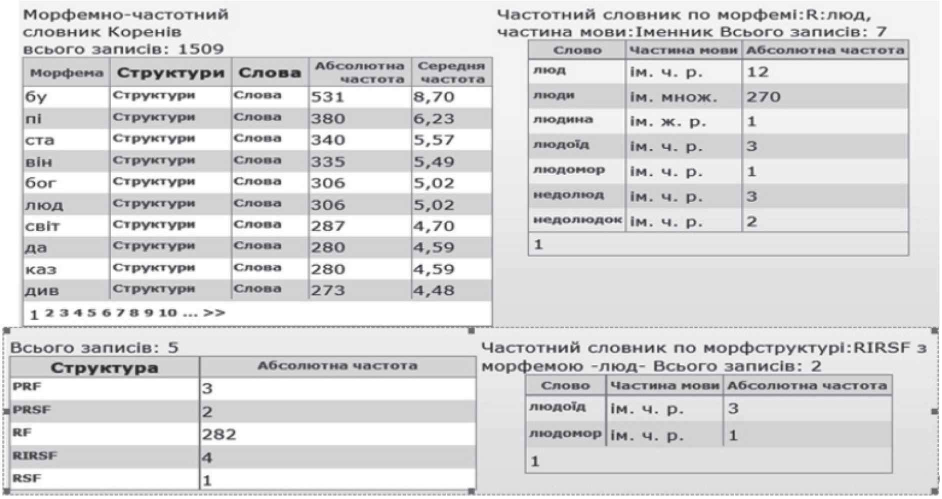


morfemid	wrdFK	mname	mseg
799	396	P	в
800	396	R	гас
801	396	S	а
802	396	F	ти
803	397	P	в
804	397	R	голос
805	398	P	в
806	398	R	город
807	398	S	и
808	398	F	ти
809	399	P	в
810	399	R	гор
811	399	S	у
812	400	P	в
813	400	R	дав
814	400	S	и
815	400	F	ти
816	401	R	вдар
817	401	S	и
818	401	F	ти

Fig. 5. Fragment of the database of morphemic segmentation of initial word forms

Different kinds of alphabetic-frequency dictionaries of morphemes and morphemic word structures are compiled on the basis of these databases according to the text samples of certain authors or style subcorpora. These dictionaries are presented in the corpus category of "Frequency Dictionaries" [4] as free-running electronic lexicographical systems. Today there are frequency dictionaries based on 13 text samples: frequency dictionaries of 9 authors, frequency dictionaries of opinion journalism, frequency dictionaries of artistic prose, frequency dictionaries of folklore texts, and frequency dictionaries of endocrinology. The dictionaries are structured into three zones combined by a comfortable navigation: 1. Inventory of units (morph structures, roots, affixes). 2. A token of a morph structure / morpheme in the text words according to the following characteristics: a) inventory and the number of words (lexemes) of each morph structure, root, affix; b) part of speech categorization of a word; c) absolute frequency of the word use (a number of lexeme tokens in texts); d) the number of texts (works), where the word is used; e) the average frequency of the word use; f) standard deviation; g) stability rate; 3. Contextual use of the analysed word (concordance).

For example, Electronic dictionary of frequency of morphemes [6], which was automatically constructed based on ≈ 80,000 tokens of T. Shevchenko's poetic speech in the Corpus of the Ukrainian language. Inventory of units for the dictionary of frequency of morphemes is compiled automatically based on a user's choice. In the dropdown lists according to the two filters: 1) a type of morpheme, 2) a part of speech; a user may choose: 1) units of analysis, in our case – the roots; 2) morphological field of sample words: all the text words or the words of a one part of speech – nouns in the demo version.



Морфемно-частотний
словник Коренів
всього записів: 1509

Морфема	Структури	Слова	Абсолютна частота	Середня частота
бу	Структури	Слова	531	8,70
пі	Структури	Слова	380	6,23
ста	Структури	Слова	340	5,57
він	Структури	Слова	335	5,49
бог	Структури	Слова	306	5,02
люд	Структури	Слова	306	5,02
світ	Структури	Слова	287	4,70
да	Структури	Слова	280	4,59
каз	Структури	Слова	280	4,59
див	Структури	Слова	273	4,48

1 2 3 4 5 6 7 8 9 10 ... >>

Частотний словник по морфемі:R:люд,
частина мови:Іменник Всього записів: 7

Слово	Частина мови	Абсолютна частота
люд	ім. ч. р.	12
люди	ім. множ.	270
людина	ім. ж. р.	1
людоїд	ім. ч. р.	3
людомор	ім. ч. р.	1
недолюд	ім. ч. р.	3
недолюдок	ім. ч. р.	2

1

Всього записів: 5

Структура	Абсолютна частота
PRF	3
PRSF	2
RF	282
RIRSF	4
RSF	1

Частотний словник по морфструктурі:RIRSF з
морфемою -люд- Всього записів: 2

Слово	Частина мови	Абсолютна частота
людоїд	ім. ч. р.	3
людомор	ім. ч. р.	1

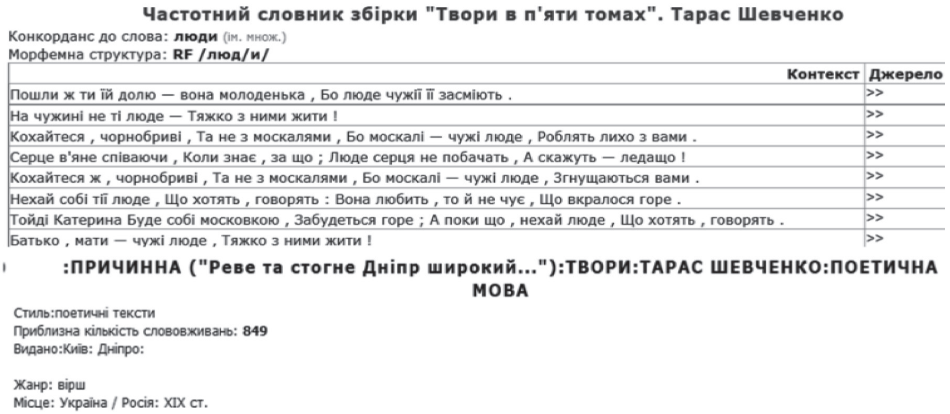
1

Fig. 6. A fragment of the Dictionary of frequency of roots

A fragment of the Dictionary of frequency of roots (Fig. 6) shows that the vocabulary of noun lexicon roots (zone 1) of T. Shevchenko's poetic speech contains 1,509 roots with information about the absolute and average frequency of each root's use in text tokens (in all text tokens regardless part of speech categorization of words). It is possible to create a root inventory according to the fall or rise of absolute frequencies.

The second zone of the dictionary presents the lexical token of specific root selected in the first zone: this fragment demonstrates the token of *–люд–*, a frequently used root (with an absolute frequency of 306 tokens) in noun tokens of the text. This root is materialized in 7 words with different productivity. The most frequent word is *люди*, which occurs in poetic speech in 270 tokens of the text.

Activation of "Structure" option in the first zone opens the list of morphemic structure models of specific root (the third zone is the bottom left table), where the root -*люд*- is materialized. Activation of the specific morph-structure model opens the window of lexical realization of the specific structure (the fourth zone is the bottom right table). The fragment demonstrates the realization of RIRSF model of the root -*люд*- in two words: *люд/о/їд/Ø/Ø, люд/о/мор/Ø/Ø*.



Частотний словник збірки "Твори в п'яти томах". Тарас Шевченко

Конкорданс до слова: **люди** (ім. множ.)
Морфемна структура: **RF /люд/и/**

Контекст	Джерело
Пошли ж ти їй долю — вона молоденька , Бо люде чужії її засміють .	>>
На чужині не ті люде — Тяжко з ними жити !	>>
Кохайтеся , чорнобриві , Та не з москалями , Бо москалі — чужі люде , Роблять лихо з вами .	>>
Серце в'яне співаючи , Коли знає , за що ; Люде серця не побачать , А скажуть — ледащо !	>>
Кохайтеся ж , чорнобриві , Та не з москалями , Бо москалі — чужі люде , Згнущаються вами .	>>
Нехай собі тії люде , Що хотять , говорять : Вона любить , то й не чує , Що вкралося горе .	>>
Тойді Катерина Буде собі московкою , Забудеться горе ; А поки що , нехай люде , Що хотять , говорять .	>>
Батько , мати — чужі люде , Тяжко з ними жити !	>>

:ПРИЧИННА ("Реве та стогне Дніпр широкий..."):ТВОРИ:ТАРАС ШЕВЧЕНКО:ПОЕТИЧНА МОВА

Стиль:поетичні тексти
Приблизна кількість слововживань: **849**
Видано:Київ: Дніпро:

Жанр: вірш
Місце: Україна / Росія: XIX ст.

Fig. 7. Text of the lexeme *люди*

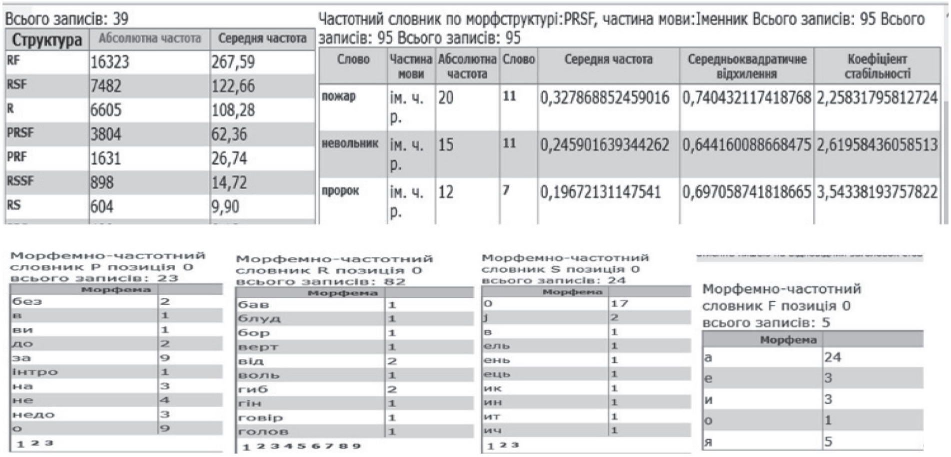
Activation of the specified word in the "Dictionary of frequency of morphemes" (upper right table) opens the window of contextual word usage (Fig. 7). Each text fragment is connected to the extralinguistic information about the work by the function "Source" in the corpus: Fig. 7 demonstrates information about the work using the first quote of the concordance.

Automatic dictionary of frequency of morphemes word structures has the same arrangement: dictionary of frequency of morph-structures, realization of the specified morph-structure in the lexicon sample, realization of the word of the specified morph-structure in texts. Fig. 8 illustrates the realization of the high-frequency morph-structure PRSF (3,804 text tokens) in 95 nouns in the texts of T. Shevchenko.

The right upper table (Fig. 8) presents the model of PRSF morphemic word structure. According to the selection of each morph's index in this model (P – prefix, R – root, S – suffix, F – ending) frequency dictionaries of morphemic are automatically compiled for those morphs, which occur in the sample vocabulary according to the model PRSF (bottom tables of the figure): prefixes, roots, suffixes, endings (only based on the initial word forms).

5 CONCLUSION

The use of ASMDA as an automatic morphemic segmentator of initial forms and a constructor of dictionaries of frequency of morphemes in the Corpus of the Ukrainian language reveals new research facilities of morphemic structure of Ukrainian words in a dictionary and in a text. The obtained statistics can be used in systemic stylistic research [12]: the comparison of frequency characteristics of morphemic units in different text samples of the corpus and in language system predicts realization expectancy of morphemic units in different texts, and formalizes the notion of statistical text structure at morphemic level as its stylometric model. Systematization of morph-structures allows to analyse the inclusion of morphemic models into the formation of new corpus vocabulary, to investigate morphemic length and intensity of words in Ukrainian texts of different styles as well as morphotactics of different morpheme types. The relation of dictionaries of frequency of morphemes to the concordance allows to analyse different aspects of morphemes and morph-structures functioning in sentences.



Всього записів: 39

Структура	Абсолютна частота	Середня частота
RF	16323	267,59
RSF	7482	122,66
R	6605	108,28
PRSF	3804	62,36
PRF	1631	26,74
RSSF	898	14,72
RS	604	9,90

Частотний словник по морфструктурі:PRSF, частина мови:Іменник Всього записів: 95 Всього записів: 95 Всього записів: 95

Слово	Частина мови	Абсолютна частота	Слово	Середня частота	Середньоквадратичне відхилення	Коефіцієнт стабільності
пожар	ім. ч. р.	20	11	0,327868852459016	0,740432117418768	2,25831795812724
невольник	ім. ч. р.	15	11	0,245901639344262	0,644160088668475	2,61958436058513
пророк	ім. ч. р.	12	7	0,19672131147541	0,697058741818665	3,54338193757822

Морфемно-частотний словник P позиція 0 всього записів: 23

Морфема	
без	2
в	1
ви	1
до	2
за	9
інтро	1
на	3
не	4
недо	3
о	9

1 2 3

Морфемно-частотний словник R позиція 0 всього записів: 82

Морфема	
бав	1
блуд	1
бор	1
верт	1
від	2
воль	1
гиб	2
гін	1
говір	1
голов	1

1 2 3 4 5 6 7 8 9

Морфемно-частотний словник S позиція 0 всього записів: 24

Морфема	
о	17
j	2
в	1
ель	1
ень	1
ець	1
ик	1
ин	1
ит	1
ич	1

1 2 3

Морфемно-частотний словник F позиція 0 всього записів: 5

Морфема	
а	24
е	3
и	3
о	1
я	5

Fig. 8. Fragment of electronic dictionary of morph-structures

The experience of carrying out the automatic morphemic analysis demonstrates that it is not obligatory to create morphemic or word-formative annotation of texts for the purpose of retrieval of relational and functional characteristics of morphemic units from the text. The methodology for morphemic annotation of lexeme dictionary of texts (the initial forms) does not reduce efficiency and immediacy in linguistic research; on the contrary, it increases resolution capacity of research due to systematization and various classification of morphemic information. A study of text organization at morphemic level by means of morphemic analysis of initial word forms instead of tokens is justified by experience in compilation frequency dictionaries of morphemic in the Corpus of the Ukrainian language. The approach is based on ontological organization of morphemic word structure of inflected languages: the morphemic structure of the word stem representing explicit word semantics remains relatively stable during inflection. The quantity of morphemes does not change, only inflectional allomorphy may appear, but it is taken into account during lemmatization and may be automatically ascribed to morphemes as a potential feature. The use of morphemic annotation methodology of dictionary of the initial word forms in the Corpus of the Ukrainian language demonstrates efficiency and optimality of this methodology: ≈ 200,000 units of the ASMDA morphemic database allow to obtain information about morphemic structure of ≈ 50 million text tokens with illustration of their contextual usage. The disadvantage of this methodology is impossibility to remove word homonymy of one part of speech in case of different morphemic segmentation of homographs: *вида-ти (to print), ви-да-ти (to hand), вид-а-ти (to see)*. In such cases morphemic segmentation is edited manually.

The development of ASMDA at this stage sets the task of automatic compilation of conjugate samples on the basis of ASMDA lexical list taking into account homonymy and allomorphy of roots, as well as compilation of common-affix samples taking into account homonymy and allomorphy of affixes. There is work underway to verify the database of homonymic and allomorphic roots and to compile a database of homonymic and allomorphic affixes. The methodology of computer modelling in the process of ASMDA construction summarizes theoretical and applied ideas of modern linguistics, which makes this system an efficient and rational tool for linguistic research.

References

[1] Alekseenko, L. A. et al. (2004). Parametrizirovannaya baza dannykh poeticheskoy rechi kak istochnik i instrument filologicheskikh studiy. In *Mezhdunarodnaya konferentsiya "Prikladnaya lingvistika bez granits". Materialy konferentsii*, pages 80–87, Sankt- Peterburg.

[2] Aleksijenko, L. A., Darčuk, N. P., and Zuban', O. N. (2001). Metodyka stvorennja avtomatyzovanoji systemy morfemno-slovotvirnoho analizu (ASMSA) sliv ukrajins'koï movy. In *Naukova spadščyna profesora S. V. Semčyns'koho. Zbirnyk naukovych prac'*, pages 38–49, Kyiv.

[3] Bogdanov, S. I. (1997). *Forma slova i morfologicheskaya forma.* Izd. Sankt-Peterburgskogo universiteta, Sankt-Peterburg.

[4] Častotni slovnyky Korpusu. Accessible at: `http://www.mova.info/article.aspx?l1=210&DID=5215`, retrieved 2017-03-15.

[5] Darčuk, N. (2013). *Komp"juterne anotuvannja tekstu: rezul'taty i perspektyvy: monohrafija*. Kyiv.

[6] Elektronnyj slovnyk movy T. Ševčenka. Accessible at: `http://www.mova.info/cfqsh.aspx`, retrieved 2017-03-15.

[7] Jacenko, I. T. (1980 – 1981). *Morfemnyj analiz: Slovnyk-dovidnyk* . Kyiv. T. 1–2.

[8] Klymenko, N. F. et al. (2014). Morfemno-slovotvirnyj fond ukrajins'koji movy jak doslidnyc'ka ta informacijno-dovidkova systema. In *Klymenko N. F. Vybrani praci*, pages 545–558, Kyiv.

[9] Korpus ukrajins'koji movy. Accessible at: `http://www.mova.info/corpus.aspx`, retrieved 2017-03-15.

[10] Kotova, N. V. and Yanakiev, M. O. (1978). O mnogoobrazii morfem v slavyanskikh yazykakh. *Slavyanskaya filologiya*, X:4–8.

[11] Solntsev, V. M. (1971). *Yazyk kak sistemno-strukturnoe obrazovanie.* Moscow.

[12] Zuban, O. (2016). Častotni morfemni slovnyky v Korpusi ukrajins'koï movy - džerelo stylemetryčnych doslidžen'. In *Acta Universitatis Palackianae Olomucensis Philologica. UCRAINICA VII: Současná ukrajinistika. Problémy jazyka, literatury a kultury*, pages 22–33, Olomouc.

[13] Zuban, O. (2015). Morphemic and derivational analysis in the corpus of the Ukrainian language. *Ukrajins'ke movoznavstvo*, 1(45):3–10.

[14] Zuban, O. N. (2016). Zadachi i metody avtomaticheskogo morfemnogo analiza v Korpuse ukrainskogo yazyka. In *Aktualnye problemy sovremennoy prikladnoy lingvistiki*, pages 122–129, Minsk.

JÁN HORECKÝ‘S APPROACH TO LANGUAGE AND THINKING

MIROSLAV ZUMRÍK
Ľudovít Štúr Institute of Linguistics, Slovak Academy of Sciences, Bratislava, Slovakia

ZUMRÍK, Miroslav: Ján Horecký‘s Approach to Language and Thinking. Journal of Linguistics, 2017, Vol. 68, No 2, pp. 426 – 431.

Abstract: The paper aims to reflect on theoretical foundations of Horecký’s approach to the relation between language (and more specifically: terms) and thinking (concepts). Reflections are devoted to Horecký’s explicit and implicit beliefs on the nature of terms and concepts and their mutual relation, as well as their relation to reality around. Definitions of both term and concept appear in some of Horecký’s major papers. The paper focuses on models of term-concept relation proposed in those papers. Finally, an attempt is made to find some convergences and divergences in theories of Horecký and the Czech logician Pavel Tichý.

Keywords: terms, concepts, philosophy, Ján Horecký, Pavel Tichý, logical spectrum, Transparent Intensional Logic

1 LANGUAGE AND THINKING

In his book *Spoločnosť a jazyk* (Society and Language), Horecký writes: *“The outer reality is always reflected in the natural language in some way, that is, it is designated by linguistic means. The relation of objective reality, thinking and language is often represented by the so called semantic triangle”* [3, p. 16].

This is basically an Aristotelian view in the sense that objective reality, thinking and language exist as three separate domains. These three nevertheless interact with each other, which means that the questions of philosophy are not restricted to the limits of either categories of mind (as for Kant) or those of language (as for Wittgenstein). The outer reality is, as Horecký states, *always* reflected in the natural language in *som*e way. This leaves, on the one hand, room for further interpretation, as long as the way of this reasoning is left without further specification. On the other hand, Horecký clearly suggests that the relation between outer reality and natural language is that of a reflection. At the same time, Horecký admits that this reflection can be labeled as designation. He introduces classic semantic triangle consisting of the object/thing (reality), the name of the thing (language), and the concept, generalization of the thing (thinking). The relation of the name of the thing and the thing itself is, contrary to name-concept, and concept-thing relations, *“immediate only during communication, that is, in cases where a given name is used to directly naming a certain thing”* [3, p. 17]. In other cases, the relation between a name (language) and the thing (object, reality) is mediated through several levels (logical, semantic, onomasiological). It is this complex and dynamic form of mediation between the language, thinking and the world around where, as I would argue,

DOI 10.1515/jazcas-2017-0052

Horecký's theory proves inspirational for further discussion on the nature of concept/term relation. In this way, the paper continues in the line of the research that seeks to uncover inspirations stemming from Horecký's linguistic and terminological achievements, such as in [5].

Horecký states that it is not only a string of characters/sounds that constitute a sign which would then stand for a given object [3]. The sign can only fulfill its function when it is linked to a given concept of the object. This means that it is both the domain of language and the domain of thinking that "stand for" the domain of reality: the sign is not unilateral, but bilateral. A sign has two aspects, one of them being the chain of sounds and the other being the concept. Let us see closer now at the way how the domains of thinking and language "stand for" the reality.

Though Horecký assumes that the "objective" reality is somehow reflected in our thinking and the language, his understanding is more subtle. However, Horecký's belief in reality existing independently on language remains intact. He distances himself from the view of L. Weisberger who claims that language actually forms certain categories independently of the objective reality [3, p. 24]. For Horecký, language is an instrument of thinking. The language, contrary to the extreme form of linguistic relativism, does not determine human thinking. As an example, Horecký mentions the cultures of Hungary and Slovakia that have much in common despite quite different features of the Hungarian and Slovak language.

"The relation between a name and its object in extralingual reality is not that simple as represented in semantic triangle. Not even the concept "concept" *is fully unproblematic*1197040685*"* [3, p. 18]. The reality for him is not just plainly reflected – projected – into thinking and language, but these three domains are rather intertwined in a dynamic and, as Horecký puts it, dialectic way. The dialectical relation (mutual dependency) of the domains is possible thanks to their structural features, that enable interaction, overlapping and correspondence between elements of the domains (objects, concepts and names/terms): *"relation between concept and term is built on the very nature of concept – that its content consists of a set of features – and on the very quality of language – that it can express at least one of these features thanks to expressive means of language"* [1, p. 101]. These structural features Horecký explores in another paper [4], where he describes system of "conceptualizers", grammatical and semantic categories that enable expressing structure of reality in the structure of language.

As already mentioned, Horecký stresses that the relation between the name and the named object is rather layered, and that even concept of a concept is not self-explaining. Here, it is possible to mention current discussions on the ontological status and further features of concepts as entities, and draw a possible parallel between these discussions and peculiarities of Horecký's approach. Amongst many participants in the discussions, I would like to name inspiring work of P. Tichý, creator of the logical system called Transparent Intensional Logic [6]. Tichý claims, that names (or terms) refer not directly to objects of reality, but rather to specific abstract entities – procedures that construct these objects [7, p. 131]. Concepts in TIL are defined as a special type of these constructions. Some hints of procedural and constructionist understanding of concepts can be found even in Horecký's approach. Horecký believes that we use concepts in order to describe and process reality. When describing an unknown object of reality, we actually describe various features of this object, we assign these features to the object and thus construct an open set of logical predicates, the logical spectrum [3, p. 18], which is a term introduced by Watanabe [10]. As an example, Horecký uses the term *parketáreň* (parquetry factory), and its logical spectrum: (1) it is a certain place or facility (this we know thanks to the suffix *-áreň*), (2) where something is produced (again, this we know because of the suffix), (3) production of this something is in some way organized (because of the suffix) and (4) this something is parquetry (this we know from the root of the word, *parket-*, stemming from "parquetry"). Since logical predicates result from mental activities, they are not objective in the way logicians like P. Tichý assign concepts and cognitions their objective existence.

We have seen that Horecký presupposes existence of certain linguistic features that enable expressing logical predicates in language. These linguistic means of expressing logical predicates he calls semantic features. To every logical predicate, there exists a semantic feature in the domain of language. In case of the term *parketáreň* (parquetry factory), he names semantic features contained therein as *miestovosť* (placeness), *výrobnosť* (produceness), *organizovanosť* (organizedness) and *predmetnosť* (objectness) [3, p. 19].

The set of logical predicates is open and the predicates can be added, removed or replaced due to changes in the actual state of (scientific) knowledge. For a given concept, there can exist several logical spectra, depending on philosophical standing point of a researcher, state-of-the-art within a given field etc. Horecký thus adheres to a "dynamic concept of a concept" that goes beyond the somehow "static" definition of a concept, defined as a generalized reflection of those aspects of a given object that are essential for human knowledge [2, p. 130]. The logical spectrum is also defined as a set of propositions that do not contradict each other, while each one expresses some feature of a given concept (ibid.). This opens for a dynamic understanding of concepts, whose content undergoes changes in the course of time. Therefore, these are in no way "eternal" Platonic objects, but cognitive constructions and instruments that enable us to think and speak of the outer, common reality.

According to Horecký, not every semantic feature expressing the logical predicates from the set of logical spectrum has to be expressed in the term itself [3, p. 19]. There is, however, always possible to discern two elementary parts, onomasiological base and onomasiological specification (*príznak*). In the example term "parquetry factory", there is semantic feature of "placeness" expressed in the onomasiological base and the name of the produced object expressed in the onomasiological specification. Which features of the onomasiological base will appear in the name of an object is decided at the onomatological level. This distinction leads to, as Horecký puts it, rather complicated scheme of reality-language relation. The scheme consists of six levels with corresponding entities (level of reality – objects, logical level – logical predicates, semantic level – semantic features, onomasiological level – morphemes/words, onomatological level – derivational morphemes/words and phonological level – morphemes, phonemes. The first three levels are separated from remaining three, which means that the thinking and objective being are not directly part of a language sign. However, language and thinking are, in Horecký's opinion, both distinct and tightly interconnected and can be described as complex dialectic unity, which then needs to be examined from multiple perspectives [3, p. 20]. Horecký introduces a system of four domains that together build an opposition to the domain of reality. These four domains are: language as a system of signs (*jazyk*, J), language as usage of the system of signs (*reč*, R), thinking (*myslenie*, M) as reflection of reality based on sensory perception, and as psychic activity of a subject. Finally, there is an exclusive domain of human thinking, the logical thinking (L), consisting of logical operations and information processing. Horecký organizes these four domains as vertices of a square, where there is direct and mutual dependence between vertices on the same side (M with J and with L, J with M and with R, R with J and with L, L with M and with R), but not across the square. This means that there is no direct connection between thinking (M) and language as usage of the system of signs (R), nor between language as sign system (J) and logical thinking/logical operations (L). This square scheme applies to the so called cognitive thinking (*poznávacie myslenie*), while there is also a communicative thinking (*komunikatívne myslenie*), whose function is to process knowledge into information, which, consequently, can be transmitted to others [3, p. 22].

2 TERMS AND CONCEPTS

Recurring definitions of term and concept can be found throughout Horecký‘s bibliography. Apart from his seminal monograph Foundations of Slovak terminology (1956), one could mention Horecký‘s papers published in various scholarly journals like *Kultúra slova*, *Jazykovedný časopis*, *Slovenské odborné názvoslovie* or *Naša reč*.

The term belongs to the domain of language, whereas concept is an entity within the domain of thinking. Horecký´s definition of term originates from terminological practice and takes into account the relation of term and concept: term is a name of a concept that is delimited by the concept‘s definition and by its position within conceptual system of a given scientific field [2, p. 130]. In an earlier version of this definition, a term is the name of a concept within a given scientific field or production area. Every term only has meaning and fulfills its function within the conceptual system of a given scientific field [1, pp. 98–99]. Concept is defined as a cognitive summary of individual objects (be it material or immaterial) on the basis of their common features. [1, p. 99].

The relation of onomatological structure of a term and the logical spectrum of a concept is not univocal: not every feature from the logical spectrum finds its expression in the structure of a term (ibid.). Normally, only the most significant (prominent) feature (*differentia specifica*) of a given concept is expressed through the structure of the term, while the other features are present implicitly, in the location of a concept/term within other concepts/terms of a given field. Which feature will be present in the structure of a term depends on existing or possible models for creating terms in a given language. The variety of such models is not unlimited, it is a “closed system” [1, p. 100].

At the same time, the relation of onomatological structure and the logical spectrum of a concept is, as Horecký puts it, "dialectical" (mutually dependent). Logical spectrum changes faster than the onomatological structure of a term, because the spectrum reflects progress in scientific knowledge [2, p. 131]. Finally, this relation is mediated, with the onomasiological and semantic structure added between the former two levels. Onomasiological structure of a term consists of the base and specification, semantic structure consists of "meaning constructions" created by human mind (ibid.).

The most prominent feature in the logical spectrum of a concept becomes the main "motivational feature", expressed in the form of a term [1, p. 100]. The motivational feature can either express inherent features of objects in question, or their relational features, which follow from the objects' relation to other objects. In each language, there are some general models of how these features are used for coining terms. In Slovak, Horecký distinguishes two models of creating terms using inherent features: one model he calls "adjective" (*adjektívny*), the second one "verbal" (*slovesný*). Other models are based on relational features. Some of these models express general relation of a feature to the entity in question, other express relations of location, purpose, material, time or similarity.

3 HORECKÝ AND TICHÝ

Finally, as hinted in Section 1, I will turn to the question whether there can be traced some similarities or differences between Horecký's and P. Tichý's approach to meaning and reference. In his paper *What do we talk about?* [8] Tichý distinguishes extensional and intensional entities and focuses on their reference. If it is true that intensional entities (propositions, individuals-in-intension, qualities, relations etc.) can be defined as functions that yield values (extensional entities like truth values, individuals, numbers or sets of individuals), do intensions exclusively refer to extensional entities? Tichý advocates quite the opposite, that by mentioning intensional entities, we actually refer to intensional entities themselves [8, p. 39]. Horecký would probably agree with Tichý, given that for Horecký, the terms are names of concepts as abstract entities (like intensions), not names of real objects (like extensions). However, Horecký adopts the view that concepts are mental generalizations located in individual and social consciousness [1], whereas Tichý believes they have an objective, albeit abstract existence, which we can and we do refer to, when we speak. Logicians like P. Tichý or H. Putnam think of terms like "planet", "crow" etc. like of functions and qualities ("to be a planet", "to be a crow"), not like of sets of individuals [8, p. 42]. Horecký believes concepts are generalizations, in other words, mental representations of sets of individuals. It is possible that Horecký would agree with Tichý, who claimed that by using terms like "planet" or "crow", we do not refer to individuals or sets thereof, but instead we refer to a certain quality understood as a function or a role named "being a planet" or "being a crow". In another paper [9], Tichý explains that the sentence "Mary's husband is dead" is not about Mary's real husband, but about the role of being person in certain relation to Mary, in other words, about the role of being husband of Mary, whoever this person might be. “Mary’s husband” is, then, not a name of a real person, but a name of a certain function, called individual role. It is obvious that such claim might sound rather counterintuitive. Where Tichý builds semantics of natural language using a logical calculus, Horecký adheres to rather Aristotelian metaphysics, albeit with faceted model of relation between reality, language and thinking.

I would like to conclude that despite differences in their philosophical standpoints, as well as substantial geographical distance between Horecký and Tichý (who exiled to New Zealand in 1970), there are some similarities in their approaches. Even without Tichý‘s refined logical system with rich procedural semantics, Horecký throughout his life reflected upon the complexities and dynamics of reality-language relation. Both Horecký and Tichý, then, were reluctant to clear-cut and simplistic models of how and with what reference we use natural language.

ACKNOWLEDGEMENTS

This article was written within the VEGA project n^{o} 2/0114/15 *Analysis of terminology work of Ján Horecký as an inspiration for the terminology management of the 21st century in Slovakia* implemented at the Slovak National Corpus Department.

References

[1] Horecký, J. (1960). Vzťah pojmu a termínu. *Jazykovedný časopis*, 11(2):97–102.

[2] Horecký, J. (1974). Základné problémy terminológie. *Kultúra slova*, 8(5):129–132.

[3] Horecký, J. (1982). *Spoločnosť a jazyk*. Veda, Bratislava.

[4] Horecký, J. (1987). Systém konceptualizátorov. *Slovenská reč*, 52(4):207–214.

[5] Šimková, M., editor (2015). Človek a jeho jazyk. 3. Inšpirácie profesora Jána Horeckého. Veda, Bratislava.

[6] Tichý, P. (1988). *The Foundations of Frege’s Logic*. De Gruyter, Berlin and New York.

[7] Tichý, P. (1996a). Konstrukce. In *O čem mluvíme? Vybrané stati k logice a sémantice*, pages 119–146, Filosofia, Praha.

[8] Tichý, P. (1996b). O čem mluvíme? In *O čem mluvíme? Vybrané stati k logice a sémantice*, pages 37–58, Filosofia, Praha.

[9] Tichý, P. (1996c). De dicto a de re. In *O čem mluvíme? Vybrané stati k logice a sémantice*, pages 79–94, Filosofia, Praha.

[10] Watanabe, S. (1969). *Knowing and guessing – a quantitative study of inference and information*. John Wiley & Sons, New York.

POKYNY PRE AUTOROV

Redakcia JAZYKOVEDNÉHO ČASOPISU uverejňujeme príspevky **bez poplatku** za publikovanie.
Akceptované jazyky: všetky slovanské jazyky, angličtina, nemčina. Súčasťou vedeckej štúdie a odborného príspevku je abstrakt v angličtine (100 – 200 slov) a zoznam kľúčových slov v angličtine (3 – 8 slov).
Súčasťou vedeckej štúdie a odborného príspevku v inom ako slovenskom alebo českom jazyku je zhrnutie v slovenčine (400 – 600 slov) – preklad do slovenčiny zabezpečí redakcia.
Posudzovanie príspevkov: vedecké príspevky sú posudzované anonymne dvoma posudzovateľmi, ostatné príspevky jedným posudzovateľom; priemerná dĺžka vypracovania posudkov je 1 mesiac. Autori dostávajú znenie posudkov bez mena posudzovateľa.
Technické a formálne zásady:

- Príspevky musia byť v elektronickej podobe (textový editor Microsoft Word, font Times New Roman, veľkosť písma 12 a riadkovanie 1,5). V prípade, že sa v texte vyskytujú zvláštne znaky, tabuľky, grafy a pod., je potrebné odovzdať príspevok aj vo verzii pdf alebo vytlačený.
- Pri mene a priezvisku autora je potrebné uviesť tituly a pracovisko.
- Text príspevku má byť zarovnaný len z ľavej strany, slová na konci riadku sa nerozdeľujú, tvrdý koniec riadku sa používa len na konci odseku.
- Odseky sa začínajú zarážkou.
- Kurzíva sa spravidla používa pri názvoch prác a pri uvádzaní príkladov.
- Polotučné písmo sa spravidla používa pri podnadpisoch a kľúčových pojmoch.
- Na literatúru sa v texte odkazuje priezviskom autora, rokom vydania a číslom strany (Horecký, 1956, s. 95).
- Zoznam použitej literatúry sa uvádza na konci príspevku (nie v poznámkovom aparáte) v abecednom poradí. Pokiaľ obsahuje viac položiek jedného autora, tie sa radia chronologicky. V príspevkoch v rubrikách Recenzie, Referáty a Kronika sa bibliografické údaje uvádzajú priamo v príspevku.

Bibliografické odkazy:

- knižná publikácia: ONDREJOVIČ, Slavomír: Jazyk, veda o jazyku, societa. Bratislava: Veda 2008. 204 s.
- slovník: JAROŠOVÁ, Alexandra – BUZÁSSYOVÁ, Klára (eds.): Slovník súčasného slovenského jazyka. H – L. [2. zv.]. Bratislava: Veda 2011. 1088 s.
- štúdia v zborníku: ĎUROVIČ, Ľubomír: Jazyk mesta a spisovné jazyky Slovákov. In: Sociolinguistica Slovaca 5. Mesto a jeho jazyk. Ed. S. Ondrejovič. Bratislava: Veda 2000, s. 111 – 117.
- štúdia v časopise: DOLNÍK, Juraj: Reálne vz. ideálne a spisovný jazyk. In: Jazykovedný časopis, 2009, roč. 60, s. 3 – 12.
- internetový zdroj: Slovenský národný korpus. Verzia prim-5.0-public.all. Bratislava: Jazykovedný ústav Ľudovíta Štúra SAV, 2010. Dostupný na: http://korpus.juls.savba.sk.

INSTRUCTION FOR AUTHORS

JOURNAL OF LINGUISTICS publishes articles **free of publication charges**.
Accepted languages: all Slavic languages, English, German. Scientific submissions should include a 100-200 word abstract in English and a list of key words in English (3-8 words).
Scientific articles in a language other than Slovak or Czech should contain a summary in Slovak (400-600 words) – translation into Slovak will be provided by the editor.
Reviewing process: scientific articles undergo a double-blind peer-review process and are reviewed by two reviewers, other articles by one reviewer; the reviewing process takes 1 month on average. The authors are provided with the reviews without the name of the reviewer.
Technical and formal directions:

- Articles must be submitted in an electronic form (text editor Microsoft Word, 12-point Times New Roman font, and 1.5 line spacing). If the text contains special symbols, tables, diagrams, pictures etc. it is also necessary to submit a pdf or printed version.
- Contributions should contain the full name of the author(s), as well as his/her title(s) and institutional affiliation(s).
- The text of the contribution should be flush left; words at the end of a line are not hyphenated; a hard return is used only at the end of a paragraph.
- Paragraphs should be indented.
- Italics is usually used for titles of works and for linguistic examples.
- Boldface is usually used for subtitles and key terms.
- References in the text (in parentheses) contain the surname of the author, the year of publication and the number(s) of the page(s): (Horecký, 1956, s. 95).
- The list of references is placed at the end of the text (not in the notes) in alphabetical order. If there are several works by the same author, they are listed chronologically. Submissions to the journal sections “Book Reviews”, “Book Notices” and “Chronicle” should have references included directly in the text.

References:

- Monograph: ONDREJOVIČ, Slavomír: Jazyk, veda o jazyku, societa. Bratislava: Veda 2008. 204 pp.
- Dictionary: JAROŠOVÁ, Alexandra – BUZÁSSYOVÁ, Klára (eds.): Slovník súčasného slovenského jazyka. H – L. [2. zv.]. Bratislava: Veda 2011. 1088 pp.
- Article in a collection: ĎUROVIČ, Ľubomír: Jazyk mesta a spisovné jazyky Slovákov. In: Sociolinguistica Slovaca 5. Mesto a jeho jazyk. Ed. S. Ondrejovič. Bratislava: Veda 2000, pp. 111 - 117.
- Article in a journal: DOLNÍK, Juraj: Reálne vz. ideálne a spisovný jazyk. In: Jazykovedný časopis, 2009, Vol. 60, pp. 3 – 12.
- Internet source: Slovenský národný korpus. Verzia prim-5.0-public.all. Bratislava: Jazykovedný ústav Ľudovíta Štúra SAV, 2010. Dostupný na: http://korpus.juls.savba.sk.

ISSN 0021-5597 (tlačená verzia/print) ISSN 1338-4287 (verzia online) MIČ 49263

JAZYKOVEDNÝ ČASOPIS

VEDECKÝ ČASOPIS PRE OTÁZKY TEÓRIE JAZYKA

JOURNAL OF LINGUISTICS

SCIENTIFIC JOURNAL FOR THE THEORY OF LANGUAGE

Objednávky a predplatné prijíma/Orders and subscriptions are processed by:
SAP – Slovak Academic Press s.r.o., Bazová 2, 821 08 Bratislava
e-mail: sap@sappress.sk

Registračné číslo 7044

Evidenčné číslo 3697/09

IČO vydavateľa 00 167 088

Ročné predplatné pre Slovensko/Annual subscription for Slovakia: 8 €, jednotlivé číslo 4 €
Časopis je v predaji v kníhkupectve Veda, Štefánikova 3, 811 06 Bratislava 1