Slovenská akadémia vied
Jazykovedný ústav Ľudovíta Štúra

Natural Language Processing, Corpus Linguistics, E-learning

Seventh International Conference
Bratislava, Slovakia, 13–15 November 2013
Proceedings

Editors
Katarína Gajdošová
Adriána Žáková

RAM-Verlag
2013

The articles have been reviewed by members of the Program Committee.

Slovak National Corpus
Ľ. Štúr Institute of Linguistics
Slovak Academy of Sciences
Bratislava, Slovakia 2013
http://korpus.juls.savba.sk/~slovko/

Table of Contents

Foreword
Mária Šimková 7

Úvod
Mária Šimková 8

A New Path to a Modern Monolingual Dictionary of Contemporary Czech: the Structure of Data in the New Dictionary Writing System
Kamil Barbierik, Martina Holcová Habrová, Pavla Kochová, Tomáš Liška, Zdeňka Opavská, and Miroslav Virius 9

Data Deduplication in Slovak Corpora
Vladimír Benko 27

Software System for Processing Bulgarian Digital Resources: Parallel Corpora and Bilingual Dictionaries
Ralitsa Dutsova and Ludmila Dimitrova 40

Slovene Corpora for Corpus Linguistics and Language Technologies
Tomaž Erjavec 51

Obstacles and Solution to Recognizing Compound Nouns in Greek: A Corpus Study
Vasiliki Foufi, Kyriaki Ioannidou, and Olympia Tsaknaki 63

From Multilingual Dictionary to Lithuanian WordNet
Radovan Garabík and Indrė Pileckytė 74

Corpora of Private Correspondence as a Source of Material Focused on a Research of Diminutives
Zdeňka Hladká 81

Identification of Idioms in Spoken Corpora
Milena Hnátková and Marie Kopřivová 92

The Corpus CzeSL in the Service of Teaching Czech for Foreigners – Errors in the Use of the Pronoun *který*
Andrea Hudousková 100

Delimitation of Participles in the Manual Morphological Annotation
Agáta Karčová 108

Corpus Based Identification of Czech Light Verbs
Václava Kettnerová, Markéta Lopatková, Eduard Bejček, Anna Vernerová, and Marie Podobová 118

Agents Expressed by Prepositionless Instrumental Modifying Czech Nouns Derived from Intransitive Verbs
Veronika Kolářová *129*

Corpus-based Online Word Formation Exercises for Advanced Learners of English – Challenges and Solutions
Grzegorz Krynicki *148*

Experimenting with Slovak Wikipedia as a Source for Language Technologies
Michal Laclavík, Štefan Dlugolinský, and Michal Blanárik *160*

Query Interface for Diverse Corpus Types
Tomáš Machálek and Michal Křen *166*

The Effect of Stop Words elimination on Sequence Patterns Extraction in Comparable Corpora
Dáša Munková, Michal Munk, and Martin Vozár *174*

Valency of Selected Primary Adjectives in the SYN2010 Corpus
Kateřina Najbrtová *183*

Event Extractor: Email Events Detection and Calendar Integration
Filip Ogurčák and Michal Laclavík *197*

Formal (Morpho)Syntax Properties of Reflexive Particles *se*, *si* as Free Morphemes in Contemporary Czech
Vladimír Petkevič *206*

Introduction to Online Learning
Katarína Pišútová *217*

Automatic Extraction of Multiword Units from Slovak Text Corpora
Ján Staš, Daniel Hládek, Jozef Juhár, and Martin Ološtiak *228*

Verb Valency and Argument Non-correspondence in a Bilingual Treebank
Jana Šindlerová, Zdeňka Urešová, and Eva Fučíková *238*

Determination of Czech BCT Prototypes on the Basis of Corpus Data
Tatiana Timoshchenko *248*

Veni, Vidi, Vici: The Language Technology Infrastructure Landscape after CESAR
Tamás Váradi *261*

Czech Subjectivity Lexicon: A Lexical Resource for Czech Polarity Classification
Kateřina Veselovská *279*

A Corpus-based Analysis of the Functionality and the Meaning of Infinitive “Frustrative Construction” in Czech and Slovak
Uliana Yazhinova *285*

Foreword

Slovko 2013 – Natural Language Processing, Corpus Linguistics, E-learning will be again held in Bratislava. The organizers – Slovak National Corpus Department of Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences are honoured to host participants from eight countries: Bulgaria, Czech Republic, Germany, Greece, Hungary, Poland, Slovakia and Slovenia.

Over three days participants will be able to benefit from 29 presentations, including 3 plenary talks. Unfortunately, one third of submitted papers on given topics has not been recommended by the Programme Committee members. We thank to all reviewers for their constructive suggestions and their help to make the conference even more successful.

The 7th edition of the biennial conference increased the presence of the linguistically-oriented (corpus-based and corpus-driven) studies. The more technically oriented papers provide information on effectiveness of the approaches applied, experimenting and innovative methods. Latest trends and tendencies in enhancing the corpus data can be found also in the papers written by Slovak authors.

We wish all participants of the conference Slovko 2013 profitable time and positive inspiration for further cooperation in the field of natural language processing, corpus linguistics and similar research.

Mária Šimková
Translated by Adriána Žáková

Úvod

Slovko 2013 – počítačové spracovanie prirodzeného jazyka, korpusová lingvistika, e-learning sa koná opäť v Bratislave a organizátori zo Slovenského národného korpusu Jazykovedného ústavu Ľudovíta Štúra Slovenskej akadémie vied vítajú na tomto podujatí účastníkov z ôsmich krajín: Bulharsko, Česká republika, Grécko, Maďarsko, Nemecko, Poľsko, Slovensko a Slovinsko.

Počas troch dní rokovania odznie celkovo 29 príspevkov, z toho 3 plenárne prednášky. Záujemcov o prezentáciu a publikovanie výsledkov svojej práce vo vymedzených tematických okruhoch bolo podstatne viac, no posudzovatelia z vedeckého výboru neodporúčali tretinu prihlásených príspevkov. Všetkým recenzentom ďakujeme za vykonanú prácu a za zvyšovanie kvality celého podujatia.

Na 7. ročníku našej bienálnej konferencie sa zvýšil podiel lingvisticky zameraných štúdií (corpus-based alebo corpus-driven), ktorých je takmer polovica. V technicky orientovaných príspevkoch prevládajú informácie o efektivite uplatňovaných postupov, experimentovanie a hľadanie inovatívnych metód. Nové oblasti výskumu a snahy o skvalitnenie korpusových dát a výstupov sa objavujú aj v príspevkoch slovenských autorov.

Všetkým účastníkom konferencie Slovko 2013 želáme užitočne strávený konferenčný čas a pozitívne prínosy z rokovaní pre ďalšie projekty v oblasti počítačového spracovania prirodzeného jazyka, korpusovej lingvistiky a súvisiacich výskumov.

Mária Šimková

A New Path to a Modern Monolingual Dictionary of Contemporary Czech: the Structure of Data in the New Dictionary Writing System

Kamil Barbierik, Martina Holcová Habrová, Pavla Kochová, Tomáš Liška, Zdeňka Opavská, and Miroslav Virius

Institute of the Czech Language of the Academy of Sciences of the CR, v. v. i., Prague, Czech Republic

Abstract. The article presents a new Dictionary Writing System (DWS) that is being developed at the Institute of the Czech Language of the Academy of Sciences of the CR, v. v. i., in connection with the preparation of a new monolingual dictionary of contemporary Czech. The new lexicographic software is being created as a specialised software platform for a modern description of contemporary Czech vocabulary in the form of a monolingual explanatory dictionary, namely both electronic (accessible via the internet, mobile communication devices) and printed. The software development has received grant support from the Ministry of Culture of the CR within the National and Cultural Identity (NAKI) applied research and development programme.

1 Introduction

At the Department of Contemporary Lexicology and Lexicography of the Institute of the Czech Language of the Academy of Sciences of the CR, v. v. i., a new monolingual dictionary of contemporary Czech has been developed since 2012 (with the working title *Akademický slovník současné češtiny* (*The Academic Dictionary of Contemporary Czech*), hereinafter as the *ASSČ*). It is a medium-sized dictionary with the expected number of 120,000–150,000 lexical units. Its aim is to capture widespread contemporary Czech vocabulary[1] used in public official and semi-official communication as well as in everyday (i.e. non-public, unofficial) communication. To a limited extent, the dictionary presents units utilised in professional and interest-group communication if their use has been extended beyond their milieu. Dialectal expressions have been included if they are common in a wider area and are used especially in oral communication or in literature.

With its size and method of treatment, the dictionary being prepared is an academic dictionary, i.e. a dictionary with an elaborated, standardised and structured explanation of the meaning of lexical units, with an adequately rich exemplification documenting the typical use of lexical units, with a sufficiently elaborated description of the basic semantic relations, mainly synonymy and antonymy, with the corresponding description of the

[1] In compliance with the definition of contemporary Czech given in [6, p. 89], we define it using the milestone of the end of the Second World War, i.e. the year 1945.

grammatical properties of lexical units and with usage labels (a description of the stylistic, temporal, spatial, frequency and pragmatic markedness) of lexical units. Unlike in previous monolingual dictionaries, grammar information in the *ASSČ* is more detailed, especially concerning the list of morphological forms and valency. In the *ASSČ*, the treatment of multi-word lexical units has been further elaborated. With most multi-word lexical units, their meaning and exemplification have been included; they are thus treated almost like one-word lexical units. As far as the dictionary macrostructure is concerned, run-on entries occur to a very limited extent, which means that certain types of the lexical units that have traditionally been included in run-on entries (relational adjectives, deadjectival adverbs and some abstract terms) have been treated in more detail. (See below for more information on the individual aspects of the dictionary macrostructure and microstructure.)

The above-mentioned information implies that the *ASSČ* in many respects closely builds on the long-term tradition of the creation of monolingual dictionaries of Czech at the department of lexicography of the Institute of the Czech Language. Nevertheless, it is also the first synchronic monolingual dictionary of this type and size during whose creation it is possible, thanks to the development of computational and corpus linguistics, corpus lexicography and computer technology, to take advantage of new technological possibilities for lexical research, both when working with lexical materials (language corpora, excerpt databases, electronic archives, texts on the internet and special software tools, esp. the Word Sketch Engine [7]) and during the actual creation of a dictionary and its publication. Unlike earlier monolingual dictionaries created in the Institute of the Czech Language in the past,[2] this dictionary has been – in accordance with the current lexicographic trend – compiled since the very beginning by means of specialised lexicographic software for dictionary creation.

2 The Dictionary Writing System (DWS) for the *ASSČ*

For the creation and publication of monolingual and bilingual dictionaries, several foreign commercial systems (e.g. TshwaneLex [32], IDM DPS [29], iLEX [30]) and open-source systems (e.g. the Mātāpuna Dictionary Writing System [31]) are available; in the Czech milieu, a DEB II [27], dictionary editor and browser has been developed. After various possibilities (mainly the issue of the purchase of commercial DWS licences and the need

[2] The following dictionaries were published in a printed form and later converted to an electronic form: *Příruční slovník jazyka českého* [12], retrieved from `http://psjc.ujc.cas.cz` [13]; *Slovník spisovného jazyka českého* [17], retrieved from `http://ssjc.ujc.cas.cz` [18]; *Slovník spisovné češtiny pro školu a veřejnost* [15] – electronically in the LEDA, s. r. o., publishing house, [16]; *Akademický slovník cizích slov* [2] – electronically in the LEDA, s. r. o., publishing house as *Velký slovník cizích slov* [26]; a new, revised edition *Nový akademický slovník cizích slov* [10] – electronically in the same publishing house under the title *Velký slovník cizích slov* [26]. These dictionaries were created by means of computer technology, in the Word text editor (but not using specialised software): e.g. the dictionaries of neologisms *Nová slova v češtině*. *Slovník neologizmů 1* [8] and *Slovník neologizmů 2* [9], the dictionary *Slovesa pro praxi* [24] and *Slovník slovesných, substantivních a adjektivních vazeb a spojení* [25]. *Příruční slovník jazyka českého*, *Slovník spisovného jazyka českého*, *Slovník spisovné češtiny pro školu a veřejnost* and *Akademický slovník cizích slov* are also part of the DEBDict application, a general dictionary browser [28].

to modify the program within the possibilities offered by the given software, be it commercial or open-source) were considered, it was decided that, because of the significant specifics of the compilation of the *ASSČ*, including requirements on the flexibility of the program and its modifications in connection with the expected component changes in the conception of the dictionary, new software would be developed for this purpose.[3] The software will therefore have both the general features of a DWS ("a text-editing interface, in which lexicographers create and edit dictionary entries; a database, in which the emerging dictionary text is stored; a set of administrative tools which support the management of the project and the publication process") as mentioned by [3], cf. [1] and specific features and functions arising from the character of the language described and from the actual concept of the *ASSČ*. A prerequisite for the software development is hence a close cooperation between the lexicographic part of the team and the programmers.[4]

The project of a new monolingual dictionary of contemporary Czech and its software provision has received funding for 2013–2016 from the Ministry of Culture of the CR within the National and Cultural Identity (NAKI) applied research and development programme – the grant project *A New Path to a Modern Monolingual Dictionary of Contemporary Czech*.

The main objectives of the project are to develop software:

1. for the modern description of contemporary Czech vocabulary in the form of a monolingual dictionary, namely both electronic (accessible via the internet, mobile communication devices, mainly tablets, mobile phones) and printed;
2. to facilitate the teamwork of the authors and editors of dictionary entries, to ensure and keep records of their communication;

[3] Cf. e.g. a similar approach by the Slovak lexicographic team preparing *Slovník súčasného slovenského jazyka* [19] (2 volumes issued so far, 2006, 2011), using their own lexicographic software for the creation of dictionary entries.

[4] In 2005–2011, the lexicographic team gained valuable experience in the development of the Praled lexicographic software within the research plan *The Creation of a Lexical Database of the Czech Language of the Beginning of the 21st Century* (AV0Z90610521) in cooperation with the Natural Language Processing Centre at the Faculty of Informatics, Masaryk University, Brno (see [5], [11]). The software was used to build the Pralex lexical database, an important source of a new monolingual Czech dictionary (on the development of Praled/Pralex, see [20], [21], [22], [23] etc.). In 2005–2006, the first outline of the conception of the future monolingual dictionary called Lexikon 21 was created (on that, see in particular [20], [14]). Although this outline was reflected in the structure of the editing items in Praled, the dictionary conception ceased to be prepared and the emphasis in the filling of the database called Pralex (a database of words, word forms and phrases) based on more concise principles of the treatment of database items was further placed especially on the description of the lexical, semantic and syntactic collocability of entry words (mainly nouns) in the subdivision into individual meanings. In terms of the DWS as described by [1] and [3], the Praled software is therefore incomplete and could not be used for the creation and publication of a monolingual dictionary without extensive program modifications, changes and complementation (what is lacking are mainly editorial and administrative tools and a tool for a print output corresponding to a traditional dictionary, because a printed form of the Pralex database was not considered).

3. for the continuous (and updated if necessary) presentation of the results of the lexicological-lexicographic research of contemporary Czech to the professional as well as wide public;
4. for the creation of special monolingual dictionaries (of neologisms, loanwords, homonyms etc.) and further for the creation of monolingual explanatory dictionaries of typologically related, especially Slavic, languages.

3 The Fundamental Conception Principles of the *ASSČ* in Terms of DWS Requirements

As indicated above, the requirements on the software being prepared arise from the concept of the *ASSČ* (i.e. from the selected dictionary macrostructure and microstructure and from the principles for the preparation of particular types of lexical units in terms of the lexical structure, word classes, semantics) and from the work needs of the lexicographic team. If there are changes made in the dictionary concept, the lexicographic software must make it possible to implement them.

3.1 The Macrostructure (List of the Entries) of the *ASSČ*

The macrostructure (list of the entries) of the *ASSČ* is formed by one-word lemmas, some multi-word expressions (multi-word loanwords and quotational loans), subword units (prefixes, prefixoids, radixoids), abbreviations and reference entries.

As far as multi-word lexical units (multi-word naming units – e.g. terms; phrasemes; multi-word grammatical function words) are concerned, they are included in the dictionary – like in earlier monolingual dictionaries – in the entry of a one-word lemma. They thus do not form separate items in the dictionary list of the entries. On the other hand, it is taken into account in their preparation that they are independent formal-semantic lexical units; therefore, most multi-word lexical units are provided with the explanation of the meaning and exemplification like one-word lemmas. These complexly presented multi-word lexemes are, unlike in the dictionary output, treated in the lexical database as independent items. This method of database treatment makes it possible not only to implement the adopted conceptual principle (giving the explanation of the meaning and exemplification as well as showing e.g. the variant forms in the lemma, usage labels etc.) but also to achieve uniformity in the treatment of the same multi-word lexical unit included in different entries (according to the adopted conceptual principle, multi-word lexemes are listed under all full-word components), to generate component lists of the entries, e.g. the treated phrasemes, and to react easily to changes in the dictionary concept. For the dictionary output, it is hence necessary to cross-reference a multi-word database with all full-word, or also other, components.

Likewise the need to add some derivatives to the lemmas as run-on entries is connected with the dictionary macrostructure. In the *ASSČ*, this lexicographic method has been selected only for very limited cases: feminine derivatives from masculine nouns, diminutives, frequentative verbs unless there is a reason to create separate entries for them.[5] In

[5] It is possible to create run-on entries for other lexical types if we decide for such a principle.

the database list of the entries, the lexical units that will be presented as run-on entries in the *ASSČ* are treated as separate items. For the dictionary output, these items need to be cross-referenced with their lemmas.

It is clear from what was written above that in the cases of multi-word lexical units and the lexical units that will be added to lemmas as run-on entries, the list of the entries of the dictionary output is not identical with the database list of the entries.

The following lexicographic requirements on software arise from the information on the *ASSČ* macrostructure (and on the macrostructure of the lexical database) provided above:

- to label in the editing form individual structural types of the lexical units (one-word, multi-word etc.); on the basis of this labeling, to be able to distinguish typographically (with a special symbol) the individual types (e.g. phrasemes, multi-word naming units) as well as to generate component lists of the entries (e.g. of the phrasemes that have been treated);
- to create an environment for the treatment of multi-word lexical units and their cross-references to a certain place in a one-word lemma entry (to the exemplification; to the end of the part of an entry covering one sense; to the end of the entry);
- to create an environment for the creation of run-on entries for selected types of derived lexical units.

In terms of software solution, the theme of the dictionary macrostructure is described in the section List of Entries Module.

3.2 The Microstructure of an Entry in the *ASSČ*

The microstructure of an entry in the *ASSČ* consists of the following parts: the lemma (including variant forms), information on homonyms, pronunciation, the etymology of the lexical unit, grammatical information (word class, morphology, valency), the usage label, the explanation of the meaning (including synonyms and antonyms, or opposites), exemplification, notes (e.g. encyclopedic information, further etymological and orthographic information) and cross-reference to (semantically, grammatically) related entries. With polysemic entries, a distinction is made between information related to the entire entry and that related to a specific sense. Some microstructure elements are provided with more detailed notes, which have arisen from consultations with the programmers on a software requirements specification.

Different lemma variant forms often contain different pronunciation, morphological, stylistic and frequency data (e.g. "**aforismus** [-izm-], **aforizmus**; **aikido** -da, 6. j. -du, **aikidó** neskl. s.; **bačkůrka**, řidč. **bačkorka**"). Since the order of the variant forms may not be clear at the beginning, it is necessary to be able to change it easily using the software tool including all the data connected with the given variant form.

A subentry is used to show which form is the only/more frequent for a specific sense of a polysemic entry; e.g. with the interjection entry **baf**, **baf baf**, the subentry for the meaning expressing the intention to scare someone says: "zprav. jen **baf**".

In the *ASSČ*, grammatical information is treated more comprehensively than in previous monolingual dictionaries. The list of morphological forms of flexible word classes is

more detailed. Morphological information consists of two to three parts: the type of morphological information (the description of the form, e.g. “1. mn.”, “2. st.”, “rozk.” etc.), the ending (resp. the entire form) including doublets and in relevant cases also a commentary on the particular form (e.g. in terms of frequency). Other grammatical information is specified when needed (e.g. “zprav. mn.” is utilised for nouns usually used in plural, “neos.” for verbs used only in the 3rd pers. sg.). In addition, as far as grammatical information is concerned, more space is given also to valency.

In the *ASSČ*, the explanation of the meaning of individual lexical units is subdivided in order to meet the requirements on the subdivision of the information provided also in terms of the typography in the resulting dictionary entry: the editing part of the form for the explanation of the meaning is subdivided into preliminary explanation (semantic note), the actual explanation of the meaning, synonyms, antonyms. Some types of entries, e.g. subword units, lexical units that will be added to lemmas as run-on entries, etc. have a special structure for the explanation of the meaning (including special fields).

For most of the information given, the editing form also offers the option to write a commentary in a special field, e.g. concerning the frequency or stylistic markedness.

The editing form of every entry must contain an internal editing (lexicographic) note which is available for the treatment of an entry but is not used for the dictionary output.

A simplified scheme of a dictionary one-word lemma entry[6]

lemma (=variant form 1) homonym number [pronunciation] morphological information word class <etymology> usage label, **lemma (=variant form 2)** homonym number [pronunciation] morphological information word class <etymology> usage label

sense number* subentry usage label (valency) (preliminary explanation) explanation of the meaning; syn. synonyms; op. opposites: *exemplification* □ **multi-word naming unit** usage label explanation of the meaning; syn. synonyms; op. opposites

Note

□ **multi-word naming unit**** usage label explanation of the meaning; syn. synonyms; op. opposites

multi-word grammatical expression grammatical information usage label explanation of the meaning: *exemplification*

◊ **phraseme** usage label explanation of the meaning; syn. synonyms; op. opposites: *exemplification*

► **run-on entry** morphological information word class usage label word-formation reference: *exemplification*

► cross-reference***

Notes on the scheme:

* In addition, each numbered sense may be complemented by morphological and word-class information if it is different from the data given for the entire entry.

** A multi-word naming unit that cannot be added to the sense.

*** A cross-reference to the verbal aspect counterpart, a diminutive treated separately etc.

In terms of the software solution, the theme of the dictionary microstructure has been described in the section Editing Module and Output Module.

[6] Certain parts of this scheme will be applicable for particular types of lexemes. Since there may be changes in the concept of the *ASSČ*, this may alter not only the order and graphic layout of individual data, but also the amount of information provided for the given type of lexical units.

For the sake of clarity, further lexicographic requirements on the DWS, both in terms of the method and administration of lexicographic work, have been described in the following text.

4 DWS

When designing the DWS, we had to choose between the web application and the standard installed standalone application. Our definite decision was to create it as a web based application; the next section explains the reason.

4.1 DWS as a Web Based Application

The web based application is software that uses a website as an interface through client software called web browser. In other words, the web application can be considered as a web page. Thus, to interact with a web based application, any web browser like Internet Explorer, Firefox, Opera, Google Chrome, etc. is sufficient. Examples of well known web based applications are Gmail, YouTube, Facebook, etc. Advantages of web based applications over standard installed ones are as follows:

1. the web application does not require any installation;
2. users of the web application do not have to worry about upgrading;
3. the web application can be accessed using any computer with an internet connectivity;
4. the web application is independent on the operating system;
5. the web application can be easily adjusted to be used on mobile platforms;
6. data generated by the web application can be easily shared;
7. users do not have to worry about backups of data generated by the web application.

In fact, the only requirement to the user is to have an URL address with appropriate login information. Since the web based application is provided as a service running on the maintained server, updates and changes are made by provider and are always reflected for all users at once. Thus, the users are always working with the latest version of the application – and they all have the same version.

The dictionary data and settings of the application are stored in the central repository, in the database on the server that is running 24 hours a day. Thus, up-to-date data are available for all users any time. It is easy to import data from similar databases and also from other ones using simple translating scripts thanks to the very precise data structure.

Since web technologies are available on different devices, it is possible to output the resulting dictionary to various devices like mobile phones or tablets, to prepare printable documents or to feed the data to a public web page.

4.2 Introduction to DWS Modules and their Cooperation

We introduce the modules of our DWS and briefly describe their functions and how they fulfil our demands in this section.

Our DWS consists of 4 main modules that are logically linked together and cooperate with each other. These modules were developed simultaneously, and then they were tested and tuned to perfectly fit our needs.

Technical details of the implementation are described in the article *The Editing Module – the Development of a Lexicographic Tool* [4].

4.3 List of Entries Module

The "List of entries module" reflects the macrostructure of the dictionary from the perspective of lexicographers that is described in the Macrostructure (List of the Entries) of the *ASSČ* section.

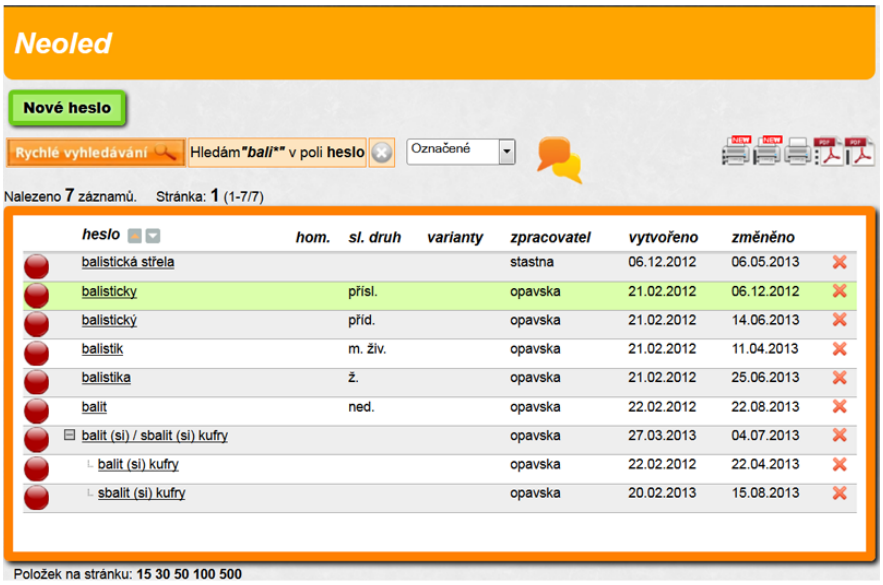


Fig. 1. Web page showing the list of entries of our DWS

The module "List of entries" (see Fig. 1) allows lexicographers to list the dictionary entries together with important information about each entry. Besides the lemma, there is information about a status of the entry (the dot on the left side) that indicates whether the entry is new (red), is semi-finished (yellow) or is finished (green). This helps lexicographers easily recognize, which entry needs an attention, to filter them out and to process them. Further, the information about the homonym is present. The next column to the right shows information about the word class of an entry. Another column contains variants of the word. Further, there is information about the processor of the particular entry, and the dates, when the entry was created and when it was last time modified. The tree structure of entries that is visible in the list is used for organizing variants of phrasemes.

Furthermore, several functions are available in this module. Besides self-evident functions like the "Create new entry" or the "Delete entry", the most important function is the "Quick search function" (see Fig. 2) that allows searching for entries that contain the search query in any specified field of the entries' microstructure. (Important microstructure information is for example the structural type of the lemma.) Another very important function is the basic filter. A few most needed filters are predefined and offered to users via the selector. These filters allow users to quickly filter out and work with the resulting subset of entries. Examples of such predefined filters are "my entries", "marked entries", "entries edited in range of dates…" etc. The Quick search function together with the filtering allows us to create subsets of entries, which helps the lexicographers to focus on the particular subset of interest.

Another important feature regarding the macrostructure is an ability to create references between particular entries. This feature is available in the Editing module and is described in the corresponding section *Cross-references* (4.4).

4.4 Editing Module

The most important module is the Editing module; this module is used mainly to input and edit the dictionary microstructure data. The cross-references between entries, which are part of the macrostructure of the dictionary, as mentioned above, are controlled by this module, too. This module is divided into the following sections:

1. Header section
2. Section of variants
3. Sense section
4. Cross-references

The description of these sections follows.

Header Section. Selected parts of the common information are available to view and edit in this section. Most of this information is neither part of the microstructure nor of the macrostructure of the dictionary. These pieces of information are rather formal data about the state of the entry, who created the entry, when it was created, when the last editing was done etc. We can set here the output in which the entry will be visible, if any. (The control

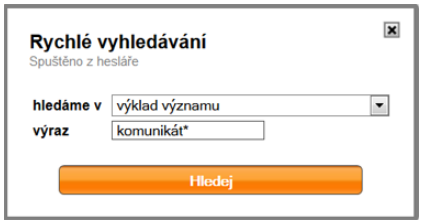


Fig. 2. The "Quick search function" in the List of entries module

over the visibility in the output is also available for some microstructure information, viz. the exemplification.) The most important field in the header section is the lexicographer's note, which enables the lexicographers to enter a note about the entry (what to adjust in the future, a message for another lexicographer etc.).

Types of input fields will be described in the following text.

Section of Variants. According to the requirements on the microstructure elements from the section 3.2, this editing section must allow adding and sorting more than one variant of lemma with all required microstructure elements of variant. The variants of variant lemmas are often equivalent in majority of values, thus we added the function "Add variant as a copy of the last one", which takes the last variant, copies their values into the new one and adds it to the lemma (see Fig. 3). Thus, only very few values have to be edited in the new variant. Consequently, working with new variants is much more efficient. It is also possible to duplicate the whole entry and then to save it as a new lemma. It saves a lot of time and lexicographer's effort when similar entries are created.

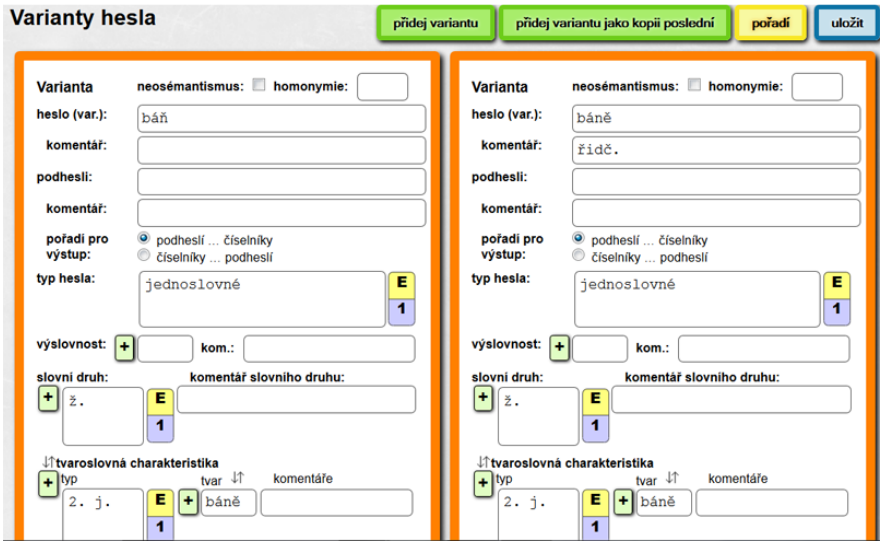


Fig. 3. Working with variants of the lemma

A lot of information may be needed to define for each variant (some pieces of this information are listed in the sec. 3.2, *The Microstructure of an Entry in the ASSČ*, describing the dictionary microstructure). Input fields for these data are of different types according to the character of data collected by this field. We use 6 different types of input fields:

1. *Single line text input* is used for data in the form of a short free text (example: a pronunciation field).

2. *Multi line text input* is used for data in the form of a long free text formatted using new lines (example: lexicographers' notes).
3. *Multi line WYSIWYG text input* is required when it is necessary to allow user to format the input text (example: the exemplification field).
4. *Select boxes* are used to choose among predefined values. The values are defined and managed in the administrative module described later. We use predefined values wherever it is possible. As several people are working with the system, it helps to avoid mistakes during updating the data and to keep values with the same meaning in the same format. In some cases, user may need to input a value, which is not predefined in the select box and therefore he or she cannot choose it. Thus, the "other" free text field is available, where user can always input the desired value.
5. *Radio buttons* are used when we want to restrict the selection among predefined values to one value (example: the output control).
6. *Check boxes* are used for values with two states, "on" and "off".

As mentioned in the dictionary microstructure description (see section 3.2, *The Microstructure of an Entry in the ASSČ*), it is necessary to input two or more pieces of information of the same type. This is provided by special form elements where an infinite number of particular data fields of the same type may be added. Of course, these multi inputs provide an option of sorting by lexicographer. An example of such information is the morphological information (see Fig. 4).

Sense Section. The sense section consists of one or more panels, where the meaning of the word is described together with other related information. The section is organized as a set of panels where every panel contains a huge form, where the information about the sense can be edited. The types of the fields in form are the same as described in the section above. User can change the order of panels what will affect the order in the dictionary printed output. The panels are numbered. They can be minimized or maximized for a clearer arrangement, according to which panel the user intends to work with. When a word has a more senses, the system allows navigating directly to the sense with the certain number using the quick navigation.

Cross-references. Another useful feature of the system is the ability to define relations between entries (see Fig. 5). Relations[7] are defined between two dictionary entries; one of the entries is considered to be the main or "master" entry, the second is the "slave"; the slave entry is connected to the main one. There are different types of relations – run-on entries, references between one-word and multi-word lexical units etc. The relation is defined in the main entry.

Additionally, the system allows creating a connection from the slave entry side (see Fig. 6). This means that user editing the slave entry may connect it to selected main entry.

Other Functions. There are several background automated processes implemented; their main purpose is help lexicographers to keep the dictionary compilation consistent.

[7] Relations in database terminology

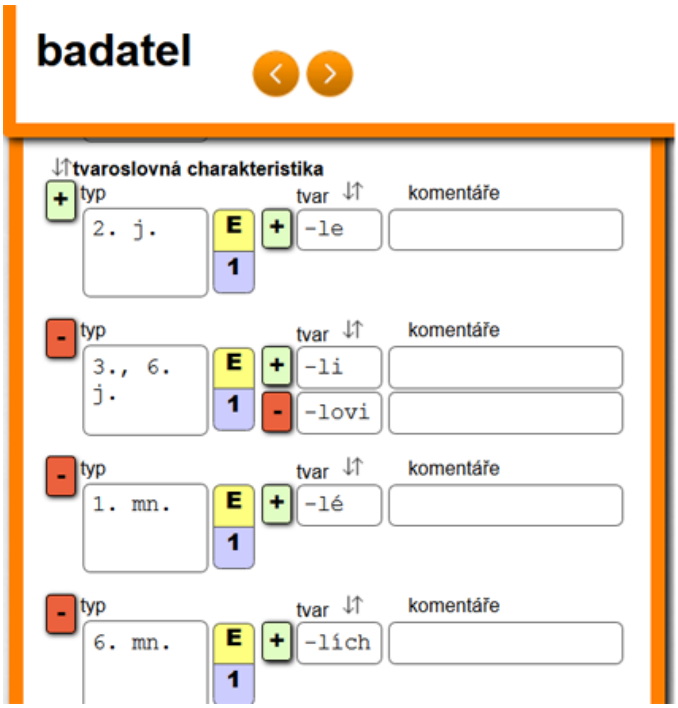


Fig. 4. Input fields for the morphological information

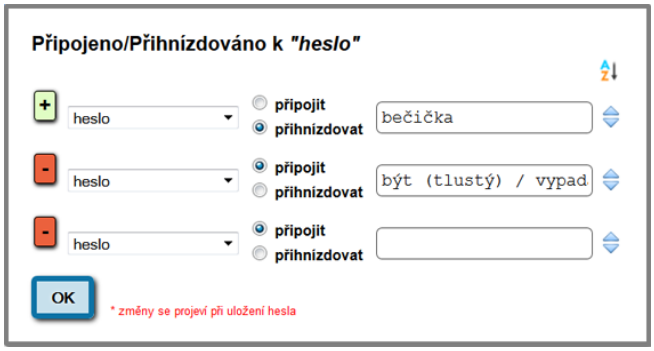


Fig. 5. Defining the relation between entries

Our DWS naturally allows multiple users to work simultaneously. However, this introduces some difficulties in the form of possible concurrent updates on the particular entry. To avoid this problem, a blocking mechanism was implemented that allows editing an entry only by one user at the time.

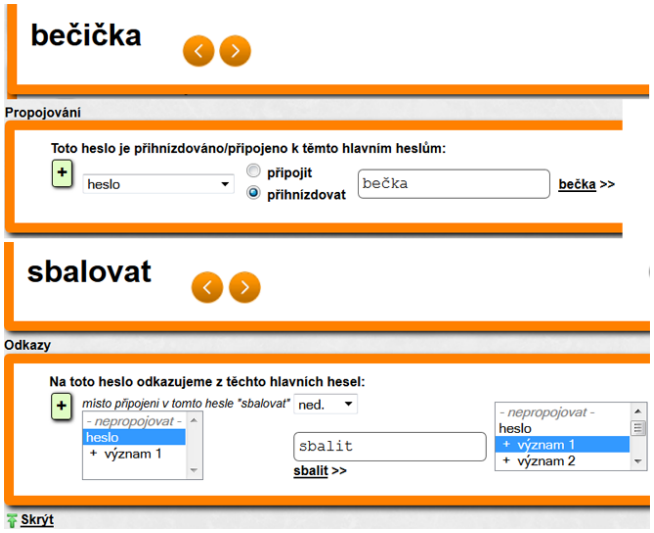


Fig. 6. Defining relation on the slave entry

Another useful function is the cross-reference automatic update when changing the order of referenced microstructure elements. For example, when some reference points to the second sense and we change the order, so that the second sense becomes first, the reference is automatically switched to the first sense.

4.5 Output Module

The Output module is used to control the printed output of the dictionary entries. The data of each entry is formatted according to specific rules. The formatting rules are independent on the data editing module, thus the editing tool described above may be optimized for the data input, while the printed output can be modified in any way to achieve exactly the desired print format. In the current stage of development, our DWS is able to export dictionary entries in the HTML format viewable in the web browser or in the PDF format. It is possible to export one particular entry or a subset of entries at once controlled by the filter or by the quick search function in the List of entries module.

4.6 Administrative Module

The Administrative module is accessible only to administrators. This module consists of two sections:

1. User administration
2. Select box administration

User Administration. The user management is provided by the user administration section, where the administrator may create a new user, grant him or her privileges, activate or deactivate his or her account, change the particular user's password. He or she may also remove users.

Select Box Administration. The Select box administration provides the management and control over predefined values in select boxes used in the Editing module (see Fig. 7).

Standard actions over the predefined select box values are editing, deleting and adding new value. Furthermore, the overview of "other" values that were inputted by lexicographers is available together with the number of usages in the Editing module. When the administrator decides that some value is used frequently, he or she can easily move this value among predefined values to allow to be selected from the select box when it is later needed.

Fig. 7. Select box administration

5 Conclusion: Proposed System Modules

The DWS is in an early phase of the development. We have already finished the core of the system: the List of entries module, the Editing module, the Output module and the first two parts of the Administration module – the User administration and the Select box administration.

Though it seems as a finished work, we are in about $^1/_3$ of our effort. The requirements on the complex DWS asked necessary for the serious lexicographic work are much larger. We already developed a comfortable tool for the everyday work with the lexicography material; on the other side, there are several planned modules to be developed in upcoming months.

5.1 Editorial Tool

The Editorial tool is a module integrated with the Output module with the dictionary output provided in the form very close to the final dictionary layout and format.[8] The editing/proofreading process should be done in this module. This tool offers a set of functions used by lexicographers:

- identification of the place of the correction and mark it exactly in the output text;
- writing the proposal of the correction, add a comment to it;
- sending the correction proposal to the record (lemma) owner.

The workflow of the Editorial tool is shown on the following diagram (see Fig. 8).

Lexicographers will get an integrated editorial tool with added features such as automatic archive, the possibility of the communication, the full control on the editing and proofreading process. All of these features will be accessible via web interface. At last but not least, we save the space in the office, papers, trees, and the environment.

5.2 Data Tracking Module

There are defined requirements for the tracking of the user behaviour and various other actions. We would like to track:

- lexicographer's work with the database entries (lemmas),
- use of the editorial tool,
- use of the dictionary output generator (web view and PDF),
- various user activities: new entries for the select boxes, user activities on entries, senses, variants,
- use of the cross-references,
- use of the revision control system, i. e. tracing the data versions, searching the differences between two revisions, merging the versions, rollback of changes.

The Data tracking module has to record the tracing of each important system activity regardless the activity nature (user initiated or the automatically triggered behaviour).

The part of the Data tracking is already implemented and it is extended every time a new module is introduced to the system. There is currently no reporting tool to consolidate the tracked data. The reporting tool is planned to be developed in the future, when there will be recorded a significant set of data of the real system use.

[8] A nice side effect of the integrated Editorial tool is the saving of the paper, thus trees and the environment. We calculated our estimation of the saved (not printed) A4 papers used for editing/proofreading based on the following facts:

- 150,000 lemmas in the final output,
- 14 lemmas per A4 page,
- the first revision: cca 11,000 pages (100% of lemmas to be printed for editing),
- the second revision: 5,000 pages (around 50% of lemmas to be printed),
- the third revision: 2,750 pages (around 25% of lemmas to be printed),
- the final proofreading revision: 11,000 (100% lemmas to be printed),
- total: 30,250 A4 pages.

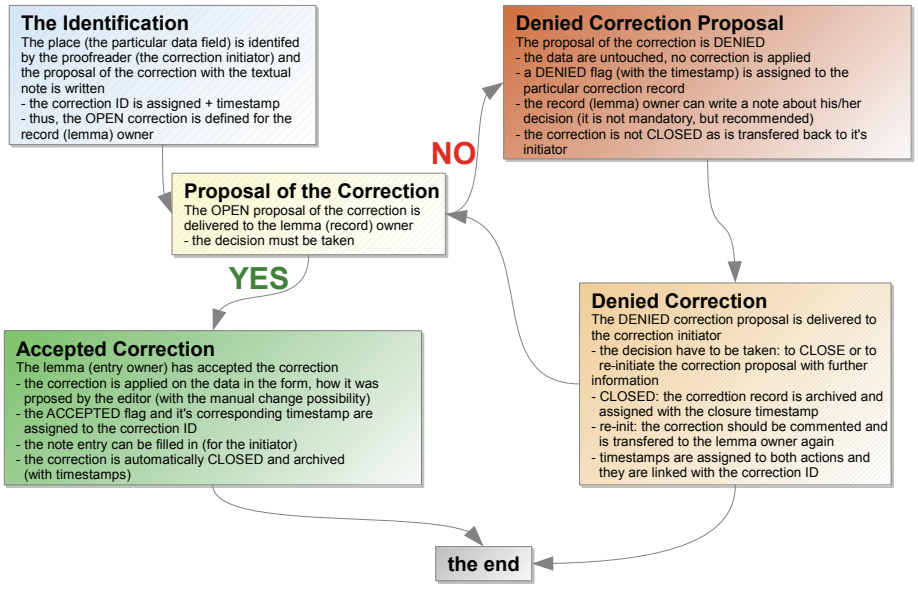


Fig. 8. Diagram of the workflow of lemma corrections

5.3 xFilter – the Complex Search Tool

The advanced search capability implemented in the module called the xFilter is the tool for those lexicographers, who like to search in the data deeply in any imaginable form. The search tool offers to do:

- the fulltext search in all text based fields,
- the exact search,
- the interval search (e.g. from-to dates),
- the subset search (data are in the subset with conditions),
- use the AND, OR, NOT logical functions in the search criteria,
- to combine any of the search criteria to build complex search queries on the whole database.

5.4 Revision Control System

The Revision control system is required mainly for to keep track of the entry editing history. Additionally, it will provide the following set of features:

- create a revision of the stored data (at the level of entry and its complex tree hierarchy of descendant data fields),
- compare selected revisions and show differences,
- rollback functionality,
- merge functionality.

5.5 Dictionary Web Interface for the Public

There is a plan to provide the whole dictionary to the public, once finished. It is proposed to open the public access on the separated database (only finalized data) in the form of:

- the web based internet application,
- the native application for mobile platforms (currently we suppose to support iOS and Android).

5.6 Automatic Lemma Processing

The routine processes are supposed to be transferred from humans to the machine. This is the case of the re-numbering and sorting in the data used in the explanation of the meaning, in the field of synonyms, antonyms, in notes and in linked entries.

Acknowledgement. The article has been written within the grant project of the National and Cultural Identity (NAKI) applied research and development programme A New Path to a Modern Monolingual Dictionary of Contemporary Czech (grant no. DF13P01OVV011).

References

[1] Abel, A. and Klosa, A. (2012). *The lexicographic working environment in theory and practice*. Fjeld, R. V. and Torjusen, J. M., editors, *Proceedings of the 15th EURALEX International Congress*, pages 1–23, University of Oslo, Oslo.
[2] *Akademický slovník cizích slov*. (1995). Academia, Prague.
[3] Atkins, B. T. S. and Rundell, M. (2008). *Oxford Guide to Practical Lexicography*. Oxford University Press, Oxford.
[4] Barbierik, K. et al. (2013). *The Editing Module — the Development of a Lexicographic Tool*. Submitted to the Federated Conference on Software Development and Object Technologies.
[5] Horák, A. and Rambousek, A. (2013). PRALED – A New Kind of Lexicographic Workstation. In *Computational Linguistics Studies in Computational Intelligence*, 458:131–141.
[6] Karlík, P., Nekula, M., and Pleskalová, J., editors, *Encyklopedický slovník češtiny*. (2002). Nakladatelství Lidové noviny, Prague.
[7] Kilgarriff, A., Rychlý, P., Smrž, P., and Tugwell, D. (2004). The Sketch Engine. In Williams, G. and Vessier, S., editors, *Proceedings of the 11th EURALEX International Congress*, pages 105–116, Université de Bretagne-Sud, Lorient.
[8] Martincová, O. et al. (1998). *Nová slova v češtině. Slovník neologizmů 1.* Academia, Prague.
[9] Martincová, O. et al. (2004). *Nová slova v češtině. Slovník neologizmů 2.* Academia, Prague.
[10] *Nový akademický slovník cizích slov*. (2005). Academia, Prague.
[11] Pala, K., Horák, A., Rambousek, A., and Rangelova, A. (2007). Nové nástroje pro českou lexikografii – DEB2. In Štícha, F. and Šimandl, J., editors, *Gramatika a korpus 2005*, pages 190–196, Ústav pro jazyk český AV ČR, v. v. i, Prague.
[12] *Příruční slovník jazyka českého*. (1935–1957). Státní pedagogické nakladatelství, Prague.
[13] *Příruční slovník jazyka českého*. URL: `http://psjc.ujc.cas.cz`, retrieved 13 May 2013.
[14] Rangelova, A. and Králík, J. (2007). Wider Framework of the Research Plan Creation of a Lexical Database of the Czech Language of the Beginning of the 21st Century. In Levická, J. and Garabík, R., editors, *Computer Treatment of Slavic and East European Languages*. Proceedings, pages 209–217, Tribun, Bratislava.

[15] *Slovník spisovné češtiny pro školu a veřejnost*. (1978, 2nd ed. 1994, 3rd ed. 2003. Academia, Prague.
[16] *Slovník spisovné češtiny pro školu a veřejnost*. (1997), (2004), (2005). [CD-ROM] LEDA, s. r. o., Voznice.
[17] *Slovník spisovného jazyka českého*. (1960–1971). Nakladatelství ČSAV, Prague.
[18] *Slovník spisovného jazyka českého*, URL: `http://ssjc.ujc.cas.cz`, retrieved 13 May 2013
[19] *Slovník súčasného slovenského jazyka.* A–G (2006), H–L (2011). Veda, Bratislava.
[20] Světlá J. (2007). The Possibilities and Limits of Lexicographical Description of the Czech Lexicon in Database Form. In Levická, J. and Garabík, R., editors, *Computer Treatment of Slavic and East European Languages*. Proceedings, pages 244–253, Tribun, Bratislava.
[21] Světlá, J. (2008). K návrhu, vývoji a funkcím lexikální databáze češtiny. In Rangelova, A., Světlá, J., and Jarošová, A., editors, *Lexikografie v kontextu informační společnosti*, pages 19–32, Ústav pro jazyk český AV ČR, v. v. i., Prague.
[22] Světlá, J. (2011). Struktura a zpracování hesel v lexikální databázi Pralex. In Světlá, J., Jarošová, A., and Rangelova, A., editors, *Česká a slovenská výkladová lexikografia na začiatku 21. storočia*, pages 9–18, Tribun EU, Brno.
[23] Světlá, J. (2012). Lexikální databáze Pralex – nástroj a základna pro výzkum a popis slovní zásoby současné češtiny. In Čmejrková, S., Hoffmannová, J., and Klímová, J., editors, *Čeština v pohledu synchronním a diachronním. Stoleté kořeny Ústavu pro jazyk český*, pages 403–408, Karolinum, Prague.
[24] Svozilová, N., Prouzová, H., and Jirsová, A. (1997). *Slovesa pro praxi. Valenční slovník nejčastějších českých sloves.* Academia, Prague.
[25] Svozilová, N., Prouzová, H., and Jirsová, A. (2005). *Slovník slovesných, substantivních a adjektivních vazeb a spojení*. Academia, Prague.
[26] *Velký slovník cizích slov*. (1999), (2005). [CD-ROM]. LEDA, s. r. o., Voznice.
[27] DEB II. URL: `http://deb.fi.muni.cz/index-cs.php`, retrieved 13 May 2013.
[28] DEBDict. URL: `http://deb.fi.muni.cz/debdict/index-cs.php`, retrieved 13 May 2013.
[29] IDM DPS. URL: `http://www.idm.fr/products/dictionary_writing_system_dps/27/`, retrieved 13 May 2013.
[30] iLEX. URL: `http://www.emp.dk/ilexweb/index.jsp`, retrieved 13 May 2013.
[31] Matapuna. URL: `http://sourceforge.net/projects/matapuna/`, retrieved 13 May 2013.
[32] TshwaneLex. URL: `http://tshwanedje.com/tshwanelex/`, retrieved 13 May 2013.

Data Deduplication in Slovak Corpora

Vladimír Benko

Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences, Bratislava, Slovakia

Abstract. Our paper describes our experience in deduplication of a Slovak corpus. Two methods of deduplication – a plain fingerprint comparison and an n-gram comparison – are presented and their results compared.

1 Introduction

Deduplication is a special technique of detection and removal of duplicate contents in digitally stored data. Motivations for such activity include a more efficient use of data storage space (duplicate data can then be stored in a single copy only), detection of plagiarism (sections of identical text without proper quoting usually indicates an author's inappropriate activity), or decreasing the size of index structures in data retrieval systems.

In text corpora, the problem of duplicate contents started to be strongly felt with the advent of web corpora. Duplicate texts distort frequencies of occurrence of lexical units and bias the statistics used to compute collocations. Expressions of low frequency found repeatedly in duplicate documents or paragraphs tend to receive very high scores of salience. In the Word Sketch Engine, that is being used at our Institute in lexicographic projects [1], such collocations appear in first place in the respective tables, causing undesirable noise. The resulting concordance then looks as follows:

Corpus: **Iura (legal-1.0.10) 142 M (#591)**
Hits: **35** (0.2 per million)
Page 1 of 2 Go Next | Last

1.0000265	12 . augusta 1949 o ochrane ***obetí*** **medzinárodných** ozbrojených konfliktov (Protokol
1.0000938	12 . augusta 1949 o ochrane ***obetí*** **medzinárodných** ozbrojených konfliktov (Protokol
1.0002059	12 . augusta 1949 o ochrane ***obetí*** **medzinárodných** ozbrojených konfliktov (Protokol
1.0006681	12 . augusta 1949 o ochrane ***obetí*** **medzinárodných** ozbrojených konfliktov (Protokol
1.0014582	12 . augusta 1949 o ochrane ***obetí*** **medzinárodných** ozbrojených konfliktov (Protokol
1.0015063	12 . augusta 1949 o ochrane ***obetí*** **medzinárodných** ozbrojených konfliktov a konfliktov
1.0015063	12 . augusta 1949 o ochrane ***obetí*** **medzinárodných** ozbrojených konfliktov a konfliktov
1.0015063	12 . augusta 1949 o ochrane ***obetí*** **medzinárodných** ozbrojených konfliktov a konfliktov
1.0015064	12 . augusta 1949 o ochrane ***obetí*** **medzinárodných** ozbrojených konfliktov a konfliktov
1.0015064	12 . augusta 1949 o ochrane ***obetí*** **medzinárodných** ozbrojených konfliktov a konfliktov
1.0015064	12 . augusta 1949 o ochrane ***obetí*** **medzinárodných** ozbrojených konfliktov a konfliktov
1.0015065	12 . augusta 1949 o ochrane ***obetí*** **medzinárodných** ozbrojených konfliktov a konfliktov
1.0015065	12 . augusta 1949 o ochrane ***obetí*** **medzinárodných** ozbrojených konfliktov a konfliktov
1.0015065	12 . augusta 1949 o ochrane ***obetí*** **medzinárodných** ozbrojených konfliktov a konfliktov
1.0015066	12 . augusta 1949 o ochrane ***obetí*** **medzinárodných** ozbrojených konfliktov a konfliktov
1.0015066	12 . augusta 1949 o ochrane ***obetí*** **medzinárodných** ozbrojených konfliktov a konfliktov
1.0015066	12 . augusta 1949 o ochrane ***obetí*** **medzinárodných** ozbrojených konfliktov a konfliktov
1.0015067	12 . augusta 1949 o ochrane ***obetí*** **medzinárodných** ozbrojených konfliktov a konfliktov
1.0015067	12 . augusta 1949 o ochrane ***obetí*** **medzinárodných** ozbrojených konfliktov a konfliktov

Subsequent analysis shows that out of 35 occurrences of the collocation *obete medzinárodných ozbrojených konfliktov* "victims of international armed conflicts", only 2 are really unique and all the rest are just repetitions of the same sentence within the same law or in its various revisions.

2 Plain Fingerprint Deduplication

Detecting duplicates by direct pair-wise text comparisons in large collections is technically not feasible as the number of comparisons grows quadratically. There exists, however, a simple method that decreases the computational complexity of this task dramatically. It is based on the idea that, for each data segment under comparison, it is sufficient to compute a "fingerprint", i.e. a short fixed-length bit pattern [2] where equal data will have an equal value of their fingerprints.[1] The fingerprints can be computed, e.g., by means of a cryptographic hashing algorithm. Duplicate fingerprints can be easily detected by their sorting and subsequent unification [3]. This method leads to detection of exact (100%) duplicates.

In reality, this method can be easily implemented by means of standard Linux utilities *sort*, *uniq* and *md5sum*, complemented by two simple filters.

In regard to paragraph-level deduplication in a text corpus, the whole procedure can look as follows. Firstly, a paragraph identifier is assigned to each paragraph. Than a filter is used that will save the contents of each paragraph into a work file and call the *md5sum* program to compute the fingerprint that will be appended (along with the paragraph's identifier) to the fingerprint file.

```
1c32b3d252f2f66207352c95e02f04f5  0000000.00000
4d7d068d0e8c37aaf76619afdb41c937  0000000.00001
41a4459ed67cf1d15cca63b8e3efac6c  0000000.00002
ae5eae7f1645cbaa0d617a2089feea89  0000000.00003
...
af22e8df9bc9bfd3930d433b9fec39c7  1500125.00827
e98135d33b38729a90c3fb0465e25b62  1500125.00828
52ab380379a284145b89cca9a3581567  1500125.00829
d19c7528f54b7c4a968899c804675b0a  1500125.00830
```

After sorting the fingerprint file according to the first column, the duplicate fingerprints will appear together (we have marked them with an asterisk).

```
0000002797f70f8e9f666fb407db5195  1499872.00389  *
0000002797f70f8e9f666fb407db5195  1499876.00388  *
000000466f8914041e68767a38f392a0  0601609.01350
00000097f3f4b3521ceb78e26c000213  1465277.00301
0000019b4aeb3b3f8bf80bef210361ed  1304808.00003
000001f224498662798071a5580c7d80  0660013.00012
00000216af425ae2ac4112994546c9ef  0089946.00013  *
00000216af425ae2ac4112994546c9ef  0091158.00012  *
00000257b12f4211909124d2b7f18fc5  0979319.00019
...
```

[1] In using hash functions, there exist situations where two unequal segments have equal hash values (so-called *collision*). Although the probability of this happening is non-zero, it is so small that (in the context of language corpora) it can be safely ignored.

As a result of unification, each different fingerprint value in the file will appear just once.

```
0000002797f70f8e9f666fb407db5195  1499872.00389  *
000000466f8914041e68767a38f392a0  0601609.01350
00000097f3f4b3521ceb78e26c000213  1465277.00301
0000019b4aeb3b3f8bf80bef210361ed  1304808.00003
000001f224498662798071a5580c7d80  0660013.00012
00000216af425ae2ac4112994546c9ef  0089946.00013  *
00000257b12f4211909124d2b7f18fc5  0979319.00019
```

In the end, the file will be sorted according to second column, which will result in the list of paragraphs that are *not* to be removed. The final simple filter will remove the duplicate paragraphs in the original source file.

2.1 Deduplication in the Slovak National Corpus

Up to Version 5.0 of the Slovak National Corpus (SNC), the data duplicity problem has not been seen as very important, as most duplicities were avoided by careful selection of the source texts. The situation, however, has rapidly changed with the advent of Version 6.0 that received a large collection of newspaper texts form the Petit Press Publishing House. These contained a large amount of data from Slovak regional weeklies where many articles were identical. The respective documents represented print pages converted from PDF format, where, due to imperfection of the conversion procedure, the paragraph breaks of identical texts were not identical.

2.2 Paragraph-level Deduplication

In the following text we shall present the results of the plain fingerprint deduplication method applied to the largest SNC corpus – *prim-6.0-juls-all* [4]. The first filter mentioned in the previous section was modified so that it would not take into account punctuation, special graphic characters, and digits. This allowed us to also identify as duplicates paragraphs

```
<p>19.30 Noviny STV</p>
<p>23.45 Noviny STV</p>
<p>1.40 Noviny STV</p>

<p>12. Marseille 14 5 4 5 13:13 19</p>
<p>15. Marseille 15 4 5 6 13:15 17</p>
<p>10. Marseille 18 6 6 6 18:17 24</p>
```

representing items of the TV schedule and the football league table, respectively, the differences of which are not lexicographically interesting.

The input source file contained 1,390,408 documents with 51,536,717 paragraphs containing 1,226,218,915 tokens. The main procedure lasted approximately 19 hours,

while the computation of fingerprints took 18 hours and 20 minutes[2] (Intel Xeon 2.83 GHz, 8 GB RAM, hardware RAID, Ubuntu 12.10 LTS). The duplicate paragraphs were not deleted from the corpus but rather just marked so that they would not be taken into account in computing the word sketches. The advantage of such an approach is that the corpus user is not deprived of the context at the boundary of duplicate and unique content.

The result of deduplication is shown in the following table.

	Paragraphs removed	**Paragraphs left**	**Total**
Paragraphs	21,251,221	3,085,496	51,536,717
Paragraphs in %	41.24	58.76	100.0
Tokens	167,743,453	1,058,475,462	1,226,218,915
Tokens in %	13.68	86.32	100.0

Now, we would like to know what kind of data have been removed. It is obvious that only a tiny fraction of the millions of removed paragraphs can be inspected "manually". We have therefore decided to perform a frequency analysis of the removed paragraphs according to their lengths (in tokens). Respecting the expected distribution, paragraphs were grouped by power of 2, i.e. group "1" contained paragraphs of 1 token, group "2" paragraphs of 2 and 3 tokens, group "4" paragraphs of 4 to 7 tokens and so on. The results are summarized in the following table:

Paragraph length	**Paragraphs removed**	**Paragraphs left**	**Total**
1	2,899,765	313,757	3,213,522
2	5,385,687	1,760,629	7,146,316
4	6,430,346	5,065,587	11,495,933
8	4,369,821	6,858,103	11,227,924
16	1,459,344	6,202,353	7,661,697
32	512,667	5,349,893	5,862,560
64	166,836	3,425,382	3,592,218
128	24,435	1,083,748	1,108,183
256	2,218	200,193	202,411
512	93	22,537	22,630
1,024	8	2,811	2,819
2,048	1	443	444
4,096	0	46	46
8,192	0	13	13
16,384	0	1	1
Total	**21,251,221**	**30,285,496**	**51,536,717**

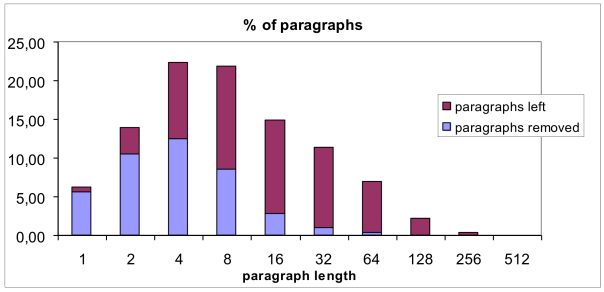
[2] It is obvious that a weak point of our implementation is the calculation of fingerprints by calling an external computationally "expensive" utility. Using a simpler hashing algorithm computed internally, it can be expected that the processing time could be significantly decreased.

The table shows that most of the paragraphs in groups 1 and 2 were removed, in groups 4 and 8 about 50% of the paragraphs were deleted, and from group 16 upwards most of the paragraphs were left.

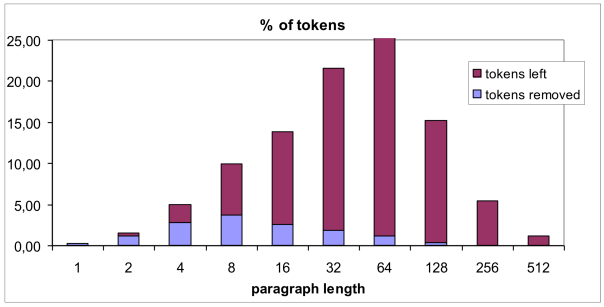
It is, however, more important to find out the token count in the deleted paragraphs.

Paragraph length	Tokens removed	Tokens left	Tokens total
1	2,899,765	313,757	3,213,522
2	13,714,814	4,557,662	18,272,476
4	33,490,542	27,787,256	61,277,798
8	45,839,898	75,856,713	121,696,611
16	30,790,283	139,059,109	169,849,392
32	22,213,759	241,725,320	263,939,079
64	14,072,053	300,236,783	314,308,836
128	3,949,428	182,717,735	186,667,163
256	700,310	66,312,816	67,013,126
512	58,514	14,489,082	14,547,596
1,024	10,412	3,835,447	3,845,859
2,048	3,675	1,169,544	1,173,219
4,096	0	250,981	250,981
8,192	0	146,055	146,055
16,384	0	17,202	17,202
Total	**167,743,453**	**1,058,475,462**	**1,226,218,915**

We can visualize the above data expressed in percentages in two graphs.



The columns in the first graph represent percentage shares of the respective paragraph groups with respect to the total number of corpus paragraphs. The light-coloured shading depicts the removed paragraphs and the dark shading indicates the paragraphs left. We can see that the first four groups contain the major portion of the removed paragraphs. The share of removed paragraphs declines sharply with the increasing length of the paragraphs. This is quite consistent with our intuition, as we can expect to have more matches in shorter paragraphs.



The second graph depicts the situation with tokens. The tendency is similar to the previous graph, and we can see that the largest contribution to the removed tokens comes from group "8". It is also quite interesting to find out that even group "64" contributes to the removed tokens considerably.

2.3 Sentence-level Deduplication

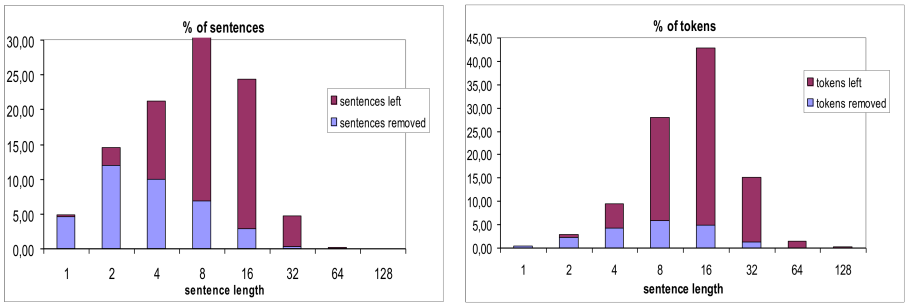
After having been deduplicated at the paragraph level, our corpus was processed by the Word Sketch Engine. Despite the removal of most duplicate concordance lines in the corpus, our lexicographers were not completely satisfied with the result. We therefore decided to repeat the whole procedure again, using the same technology at the sentence level.

The assignment of sentence identifiers revealed that there were 100,915,602 sentences in the corpus, which was roughly twice as many as the number of paragraphs. The deduplication was performed on a different computer (Intel Core i5 3.2 GHz, 12 GB RAM, software RAID, Ubuntu 12.10) and lasted approximately 55 hours[3]. The results were evaluated in a similar way as those of the paragraphs.

	Removed	**Left**	**Total**
Sentences	36,704,850	64,210,752	100,915,602
Sentences in %	36.37	63.63	100.00
Tokens	231,847,624	994,371,291	1,226,218,915
Tokens in %	18.91	81.09	100.00

If we compare the share of removed tokens by sentence-level deduplication with that of paragraph-level, we shall see that the number of removed tokens has increased 1.38 times and it represents almost 19% of all corpus tokens. The following graphs visualize the distribution of removed sentences and tokens by sentence length. The tendency shown in the graphs is similar to that of paragraph deduplication, with the difference being the greater contribution of shorter deleted segments (sentences).

[3] The computing time compared to the previous run was unexpectedly long. It is not clear what was the cause of this behaviour as all parameters of the computer used were higher (with the exception of the software RAID).



After this second deduplication phase the number of duplicate concordances observed by our users dropped to a minimum. We have decided to use this method also for other corpora of the SNC collection.

3 Detecting Near-duplicate Contents

As an alternate tool we decided to use the recently released open-source utility Onion[4] designed to detect near-duplicate contents in language corpora. This program was created within the framework of the PhD research of Jan Pomikálek at Masaryk University in Brno [5].

Onion ("One Instance Only") is also based on fingerprints but it does not compare whole segments but rather just n-grams of selectable length (7 by default). The input file is expected to be in one-column vertical format and it is processed in one pass. In the default mode, the deduplication is performed at the level of paragraphs marked by the ** ** ... ** ** tags. A paragraph is considered duplicate if it contains more than the threshold level of n-grams already encountered in previous text. The similarity threshold is a value in the range between 0 and 1, where 1 means a 100% match. The user can select deduplication at the level of segments marked by any pair of tags, with the most obvious values being documents and sentences. The duplicate segments can either be removed completely or indicated by a special mark in the first column of the output file. Implementation is optimized for speed (the fingerprints are computed by a computationally "cheap" routine *BUZ Hash* [6] and all data structures are stored in main memory) so the size of the corpus processed is limited by the size of available RAM only. Memory requirements can be substantially decreased by an alternate mode of program operation, where all computed fingerprints are saved into a file and deduplicated first. In the second pass it is necessary to keep only the duplicate fingerprints in the memory. According to information provided by the author, under typical conditions memory use can drop to only 10%. In this alternate mode of operation, the saved fingerprints can also be reused in subsequent experimentation with different values of similarity threshold and/or different levels of deduplication.

[4] URL: `http://code.google.com/p/onion/`

3.1 The Onion Experiment

To get an idea of the number of near-duplicates detected by n-grams of tokens, we decided to run an experiment with Onion applied to the corpus mentioned in the previous section. As Onion expects to get the input data in one column, at the beginning of our experiment the columns containing morphological annotation (Lemma, Tag) were removed from the source file.

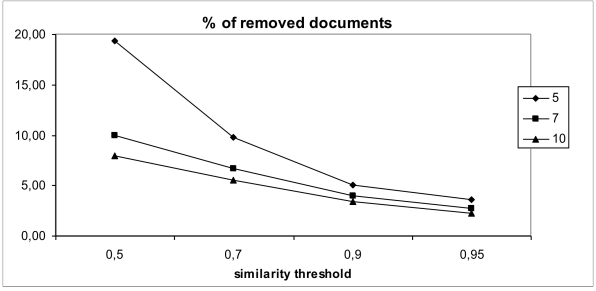
3.2 Document-level Deduplication

As a first step, we decided to observe the level of deduplication at the document level by means of 5-, 7- and 10-grams with four values of similarity threshold (0.5, 0.7, 0.9, and 0.95, respectively). For each value of n-gram we let Onion pre-compute the fingerprints first, which were subsequently used for deduplication with different values of similarity threshold. The computation of fingerprints lasted on average 27 minutes and the respective deduplication passes lasted typically 32 minutes.

The results of the deduplication are summarized in the following tables. The first one shows the numbers of removed documents with different values of n-grams and threshold levels.

Similarity threshold	5-grams	7-grams	10-grams
0.5	269,076	137,780	110,108
0.7	136,158	92,215	77,183
0.9	69,572	54,864	47,381
0.95	49,498	38,140	31,098

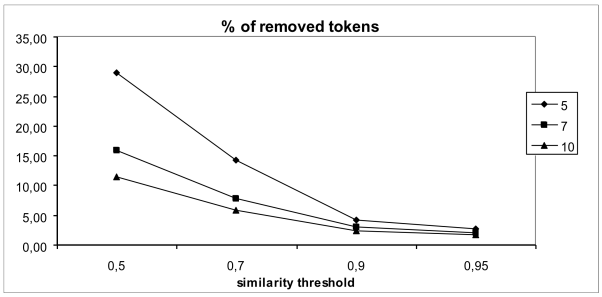
The above values expressed in per cents can be visualized as follows:



The next table indicates how the various deduplication parameters influence the numbers of removed tokens.

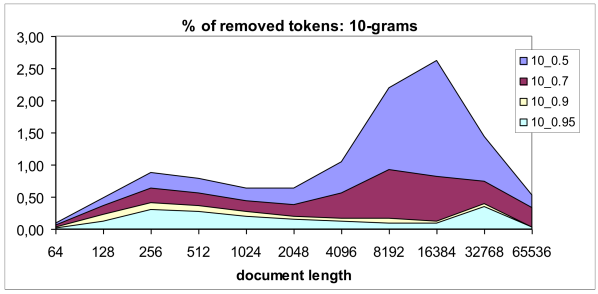
Similarity threshold	5-grams	7-grams	10-grams
0.5	354,704,820	194,226,021	139,317,046
0.7	174,064,712	94,776,159	71,304,396
0.9	50,434,177	36,612,604	29,284,459
0.95	32,465,363	24,242,558	21,209,472

Again, the situation expressed in percentages can be visualized by a graph.



The graphs show clearly that with a low similarity threshold (0.5), the share of removed texts and tokens is strongly dependent on the value of n-grams. On the other hand, with a "conservative" setting of the threshold (0.9, 0.95) the value of the n-grams has only a limited influence.

In the end we show the frequency distribution of the removed tokens by the length of documents (for 10-grams).



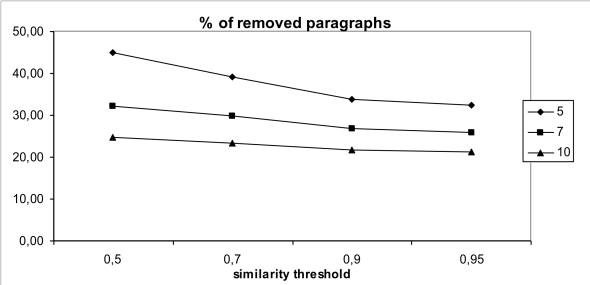
An interesting observation is the rapid increase in the number of deleted tokens with the low frequency threshold within the longer documents. This phenomenon deserves further inspection.

3.3 Paragraph-level Deduplication

The second experiment aimed at deduplicating paragraphs was performed with identical settings. Onion was run in the "no smoothing" mode[5]. The following tables report the numbers of removed paragraphs.

Similarity threshold	5-grams	7-grams	10-grams
0.5	23,119,807	16,541,810	12,697,603
0.7	20,164,592	15,294,473	12,027,316
0.9	17,353,000	13,738,966	11,187,422
0.95	16,652,531	13,285,899	10,890,177

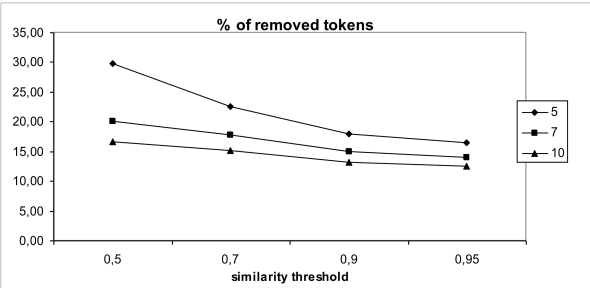
And the graph expressing this in percentages.



The situation with tokens looks like this.

Similarity threshold	5-grams	7-grams	10-grams
0.5	364,942,345	245,171,349	203,061,064
0.7	276,284,030	217,137,998	184,798,690
0.9	218,857,869	184,131,198	161,354,010
0.95	201,514,920	171,852,183	152,246,068

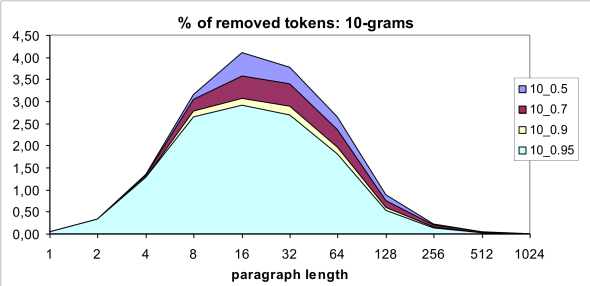
The last graph shows the percentage of removed tokens.



[5] In the "smoothing" mode Onion removes also short non-duplicate paragraphs between two duplicate ones.

We can see that with "aggressive" parameter settings, the deduplication procedure would remove 45% of paragraphs containing 30% of tokens. With the more conservative settings the respective curves approach the 15% level, which is quite similar to the 13.7% achieved by the plain fingerprint method.

We also show the frequency distribution of removed tokens (for 10-grams).

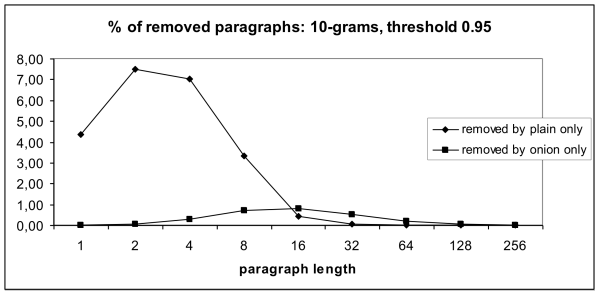


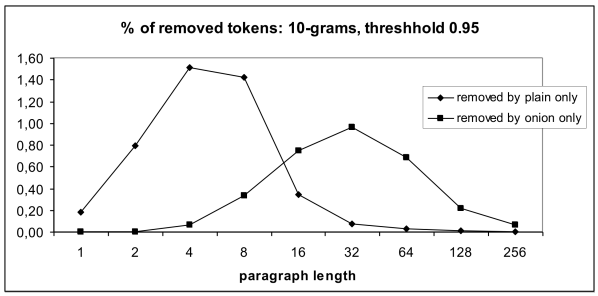
If we compare this graph with the similar one from the plain fingerprint method, we can see that Onion prefers removing paragraphs of medium length where partial match is more likely to happen. Based on these findings we decided that Onion will not be used for deduplication on the sentence level.

3.4 What Has not Been Removed by Onion

The last interesting question is which paragraphs were removed by the plain fingerprint method but remained undetected by Onion. Our analysis was performed just with the most conservative values of Onion settings (10-grams, threshold-level of 0.95), where the results were expected to be most similar.

The frequency distributions of the removed paragraphs and tokens are depicted in the following graphs.





We can see that Onion's weak point is the ignorance of duplicates in short paragraphs. According to our analysis, this is mainly caused by the fact that paragraphs shorter than the length of the n-gram are considered duplicate only in the case when the n-grams also match the respective tokens from the end of the previous paragraphs. With the shorter paragraphs, there is also a greater chance of partial match with ignored punctuation and digits implemented in our simple method.

4 Conclusion and Further Work

In our paper, we have compared the results of deduplication achieved by two methods – with a plain fingerprint method and by means of Onion. While in detecting exact duplicates the situation is fairly simple, in the detection of near-duplicates there is always a trade-off between the amount of "good" text to be lost and the amount of duplicate contents that will remain in the corpus.

Onion is a very fast and versatile tool that can be conveniently used to detect near-duplicates both at the document and the paragraph level. Its main deficiency is the inability to detect duplicates in short paragraphs. Our suggestion for corpus deduplication is therefore based on a combination of both tools. The whole process would consist of three stages. In the first stage the corpus is deduplicated by Onion at the document level with conservative levels of the parameters (duplicates are removed). In the second stage Onion deduplicates paragraphs (duplicates are marked). And in the last stage the short duplicates are "cleaned" by the plain fingerprint method at the sentence level.

In the future we want to optimize computation of fingerprints in the plain method and apply the results of our research to the whole SNC collection of corpora, as well as to the newly created Slovak web corpus.

References

[1] Benko, V. (2010). *Optimizing Word Sketches for a large-scale lexicographic project*. Invited lecture. URL: `http://videolectures.net/korpusi2010_benko_ows`.

[2] Rabin, M. O. (1981). *Fingerprinting by Random Polynomials. Center for Research in Computing Technology*. Harvard University. Tech Report TR-CSE-03-01. URL: `http://www.xmailserver.org/rabin.pdf`, retrieved 10 May 2013.

[3] Broder, A. Z. (1993). Some applications of Rabin's fingerprinting method. In *Sequences II: Methods in Communications, Security, and Computer Science*. Springer-Verlag. URL: http://xmail.eye-catcher.com/rabin_apps.pdf, retrieved: 28 April 2013.

[4] Slovak National Corpus – prim-6.0-juls-all. (2013). Bratislava: Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences. Accessible at: http://korpus.juls.savba.sk.

[5] Pomikálek, J. (2011). *Removing Boilerplate and Duplicate Content from Web Corpora*. Ph.D. Thesis, Faculty of Informatics, Masaryk University in Brno. URL: http://is.muni.cz/th/45523/fi_d/phdthesis.pdf, retrieved 14 June 2012.

[6] Uzgalis, R. (1995). *Random Numbers, Encryption, and Hashing*. Lecture Notes. Computer Science Department, University of Auckland. URL: http://www.serve.net/buz/Notes.1st.year/HTML/C6/rand.012.html, retrieved 20 April 2013.

Software System for Processing Bulgarian Digital Resources: Parallel Corpora and Bilingual Dictionaries

Ralitsa Dutsova and Ludmila Dimitrova

Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Sofia, Bulgaria

Abstract. The paper presents a software system for processing and web-presentation of Bulgarian digital resources – parallel corpora and bilingual dictionaries. The main components of the system's current version and its functionalities are briefly described. The focus is given to the description of the new modules "Search" (a tool for information retrieval and data extraction from a bilingual dictionary) and "Connection" (a component that links two other modules: "Dictionary" and "Corpus"). Several examples illustrate the usage of the system's web-applications.

1 Introduction

In this article we describe a software system used for processing bilingual digital resources with Bulgarian, developed at the Institute of Mathematics and Informatics of Bulgarian Academy of Sciences. The system will provide web services for adequate presentation and usage of these resources via the Internet. Our idea is to create a system that allows all users (including the super administrator) to work with and use all of our digital resources, available at the time, and the newly-developed web-based services for open access to the language resources. The structure of the system is designed and developed on a modular principle: the system comprises separate, but connected with special links, autonomous software tools (modules). Each module has its own database and its own user interface. At the first stage of the development of our project, only one software tool, an online Bulgarian-Polish dictionary [2], [3], was realized. In a second phase, a software tool – a web-application for the presentation of bilingual aligned corpora with Bulgarian as one of the paired languages – was designed and developed, its implementation is upcoming [1]. The third step consists of reprogramming of the software package for the bilingual online dictionary. The recent developments include new modules:

1. "Search" – a web-application for information retrieval and data extraction from a bilingual dictionary with Bulgarian as a source language,
2. "Connection" – a software tool for the realization of the connection between both modules: module "Dictionary" (web-application for supporting Bulgarian-Lang2 online dictionary) and module "Corpus" (web-application for the presentation of bilingual aligned corpora with Bulgarian as one of the paired languages).

The **software system** for processing of Bulgarian digital resources is aimed at two user groups. The first user group includes the so-called "administrators", including super administrators – people who have designed and developed, and manage the system, and the second, the so-called "end-users" (or casual users) – people who use it. Depending on user type, we have allocated tasks and services in two sets: for "administrators" and for "end-users". Thus, each module of the system – web-based application – consists of

two main software packages: an **"administrative (control)" panel** intended for "administrators" and an **"end-user" part** of the website intended for "end-users".

Tasks and services allocated to the "administrative (control)" panel: to create a special kind of user – "super administrator" – who will manage the web-based application; to give access to the "administrators" (to register a new "administrator" and delete an existing one); and to receive messages with information reported by the "end-user" and to store these messages for future processing.

The "administrative (control)" panel does not require additional "administrator" training, it has a very simple interface and offers the possibility for the user to add to, edit, delete from and search the database of each web-application. After the administrator has logged in to the module, he/she is redirected to the corresponding module page. After a search has been performed, the user has the ability to edit the database if corrections are needed or delete the listed records. The database supporting the web presentation of the software system is Relational Database (RDB). The main function of the RDB is to store the data for the correspondent modules and to support the process of searching for and extracting requested data. There are several advantages for the usage of RDB to store content, well structured language knowledge: maintenance of data integrity, ensuring data security and independence, quickly and efficiently search and data retrieval, data upload and update. The RDB is realized in MySQL – the world's most widely used open source relational database management system that runs as a server providing multi-user access to the database. The modules "Dictionary" and "Search" use the same RDB, but "Coprus" uses another specific RDB to store corpora segments (point 4.1).

Tasks and services allocated to the "end-user" part of the website: to create a user-friendly interface in both languages – Bulgarian and Lang2; to ensure quick search to the module database; and to provide accurate and up-to-date information to "end-users".

Web-interface: the web-based user interface is multilingual. The user can switch between two or more languages. If we specify the second language from the Bulgarian-Lang2 paired database we can allow the new language to be optional and to display the results in this language. The user can use Bulgarian as input language. A virtual keyboard is implemented to help the "end-user" if no Bulgarian alphabet is installed on the user's computer. The search is performed according to the primary language selected from the user. Currently the "end-user" can switch between English and Bulgarian language.

2 Module "Dictionary" – Web-application for the Presentation of Bilingual Dictionaries (with Bulgarian as the Source Language)

Module "Dictionary" is a multifunctional software tool for the creation and web-presentation of bilingual online dictionaries with Bulgarian as the source language and unspecified changeable Lang2 as the translation language, in other words, the dictionary is independent of the target language. Our goal is to design a bilingual Lexical Database (LDB) independent of the target language (Lang2).

2.1 Web-application's Databases

The database of this Web-application is a specialized bilingual LDB of a bilingual dictionary with Bulgarian fixed as a first language in the pair, and Lang2 as the second. The

current version uses a digital Bulgarian-Polish LDB [6]. This bilingual LDB was designed for supporting the first Bulgarian-Polish experimental online dictionary [5]. The formal model of the bilingual LDB follows the CONCEDE model for monolingual dictionary encoding [7], with extensions and modifications aiming to cover more of the specific features of Bulgarian. The hierarchical tree structure of dictionary entries corresponds to the TEI standards for encoding dictionaries [8]. The functions of the LDB is to store the dictionary entries and to serve as an entry point to the RDB. The implemented bilingual LDB makes it possible to design a RDB of a Bulgarian-Lang2 online information retrieval tool. The RDB is supported by tables, containing dictionary entries contents, and links between the tables. An XML parser performs the transformation of LDB into RDB. The aim of this syntactic analyzer is to initialize the RDB serving as a basis of the bilingual dictionary so that the entries, saved in the RDB, can then be easily edited. The parser is programmed in Java, so it can be run on different platforms independent of the architecture or the operating system.

2.2 Specific Web-services

After the user-name and password have been entered and verified, the "administrator" user is redirected to the "administrative (control)" panel – Fig. 1.

The "administrative (control)" panel creates a web-based Bulgarian-Lang2 dictionary, ensures an easy usage of the tool and provides functionality for updating the dictionary content and possibilities to store the information about missing words reported by the "end-users" for a future processing. The software tool offers a user-friendly interface for word addition, editing, deletion and search.

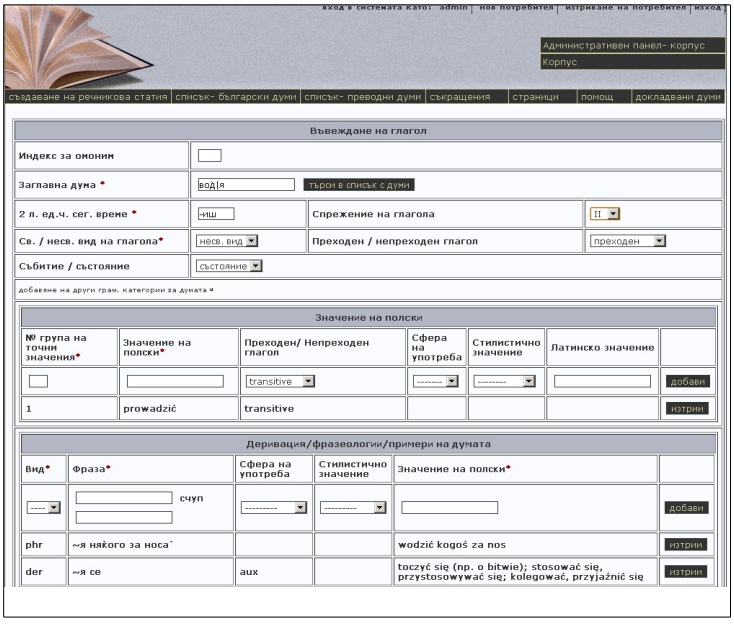


Fig. 1. "Administrative (control)" panel – adding of the grammatical characteristics of the Bulgarian verb "***водя***" */lead, conduct, guide, wage/*

***How the "administrative (control)" panel works*:** The access to this panel is restricted only to authorized people. The panel consists of several sections: for uploading a new word, for searching Bulgarian or Lang2 word, for translation settings, etc. The user must choose from a drop-down menu what he/she wants to upload: a noun, a verb, an adjective or any other part of speech (pronouns, conjunctions, adverbs). The fields displayed are only the ones necessary for adding the chosen part of speech. All other fields needed for other parts of speech are hidden from the user. When all the information is filled out and the user presses the button "Save", the word is stored in the database, and it will be possible for that word to be searched and displayed on the "end-user" screen.

Tasks and services allocated to the "end-user" part of the website: to ensure quick search of words in the online dictionary LDB, and to provide the ability for translation in both directions. There is a possibility to search for a translation in both directions: from Bulgarian to Lang2 or from Lang2 to Bulgarian. The translation from Bulgarian to Lang2 will display the whole information existing in the LDB for the searched word. The translation from Lang2 to Bulgarian will be composed only of the main meaning of the Bulgarian headwords.

3 Module "Search" – Web-application for Information Retrieval and Data Extraction from a Bilingual Dictionary

The main functions of any information retrieval tool are: (1) to process user requests, i.e. to check the validity of the request and then to search for the requested data, (2) to produce the results i.e. to extract requested data and to show them to the user's screen. This web-tool will search for, extract and facilitate access to information, which is already well systematized and stored in digital form as a set of dictionary entries. Our web-tool uses the simplest (common) method of data search and data extraction from a dictionary – the search by a pattern and extraction of encoded information.

3.1 Search Organization

Since our software tool uses an implemented RDB for bilingual dictionary with Bulgarian, we need only to develop a module oriented to the end (i.e. casual) user. This end-user module must provide an effective search in the Web-application database, based on different criteria given by the users via their requests, to filter the available data and ensure adequate output.

The end-user module is generally accessible to the casual users, but the user can register by filling in the registration form. The tool enables registered users to save different search criteria and filters (most preferable or usable), so that the user can use them without entering them again. Depending on the user requirements multiple criteria for search are permissible for the search procedure.

For example, the search procedure can find all nouns and verbs that have the same beginning. The search is performed according to the string inserted in the text field of the user request. The result is filtered according to the tag search criteria. First of all the search procedure checks the type of search, namely if the user uses criteria for lemma search. Otherwise the procedure checks the additional tag search criteria. As we already mentioned, the user can insert more lemmas in the text field separated by semicolon (";").

In this case a small parser creates substrings from the separate words and the web-tool retrieves the available information for each substring which is an actual lemma or part of a lemma. If lemma search criteria are fixed, the tool will retrieve the whole information available for the corresponding lemma, which is equivalent to the dictionary entry.

The tool has the possibility to perform a search only by tag or only by lemma. If the user does not insert a string in the text field, the program will retrieve only a list of available words which fulfil the user's tag search criteria. Then the search is performed according to the part-of-speech-criteria and its specific characteristics, for example: to retrieve all verbs with conjugation type=III, which are intransitive. The request form doesn't allow the user to insert discrepant information.

3.2 Types of Search Requests

The program tool executes search requests according to the information given by the user: 1) tag search, 2) lemma search, and 3) combination of both.

Tag search is carried out when the user enters only one characteristic of a group of words, specified by language or linguistics information, for example, "phrase", "derivation", "transitive/intransitive" for verbs, etc. It is also allowed for user to enter an expression that join two characteristics, for example, "phrase" and a list of words, "phrase" and a given "POS".

Lemma search is carried out when the user enters a given lemma or its "part" (some sub-string, for example, a syllable). If the user enters an initial syllable or a final syllable of a given lemma (so called "rhyme"), a Rhymer procedure will produce result as a dictionary of "rhymes".

In this case the Rhymer procedure retrieves information for the rhymes of a corresponding word. We recognize two types of rhymes: head-rhyme and end-rhymes. Words with head-rhyme have the same initial syllable. Words with end-rhyme have the same final syllable. For example, if the user enters the word ***"вятър"*** */wind/* under this option, Rhymer retrieves a list of words ending the same way (e.g. ***"пъстър"*** */motley/*, ***"театър"*** */theatre/*, ***"филтър"*** */filter/*, ***"хитър"*** */sly/*, etc.). This option lets easily find exact rhymes (words in which the final syllables are the same). The system can narrow the user search if he/she specifies a list of rhymes: to show the rhymes for a correspondent word which are only verbs, for example. The user can also enter a list of lemmata. The lemmata should be separated by semicolon (";"). In these cases, the search will be performed separately for each lemma and all the available information will be displayed.

3.3 Input Data

When the casual user loads the web-application to work with, a web form is loaded: the user can specify there the search type. In order to check the validity of the user requests some control functions in the search procedure are added. In the text field the user can insert a lemma, or part of a lemma, or a list of several lemmata separated by semicolon. The displayed results can be narrowed by choosing the additional criteria in the web form of the request.

The user can specify his/her requirements concerning the words (the lemmata listed in the text field) by clicking selected menu buttons of the web form. For example, performing the user request shown on the Fig. 2, the web-tool will display on the screen only transitive verbs, conjugation type II, expressing state, and appearing in phrases and examples in dictionary entries.

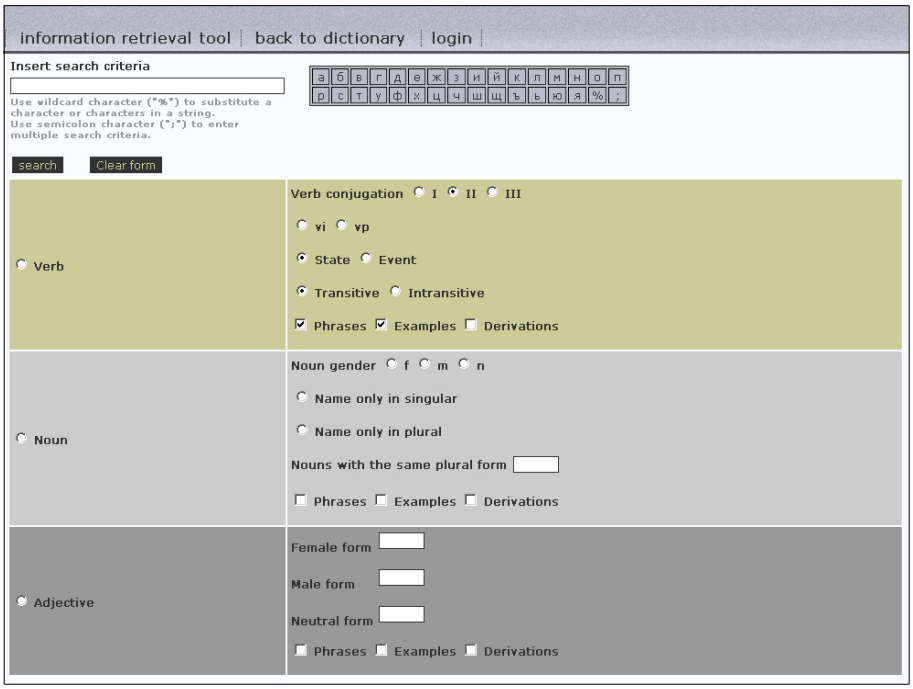


Fig. 2. User request form for word search and extraction with additional criteria

3.4 Output Data

The system is designed to produce output data that can be visualized as sequences of dictionary entries or word lists with special characteristics, specified by the user request. If the user inserts only a headword in the search criteria field, all the available information for this headword is displayed, i.e. the displayed result is the dictionary entry.

If the user wants to retrieve words, with specific characteristics (for example, all verbs expressing "state") the web-tool will display a list of the verbs that meet the search criteria. The listed words are hyper-links redirecting the user to the corresponding dictionary entries.

If the user is interested in the usage of a given word in any phrases, derivations or examples, the web-tool will display the whole list of the corresponding headwords and all the available phrases or/and examples. Such narrowed output helps the user understand easily the contextual usage of the word.

The Fig. 3 shows the screen after the user request has been processed: a part of resulting list of transitive conjugation II verbs, expressing state, appearing in phrases and examples of dictionary entries, and their translations in Polish. The user has chosen to see full content saved in entry with headword Bulgarian verb ***"водя"*** */lead, conduct, guide, wage/*.

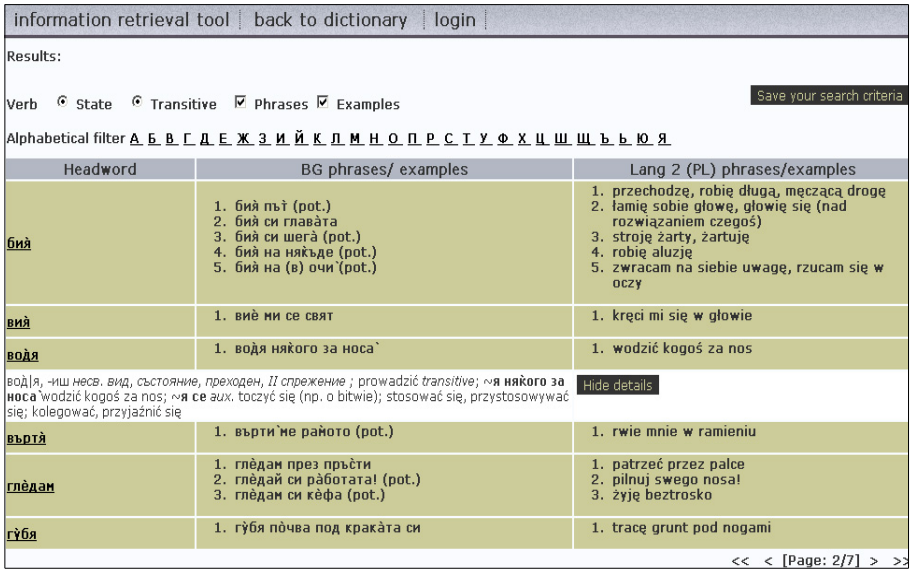


Fig. 3. Results (second page) received after processing of user request from Fig. 2

4 Module "Corpus" – Web-application for Aligned Corpora Presentation

Module "Corpus" is a technological tool implemented as a web-based application for the presentation of bilingual aligned corpora with Bulgarian as one the two paired languages. An aligned corpus is a parallel corpus containing relations between corresponding excerpts of text in multiple languages [4], [9]. The texts in the ongoing version of the corpora are automatically aligned at the sentence level. We use language-independent freely-available software tools to align bilingual corpora employing Bulgarian: the MT2007 Memory Translation computer aided tool (TextAlign), and the Bitext Aligner/Converter (Bitext2tmx aligner). The resulting aligned texts (usually called bi-texts) are similar. Both software packages align bilingual texts without bilingual dictionaries, but human editing is obligatory.

4.1 Relational Database, Supporting Web-application

The module "Corpus" is based on a RDB of bilingual Bulgarian-Lang2 corpora (Bulgarian-Polish in the current version). The relational model is supported by tables containing core information of the corpora entries and the links established between them. We pay special attention to building the database that supports the web presentation of bilingual corpora in order to address the following computational complexities. Searching a large text can be a costly time-consuming operation. The RDB structure was therefore designed in a way to provide easy and fast search capabilities for the end-users of the bilingual web corpora (Fig. 4).

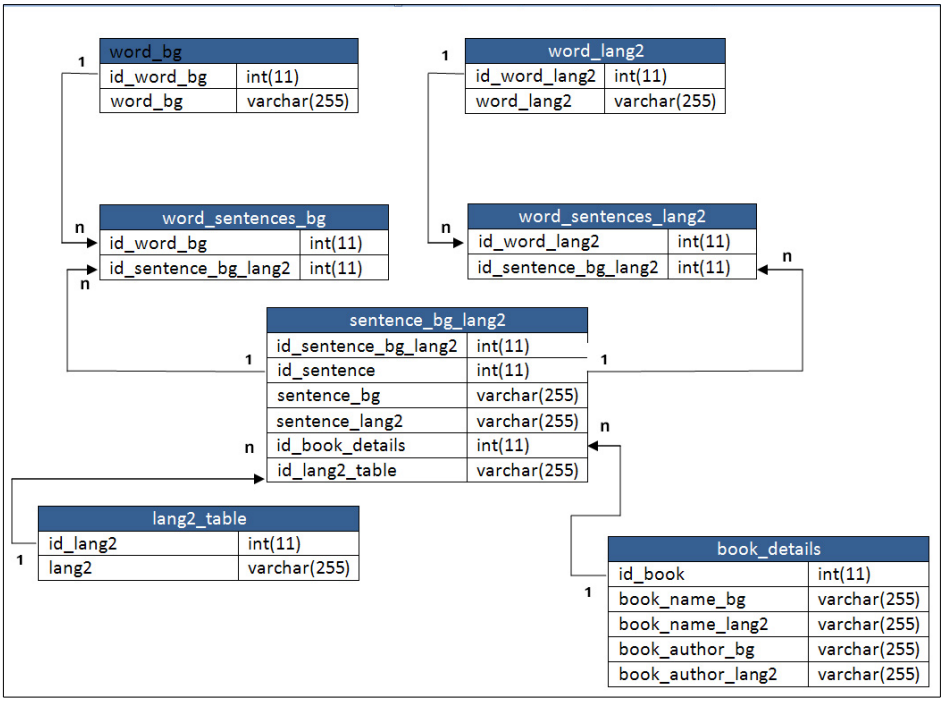


Fig. 4. Structure of the RDB

When a user inserts a new record in the RDB through the "administrative (control)" panel, a back-end text parser program takes the input text and simplifies it to its separate constituent words. The different words are then saved in different fields in an index table of the database, and for each word a link is kept to another table where the full text of the aligned pair is saved. This parsing is done for both Lang2 and Bulgarian texts. In this way, we achieve a good search performance and only a small delay while inserting new records in the RDB. The delay is not so sensible and the administrator will not pay a big attention to it, because of the possibility to add the new aligned pairs only one by one. The "administrative (control)" panel provides a simple web-form where the user can insert a new pair of aligned texts.

4.2 End-user Web-interface

Only a ***search-by-word*** capability is provided to the end user. All pairs of aligned text where the searched word has been found are listed in a table. In order to show the word in a better context, together with the target pair we display the previous and next pair as well. If the search results exceed more than 15 records, paging is provided.

5 Module "Connection" – Web-application for Corpus and Dictionary Connection

5.1 Module Main Functions

Administrator & Super Administrator Functions: The administrator of any module has to have access to the "administrative (control)" panel and "end-user" part of the website of the other module: that is, access has to be provided from the "administrative (control)" panel of "Dictionary" module to the "end-user" part of the website and "administrative (control)" panel of "Corpus" panel and vice versa. The administrator has to be able to change personal data for access to the "administrative (control)" panel – password and email. If the administrator has access rights for both "Dictionary" and "Corpus" modules, any similar changes have to be noted in both places. The super administrator has to determine access rights for a given user. From the "administrative (control)" panel of "Dictionary" or "Corpus" module it must be possible to create new users (including administrators) who have simultaneous access to "administrative (control)" panels in both modules.

User Interface Functions: To allow users to search in parallel or in circuit in the dictionary and/or corpus, to display the search results in a synthesised way, to facilitate the user in an effective use, and to provide user (by requested registration) access to additional functionalities, some of which could be developed in the future.

5.2 Web-application for Corpus and Dictionary Connection

The web-application developed to unite the use of "Dictionary" and "Corpus" modules was easily realized because autonomous user interfaces are developed for every module. Every module is accessible separately by its own internet address. The need for developing a common user interface arose with the idea of creating a common system which processes digital bilingual resources with Bulgarian. The "end-user" part of the website in the linking modules has a common access, and users are able to search with it in both dictionary and corpus databases. The search is bilingual, for Bulgarian and Lang2 words. This module has a relatively simple structure: mirror Bulgarian and Lang2 versions, hyperlinks to the "end-user" part of the dictionary and corpus website and several sections: "about the project", "maintenance", and "entry". The module's "home"-page consists of a query form with text field, where the user can enter the word for a search and redirects this search via a check-box. If the user searches for the translation correspondence of the word entered (in the dictionary database), the screen displays the dictionary entry whose headword is this given word. If the user searches the given word in the corpus database, the screen displays the concordance of the given word. A dual search option is also provided – that will display on the screen the information present in the dictionary and corpus databases: dictionary entry plus pairs of aligned text where the word occurs. Since the user interface of "Dictionary" and "Corpus" has a two-way connection for switching between systems, the user is provided with the following possibility: if the query result in any module is "nil", the user has the possibility to start an analogical search in the other module by a button click. A small sub-window appears displaying the results of the second search, for example, if the first search was in the dictionary, the sub-window displays the results from the secondary search in the corpus and vice versa.

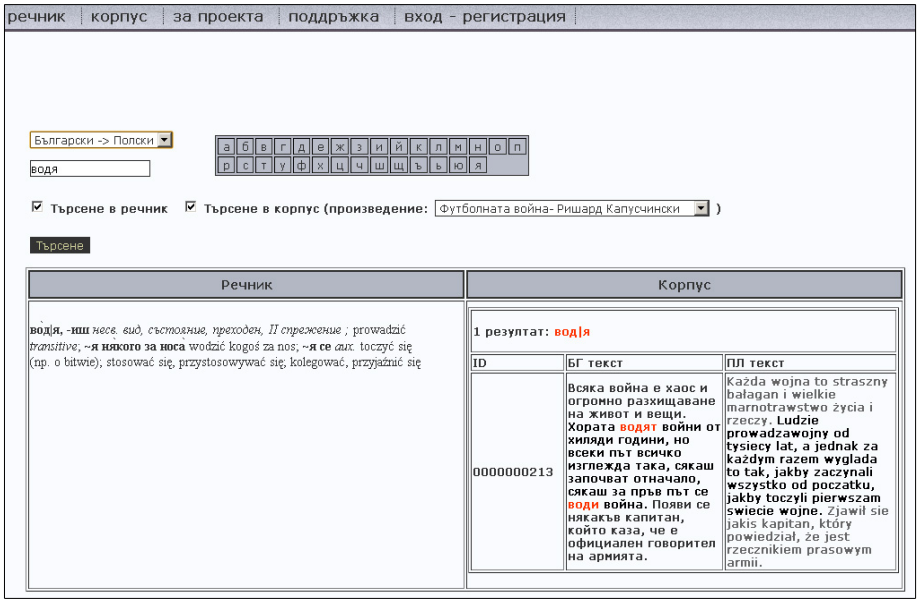


Fig. 5. Result displayed after search of Bulgarian word ***"водя"*** */lead, conduct, guide, wage/* in both modules "Corpus" and "Dictionary"

The "administrative (control)" panel of a new module is not envisaged: both tools "Dictionary" and "Corpus" has different structures and specifications, so that joining them into a single "administrative (control)" panel would create a complex structure accessible via a complex interface and create difficulties for the user. The only common part between both tools is the ***login*** page. When the user loads the system through a web-browser, a ***login*** form appears. The ***login*** form provides a possibility for the user to enter ***user-name*** and ***password***, and then to choose which tool to enter by clicking a radio-button. After a verification of access rights, the system redirects the user to the "administrative (control)" panel of the "Dictionary" or to the "administrative (control)" panel of the "Corpus". The user has access to both parts with the same password. However, the "administrative (control)" panels of "Dictionary" and "Corpus" can be used to create users with different access rights: those with access to the dictionary only, and those with access to the corpus only. There may be users with no common simultaneous access to both systems. After the ***login*** prompt the system recognises whether the user is an administrator with full rights and loads only the sections accessible to the user. The "administrative (control)" panel of each module has a link to "administrative (control)" panel and "end-user" part of the website of the other module. If the user wishes to enter the "administrative (control)" panel of the other module, his/her rights are checked first. If these access rights exist, the user is redirected to the "administrative (control)" panel of the other module and his access rights to the other module are verified. If everything is OK!, a link to the "administrative (control)" panel loads. Thus the user can access the "administrative (control)" panel of the second module without a repeated verification of

the access rights. If the user has no access rights to the second module, no link appears. The link to the user interface always loads regardless of the module the user is in and independent of the user access rights.

6 Conclusion and Further Works

The paper presents briefly a system for processing and web-presentation of Bulgarian digital resources (parallel corpora and bilingual dictionaries), still an experimental tool with flexible structure in state of development. Changes in the system are possible during the ongoing implementation. Some possibilities for inclusion of new tools for extending the service capabilities are envisaged. In our opinion, the module "Search" – an web-application for information retrieval and data extraction from a bilingual dictionary – is very useful tool with a broad range of applications in contrastive studies and education, especially in language learning. The implementation of various parameters of a search, namely, morphosyntactic, derivative data, rhyme search – is an advantage of this software tool. As a whole the system will be widely applicable for research purposes in digital humanities, in a system for human and machine translation systems, as well as for the development of bi- and multilingual lexical databases and different kinds of digital dictionaries, and in everyday life (human communication).

References

[1] Dimitrova, L. and Dutsova, R. (2013). Web-Application for the Presentation of Bilingual Corpora (Focusing on Bulgarian as One of the Paired Languages). *J. Cognitive Studies/Études Cognitives*. 13, SOW, Warsaw. In press.

[2] Dimitrova, L. and Dutsova, R. (2012). Implementation of the Bulgarian-Polish Online Dictionary. *J. Cognitive Studies/Études Cognitives*. 12, SOW, pages 219–229, Warsaw.

[3] Dimitrova, L., Dutsova, R., and Panova, R. (2011). Survey on Current State of Bulgarian-Polish Online Dictionary. In *Proc. of the International Workshop "Language Technology for Digital Humanities and Cultural Heritage" within RANLP'2011*, pages 43–50, Hissar.

[4] Dimitrova, L. and Garabík, R. (2011). Bulgarian-Slovak Parallel Corpus. In *Proc. of the 6th International Conference NLP, Multilinguality.* SLOVKO 2011, pages 44–50, Slovak National Corpus, Bratislava.

[5] Dimitrova, L., Koseska, V., Dutsova, R., and Panova, R. (2009). Bulgarian-Polish online Dictionary – Design and Development. In *Proc. of the MONDILEX Fourth Open Workshop*, SOW, pages 76–88.

[6] Dimitrova, L., Panova, R., and Dutsova, R. (2009). Lexical Database of the Experimental Bulgarian-Polish online Dictionary. In *Proc. of the MONDILEX Third Open Workshop, 15 – 16 April, 2009, Bratislava,* pages 36–47.

[7] Erjavec, T., Evans, R., Ide, N., and Kilgarriff, A. (2000). The Concede Model for Lexical Databases. In *Proc. of the Second International Conference on Language Resources and Evaluation, LREC'00*, pages 355–362, ELRA, Paris.

[8] Ide, N. and Véronis, J. (1995). *Encoding dictionaries*. In Ide, N. and Veronis, J., editors, *The Text Encoding Initiative: Background and Context,* pages 167–179, Kluwer Academic Publishers, Dordrech.

[9] Rosen, A. (2005). In search of the best method for sentence alignment in parallel texts. In *Proc. of the Computer Treatment of Slavic and East European Languages: Third International Seminar*, Bratislava, pages 174–185.

Slovene Corpora for Corpus Linguistics and Language Technologies

Tomaž Erjavec

Jožef Stefan Institute, Ljubljana, Slovenia

Abstract. The paper introduces annotated Slovene reference, specialised and parallel corpora meant for use in corpus linguistics and language technology research. The corpora are (also) available on the Natural Language server at the Jožef Stefan Institute `nl.ijs.si` via two concordancers, typically not requiring authentication. Some corpora, esp. manually annotated ones, are also available for download under one of the Creative Commons licences. The paper then addresses aspects of accessibility, the technical ones of encoding, access to concordancer and sustainability, and the legal ones of copyright and of personal data protection.

1 Introduction

Computer corpora are the basic language resource for, on one hand, empirical language studies, i.e. corpus linguistics and, on the other, human language technology (HLT) research and development. The former includes not only theoretical, but also many fields of applied linguistics, esp. lexicography, terminology and language teaching, while the latter uses corpora as training and testing datasets for language technology tools, e.g. part-of-speech tagging, lemmatisation, normalisation (for non-standard language), named entity recognition, syntactic parsing, word-sense disambiguation, etc.

The two types of research, in general, require different availability of the corpora. While linguistic research is most often performed through a concordancer, while in human language technologies the complete dataset needs to available for download. This brings with it more demands on the corpus compiler, from ensuring a well-understood and documented encoding, to taking care of legal issues connected with distributing the corpus texts.

For each language, corpora must be developed more or less separately, including the identification and collections of base texts and manual or automatic linguistic annotation. While the range of possible corpora is almost unlimited, a "resourced" language and should have large, monolingual reference corpora, hand-annotated gold-standard corpora, specialised corpora of various domains and text types (including non-standard langauges), as well as multilingual parallel corpora for translation studies and machine translation research.

In this paper we introduce the corpora of (mostly) Slovene language, which have been developed in over twenty years, but have been recently re-encoded in line with current encoding encoding practices, and, in many cases, re-annotated [9] and all made available through two web concordancers. Some corpora with less restrictions on availability are also avaliable for download, which includes the two linguistically hand-annotated corpora of Slovene. The corpora are available from the "Natural language server" `nl.ijs.si` at the Jožef Stefan Institute, which has also been operational for about two decades. In adition to corpora, the server also offers other langauge resources (dictionaries and digital libraries), as well as various services, e.g. tagging and lemmatising Slovene texts.

The rest of this paper is structured as follows: Section 2 lists the corpora available on the server, concentrating on their text type, size, temporal scope, and how and what linguistic information they encode; Section 3 addresses technical aspects of their availability, in particular our encoding of the annotated corpora, the provision of concordancers, and the sustainability of these valuable language resources; Section 4 remarks on legal aspects of availability, which can make it difficult to share the resources, namely copyright and personal data protection; and Section 5 concludes with some directions for further research.

2 Slovene Language Corpora

In this section we present the corpora where the Jožef Stefan Institute was a partner in their development and that are available – either via concordancers or for download – on the JSI natural language server.

2.1 Corpus Annotation

The corpora mentioned below contain various structural elements, such as paragraphs or page breaks. The structural elements are associated with meta-data, e.g. for a text, its title, author, year of publication, and publisher, or for page breaks the pointer to the facsimile (image) of the particular page.

All the corpora are also linguistically annotated. At the most basic level this process involves tokenisation, tagging each word token with its context disambiguated morphosyntactic description (MSD) and lemmatising it. The MSDs (commonly known as part-of-speech tags) all follow the scheme developed in MULTEXT-East [5], where each MSD is a string, e.g. *Ncmsn*, a valid MSD for Slovene, which can be decomposed into a set of attribute-value pairs, in this case to *Noun, Type=common, Gender=masculine, Number=singular, Case=nominative*. The MSDs can also be localised, e.g. *Ncmsn* is equivalent to the Slovene *Somei* or *samostalnik, vrsta=občno_ime, spol=moški, število=ednina, sklon=imenovalnik*. Most corpora discussed below use the set of MSDs (and features) defined in the morphosyntactic specifications of the JOS project [12] (equivalent to the MULTEXT-East specifications for Slovene), which defines a detailed tag set comprising almost 2,000 distinct MSDs.

The majority of the corpora below have been annotated automatically, although a couple do have manual annotations. For automatic annotation we mostly used ToTaLe [11], which performs tokenisation, morphosyntactic tagging and lemmatisation.

2.2 Reference Corpora of Slovene

We first introduce the available reference corpora of Slovene, which are try to be representative and balanced for the language. They are also the results of relatively large projects, so they are likely to be larger and more carefully composed than the other specialised corpora.

The largest available corpus of Slovene is Gigafida, the reference corpus of contemporary Slovene (1995–2011), with over a billion words [20]. The corpus was MSD tagged and lemmatised with the Obeliks tool [13]. The ten times smaller balanced corpus **KRES** [20] with 100 million words has been paragraph-sampled from Gigafida and was compiled with to offer a better distribution of text types than that of Gigafida, so, for example, word frequency information obtained from this corpus is more in line with

"typical" Slovene. While both corpora are freely available only via concordancers, they have ten times smaller variants (again, paragraph-sampled), called **ccGigafida** and **ccKRES** [20], which are available for download.

The corpus **ssj500k** [1] comprises half a million words and contains manually annotated sampled paragraphs from Gigafida and is meant for training (and testing) HLT tools. In addition to MSD tags and lemmas, a portion of the corpus is also manually annotated with named entities and with syntactic dependency structures; the corpus is also available for download.

The corpus of spoken Slovene **Gos** [31] contains just over one million words. The corpus on the nl.ijs.si server offers transcriptions only, as the speech files are not publicly downloadable. The verbatim transcription of each word token in the corpus is annotated with its normalised form, as well as its MSD and lemma. The corpus of the transcriptions is available via concordancers as well as for download.

All the above mentioned corpora have been developed in the scope of the large Slovene project, "Communication in Slovene", and are available (for concordancing and, where appropriate, for download) also from the home page of this project at `www.slovenscina.eu`.

The **IMP** corpus of historical Slovene (1750–1918) contains hand-corrected transcriptions of (mostly) complete books and some newspapers, and comprises 650 units with almost 15 million words [7]. The pages in the corpus are connected to their facsimile in the IMP digital library and the words in the corpus are annotated with their modern equivalent, lemma and MSD tag, where the annotation was performed automatically with the ToTrTaLe tool, [4], a variant of ToTaLe, which, after tokenisation, performs transcription, i.e. it modernises historical word-forms, and then proceeds with MSD tagging, using the modernised forms.

The reference corpus of historical Slovene **goo300k** contains 1,100 pages (about 300,000 tokens), which were page-sampled from the IMP corpus [6]. The transcriptions of this corpus have been additionally corrected and the linguistic annotations hand-verified. The manual annotation focused on word modernisation and lemmatisation, rather than on morphosyntactic tagging, which is, due to the large number of tags, quite labour intensive and difficult to perform without errors. We therefore developed a simplified MSDs scheme, defined in the IMP morphosyntactic specifications, a subset of the JOS specifications. The IMP MSDs do not code the inflectional properties of words, such as case or person, and they also simplify their lexical features, so that instead of the almost 2,000 JOS MSDs the IMP specifications define only 32.

The goo300k corpus is available both via the concordancers and for download, and is meant for developing HLT tools for processing historical Slovene, e.g. for full-text search in cultural heritage libraries.

2.3 Specialised Corpora

The specialised monolingual corpora of Slovene are focused on a particular sub-language or text type and were compiled for the purposes of a particular investigation, e.g. as the basis for a (terminological) dictionary or as a teaching aid or are simply opportunistic: if interesting language data was available, we compiled it into a corpus in the hope that they will be, eventually, useful.

The **VAYNA** corpus is, in terms of when it was developed, the oldest corpus on the server, and one of the oldest computer corpora of Slovene. It was compiled in the late '80 to (dis)prove the thesis, that the Slovene media were attacking the Yugoslav National Army [30]. It is, by today's standards, rather small with 220 thousand words of relevant texts from periodicals of that period, just before the secession of Slovenia from Yugoslavia, and offers an interesting glimpse into this important moment of Slovenian history.

The **DSI** corpus [24] is meant primarily as support for the development of the on-line iSlovar, `www.islovar.org`, a terminological dictionary of informatics, and contains the proceedings of the annual Slovene Conference on Applied Informatics, as well as selected issues of the Applied Informatics journal. DSI could be considered a monitor corpus, as new materials are added each year – it currently contains nine volumes of the proceedings (2003–2012), together with the journal, over 1,300 papers or more than 3 million words.

The **SDJT** corpus [27] is much smaller (183 papers, 280 thousand words) and contains the proceedings (1998–2010) of the biennial Slovene Conference on Language Technologies, making it a good resource for studying the terminology of our field.

The **KoRP** corpus was developed at the Faculty for Social Sciences at Ljubljana University, in the scope of a doctoral dissertation [19] and contains around 2 million words of texts centred on field of public relations.

Two corpora were made in the scope of MSc studies, **KONJI** [23] and **FILMI** [17], and served as the basis for terminological dictionaries from the fields of equestrianism (horse riding, care, breeding, etc.) and film reviews respectively. The former contains 400 thousand words, and the latter almost 800 thousand words.

Given the current interest in social networks, maybe the most interesting specialised corpus is **Tweet-sl**, which contains 360 thousand Slovene tweets or 5 million words from the period 2007–2011 [10]. It should be noted that, as the others, this corpus has been automatically MSD tagged and lemmatised, but the quality of the annotations is currently much worse than for standard Slovene because tweets contain many non-standard words. Normalising them, along the lines of what was done for historical Slovene, remains further work.

Finally, the server also offers the **slWaC** corpus of web pages, which currently contains about 500 million words (almost 2 million HTML documents), which were crawled from the Slovene (`.si`) web, cleaned in a number of steps (boilerplate removal, deduplicaiton, language and character set detection, etc.) and then annotated with ToTaLe [10].

2.4 Parallel Corpora

Parallel corpora contain texts together with their translation(s), and are typically sentence-aligned. All the corpora on the server have Slovene as one of the languages, with hand-validated sentece alignments to the other language. Aligned parallel corpora are very useful for translators and are also a core resource for training and testing machine translation systems. Making them, however, is much more difficult than monolingual ones, so they are typically much smaller.

The linguistic annotation for the foreign language of most corpora listed below was done with TreeTagger [26]. The TreeTagger tagsets, however, differ considerably between the languages, so we developed a mapping from them to a harmonised tagset,

called SPOOK, which, in addition to Slovene, also cover English, French, German and Italian [8].

The English-Slovene corpus **TRANS5** contains 1.3 million words (in this section we give the sizes for the Slovene part) and is composed of various, quite varied, sources, from a book on the Linux operating system, to the Slovene constitution. In its compilation we were not guided by any linguistic criteria, rather, the goal was to make a Slovene-English parallel corpus which would be as large and varied as possible. The corpus is available via the concordancer, and, upon request, for download.

The multi-lingual **SPOOK** corpus [32] was developed in the scope of the eponymous project, with the goal of enabling translation-oriented research. It consists of 35 novels in English, French, German and Italian and their translations into Slovene, and, additionally, of 25 Slovene novels (without translations), together containing 4 million words. Because of copyright issues, this is one of a few corpora offered, which is not publicly available.

Connected with the SPOOK project is the **LeMonde** corpus [22], containing 300 articles (just over half a million words) from the years 2006–2009 of the LeMonde weekly in the original French and its translation into Slovene, which was published in the Slovene daily Delo; this corpus is also freely available.

The corpus **EU-DGT** contains the translation memory which was compiled on the process of translating the European legal documents and contains 29 million words. The freely available source corpus JRC DGT [29] contains 22 languages but we included in our corpus only the five SPOOK languages, as we had for these already the developed annotation methodology.

The Japanese-Slovene **jaSlo** corpus [15] is used as support for the jaSlo learner's dictionary (`nl.ijs.si/jaslo`) and contains novels, web texts, lecture handouts etc., just over half a million words from 132 sources. The Japanese texts were tagged and lemmatised with the program Chasen [21], and the Japanese tags translated to equivalent English codes.

2.5 Foreign Language Corpora

Although the purpose of the server is to offer Slovene language resources, we have also compiled several corpora of other languages, mostly to support translation studies to and from Slovene, by offering large monolingual corpora in addition to the parallel ones.

The largest group is that of Web corpora, i.e. corpora which were made from web pages, similar to slWaC. The **jpWaC-L** corpus is large corpus (300 million words, 50 thousand pages) automatically collected from the Web [28]. It was tagged and lemmatised with Chasen and, additionally, each word in the corpus was annotated by its difficulty level [14], making it a useful corpus for learning Japanese as a foreign language. Using the same principles as for slWaC, we have also compiled **hrWaC** [18], a web corpus of Croatian, which currently contains about 800 thousand words or over 2 million web pages from the .hr domain. Finally, we have taken Web corpora of French, Italian and German (**frWaC, itWaC, deWaC**, all over a billion words) collected and made available by the WaCkY initiative [2] (`wacky.sslmit.unibo.it`), which we have cleaned, tagged and lemmatised using the SPOOK methodology.

The final corpus is **ELIZA**, is a very specialised but nevertheless quite large collection, which contains almost 6 million (English) responses or over 22 million words (2002–2007), to the well known ELIZA chatbot program [33], as available on `www-ai.ijs.si/eliza`. Access to this corpus requires authentication, but the password is available on request.

3 Technical Requirements: Standards and Sustainability

Processing of the large collection of corpora described above requires that these language resources are uniformly encoded. A common and well-document encoding is even more important where language resources (on the nl.ijs.si server including machine readable dictionaries and digital libraries), are made downloadable, as the users must be able to understand and process the encoded files. In this section we sketch the encoding standards used for most of the resources on the server, and also mention other aspects of open and permanent access to language resources.

3.1 Text Encoding

The presented corpora are varied in terms of encoding they contain, which spans diverse meta-data on the corpus texts, various structural elements (e.g. utterances, verses, etc.), to linguistic elements such as sentences and named entities, with further linguistic annotation including word-level features (e.g. MSDs), and sentence-level syntactic annotation.

For the common encoding infrastructure we use the Text Encoding Initiative Guidelines (`www.tei-c.org`) , an open set of recommendations for encoding various type of texts, including annotated computer corpora. The TEI defines several hundred elements for various text types and types of analysis, and makes it possible to generate XML schemas to use in particular projects. Such a schema makes is possible to formally validate a particular TEI document, e.g. a corpus, while the Guidelines serve to document the meanings of the elements used.

While the Guidelines are the most comprehensive set of text encoding recommendations, it should be noted that in the context of natural language processing, they have the drawback of being more geared towards humanities, rather than to linguistic or HLT research: they are not prescriptive enough, and often one analysis can be encoded in a multitude of ways. For HLT, there has been recently much work in defining encoding standards in the scope of ISO (in particular, ISO TC 37, Technical committee for terminology and other language resources), but quite a few of them are still under development, and lack good practices, extensive documentation, software support, which are all attributes of the TEI.

3.2 The Linguist's Workbench

While it is important to be able to offer corpora and other language resources for download, linguists, lexicographers, students and teacher will most often not be interested in processing the files themselves. Rather, they expect a concordancer with which to be able to explore the language(s) and this has been in fact the usual situation with most (say, national reference) corpora, which are available only through a (often custom built) concordancer.

The corpora on the nl.ijs.si server are available via two concordancers, noSketchEngine [25] and CUWI [8]. The noSketchEngine is the open source versions of the well-known and powerful (but commercial) SketchEngine. The noSketchEngine does not offer all the functionality of SketchEngine, but does support all the standard functions for searching and displaying concordances, frequency lexicons, keywords and collocations, use and display of structural and positional (word-level linguistic) attributes, saving the results locally, etc.

CUWI is our own Web front-end, which uses the well-known open source CWB back-end [3]. The CUWI interface is less polished than the noSketchEngine one and is geared more towards corpus development and debugging than towards linguistic use.

3.3 Towards a Language Resource Infrastructure

Given the large amount of work invested in the compilation of the described corpora and in making them interchangeable and available, it is also sensible to make them proof against loss. While the server is regularly backed-up, it nevertheless represents the single point of failure for most of the corpora stored on it. This, however, does not apply to the corpora of the SSJ project, i.e. (cc)Gigafida, (cc)KRES, ssj500k, and Gos, which are available from their home pages on the SSJ project site.

Indeed, such redundancy in access points also seems to be the best way of ensuring robust and wide dissemination of language resources. While we have not implemented this approach yet, we have plans to make all the downloadable corpora also available on public repositories. There are European initiatives that aim to provide such services for language resources, in particular CLARIN (`www.clarin.eu`), to serve the needs of humanities researchers, and META-SHARE (`www.meta-share.eu`) for sharing HLT datasets. But while these centres are being set up, there are quite usable alternatives, e.g. Wikisource (`wikisource.org`) and esp. Github (`github.com`), the public GIT repository. GIT is a very popular version control and source code management system, which, however, can also be used for textual data. The advantage of using GIT is primarily in that it allows effectively storing all version of a resource, which enables replication (and extension) of experiments performed on a particular data set, making it easier to evaluate and compare developed HLT methods.

If storing the language resource files (TEI P5, and possibly various derivatives) on repositories helps in protecting the data and making it better available for HLT, a similar approach can also be envisioned for corpus linguistics. Now, many national centres (e.g. national academies) offer concordancers which the local linguists are familiar with, but these concordancers work over just one or, at best, a small number of corpora. By converting the TEI into the format required by such concordancers it should not be too difficult to mount other, say Slovene, corpora, on these concordancers, offering those interested in foreign languages access to them in an environment that they are familiar with – and, again, it makes the data better insured against loss.

4 Legal Issues

While the preceding section has discussed some technical considerations in making (Slovene) corpora available, there is another obstacle to wider availability of language

resources, namely copyright and protection of personal data. In this section we briefly sketch these problems and our solutions, where we first list the types of access that are, in general, available for language resources.

4.1 Types of Accessibility

We have already made the distinction between allowing access via a concordancer and allowing downloading of the full resource – the main difference between the two is, of course, that via a concordancer it is, in general, possible to obtain only small fragments of the resource, and, depending on the concordancer, maybe only in HTML format, set for the screen. This is the preferred mode of allowing access to corpora, as many institutions either do not wish for the complete resource (say, corpus) to be made available to either protect their investment (and monopoly) or, more commonly, they do not have the rights over the original texts that would enable them to disseminate them further in their entirety. Some corpus providers lock the corpora further by allowing access only to registered – this has the (somewhat doubtful) advantage of allowing use only to researcher that have agreed to certain conditions, but has the distinct disadvantage of driving away casual visitors, who are the "long tail" of the web users.

The possibility of downloading a complete corpus is offered quite rarely, esp. for reference or manually annotated corpora, a practice we are trying to change, at least in Slovenia. While many corpora on nl.ijs.si cannot be made openly downloadable, those that can, are, and they cover a large spectrum of sizes and text types, e.g. slWaC (web, 500 million words), ccGigafida (reference, 100 million), IMP (historical, 15 million), Gos (speech transcriptions, 1 million), and the two manually annotated ssj500k and goo300k.

These corpora are available under one of the Creative Commons licences, which gives the right to redistribute the resources, provided certain conditions are met, dependent on the type of the CC licence. None of the downloadable corpora specify No Derivatives (CC-ND), i.e. it is permitted to make derived versions of the corpus, say correcting tagging errors, and re-distributing that. Some resources (e.g. those of the SSJ project) specify Non-Commercial use only (CC-NC), but not all do (e.g. the IMP corpora). And all corpora require Attribution (CC-BY) i.e. that the use of the resource is suitably acknowledged, which, in scientific publications, means citing one of the papers describing the corpus. We argue that such acknowledgements are, for academic circles, the appropriate payment for the work that was invested in making them, but this courtesy should be, by the researchers using the corpora, observed strictly.

4.2 Copyright

With corpora, copyright over the source texts has always presented the greatest barrier towards their free distribution. With reference, usually national corpora, this problem was typically dealt with by obtaining signed agreements by all the text providers (authors or publishing houses) that allow the use of the texts in the corpus, under certain conditions – usually, that only portions of their texts will be made available, and only for non-commercial purposes. This then allows for making the corpus available for searching via a concordancer, possibly requesting more or less rigorous sign-in.

The procedure of having signed agreements was followed for (copyrighted) texts in the reference Gigafida corpus, and no log-in is needed for its use. As the agreement with the text providers stipulates that up to 10% of each text can be made publicly available,

this also enabled us to produce freely (CC-NC-BY) downloadable corpora (ccGigafida, ccKRES, ssj500k) which have been paragraph-sampled from Gigafida, taking care that at most one tenth of a text is present in these derived corpora.

Another case arises with the IMP corpus of historical Slovene, where the situation is different, as the texts are out of copyright – although, even this might not be the case for all the texts, as the most recent ones in IMP are from 1918, and Slovene law states that a text is subject to copyright until 70 years have passed since the death of the author. However, copyright over the texts themselves is not the only barrier to further distribution – for historical texts, institutions that have digitised the texts can (and do) claim copyright over the scans (facsimiles) and, possibly, the transcriptions. For IMP, all the providers agreed that both the facsimile and text can be made freely available – with most we simply came to a verbal agreement, except for the National and University Library, where a written agreement was signed.

In fact, a verbal agreement is also all that we have for most of the other corpora containing previously printed text, as written agreements are much more difficult to obtain. However, being overly cautious is probably counter-productive: we've yet to encounter a case, where the copyright holder would object to their work being made available via a concordancer.

A special case are corpora containing materials published digitally – in particular, the WaC corpora and the Twitter corpus. Here it would be impossible to ask permission from every author, so we take the view that we make these texts available via the concordancer, but authors are free to request removal of their texts from the corpora, which has, in fact not yet happened either. Download of such corpora could be more problematic, even though the WaCky web site does offer downloadable corpora, but does require potential users to first specify what the corpora will be used for.

4.3 Personal Data Protection

Apart from copyright, the other barrier to distributing corpus data is that of personal data protection. This aspect is well known from speech corpora, esp. those containing non-public conversations, and even more so if they involve minors. Indeed, to comply with such requirements, signed agreements were collected for the Gos speech corpus from all the participants. Furthermore all the personal names appearing in the discourses were anonymised – in the transcription they are substituted by codes, while the recording covers them with a beep; even more, in private discourse, the frequency of the recording was changed, to prevent the recognition of the speaker from their voice [31], and, as mentioned, the speech files are not available for download. In our opinion, the protection of personal data has here been taken to an extreme, as even public discourses (say radio news) have been anonymised.

With previously published text data the issue personal data protection is less known, nevertheless, at least in Slovenia, it turns out to be rather severe as well. In 2012 the Information Commissioner of the Republic of Slovenia reacted to complaint in connection with access to the "Nova Beseda" corpus [16] (`bos.zrc-sazu.si`) at the Scientific Research Centre at the Slovene Academy of Sciences and Arts. The complaint was connected to the "Right to forget", i.e. that sins from the past should not haunt one their whole life. Namely, Nova Beseda contains newspaper articles going back to the '90, to the time before the internet, and the concordancer was, furthermore, open for indexing

by robots. So, for the person in question, the first Google hit on their name gave the concordances from the twenty-year old crime news section of a major Slovene daily. In reaction to this complaint, the commissioner ruled that searching for personal names in the corpus be blocked. The Centre complied with this ruling by disallowing searchers of the form "possible name + possible surname", which raised a storm of protest among linguists, as this also prevents searches for people long dead and of public importance, e.g. artists, politicians, or for combinations where one or both of the words can be a name but can also be a common noun. Nevertheless, the ruling still stands, but, to the relief of other corpus providers in Slovenia, it applies only to one corpus mounted on one server.

Finally, there is one corpus that deserves here a special mention, namely the ELIZA corpus, which, as mentioned, contains logs of conversions with a Web service mimicking a (quite stupid) psychiatrist. As the web page does not display any warnings that the conversations could be further published and because the conversations might contain personal data, access to the corpus is password protected; however, on application, researchers with a legitimate reason can receive the password.

5 Conclusions

In this paper we have presented the corpora available from the nl.ijs.si server and discussed several issues to do with their availability: technical ones of encoding, viewing and downloading, and legal ones of copyright and protection of personal data. As the paper shows, we advocate the use of open standards and recommendations, in particular the Text Encoding Initiative Guidelines, use of open source software solutions, as are TreeTagger, and the CUWI and noSketchEngine concordancers, and free availability of produced corpora. Such language resources are costly and time consuming to make, and if, as is often the case, they have been produced with public money, then the corpus compilers have, in our view, an obligation to maximise their impact by further disseminating them as freely as possible. While they can be limited in this by legal issues, we take here a somewhat relaxed position: as long as a text has been already published, even more so if this has been done on the internet, then, given that we don't run a commercial enterprise, but offer the corpora for free, it is fair use of the materials to make them further available, at least via the concordancers, and, in more clear cut cases, for download as well. However, if an author or, in general, copyright holder requests the removal of their text from a corpus, we will respect their request. As regards further work, we will continue work on corpora, striving towards larger, better annotated and (redundantly) available corpora, and other language resources for Slovene.

References

[1] Arhar, Š. (2009). Učni korpus SSJ in leksikon besednih oblik za slovenščino. (The SSJ training corpus and word-form lexicon for Slovene). *Jezik in slovstvo*, 54(3–4):43–56.

[2] Baroni, M., Bernardini, S., Ferraresi, A., and Zanchetta, E. (2009). The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation*, 43(3):209–226.

[3] Christ, O. (1994). A modular and flexible architecture for an integrated corpus query system. In *Proceedings of the Conference in Computational Lexicography, COMPLEX'94*, pages 23–32, Hungarian Academy of Sciences, Budapest.

[4] Erjavec, T. (2011). Automatic linguistic annotation of historical language: ToTrTaLe and XIX century Slovene. In *Proceedings of the 5th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, LaTeCH 2011*, pages 33–38, Association for Computational Linguistics, Portland.

[5] Erjavec, T. (2012). MULTEXT-East: morphosyntactic resources for Central and Eastern European languages. *Language resources and evaluation*, 46(1):131–142.

[6] Erjavec, T. (2012). The goo300k corpus of historical Slovene. In *8th International Conference on Language Resources and Evaluation, LREC 2012: proceedings,* pages 2257–2260, ELRA.

[7] Erjavec, T. (2012). Jezikovni viri starejše slovenščine IMP: zbirka besedil, korpus, slovar. (The IMP language resources of historical Slovene: text collection, corpus, lexicon). In *Proceedings of the Eighth Conference on Language Technologies*, pages 52–56, Jožef Stefan Institute, Ljubljana.

[8] Erjavec, T. (2013). Vzporedni korpus SPOOK: označevanje, zapis in iskanje. (The SPOOK parallel corpus: annotation, encoding and searching). In Vintar, Š., editor, *Slovenski prevodi skozi korpusno prizmo*, pages 14–31, Znanstvena založba Filozofske fakultete Univerze v Ljubljani.

[9] Erjavec, T. (2013). Korpusi in konkordančniki na strežniku nl.ijs.si. (Corpora and concordancers on the nl.ijs.si server). *Slovenščina 2.0*, 1/1:24–49.

[10] Erjavec, T. and Fišer, D. (2013). Jezik slovenskih tvitov: korpusna raziskava. (The language of Slovene tweets: a corpus based study). In *The 32nd Symposium "Obdobja"*, Znanstvena založba Filozofske fakultete, Ljubljana. In press.

[11] Erjavec, T., Ignat, C., Pouliquen, B., and Steinberger, R. (2005). Massive Multi-Lingual Corpus Compilation: Acquis Communautaire and ToTaLe. In *Proceedings of the 2nd Language & Technology Conference,* pages 32–36, Poznan, Poland.

[12] Erjavec, T. and Krek, S. (2008). Oblikoskladenjska priporočila in označeni korpusi JOS. (The JOS morphosyntactic specifications and annotated corpora). In *Proceedings of the Sixth Conference on Language Technologies*, pages 49–53, Jožef Stefan Institute, Ljubljana.

[13] Grčar, M., Krek, S., and Dobrovoljc, K. (2012). Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik (Obeliks: a statistical morphosyntactic tagger and lemmatiser for Slovene). In *Proceedings of the Eighth Conference on Language Technologies*, pages 89–94, Jožef Stefan Institute, Ljubljana.

[14] Hmeljak Sangawa, K., Erjavec, T., and Kawamura, Y. (2010). Automated collection of Japanese word usage examples from a parallel and a monolingual corpus. In *Proceedings of eLex »eLexicography in the 21st century: new challenges, new applications«,* pages 137–147, Presses Universitaires de Louvain, Louvain.

[15] Hmeljak Sangawa, K. and Erjavec, T. (2008). A low cost approach to building a Japanese-Slovene parallel corpus. *Denshi Jōhō Tsūshin Gakkai gijutsu kenkyū hōkoku*, 108:7–10.

[16] Jakopin, P. and Michelizza, P. (2007). Besedilni korpus Nova beseda (The Nova beseda text corpus). *Mostovi*, 41(1-2):165–176.

[17] Košir, M. (2010). Slovenska filmska terminologija v korpusu filmskih kritik. (Slovene film terminology in the corpus of film reviews). Master's Thesis. University of Nova Gorica.

[18] Ljubešić, N. and Erjavec, T. (2011). hrWac and slWac: compiling web corpora for Croatian and Slovene. *Lecture notes in computer science*, 9743:395–402, Springer.

[19] Logar Berginc, N. (2007). Korpusni pristop k pridobivanju in predstavitvi jezikovnih podatkov v terminoloških slovarjih in terminoloških podatkovnih zbirkah. (A corpus-based approach to acquiring and presenting language data in terminological dictionaries and terminological databases). Doctoral dissertation, University of Ljubljana.

[20] Logar Berginc, N., Grčar, M., Brakus, M., Erjavec, T., Arhar Holdt, Š., and Krek, S. (2012). Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba. (Corpora of Slovene: Gigafida, KRES, ccGigafida and ccKRES: compilation, content and use). Trojina, Ljubljana.

[21] Matsumoto, Y. (2000). Japanese Morphological Analysis System ChaSen. *JPSJ Magazine*, 41:1208–1214.

[22] Mezeg, A. (2011). *Korpusno podprta analiza francoskih polstavkov in njihovih prevedkov v slovenščini*. PhD dissertation, University of Ljubljana.

[23] Plahuta, H. (2010). Korpusne metode v jezikoslovju pri izdelavi osnutka konjeniškega terminološkega slovarja. (Corpus methods in linguistics in compiling a draft of an equestrianism terminological dictionary). Master's Thesis, University of Nova Gorica.

[24] Puc, K. and Erjavec, T. (2006). Uporaba korpusa pri urejanju spletnega terminološkega slovarja. (Using a corpus in editing a web-based terminological dictionary). In *Proceedings of the Fifth Conference on Language Technologies*, pages 156–161, Jožef Stefan Institute, Ljubljana.

[25] Rychlý, P. (2007). Manatee/Bonito – A Modular Corpus Manager. In *1st Workshop on Recent Advances in Slavonic Natural Language Processing,* pages 65–70, Masaryk University, Brno.

[26] Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing*, Manchester.

[27] Smailović, J. and Pollak, S. (2011). Semi-automated construction of a topic ontology from research papers in the domain of language technologies. In *5th Language & Technology Conference, LTC'11,* pages 121–125*,* Poznań, Poland.

[28] Srdanović, I., Erjavec, T., and Kilgarriff, A. (2008). A web corpus and word sketches for Japanese. *Shizen gengo shori*, 15(2):137–159.

[29] Steinberger, R., Eisele, A., Klocek, S. Pilos, S., and Schlüter, P. (2012). DGT-TM: A Freely Available Translation Memory in 22 Languages. In *Proceedings of the 8th international conference on Language Resources and Evaluation (LREC'2012)*, pages 454–459.

[30] Tancig, P. and Žagar, I. (1989). Računalniško podprta analiza velikih tekstualnih baz podatkov: Primer napadov na JNA. (A computer supported analysis of large textual databases: the case of the attacks on the Yugoslav national army) In *Proceedings of the Fifth conference of the Network of Societies of Applied Linguistics of Yugoslavia*, pages 51–56, Ljubljana.

[31] Verdonik, D. and Zwitter Vitez, A. (2011). *Slovenski govorni korpus Gos.* (The Slovene speech corpus Gos). Trojina, Ljubljana.

[32] Vintar, Š., editor, (2013): *Slovenski prevodi skozi korpusno prizmo*. (Slovene translations through a corpus prism). Znanstvena založba Filozofske fakultete Univerze v Ljubljani.

[33] Weizenbaum, J. (1966). ELIZA – A Computer Program For the Study of Natural Language Communication Between Man and Machine. *Communications of the ACM* 9:36–45.

Obstacles and Solution to Recognizing Compound Nouns in Greek: A Corpus Study

Vasiliki Foufi, Kyriaki Ioannidou, and Olympia Tsaknaki

School of French, Aristotle University of Thessaloniki, Greece

Abstract. This paper, part of a multi-faceted research on compound nouns carried out in the Laboratory of Translation and Language Processing, School of French, Aristotle University of Thessaloniki[1], deals with the recognition of compound nouns of the structure Adjective + Noun often characterized by the existence of external inserting materials among their constituents. To locate and eliminate these barriers and then recognize the compound nouns, we use large-scale corpora and the method of finite-state transducers. We finally calculate precision and recall rates to verify the efficiency of our method.

1 Introduction

Multiword expressions cover a wide range of linguistic constructions: idioms, multiword compound nouns, compound verbs, frozen or semi-frozen expressions, proper nouns, time expressions, etc. In this article, we focus our research on Greek compound nouns of the structure Adjective + Noun [1], [4].

Following Gaston Gross' criteria [6], we define as 'compound noun' a multiword noun whose at least one of the semantic, syntactic or distributional properties cannot be deduced by the properties of each of its constituents.

Texts abound with compound nouns, which reveals the need for their recognition by the Natural Language Processing systems in order to improve their applications (summarization, automatic translation, term extraction, etc.). Due to their non-compositional meaning, compound nouns have to be grouped together and lemmatized in electronic dictionaries. For instance, in the following sentence, there are two compound nouns *καταναλωτικό δάνειο/consumer loan* and *πιστωτική κάρτα/credit card* which will be recognized provided that they are listed in a dictionary:

> *Η πλειονότητα του λαού πήρε καταναλωτικό δάνειο ή πιστωτική κάρτα το 2003*
> (The majority of the population took a consumer loan or a credit card in 2003)

However, the automatic recognition of Adjective + Noun multiword compound nouns in Greek texts can be hindered to a large extent because of the presence of various linguistic or extra-linguistic elements among their constituents. Those inserting materials may be of different kinds and will be subsequently presented. For instance:

> *Εκατομμύρια άστρα περιβάλλουν το ηλιακό* ***μας*** *σύστημα*
> (Millions of stars surround **our** solar system [solar **our** system])

[1] http://www.frl.auth.gr/index.php/gr/structure-gr/laboratories-gr/laboratory-translation-language-gr

In the Greek example, we can observe that the possessive determiner *μας/our* comes between the adjective *ηλιακό/solar* and the noun *σύστημα/system*.

It is also important to point out that the adjectives which form compound nouns - usually called pseudoadjectives [3], [1], [11] can be linked to other pseudoadjectives susceptible to form a compound noun of the same structure. It must be noted that these adjectives cannot be descriptive, e.g.:

γυναικολογική ***και*** *μαιευτική κλινική*
(gynecology **and** maternity clinic)

In this example, neither of the two adjectives can be replaced by a descriptive one:

γυναικολογική* ***και* *καθαρή κλινική*
*(gynecology **and** clean clinic)

καθαρή* ***και* *μαευτική κλινική*
*(clean **and** maternity clinic)

To proceed to our research, we used large-scale lexical resources and the method of finite-state transducers (FSTs).

2 Description of the Corpus

Our corpus of approximately 6,000,000 words consists of Journalistic and Educational Discourse. It was built in the frame of the research project "Portal for the Greek Language" in cooperation with the Centre for the Greek Language and the Laboratory of Translation and Language Processing (2005–2006) and it is available on the website `http://www.greek-language.gr/greekLang/modern_greek/tools/corpora/index.html`.

The text corpora that originate from the field of journalism contain, in electronic format, about 5,000,000 words published in the newspapers *Makedonia* (3,000,000 words) and *Ta Nea* (2,000,000 words). The material is grouped into thematic units and classified by genre (short news, social reporting, etc.). Precisely, the corpus from the journal *Makedonia* contains 143,500 simple structures and 14,808 compound structures (compound nouns, phraseology etc.). The corpus from the journal *Ta Nea* contains 138,056 simple structures as well as 14,902 compound structures.

The text corpora from the field of educational writing are classified by genre (narrative, description, instructions, process analysis and argumentation). They contain 2,000,000 words from which 120,271 are simple structures and 9,971 are compound structures.

A representative extract of our journalistic corpus is cited below:

BeginArticle
Σύντομη είδηση
ΚΟΣΜΟΣ
{S}ΣΕΡΙ ΜΠΛΕΡ
{S}Θέλουν να την απελάσουν!
{S}Η Σερί.{S} Στην πρώτη θέση με 31%

{S}Η σύζυγος του Βρετανού πρωθυπουργού Τόνι Μπλερ είναι το πρόσωπο που,

περισσότερο από όλα τα άλλα, θα ήθελαν οι Βρετανοί να "απελαθεί" από τη χώρα.

{S}Αυτό προκύπτει από δημοσκόπηση του BBC σε 15.000 Βρετανούς.{S} Στη λίστα των απελάσεων, η Σερί Μπλερ ήλθε στην πρώτη θέση με 31% των ψήφων, με δεύτερο τον σεΐχη Αμπού Χάμζα αλ-Μάσρι, που έχει κατηγορηθεί για σχέσεις με τον Οσάμα μπιν Λάντεν.{S} Η δημοτικότητα της συζύγου του πρωθυπουργού μειώθηκε δραστικά μετά τις αποκαλύψεις για σχέσεις με έναν καταδικασμένο απατεώνα.

3 Categorization of the Inserting Materials

The inserting materials can belong to classes such as determiners, conjunctions and adverbs. Additionally, punctuation marks, in particular the comma, can disturb the recognition of the sequence.

3.1 Determiners

In many cases, possessive determiners are inserted between the adjective and the noun of the compound structure. Some representative examples are cited below:

> *Είναι υπεύθυνοι του γενετικού **της** προφίλ*
> (They are responsible for **her** genetic profile [genetic **her** profile])
>
> *Τα στοιχεία της πιστωτικής **του** κάρτας βρέθηκαν χάρη σε έρευνες του FBI*
> (The data of **his** credit card [credit **his** card] were found due to FBI's investigations)
>
> *Οι υποψήφιοι πρωθυπουργοί προετοιμάζονται για την προεκλογική **τους** εκστρατεία*
> (The candidate prime ministers are being prepared for **their** electoral campaign [electoral **their** campaign])

Apart from possessive determiners, we could also find demonstrative determiners between the two constituents:

> *Είναι μέλη των επιχειρηματικών **αυτών** οικογενειών*
> (They are members of **these** business families [business **these** families])
>
> *Έπρεπε να επιλέξουν τις αγροτικές **εκείνες** περιοχές όπου θα καλλιεργούσαν σιτηρά*
> (They had to choose **those** agricultural areas [agricultural **those** areas] where they would cultivate cereals)

3.2 Conjunctions

Conjunctions such as *και/and* are frequently inserted among the constituents of compound nouns. In the following example, there are two compound nouns (*δημόσιου τομέα/public sector, ιδιωτικού τομέα/private sector*) connected by *και/and* but the noun accompanies only the second adjective:

> *Ένας νόμος σχετικά με τους κανόνες της συνύπαρξης δημόσιου* ***και*** *ιδιωτικού τομέα*
> (A law about the rules of coexistence of the public **and** private sector)

The conjunction *και/and* can be followed by the negative particles *όχι/not*, as it is shown in the following examples:

> *Θα λειτουργούν με ηλεκτρομηχανικούς* ***και όχι*** *υδραυλικούς μηχανισμούς*
> (They will run on electromechanical **and not** hydraulic machinery)

Even though in most cases the conjunction *και/and* links the first constituents of two compound nouns of the structure Adjective + Noun, it may also be inserted between the first and the second constituent of one compound. Particularly, in the following example, between the adjective *ηλεκτρονικές/electronic* and the noun *γνώσεις/ knowledge* of the compound noun *ηλεκτρονικές γνώσεις/electronic knowledge*, the conjunction *και/and* is followed by *όχι μόνο/not only*:

> *Ο πρόεδρος αποχώρησε από το δικαστήριο με ηλεκτρονικές* ***και όχι μόνο*** *γνώσεις*
> (The president left the court with **not only** electronic knowledge [electronic **and not only** knowledge])

Moreover, the use of *και/and* followed by an adverb can be noted between the adjectives of two compound nouns. An example with the adverb *κυρίως/mainly* is shown below:

> *Η οικονομική* ***και κυρίως*** *στρατιωτική υπεροχή της χώρας*
> (The economic **and mainly** military predominance of the country)

Furthermore, *και/and* can be followed by indefinite determiners. In the following example, the determiner *άλλο/other* appears between the two adjectives *φωτογραφικό/ photographic* and *λαογραφικό/folklore*:

> *Μπορείτε να συγκεντρώσετε μαρτυρίες, φωτογραφικό* ***και άλλο*** *λαογραφικό υλικό*
> (You can collect testimonies, photographic **and other** folklore material)

Usually, the first constituent of a compound noun is preceded by an article, e.g. ***ο*** *γαλλικός και διεθνής κινηματογράφος****/the*** *French and international cinema*. In some cases, the article is repeated before the second adjective (first constituent of the

second compound noun). In particular, in the following example, the conjunction *και/and* and the definite article *ο/the* are placed between the two adjectives *γαλλικός/French* and *διεθνής/international*:

> ***Ο*** *γαλλικός* ***και ο*** *διεθνής κινηματογράφος έχασαν ένα σημαντικό σκηνοθέτη*
> (**The** French **and the** international cinema have lost an important director)

Moreover, the adjective of the second compound could be preceded by an article and followed by a possessive determiner, like in the following example:

> *Λόγω* ***της*** *γεωστρατηγικής και* ***της*** *γεωγραφικής* ***της*** *θέσης, η χώρα είναι ευάλωτη σε εξωτερικές πιέσεις*
> (Because of **her** geostrategic and geographical location [**the** geostrategic and **the** geographical **her** location], the country is vulnerable to external pressures)

An additional remark should be added to the case mentioned above. A preposition that figures before the first adjective can also occur before the adjective of the second compound as it can be seen in the example below:

> *Οι αποφάσεις πρέπει να λαμβάνονται* ***σε*** *ευρωπαϊκό και όχι* ***σε*** *εθνικό επίπεδο*
> (The decisions must be taken **at** European and not **at** national level)

One thing we must underline here is that very few examples of compound nouns whose constituents are connected with *κι/and*, a variant of *και/and* used when the following word starts with a vowel, can be found in our corpus. Therefore, we included that case in our FST.

Except for *και/and* and *κι/and*, we could also find other conjunctions like *όμως/but, ή/or, λοιπόν/therefore, είτε... είτε/either... or, ή... ή/either... or, ούτε... ούτε/neither... nor* and others:

> *[...] ανάλογη δημόσια* ***όμως*** *έκκληση είχε κάνει και ο γενικός γραμματέας των Ηνωμένων Εθνών*
> ([...] **but** similar public appeal [similar public **but** appeal] was also made by the Secretary-General of the United Nations)

> *Κάθε διαγνωστικός* ***ή*** *ιατρικός χώρος πρέπει να διατηρεί ενημερωμένο αρχείο*
> (Each diagnostic **or** medical centre must keep an informed record)

> *ενώπιον στρατιωτικών* ***ή*** *δημοτικών* ***ή*** *κοινοτικών αρχών*
> (in the presence of military **or** city **or** municipal authorities)

> *Δεν έχετε* ***ούτε*** *νόμιμο* ***ούτε*** *ηθικό δικαίωμα*
> (You have **neither** legal **nor** moral right)

Moreover, three compound nouns of the type Adjective + Noun, which have in common the noun, can be successively linked with a comma and a conjunction in the same sequence:

> *Πολλοί υποφέρουν από αναπνευστικές, καρδιοαγγειακές **ή** ηπατικές νόσους*
> (Many people suffer from respiratory**,** cardiovascular **or** liver diseases)

> *Οι πολιτικές, επιχειρηματικές **και** κοινωνικές δυνάμεις του νομού κινητοποιήθηκαν*
> (The political**,** business **and** social forces of the prefecture were mobilized)

In the example above, we can remark the presence of three different compound nouns *πολιτικές δυνάμεις/political forces*, *επιχειρηματικές δυνάμεις/business forces* and *κοινωνικές δυνάμεις/social forces* whose first constituents (adjectives) are linked with a comma and the conjunction *και/and* given that their noun is common.

Conjunctions and punctuation marks like the comma can also join up to four and five different adjectives which form compounds with the same noun. The example cited below illustrates this remark:

> *ερευνητικές, επιστημονικές, φιλανθρωπικές **ή** καλλιτεχνικές δραστηριότητες*
> (research**,** scientific**,** charitable **or** artistic activities)

> *εμπορικές, βιομηχανικές, βιοτεχνικές, μεταλλευτικές **και** λατομικές επιχειρήσεις*
> (the commercial**,** industrial**,** craft**,** mining **and** quarrying businesses)

3.3 Adverbs

According to Anastassiadis-Symeonidis [1, p. 153], pseudoadjectives are modified by restrictive adverbs such as: *ιδιαίτερα/particularly, αυστηρά/strictly, αποκλειστικά/ exclusively, κυρίως/mainly, κατεξοχήν/primarily, βασικά/basically, αληθινά/really*, etc. Such remarks have also been highlighted by [9, p. 80], [10, p. 30), who studied the Noun + Adjective compound nouns for the French language:

> *Το σιδηροδρομικό δίκτυο εξυπηρέτησε τη διακίνηση αγροτικών **κυρίως** προϊόντων*
> (The railway network served the transport of **mainly** agricultural products [agricultural **mainly** products])

Furthermore, adverbs of time can be inserted between the adjective and the noun of a compound structure. In the following example the adverb *ακόμη/still* is situated between the adjective *σοβιετική/soviet* and the noun *εποχή/times*:

> *Δύο εκ των οποίων είναι από τη σοβιετική **ακόμη** εποχή*
> (Two of those **still** come from the soviet times [soviet **still** times])

4 Automatic Processing of the Inserting Materials

Our aim was to use the FST method in combination with linguistic resources, that is the electronic morphological dictionaries developed in the Laboratory of Translation and Language Processing. To extract our data, we used the open-source system, *Unitex*[2]. This corpus processing system is based on automata-oriented technology. By means of this tool, electronic resources such as electronic dictionaries and local grammars can be handled. Researchers can work at the levels of morphology, lexicon and syntax.

In particular, in order to eliminate the inserting materials mentioned above and improve the results of the automatic recognition of compound nouns in corpora, we constructed the following FST (Fig.1) which recalls 24 sub-graphs like the one cited below (Fig. 2):

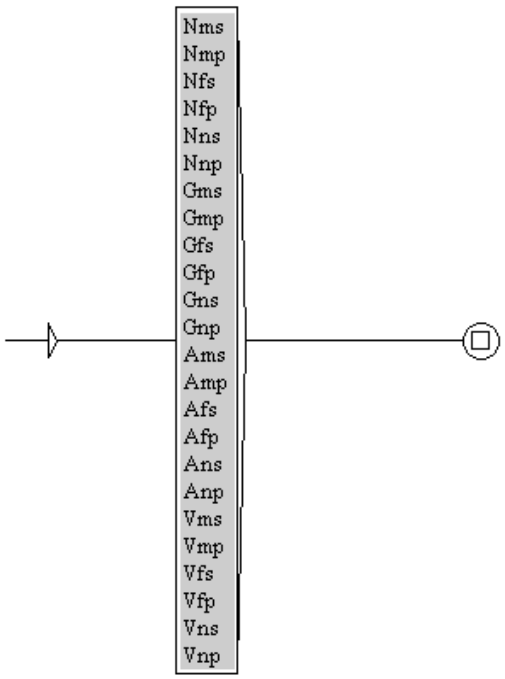


Fig. 1. Main FST

[2] http://www-igm.univ-mlv.fr/~unitex/

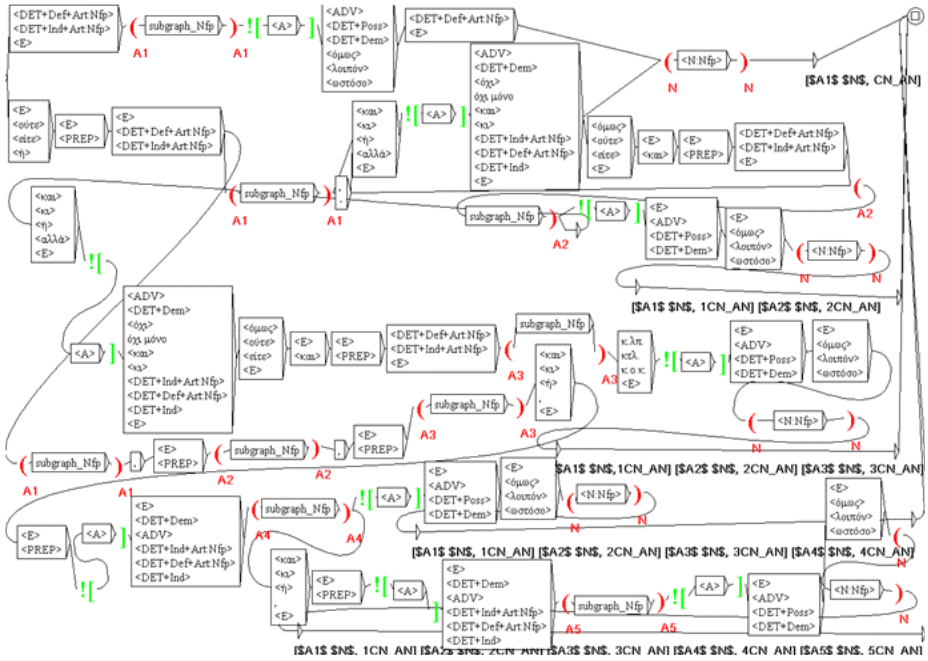


Fig. 2. Sub-graph of feminine compound nouns in the nominative plural

4.1 Description of a Sub-graph

Each sub-graph, like the one presented in Fig. 2, represents the remarks made above. In particular, the above sub-graph presents the possible cases of barriers that can be intruded among the constituents of a feminine compound noun in the nominative plural.

The sub-graph (subgraph_Nfp) that corresponds to adjectives contains all possible pseudoadjectives followed by morphological information about the gender (feminine), number (plural) and case (nominative). As it has been already reported, between the adjective(s) (A1, A2, A3...) and the noun (N) of a compound noun (CN), can be found conjunctions, determiners (DET), adverbs (ADV) and punctuation marks. With regard to conjunctions, it should be noted that we did not include in our FST all the conjunctions listed in the electronic dictionary but only those are related to our research. Additionally, *κ.λπ./etc.* before the noun, like in the example *συγκοινωνιακά, περιβαλλοντικά, αναπτυξιακά κ.λπ. θέματα/transport, environmental, development etc. issues* will be also recognized. Furthermore, in the beginning of each path, we have predicted the possible existence of an article (<DET+Def+Art:Nfp>, <DET+Ind+Art:Nfp>, <E>) or a preposition (<PREP>). The green square brackets before each box containing an adverb (ADV) aim at avoiding morphological ambiguity between an adjective and an adverb.

The defined variables are also used as an output of the FST. For example, the variable $A1$, followed by the variable N, are reused at the output produced ([$A1$ N, CN_AN]). In this FST, there are five outputs which represent the existence of one up to five compound nouns in the same sentence.

4.2 Results

After the creation of the FST and having taken into consideration all possible cases in order to ensure its effectiveness, the next step of our research was the application of the FST to the corpus. An extract of the tagged concordances is presented below:

ταυτοποιήθηκαν δακτυλικά του αποτυπώματα[δακτυλικά αποτυπώματα, CN_AN]

διατηρούν τη βουλευτική τους ιδιότητα[βουλευτική ιδιότητα, CN_AN].

άλλα σημαντικά θέματα κοινωνικής, οικονομικής και εξωτερικής πολιτικής [κοινωνικής πολιτικής,1CN_AN] [οικονομικής πολιτικής, 2CN_AN] [εξωτερικής πολιτικής, 3CN_AN]

ένα μεγάλο παιδικό, πρωινό πρόγραμμα [παιδικό πρόγραμμα,1CN_AN] [πρωινό πρόγραμμα, 2CN_AN],

εντατικά, χειμερινά μαθήματα [εντατικά μαθήματα,1CN_AN] [χειμερινά μαθήματα, 2CN_AN]

τα Χανιά κατά τους θερινούς κυρίως μήνες[θερινούς μήνες, CN_*AN*].

Πλουσιότερος με ηλεκτρονικές και όχι μόνο γνώσεις[ηλεκτρονικές γνώσεις, CN_AN]

Ανάλογη δημόσια όμως έκκληση[δημόσια έκκληση, CN_*AN*] *είχε κάνει σε συγκοινωνιακά, περιβαλλοντικά, αναπτυξιακά κ.λπ. θέματα [συγκοινωνιακά θέματα,1CN_AN] [περιβαλλοντικά θέματα, 2CN_AN] [αναπτυξιακά θέματα, 3CN_AN]*

εξαιρούνται ερευνητικές, επιστημονικές, φιλανθρωπικές ή άλλες κοινωφελείς δραστηριότητες [ερευνητικές δραστηριότητες, 1CN_AN] [επιστημονικές δραστηριότητες, 2CN_AN] [φιλανθρωπικές δραστηριότητες, 3CN_AN] [κοινωφελείς δραστηριότητες, 4CN_AN]

και οι εμπορικές, βιομηχανικές, βιοτεχνικές, μεταλλευτικές και λατομικές επιχειρήσεις [εμπορικές επιχειρήσεις, 1CN_AN] [βιομηχανικές επιχειρήσεις, 2CN_AN] [βιοτεχνικές επιχειρήσεις, 3CN_AN] [μεταλλευτικές επιχειρήσεις, 4CN_AN] [λατομικές επιχειρήσεις, 5CN_AN]

(CN=Compound Noun, 1CN=First Compound Noun, 2CN=Second Compound Noun, 3CN=Third Compound Noun, 4CN=Fourth Compound Noun, 5CN=Fifth Compound Noun, AN=Adjective + Noun)

Given the large size of our corpus, the results of our method have been evaluated upon a representative sample text of 2 170 Kb (157,284 words). Thanks to the FST method, which can guarantee great precision, our precision rate was 94.26% and our recall rate was 100%.

Despite the satisfactory results, the lower precision value obtained is due to ambiguities. Ambiguities in Greek can be found in the morphological, lexical and syntactic level, often raising numerous obstacles. Firstly, as regards ambiguity due to different parts of speech, adjectives and nouns may present the same form: *πολιτική/politics* as a noun and *πολιτική/political* as an adjective (*πολιτική εξουσία/ political power*). Consequently, the structure *της πολιτικής της κυβέρνησης/the government's policy* was incorrectly recognized. Secondly, some pseudoadjectives exist as qualitative adjectives as well. For instance, the adjective *ελεύθερος/free* forms the compound noun *ελεύθερο εμπόριο/free trade.* Therefore, it was integrated in our FST. However, in the free construction *στην ελεύθερη τώρα χώρα/in the free nowadays country,* it operates as a qualitative adjective.

5 Open Issues and Conclusion

The following cases that occur in the corpus did not form part of this particular research. Nevertheless, they are challenges to be tackled in a future research. Firstly, we did not treat cases like *πολιτιστικοί [χώροι] και χώροι καλλιτεχνικών εκδηλώσεων/cultural [spaces] and spaces for artistic events* and *των 50 κεντρικών [θεάτρων], 41 περιφερειακών [θεάτρων] και 38 θεάτρων-καμπαρέ/of the 50 central [theaters], 41 peripheral [theaters] and 38 theater-cabarets* which consist of two different categories of compound nouns: Adjective + Noun (*πολιτιστικοί χώροι/cultural spaces*) followed by Noun + Adjective + Noun in genitive (*χώροι καλλιτεχνικών εκδηλώσεων/spaces for artistic events*) and Adjective + Noun (*κεντρικών θεάτρων/central theaters, περιφερειακών θεάτρων/peripheral theaters*) followed by Noun + Noun (*θεάτρων-καμπαρέ/theater-cabarets*) respectively.

Secondly, we did not take under consideration the phenomenon of nominal subdeletion [5], [2], e.g. *σε όλα τα επίπεδα: φιλοσοφικό, λογοτεχνικό, πολιτικό, ιδεολογικό και κυρίως καλλιτεχνικό/in all levels: philosophical, literary, political, ideological and mainly artistic*.

We also left out very few examples found in our corpus due to the writer's stylistic choices like ellipsis points between the adjective and the noun (e.g. *καλλιτεχνική... δημιουργία/artistic... creation*). Finally, we excluded from our research the example *διαχωρίζουν* ***την*** *επαγγελματική* ***από την καθαρά*** *πνευματική δραστηριότητα/they distinguish* ***the*** *professional* ***from the merely*** *intellectual activity* which presents a single occurrence in the whole corpus.

This research showcases the importance and utility of a corpus approach in order to achieve the recognition of compound nouns. We combined large-scale lexical resources and the FST method to the benefit of the recognition of compound nouns of the structure Adjective + Noun whose structure is discontinued by inserting materials among their constituents. After the application to our corpus of the FST constructed for

this task, it is visible that the results were satisfactory. However, because of the possibility for revision of an FST, the improvement and enrichment with new elements of the FST built needs to be done for other structures of compound nouns.

References

[1] Anastassiadis-Symeonidis, A. (1986). *Neology in Modern Greek* (in Greek). PhD Thesis, Aristotle University of Thessaloniki, Thessaloniki.

[2] Anastassiadis-Symeonidis, A., Efthimiou, A., and Faliatouras, A. (2003). Phenomena of substantivation in Modern Greek (in Greek). In *Proceedings of the 6th International Conference "Intercultural Education-Greek as a second/foreign language",* pages 385–402, University of Patra, Patra.

[3] Bartning, I. (1976). *Remarques sur la syntaxe et la sémantique des pseudo-adjectifs dénominaux en français*. Stockholm.

[4] Foufi, V. (2012). *Morphological, semantic and syntactic description of Greek compound nouns in the form of Adjective+Noun. Application to teaching of Greek as a second/foreign language*. (in Greek) PhD Thesis, Aristotle University of Thessaloniki, Thessaloniki.

[5] Giannakidou, A. and Stavrou, M. (1999). Nominalization and Ellipsis in the Greek DP. *The Linguistic Review*, 16:295–331.

[6] Gross, G. (1996). *Les expressions figées en français: Noms composés et autres locutions*. Ophrys, Paris.

[7] Ioannidou, K., Michailidis, I., Politis, P., and Voyatzi, S. (2005). « Paramétrage du corpus grec numérique par genre textuel: le discours journalistique », 25e colloque international sur le lexique et la grammaire, Liverpool (oral presentation).

[8] Ioannidou, K. (2013). *Noun phrases in Modern Greek: automatic recognition and morphological disambiguation in automatic text processing, some suggestions for possible applications in translation*. (in Greek). PhD Thesis, Aristotle University of Thessaloniki, Thessaloniki.

[9] Monceaux, A. (1992). Un exemple de formation productive de composés de structure Nom Adjectif, *Langue Française* 96:74–87.

[10] Monceaux, A. (1993). *La formation des noms composés de structure Nom Adjectif. Elaboration d'un lexique électronique*. PhD Thesis, Université Paris VII., Paris.

[11] Pedreira, N.-R. (2000). *Adjectifs qualificatifs et adjectifs relationnels: étude sémantique et approche pragmatique*. PhD Thesis, University of Santiago de Compostela, Santiago de Compostela.

From Multilingual Dictionary to Lithuanian WordNet

Radovan Garabík and Indrė Pileckytė

Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences, Bratislava, Slovakia

Abstract. In this paper we describe the motivation for building a small Lithuanian language WordNet out of a bilingual dictionary. The WordNet contains semantic relations for nouns, adjectives, adverbs and verbs, with mapping of synsets to the Princeton WordNet and the Slovak language WordNet. The process of combining various dictionaries to create an initial version and subsequent manual proofreading is described; the first preliminary version of the WordNet has been released.

1 Introduction

For the English language, the Princeton WordNet [7] became de facto a foundation for semantic analysis and annotation, and it inspired WordNets in many other languages. Such projects frequently use the same structure of semantic relations (often augmented by relations specific for the given language).

Considering their interconnections, WordNet projects can be divided into two groups: monolingual and multilingual ones. Monolingual WordNets capture semantic relationship for the given language without any out-of-the language links, while multilingual ones have some way of relating their synsets to other WordNets (most likely the English one).

This relation can form the main design principle of a WordNet, sometimes even to the point of the WordNet being created as a translation of the English one [6]; or the relations are included additionally [9]. There are several different projects that specifically aim to provide multilingual synonym databases, like BalkaNet [10], EuroWordNet [11], or WordNet Grid [8].

The described Lithuanian WordNet database started its life as a (small) multilingual dictionary[1] for students of Slovak as a foreign language.[2] The dictionary used Slovak as the pilot language, and the English part of it has been based on Princeton WordNet v. 3.0. Our database has been therefore created with the semantic relations in mind, and later we specifically extended the contents with the aim of creating bases for Slovak and Lithuanian WordNets.

2 Automatic Synset Generation

The database has been bootstrapped by an automatic synset generation. The method used for the Slovak synsets is described in [3] – in a nutshell, the method is based on translating synsets, hypernyms and hyponyms according to an existing bilingual dictionary and then taking an intersection of various combinations of the translations. The initial database

[1] The term 'dictionary' is perhaps a little ambitious, 'glossary' would be more appropriate.

[2] The dictionary also includes other languages, in particular German and Polish, but since they are not germane to the Lithuanian WordNet, we will not describe them here.

has been filled with Slovak synsets generated by a union of all the four methods (A, B, C, D) described therein. This database has been then manually proofread and extended, with the synsets being mapped to their equivalent English synsets, with the aim to cover (as a minimum) hypernyms for each Slovak synset – thus creating a complete semantic chains up to the top-level categories.

Since we lacked a computer readable English-Lithuanian dictionary, the Lithuanian part of the database has been generated differently – first we obtained a rough Slovak-Lithuanian dictionary based on Slovak-Esperanto and Esperanto-Lithuanian dictionaries provided by the `lernu.net`[3] portal. Using Esperanto as a pivot language had several advantages:

- Word suffixes in Esperanto denote unambiguously part of speech, therefore we obtained highly reliable separation of synsets into nouns, adjectives, adverbs and verbs.
- There is a very low amount of homonymy (although it does exist) in Esperanto [4], which limits the risk of carrying improper semantic chain into a given synset.
- Bilingual Esperanto-Slovak and Esperanto-Lithuanian dictionaries were available and we obtained a copyright agreement allowing us to use them for this purpose.

The dictionary entry consisted of one Esperanto word and its one or several translations. The size of Esperanto-Lithuanian dictionary was 11 529 entries or 16 268 words, Esperanto-Slovak 7 116 entries or 8 130 words. By combining the dictionaries, we obtained a simple Slovak-Lithuanian dictionary of 3 977 entries (one entry corresponds to one Slovak word and its possible Lithuanian translations), or 10 048 Lithuanian words – we can see that there was a substantial ambiguity in the translations.

The dictionary has been then manually proofread and corrected, with the emphasis on keeping 'precision' – i. e. the proofreaders were instructed to predominantly delete incorrect translations, in order to keep down the time needed to complete the task.

This proofread dictionary has been then used to automatically assign Lithuanian synsets to the Slovak ones via a simple substitution of Slovak literals with Lithuanian equivalents.

3 Database Structure

One entry in the database corresponds to one synset. In addition to the synset itself it contains optional definition (not used much), a link to one (or more) English synsets and an optional links to one or several Slovak language synsets. Generally, the relations in the database are $L{:}\,M{:}\,N$, where L is the number of English language synsets, M the number of Slovak language synsets and N that of Lithuanian language synsets – i. e. any number of synsets from any of the languages can be connected to any other number of synsets in the other languages, although in practise the relation is usually split into $L{:}\,M$ where $L = 1 \vee M = 1$; $M{:}\,N$ where $M = 1 \vee N = 1$; $L{:}\,M$ where $L = 1 \vee N = 1$; that is, we try to refrain from introducing complicated and hard to read connections and try to use simple, at most one-to-many relations between two languages. However, most of the entries are simple one-to-one.

[3] `http://lernu.net/`

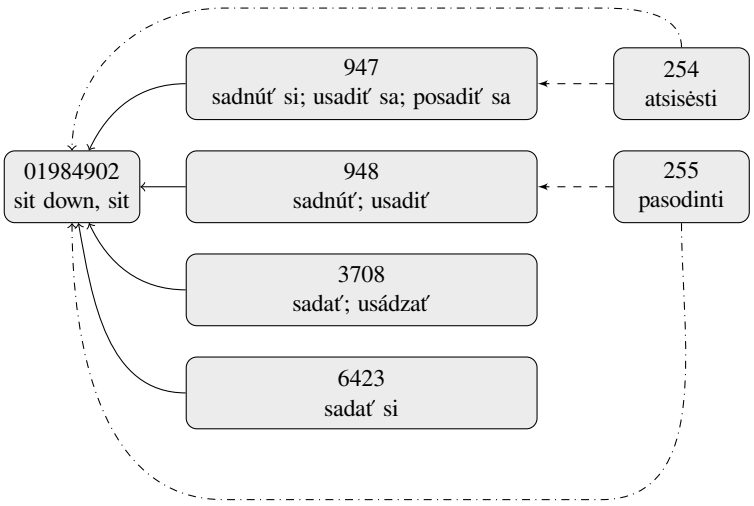


Fig. 1. Example of the interconnection between synsets. The four Slovak synsets correspond to reflexive (intransitive) perfective, transitive perfective, transitive imperfective and reflexive (intransitive) imperfective. Two Lithuanian lexemes for transitive and intransitive are linked to respective Slovak synsets.

3.1 Synset Microformat

The synsets in our database (both Lithuanian and Slovak) are written down using formalized formatting rules in order to ease further automated use and to include additional information (see Figure 2 for the syntax diagram). To put it informally, each synset consists of one or several literals separated by a semicolon; this allows us to include embedded clauses separated by a colon and a relative pronoun (such use is discouraged, but it is necessary to cover those English synsets that do not have direct Slovak or Lithuanian equivalents). Thus the literals can be multiword, simple two-word constructions (adjective+noun) are quite common.

A literal can have an optional annotation character '+' in front of it, this denotes that the literal is semantically 'most important' in the synset, i. e. this is *the* word that is usually used to express the meaning. Another optional annotation is formed by an optional gloss in parentheses, explaining or clarifying the literal in case its inclusion in the synset not obvious to the user, usually in the case of surprising homonymy or a rarely used meaning.

There are also two synset-wide annotations – a minus character and a question mark. Minus character in front of the synset means that the linked Princeton WordNet synset cannot be expressed clearly in the target language (i. e. the semantic meaning is too wide or too narrow, or it covers specific English-culture term that does not have a direct equivalent, or – rarely – there is an outright semantic lacuna in one language). This appears almost exclusively when trying to cover hypernyms of an already existing synset.

A question mark means that the annotator is not sure about the synset – either the synset itself, or its relation to other languages. In theory, this means that we should try to resolve the problems later and the annotation helps to keep the track of such problems.

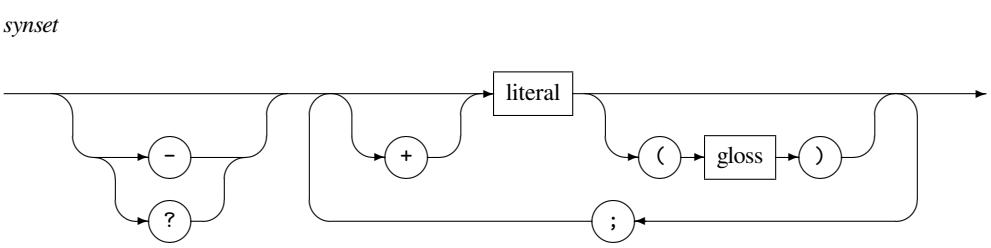


Fig. 2. Syntax diagram of synset definition. Definitions of *literal*, *gloss* and optional whitespace are not included for the sake of brevity

4 Nouns, Adjectives and Adverbs

The mapping for nouns, adjectives and adverbs is often straightforward. Sometimes one English synset is linked to two Slovak or Lithuanian synsets – this often appears when referring to gender distinct nouns that have only the gender neutral form in English (e. g. *10020890 doctor, doc, physician, MD, Dr., medico* is connected to two Lithuanian synsets, *2204 daktaras; gydytojas* (masc.) and *4914 daktarė; gydytoja* (fem.). Since the masculine/feminine gender distinctions in Lithuanian and Slovak are rather compatible, the mapping from Lithuanian synsets to Slovak ones is one-to-one.

Other kind of multiple correspondence is very rare, and although there is a number of homonyms and polysemous words in Lithuanian, we were able to disambiguate them quite clearly based on the English WordNet meanings – sometimes the distinction was even too detailed (e. g. Lithuanian word *veiksmas* appears in 6 different synsets, ranging from *14006945 action, activity, activeness (the state of being active)* to *07009640 act (a subdivision of a play or opera or ballet)*.

5 Verbs

5.1 Aspect

Both Lithuanian and Slovak verbs distinguish two aspects, perfective and imperfective, not as a morphological category, but as separate lexemes (though there are often derivation patterns converting between the two).

In Slovak, the base form is either perfective (e.g. *dať*) and the imperfective is derived semi-regularly with the *-va* morpheme (*dávať*), or the basic form is imperfective and there is a sizeable set of verbal prefixes turning it into perfective, often with considerable semantic differences (*robiť* $\rightarrow$ {*u,vy,za,pre,do,na,od*}*robiť*). There is also a class of ambivalent aspect verbs, but these can be though of as close homonyms. In the Slovak WordNet, we consistently keep both forms (if they exist) as separate synsets, linked to the same English synset.

In Lithuanian language, most of the verbs without any prefix are imperfective, whereas prefixed verbs denote a finished action. There are some exceptions, directional movement verbs are imperfective in the present tense – e.g. *atvykti* (inf., 'to come'), *atvyksta* (present tense, '[he] is coming'), however, in the past simple tense they denote a finished event

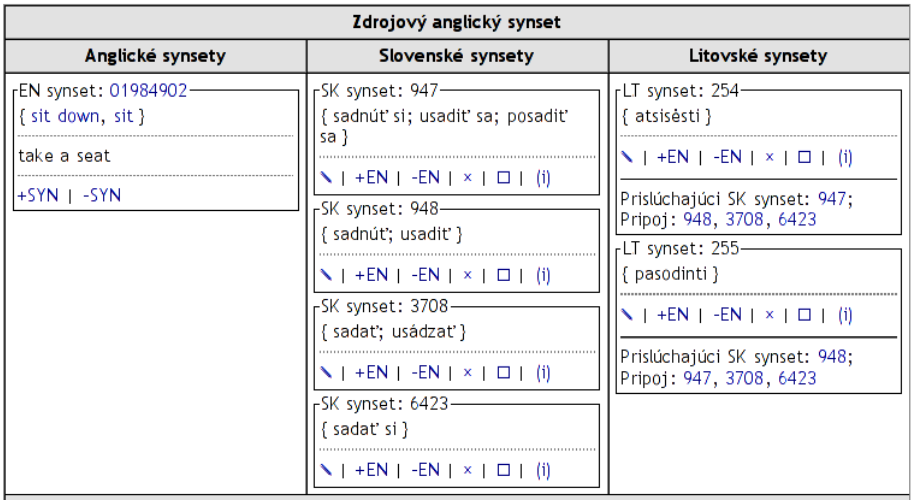


Fig. 3. Screenshot of the WordNet interface, with a view of corresponding English, Slovak and Lithuanian synsets

(*atvyko*, '[he] came'). Then there is a class of verbs neutral with respect to the aspect – e.g. *mirti 'to die'*.

In general, we preferred imperfective aspect to the perfective for simplicity, but we try to cover both aspects of the Slovak synset, if the same semantic meaning is preserved in Lithuanian.

Another exception are verbs indicating momentous (very short or abrupt) actions with suffixes *-el(ė)ti* and *-er(ė)ti*. In general, these forms were avoided in the WordNet, but they are included in cases where they tend to have a specific meaning – e.g. *gūžtelėti (pečiais)* 'to shrug (shoulders)'.

5.2 Reflexive verbs

Both Slovak and Lithuanian languages contain reflexive verbs, with approximately similar semantic behaviour. In Slovak, reflexivity is expressed by a separate reflexive pronoun/particle *sa* or *si*, which is nonetheless considered a part of the lexeme and we treat reflexive verbs as single units (literals including a space and the reflexive pronoun).

In Lithuanian, reflexive verbs have a reflexive affix *-si* or *-s*, which is attached to the end of the stem as an affix for prefixless verbs, but it is put as an infix after the prefix morpheme – e. g. *sukti* → *sukti**s***, but *nuprausti* → *nu**si**prausti*.

Syntactic reflexivity can express various semantic meanings, ranging from true reflexivity (action reflected towards oneself) through reciprocal, to pronominal reflexivity (where the reflexive status is obligatory but has no inherent meaning). There is often a ough conflation between reflexive and intransitive categories, and the non-reflexive and transitive ones.

In the Slovak WordNet we try to cover both reflexive and non-reflexive variants of the verb (if they both exist) in two separate synsets. In case where the reflexivity overlaps with transitivity, both synsets are mapped to the same English language synset (unless there are

separate transitive and intransitive English synsets). Lithuanian synsets are then mapped to the Slovak ones (not necessarily only related verbs, see Figure 1 for an example) if they cover the same meaning.

6 Manual Proofreading

The proofreading of both Slovak and Lithuanian parts was done almost simultaneously – the Slovak synsets have been proofread in two step process, first proofreading by one annotator and then a second proofreading by an independent one. Each step in itself consisted of two actions – verifying the completeness and correctness of literals in each synset, and verifying the synset position in the ontological hierarchy (i. e. its connection to the Princeton WordNet synset, its hypernyms and – if existing – hyponyms).

As the Slovak synsets acquired the "verified" status, corresponding Lithuanian ones have been proofread and edited as well, with paying attention to its interconnection to both the English and Slovak synsets.

The main Lithuanian language resources used for the proofreading were Modern Lithuanian Dictionary[4], Dictionary of International Words [1], Terminology Database of Lithuanian Republic[5], the website of the State Language Commission[6] and an encyclopaedic dictionary of computer science [2].

Only the terms approved by the Lithuanian Language Commission or present in one of the recommended (by the Commission) language resources were added to the Lithuanian WordNet. Therefore, colloquial expressions, neologisms and frowned-upon words were avoided at this phase of the proofreading (this however does not mean we are against their inclusion in the future).

7 Current Status

The Lithuanian WordNet started its life as a multilingual glossary, but it has grown up to be a small WordNet, with semantic hierarchy provided by Princeton WordNet. At the time of writing, the database composition is 7 874 noun synsets, 2 099 adjective synsets, 682 adverbial synsets and 533 verbal Lithuanian synsets. All of them are connected to the Slovak and English equivalents and the nouns, adjectives and adverbs are (once) manually proofread. Current work includes proofreading the verbs and extending existing word coverage. Once the database coverage and accuracy reaches satisfactory levels, its conversion into VisDic/DEBVisDic [5] could be considered, however the database still contains too many errors and omissions. Nevertheless, a preliminary version has been released[7] under GNU Affero General Public License, v. 3[8]; Creative Commons Attribution-ShareAlike 3.0 Unported License[9]; and Open Database License (ODbL) v1.0[10].

[4] Dabartinis lietuvių kalbos žodynas, `http://dz.lki.lt`
[5] Lietuvos Respublikos terminų bankas, `http://terminai.vlkk.lt/pls/tb/tb.search`
[6] Valstybinė lietuvių kalbos komisija, `http://vlkk.lt/`
[7] `http://korpus.sk/ltskwn_lt.html`
[8] `http://www.gnu.org/licenses/`
[9] `http://www.creativecommons.org/`
[10] `http://opendatacommons.org/licenses/`

Acknowledgments

The original multilingual dictionary was funded by the Slovak Online project[11]. Automatic synset generation and web based editing interface was provided by Faculty of Electrical Engineering and Informatics, Technical University of Košice. We thank Ján Genči and Ondrej Dzurjuv for their help.

References

[1] Bogušienė, V. and Bendorienė, A. (2008). *Tarptautnių žodžių žodynas.* Alma littera, Vilnius, Lithuania.

[2] Dagienė, V., Grigas, G., and Jevsikova, T. Anglų-lietuvių kalbų kompiuterijos žodynėlis. Retrieved from `http://www.likit.lt/en-lt/angl.html` on 17th October 2013.

[3] Dzurjuv, O., Genči, J., and Garabík, R. (2011). Generating sets of synonyms between languages. In *Natural Language Processing, Multilinguality. Proceedings of the 6th International Conference SLOVKO 2011*, Modra, Slovakia.

[4] Hana, J. (1998). Two Level Morphology of Esperanto. Master's thesis, Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic.

[5] 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, Brno, Czech Republic.

[6] Lindén, K. and Carlson, L. (2010). FinnWordNet – WordNet på finska via översättning. *LexicoNordica – Nordic Journal of Lexicography*, (17):119–140.

[7] Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. J. (1990). Introduction to WordNet: An On-line Lexical Database. *International Journal of Lexicography*, 3(4):235–244.

[8] Pease, A., Fellbaum, C., and Vossen, P. (2008). Building the Global WordNet Grid. In *Proceedings of the 18th International Congress of Linguists (CIL18)*, Seoul, Republic of Korea.

[9] Rudnicka, E., Maziarz, M., Piasecki, M., and Szpakowicz, S. (2012). A Strategy of Mapping Polish WordNet onto Princeton WordNet. In Kay, M. and Boitet, C., editors, *COLING (Posters)*, pages 1039–1048. Indian Institute of Technology Bombay.

[10] Stamou, S., Oflazer, K., Pala, K., Christoudoulakis, D., Cristea, D., Tufiş, D., Koeva, S., Totkov, G., Dutoit, D., and Grigoriadou, M. (2002). BALKANET: A Multilingual Semantic Network for the Balkan Languages. In *Proceedings of the 1st Global WordNet Association conference*.

[11] Vossen, P. (1997). EuroWordNet: a multilingual database for information retrieval. In *Proceedings of the DELOS workshop on Cross-language Information Retrieval*, pages 5–7.

[11] `http://slovake.eu/`

Corpora of Private Correspondence as a Source of Material Focused on a Research of Diminutives[1]

Zdeňka Hladká

Faculty of Arts, Masaryk University, Brno, Czech Republic

Abstract. Specialized corpora of private correspondence have been created at the Masaryk University in Brno for many years. The contribution provides a brief information concerning their characteristics, presenting them as an almost irreplaceable source of linguistic data. The contribution is focused especially on marked items from the sphere of vocabulary, namely on diminutives (especially on the pragmatic ones), whose high number can be found in the material – in accordance with its character. This study does not strive to contribute to solving the complex issues relating to diminutives, already reflected by scholarly sources form many viewpoints, but merely to turn attention to a relevant source, suitable for investigations of this part of vocabulary. The contribution is based on material gathered from the corpus of 2,000 letters included in the corpus KSKdopisy and in detail on the material excerpted from 300 letters of young people. Types, functions and specific features of diminutives occurring in private correspondence are analysed, demonstrating the creativity in using them as well. The Appendix provides a complete list of the excerpted diminutives (including the lexicalized forms), accompanied by the data relating to frequency and by elementary data relating to their lexicographical reflection.

1 Introduction

Alongside large, representative national corpora of written/spoken language, there gradually emerge minor corpora, specific with regard to the nature of collected material, methods of its processing and possibilities of using this material, such as corpora of correspondence created during the past decade at the Masaryk University in Brno.[2] The

[1] This contribution was created within a grant project of specific research Czech Language in the Unity of Synchronicity and Diachronicity (Čeština v jednotě synchronie a diachronie) – 2013 (MUNI/A/0705/2012).

[2] It is especially the corpus **KSKdopisy** (942,573 positions, 758,513 words) containing 2,000 traditional letters written by hand between 1990 and 2004, published in the book *Čeština v současné soukromé korespondenci. Dopisy, e-maily, SMS* [1] and in 2006 incorporated into the Czech National Corpus (KSK-dopisy). The same methodology was used to create the corpus **KSK111** (856,060 positions, 705,659 words) containing 2,000 letters that originated between 1902 and 2012, published in the book *111 let českého dopisu v korpusovém zpracování* [2]; this corpus is to be incorporated into the Czech National Corpus as well. Each of the above-mentioned corpora represents 2,000 idiolects, authors come from the whole territory where Czech is used. The corpora are tagged form sociolinguistic, dialectal and partly morphological viewpoints, corpus processing being accompanied by full-text versions of the letters and photocopies or original, which appears to be a very suitable combination for many kinds of research work. Besides the two corpora, a corpus containing 1,000 e-mails was created and published in Brno (**KSKe-maily** in [1]) and in 2009, a hitherto unpublished corpus was created **Korpus odeslané a přijaté česky psané korespondence Bedřicha Smetany** (focused

aim set by authors of these corpora is to make available language data obtained from correspondence analysis, since such data are otherwise difficult to acquire and to promote research of this material.[3]

A high concentration of all kinds of marked vocabulary can be found in private correspondence, this material therefore provides a lot of stimuli especially as far as lexicology/lexicographic reflection is concerned (see a probe in lexicography by [4]. This contribution is focused on diminutives – diminutives belong to items that can be found in the examined material in high concentration and rich variety. The number of diminutives in style of correspondence is related to the fact that private correspondence is open to elements of common language, to emotionally marked means of expression and to innovations of all kinds. In the case of correspondence material examined conditions of the occurrence of diminutives are augmented by the fact that it is a correspondence of the young generation, who enjoy expressive and creative ways of expression. Moreover, letters by girls prevail in the sample examined, a higher extent of intimacy and emotionality is characteristic of communication strategies used by girls.

This study does not strive to contribute to solving the complex issues relating to diminutives, already reflected by scholarly sources form many viewpoints, but merely to turn attention to a relevant source, suitable for investigations of this part of vocabulary.

2 Material Basis

The contribution is based on the KSKdopisy corpus (see Note 2), namely on a sub-corpus of 300 letters chosen at random, the only criterion being the age of authors (under 30 years),[4] 71% of the letters were written by women; women were also the most frequent addressees. Authors and addressees were mostly students of secondary schools. 73% of the letters were letters to friends, 13% letters to family members, 10% letters to partners and the rest letters by friends (4%). A relatively complete excerption of all diminutives, **including the lexicalized units**, was carried out in this sub-corpus, as well as excerpting (rather dubious) items belonging to other parts of speech. Additional excerption from the whole corpus of 2,000 letters was focused only on diminutives that are not included in SSČ (although this limitation has no relevance to the system, it helps to accentuate less usual or occasional items).

Diminutives within hypocoristic forms derived from personal proper names were left aside, because this very topic was already dealt with using material from KSKdopisy (see Note 3). Studies of this kind carried out up no now on the corpus data prove that diminution participates on hypocoristic forms derived from personal proper names by 42%, often as a part of composed suffixes (just as an illustration, there is a KSKdopisy based list of diminutive derivations relating to hypocoristic forms containing the suffix *-ul(a)/-ul(e):* feminines: ***-ul-k(a)****: Andulka, Ditulka, Evulka, Janulka, Jaňulka, Jitulka,*

on the Czech correspondence of Bedřich Smetana).

[3] KSKdopisy provided e.g. a unique material source (ca. 1,000 lemmas representing 7,000 occurrences) for a research in word-formation of Czech hypocoristic forms derived from personal proper names [3], [5].

[4] Letters from the end of KSKdopisy (between No. 1,624 and 2,000) matching the selected age criterion were selected.

Kájulka, Katulka, Marťulka, Miťulka, Moňulka, Naďulka, Peťulka, Raďulka, Zdendulka, Zdenulka, ***-ul-ink(a)****: Barulinka, Věrulinka,* other: *Jaňulilinka, Jituličečinečička, Staňulínečka;* masculines: *Jiřulka, Peťulínek, Vítťulka*).

From the total of 300 letters, 358 various diminutive forms were excerpted, occurring 779 times (for full specifications, see the Appendix), among them appellative nous 319/704 (among them diminutives formed by primary suffixes 224/443), proper nouns 3/3 (bionyms are not included, see above), adjectives 15/20, verbs 2/3, adverbs 7/18, particles functioning as greetings 12/31. By mechanic calculations there is an average of 2.6 diminutives per letter, real distribution in letters (reflecting also the relation of lexicalized/non-lexicalized diminutives): 167 letters contained at least one non-lexicalized, usually qualitative diminutive form. Other 19 letters contained (from the non-lexicalized items) only particles serving as greetings, formally close to diminutives, 49 letters contained only diminutives that were lexicalized or petrified by their occurrence in phraseological units, 65 letters contained no diminutives at all. On the other hand, in other letters the occurrence of diminutives was relatively high: 56 texts contained 5 and more diminutive forms, 17 of them contained even ca. 10 diminutives. As far as sociological parameter of authors' gender is concerned, the material proved that letters written by men contain less (especially qualitative diminutives), women use by ca. 30% more diminutives (both genders are equal only in love letters, in this case men even slightly prevail in using diminutives).

3 Diminutives

The basic meaning of diminutives usually is (or used to be, especially in the structuralistic theory) a dimension beyond norms, a dimension smaller compared with the norms. Besides a quantitative modification of the meaning or alongside it, diminutives can express a qualitative modification, on a whole scale of pragmatic tones. Most frequently it is an expression of a subjective evaluation, usually positive, but also negative, the meaning of the basic word can be in the diminutive form intensified or weakened. The most autonomous meaning is usually the quantitative one, qualitative meanings strongly depend on the context (on the immediate context, expressed by an attribute, as well as on the wider one), on the semantic nature of the basis, on the nature of the speech act etc. [11].

Some diminutives got lexicalized or were primarily formed as specifying naming units – i.e. they only have a diminutive form, but do not bear a quantitative nor qualitative modifying feature. A classification of these forms was attempted by [13].

Diminutives are usually regarded only as a category of nouns. However, some of their categorial means of derivation and partly also semantic functions spread to other parts of speech as well. This usually applies to adjectives (*malinký, slaďoučký*), adverbs (*malinko, trošičku*) and verbs (*papinkat, spinkat*).[5] It is possible to assume that particles used as greetings and containing the *-k-* formant (*ahojky,*[6] *čauik, čusík, nazdárek*) get close to that type also by their expressive function, for this reason they are in margine included to the excerpted items.

[5] Neščimenko [8, p. 14] referring to Horecký points out that it is absolutely unusual to transpose means of derivation across the borders of parts of speech.

[6] Possibilities to explain the *-k-* formant in greeting such as *ahojky* are mentioned by [6, p. 155].

4 Diminutives in Private Letters

In private letters, diminutives of the following types were found: quantitative diminutives,[7] qualitative diminutives, diminutives containing both features as well as lexicalized diminutives. Qualitative diminutives prevailed, often those derived from words that (because of the particular meaning) do not make quantitative modification possible (*bodík, jaříčko, kulatinky, počásko/počasíčko; blonďáček, doktůrek, negřík*). Lexicalized diminutives occurred relatively often – they were included in the excerpted material to gain a general overview and because most of them belonged to the types retaining certain quantitative or qualitative connotations as far as diminution is concerned (using Štícha's classification, these are lexicalized diminutives belonging to the types "knížka": *básnička, prstýnek, ukazováček;* "slovíčko": *dárek, domek, lavička;* partly also "lopatka": *stolek* – u postele, *krabička, lístek;* "sýček": *hrobařík, páteříček*). Similarly, petrified diminutive forms used in phraseological units occurred as well.[8]

Correspondence texts are an attractive material for research work relating to diminutives, because they contain a large number of qualitative diminutives. From the formal viewpoint, diminutives containing primary suffixes prevailed in the selected sample; extended suffixes (especially *-ička* and *-íčko*) could be found especially where for formal or other reasons they performed the function of basic suffixes (*mašlička, pitíčko*). However, the occurrence of other types of diminutives is not negligible as well, namely of diminutives where secondary extended suffixes expressed a higher degree of diminution (*myšička, pokojíček*) and also diminutives with strongly expressive extended suffixes (*broučínek, pejsánek*). Expressive using of diminutives in KSKdopisy is also supported by a relatively frequent occurrence of the primary suffix *-ík* in relation to the basic suffix *-ek*.[9]

---

[7] It is worth mentioning that among all diminutives, diminutives with the function of simple diminishing were found in the least extent, despite the fact that this function is regarded as primary for the particular category. Quantitatively used diminutives were in many cases accompanied by an attribute repeating the feature of being small (*malý buldoček, malé hroudky, malý kamínek*). Scholarly works usually treat this phenomenon as an amplification, stressing the feature, but it is also possible to assume that there is a need to express this feature explicitly. With regard to existence of many lexicalized diminutives of various types and degrees of lexicalization (covering also the need to name the whole classes of small objects, within which the differentiation by size need not be relevant) and on the background of a frequent occurrence of diminutives in the function of a qualitative modification it seems that the primary function of categorial means of diminution from the viewpoint of word-formation is disappearing, i.e. the function to express quantity beyond norms, pragmatic functions coming into the foreground. Cf. also [7], who, with regard to using the attribute *malý,* suggests the possibility of grammaticalization of analytic diminution in Czech.

[8] On diminutives in phraseological units see [12]. In the analysed sample of correspondence, diminutives appeared also in phraseological units that usually contain the non-diminutive form (*držet palečky; porce jak pro vrabečky*).

[9] In the analysed sample, the occurrence of the suffix *-ík* (in relation to the suffix *-ek*) is 22.6%. Just for an illustration, in the material of hypocoristic forms of proper names excerpted from KSKdopisy the occurrence of the suffix *-ík* is 31.8%.

At the beginning of the contribution it was stated why style of private letters opens a way to a large number of diminutives. It is also necessary to mention explicitly main functions of private correspondence, namely contact functions. The authors usually wish to keep positive relationships with their partners in communication, therefore they express their emotions and try to look attractive not only as far as the contents of the letter is concerned, but also by language creativity. Both of these factors influence also using qualitative diminutives. Introductory and final sections of letters prove this statement most convincingly, as they are especially rich in diminutives. Introductory passages contain several types of diminutives – besides hypocoristic forms of proper nouns, hypocoristic forms of the *miláček, drahoušek* type (they are left aside in this study) and hypocoristic forms derived from names of family members there are especially metaphorical diminutives, those that are more or less common is such functions (*andílek, brouček, kočička, pusinka, sluníčko, zlatíčko*), as well as original items (*Ahoj mé drahé šuflátko*). Forming chains, at least chains with other diminutives is common here[10] (*Holčičko moje, broučku; Ahoj Kytičko! Chybíš mi, a tak Ti píšu či snad píši dopisek … Jsi moje hvězdička na obloze. Nebo snad kytička na mýtince?; Simonko moje, pusinko, miláčku, čumáčku, šňupáčku, kočičko*). Diminutives of this type can be found especially in love correspondence, they frequently occur in letters among female friends. Diminutive forms can usually be found also in concluding sections of letters (*pusu na čelíčko; hlavně hodně zdravíčka, štěstíčka; všecky moc, moc, mocinky pozdravuj; Posílám pusinku. Tvůj broučínek*), especially in girls' letters they appear also a a part of vatious rhymed sayings and clichés (*Pac a pusu na čumáček posílá Ti tvůj miláček; Usměvavý sluníčko pro Tvý hodný srdíčko*). Also people who are asked to be remembered are referred to in a diminutive form (*Pa mé milé sluníčko a pusu bráškovi; Pozdravuj všechny známý lidičky a měj se suprovoučce*).

Qualitative diminutives appear in all other parts of private letters, most frequently being bearers of a positive feature (in accordance with general characteristics of private letters). They usually express authors' positive evaluative attitude towards the object described (*narozeninová oslavička; takovej svetřík přijde vhod*), in many cases moreover strengthened and specified by an attribute in the form of an adjective (*podařená akcička, nové bydleníčko, dobrý čajík, vtipný časopísek, milý dopísek, nejlepší strýčínek*). However, functions of diminutives need not always be in direct relationships to the reality described, diminutives can also give indirect hints as to authors' compassion with the addressee (*Takže se brzo uzdrav, kdyby to šlo pošlu Ti vitamínky*), to express the author's mood (*je neděle ráno … napapinkali jsme se; Teď sedím před naším bytečkem, poslouchám Lenyho, píšu Ti dopis, piju pivko*) etc.

Diminutives in correspondence of young people express negative attitudes as well, usually accompanied by irony (*z úst mé drahé tetičky a Tvé ctěné matičky; přehnaně inteligentní typ s brejličkama; Ale navečer mi doktůrek slíbil, že mi s tou nožkou něco provede*). Also in such cases, the pragmatic features need not explicitly relate to the topics described, a correct interpretation is possible only within a wider context[11] (*Zato teď všude chodím pozdě, lidem – zvlášť těm pošahaným a netrpělivým důchodcům jsem rozhodil se svým zpožděním jejich harmonogram; nedostávaj důchodek v 10 dopoledne,*

[10] Quotations are given in the original ortographic form.

[11] For a contextual interpretation of diminutives, see e.g. [9].

ale v jednu po obědě a to ještě na ně musím zvonit vo sto šest, poněvadž si dou po obídku spočnout.).

The scale of pragmatic features of diminutives occurring in correspondence is quite large; because of formal extent limitations at least one function, namely the one relating to euphemisms, will be mentioned in a more detailed way. A tendency towards euphemised ways of expression can relate to the reality described (*za půl hoďky valím za svým dlouhodobáčkem; opaloval jsem si právě špíčky na sluncem zalité pláži*), to the form of the basic word, if it is a rude one (*chujovinka, kokotek, kriplíček, šulínek*) or to both (*Já bych totiž ještě pořád chtěla toho hajzlíka; s fagánkama vyrazili na kolech do lesa; Můj milý slepounku*). In some cases, less polite expressions in a diminutive form even express positive attitudes (*Ahojte mršky, předem mého dopisu Vás velice zdravím; už se moc těším až se vrátíš, potvůrko*). On the other hand, some rude expressions in a diminutive form express an intensified negative and derogatory attitude (*opovrhuje davem nesvéprávných idiotků; co plácá ten blbeček*).

A reason to use diminutive forms in letters of young people (especially of girls) can also be a tendency to a hyperbolic expression. There is a number of various intensifiers in correspondence, including adjectives and adverbs with diminutive suffixes (*krať oučký, kratilinký, kratilinkatý, malilinký, malinečičký, malinkatý, nejrůžovoučký* sic!, *malililinko, mocinky*).[12]

At the end of this contribution, it is necessary to repeat and emphasize that the extent to which diminutives occur in private correspondence is not linked only with a higher level of emotionality in epistolary style, but also with authors' striving to express themselves in a creative way, toying with their language. Such striving is manifested e.g. by forming diminutives derived from words whose meaning is not usually linked with any diminution, not even with a qualitative one (*drbík, nudička, proudík, žvatlalík*), variations in the usual for of the diminutive (*bořík, bytek, ohník, vůzek*),[13] in unexpected contextual combinations (*nožička mě bolí … Tak jsem tu tlapu rozchodil*), in frequent formation of chains of diminutives (see above), in using rhymes (*teplíčko – žitíčko; Adelka – prdelka*), in using diminutives in "high-style" sections of the text (*Topoly se chvějí, koníci se pasou, i když teď se nejspíš hřejí ve svých vyhřátých pelíšcích*) etc. The following verses by a sixteen-year-old girl (linking lexicalized as well as non-lexicalized forms) can serve both as an example of using diminutives in a creative way and as an adequate invitation to further research work in the field of diminutives in correspondence.

Óda pro milovanou bytost (ze by pro Jirku ?)

Ty můj malý broučku # roztomilý kloučku # ty můj sladký tesaříčku # černoskvrnný kovaříčku # rozesmátý hrobaříku # kozlíčku dazule # (co to bylo minule ?) # potemníčku moučný # zlatohlávku Goliáši # (jak mě tvoje matka snáší ?) # mandelinko bramborová # (sejdeme se zítra znova ?) # chrobáčku, hnojníčku # (půjdeš ven chviličku ?) #

[12] Examples were selected from the whole KSKdopisy. Cf. also other intensifying items from the same material: *děsný, šílený, úžasný; obrovitánský, přenádherný, megachytrý, hypersuperabstraktní*. For intensification expressed by diminutives see e.g. [10], the topic is dealt with also by [7] etc.

[13] Language creativity and toying with language can also be observed in cases where the diminutive feature is removed (krabka, travka, světlucha).

ponravičko moje # cos vylezla z hnoje # páteříčku sněhový # (nikdo se nic nedoví !) # pestrokrovečníku # (zítra na seníku !) # květopase jabloňový # (jak to seno pěkně voní …) # vrbounku posvátný # (ať je náš den památný !)

5 Conclusions

The presented probe was aimed at indicating the possibility to use corpora of private correspondence to carry out the research of diminutives, to find out potentialities of this rich and open category and to study contextual linkage of these particular items as well as mapping their pragmatic functions.

References

[1] Hladká, Z. et al. (2005). *Čeština v současné soukromné korespondenci. Dopisy, e-maily, SMS*. Masarykova univerzita, Brno.

[2] Hladká, Z. et al. (2013). *111 let českého dopisu v korpusovém zpracování*. Masarykova univerzita, Brno.

[3] Hladká, Z. and Machalová, J. (2011). Leni, Lenčo, Leničko, ty moje sladká písničko. Osobní dopisy jako zdroj poznání hypokoristických obměn rodných jmen. In Rusinová, E., editor, *Přednášky a besedy ze XLIV. běhu LŠSS*, pages 40–51, Masarykova univerzita, Brno.

[4] Hladká, Z. and Martincová, O. (2012). *Slova v soukromných dopisech. Lexikografická sonda*. Masarykova univerzita, Brno.

[5] Machalová, J. and Osolsobě, K. (2013). Hypokoristika z rodných jmen v Korpusu soukromné korespondence. In Hladká, Z. et al., editors, *Soukromá korespondence jako lingvistický pramen*, pages 33–59, Masarykova univerzita, Brno.

[6] Nekula, M. (2003). System und Funktionen der Diminutive. Kontrastiver Vergleich des Deutschen und Tschechischen. In *brücken – Germanistisches Jahrbuch Tschechien – Slowakei*, Neue Folge, 11, pages 145–188, Nakladatelství Lidové noviny, Praha.

[7] Nekula, M. (2010). Deminutiva a augmentativa v češtině z typologického hlediska. In Bičan, A. et al., editors, *Karlík a továrna na lingvistiku (Prof. Petru Karlíkovi k šedesátým narozeninám)*, pages 304–315, Masarykova univerzita, Host, Brno.

[8] Neščimenko, G. (1980). Očerk deminutivnoj derivacionnoj sistemy v istorii češskogoliteraturnogo jazyka (konec XIII – seredina XX vv.). Academia, Praha.

[9] Rusínová, Z. (1995). Deminutivní modifikace z hlediska pragmalingvistického. In Karlík, P. et al., editors, *Pocta Dušanu Šlosarovi*, pages 187–193, Albert, Boskovice.

[10] Rusínová, Z. (1996). Deminutivní modifikace z hlediska pragmalingvistického. Intenzifikace. *SPFFBU* A 44:91–95.

[11] Rusínová, Z. (1997). Významy deminutiv v komunikaci.In Rusinová, E., editor, *Přednášky a besedy ze XXX. běhu LŠSS,* pages 112–120, Masarykova univerzita, Brno.

[12] Rusínová, Z. (1998). Deminutiva a frazeologie. In Karlík, P., Krčmová, M., editors, *Jazyk a kultura vyjadřování (Milanu Jelínkovi k pětasedmdesátinám)*, pages 113–118, Masarykova univerzita, Brno.

[13] Štícha, F. (1978). Substantiva deminutivní formy s lexikalizovaným významem. *Naše řeč*, 61:113–127.

Appendix: An Overview of the Excerpted Diminutives

The materials are organized as follows: for diminutives derived from nouns, list of forms excerpted from **300** letters for each of the suffixes is given. Frequency of occurrence, as well as items not included in SSČ are given. Differences between lexicalized and non-lexicalized items are not followed, as well as meanings or differences in meaning (with some exceptions), they can be traced in KSKdopisy.

The sign + is followed by a list of forms excerpted from the remaining 1,700 letters of KSKdopisy. Only items not listed in SSČ, without frequencies, are given. A list of diminutives derived from nouns is followed by an overview of diminutives forms derived from other parts of speech, in all cases from the 300 letters only.)

Nouns[14]
containing primary suffixes

-ek

(300 letters): andílek (3), balíček (4), baťůžek, blbeček (3), boreček, brouček (8), budíček (2), citrónek, part, černoušek (2), čumáček, dárek (7), doktůrek, domek, dopisek/dopísek (18), důchodek, fiátek, háček, hlásek, hnojníček, hranolek, hrášek, hrneček (2), hříbek, chlapeček (5), chrobáček, chudáček, jahelníček, kamínek, klídek, klouček, kocourek, kohoutek, koneček, kopeček, korálek, kousek (13), koutek (2), krámek (2), krček, kroužek, křížek (3), lístek (7), magorek, medvídek, měsíček, mlýnek, motoráček, obídek, obleček, obrázek (17), oříšek, paleček (2), párek, pásek, pejsek (4), pelíšek, plyšáček, podtácek, polštářek (3), potemníček, potůček (2), prášek (2), proutek, provázek (3), prstýnek (4), ptáček, řemínek, řízeček, schodek, sklípek, slavíček, slepounek, smrček, soudek, stánek (3), stolek, strejček/strýček (2), strojek, synek (2), šukáček, šulínek, tácek, telefonek, tlouček (2), traktůrek, ukazováček, váček, vitamínek, vlásek, vrbounek, vtípek, vůzek, zadeček, zlatohlávek, zoubek; plt. penízky (4)

+ (1,700 letters): adresníček, análek, ateliérek, bloček, blonďáček, brífek, buldoček, bytek, cancáček, citátek, časopísek, deníček, dlouhodobáček, fagánek, fotbálek, idiotek, idolek, jebáček, jogurtek, kašpárek, kaštánek, kilásek, klípek, klokánek, knůtek, kočičáček, kokotek, obchůdek, ogárek, otazníček, parníček, paškvírek, puťáček, románek, rožek, smrádek, soukromníček, staroušek, šetřílek, šneček, šňupáček, špíček, šufánek, táborek, týpek, úkolníček, vlčáček, vodáček, vojáček, vrabeček, výslešek, zobáček, žrádelníček

[14] The list of diminutive nous is virtually complete, only some forms with an unclear meaning were eliminated (e.g. haluzáček, šlukovečka); the same applies to items standing for various reasons on the borderline of categories (e.g. světluška, borůvka, čuník, papoušek, mandelinka, mamka, taťka, miláček). On the other hand, some other forms were added – those for which there need not exist the basis in the form of a noun derived from adjectives or verbs (dlouhodobáček, šukáček) or those where the path to the diminutive form is not "pure" from the viewpoint of word-formation (škudlík). The classification of diminutives by suffixes is provided just for a better orientation, from other viewpoints it is possible to put some forms into other groups.

-ík
(300 letters): bobřík (2), čajík, ďáblík, džemík, fotřík, hajzlík, hrobařík, chlapík, koncertík, koník, košík (4), mejlík (4), negřík, oslík, pokojík (4), pytlík (2), studentík, svetřík, vozík
+ (1,700 letters): *bodík, bořík, človrdík, dortík, drbík, hrošík, ímejlík, ksichtík, lettřík, mailik, nehtík, ohník, proudík, smajlík, soklík, sokolík, sportík, srabík, šéfík/šefík, škudlík, šmoulík, testík, textík, žvatlalík*
-ka
(300 letters): bavlnka (2), *bolístka,* botka, branka, bublinka, budka (2), cedulka (3), čárka (2), čepička, dečka, dírka, *drobnůstka, družinka* (3), *elektroskříňka,* hadička, hodinka (2), hospůdka, hroudka, chaloupka, chatka (3), *chujovinka,* chvilka (24), jahůdka, jiskérka, kačenka, knížka (20), konvička, košilka, krabička (3), kuchyňka, lavička (5), lednička, *místnůstka,* mrška (3), muška, myška, *mýtinka,* novinka (6), nožka, opička, osůbka (2), otýpka, ovečka, plínka, *plošinka,* postýlka (2), potvůrka (4), *prdelka/prdýlka* (2), *rodinka* (4), rukavička (2), rybka, řádka (2), říčka, schůzka (5), síťka, sklenka, skříňka, *specialitka, spirálka,* stránka (3), světnička, *šikulka,* školka (2), tabletka, tkanička, tůňka, ulička, *včelka, vesnička* (2), vodka, zah*rádka* (2), závorka, *zeleninka, zrůdka* (2); masc.: *bráška;* plt.: hodinky (2), *narozeninky* (2)
+ (1,700 letters): *adreska, blbůstka, brigádka, cérka, cigaretka, dobrůtka, fontánka, chemijka, robotnička, kadibudka, kánojka, klobáska, kocovinka, kolegyňka, krasotinka, kravka, kytárka, lysinka, nezbytnůstka, papřička, petunka, pičovinka, pitominka, plachetka, prasečinka, přehrádka, rostlinka, slaninka, svačinka, šešulka, švadlenka, televizka;* plt. *kulatinky, prázdninky*
-ko
(300 letters): *brčko,* bříško, *dílko,* dítko (3), *jídélko,* křidýlko, *lehárko,* očko, okýnko, ouško, pírko, písmenko (3), pivko (3), *počásko,* polínko, ramínko, *ranko,* semínko, stehýnko, světýlko (2), tričko (4), víčko, *vínko;* plt. jatýrka, kolínka (=pasta) (2), vrátka
+ (1,700 letters): *čísílko, divadélko, letadýlko, plátýnko, sáčko, štísko, témátko*
-átko
(300 letters): děťátko, koťátko, prasátko (4), *šuflátko,* zvířátko (4)
+ (1,700 letters): *čuňátko,* (*rajčítko*)

containing enlarged suffixes

-eček
(300 letters): *byteček* (5), *dáreček* (5), dědeček (3), domeček, *filmeček,* hrobeček, *krámeček,* lísteček, páreček, stoleček (2), stromeček
+ (1,700 letters): *dopiseček/dopíseček, kouteček, lemureček, obrázeček, státeček, stroječek*
-íček
(300 letters): bratříček (3), *čajíček,* človíček (4), koníček, *kovaříček, kozlíček,* mazlíček, *olejíček, páteříček, pohledíček,* pokojíček (6), prstíček, *tesaříček*
+ (1,700 letters): *cukříček, kriplíček, křížíček, ksichtíček, (v)obličejíček, páníček, sexíček, šmoulíček, úkolíček*

-ička
(300 letters): babička (29), básnička (8), *brigádička, broskvička, dušič*ka, hlavička (4), holčička (3), hvězdička (3), chvilička (3), kartička (2), kočička, kulička, kůlnička, kytička (5), lahvička, mašlička (2), matička, myšička, nožička (4), perlička, písnička (8), *pohodička* (4), *ponravička,* rakvička, ručička, rybička (3), sestřička (5), *srandička,* svíčička, *šatnička, školička,* taštička (2), tetička, vanička, větvička, židlička (2); plt.: *brejličky,* dětičky, lidičky (5), *mušličky* (= pasta)
+ (1,700 letters): *akcička, bundička, cukrárnička, čokoládíčka, dílnička, flétnička, kachlička, kresbička, mastička, notička, nudička, nymfička, oslavička, partička, pastička, pikslička, skladbička, stopička, travička, větička, vyžlička, zábavička, zemička, zprávička, žárovička, žirafička*
-ečka
(300 letters): *fotečka* (2)
+ (1,700 letters): *cérečka/dcerečka, kamarádečka, knížečka, otázečka, složečka, sponečka, trnečka, zmínečka, známečka*
-íčko/(-ičko)
(300 letters): *albíčko*, autíčko (2), *čelíčko* (2), jablíčko (2), náměstíčko, nebíčko, *pitíčko, přáníčko, seníčko,* sluníčko (22), srdíčko (3), *šitíčko,* tělíčko, *údolíčko,* vajíčko, *zdravíčko,* zlatíčko (2); plt.: prsíčka, *zádička*
+ (1,700 letters): *bydleníčko, jaříčko, očičko, papáníčko, písmíčko, počasíčko, pozdraveníčko, štěstíčko,* teplíčko, *učeníčko, vystoupeníčko, žitíčko*
-ečko/-éčko
(300 letters): kolečko (3), městečko (2), pivečko
+ (1,700 letters): *kupéčko, srdéčko*
-ánek
(300 letters): copánek (2), *pejsánek*
+ (1,700 letters): *chudánek*
-ínek
(300 letters): *broučínek,* tatínek (8)
+ (1,700 letters): *strýčínek*
-enka
(300 letters): dívenka
+ (1,700 letters): *kočenka*
-inka
(300 letters): maminka (21), *prcinka* (2), pusinka (10)
+ (1,700 letters): *dušinka, hlavinka, chvilinka, matinka, neteřinka, tetinka*

combination of suffixes *meďáneček*
Proper nouns: *Bristůlek, Budvárek, Nirvánka*
Other parts of speech (only from 300 letters):

Adjectives
-ičký každičký, kratičký, maličký
-oučký *kraťoučký, nejrůžovoučký, slaďoučký,* žluťoučký

-inký malinký (5), *samotinký*
-**ounký** hezounký, *slaďounký*
Containing other composed suffixes and reduplication: *malinkatý, malilinkatý, nemocinkatý, nemocinkaný* (2)

Verbs
*(na)*papinkat, spinkat (2)

Adverbs (only in alphabetical order; including the cases where the diminutive feature is merely transposed from the funding adjective)
drobátko, malinko (2), *mocinky, suprovoučce, trošičku* (5), *trošinku* (2), trošku (6)

Particles/interjections (only in aphabetical order)
ahojky (9), *culisek, čauik, čauka, čauky* (10), *čauverek, čusík busík, nazdárek* (2), *nazdarky, papáček, sbohemky, zdárek* (2)

Identification of Idioms in Spoken Corpora[1]

Milena Hnátková and Marie Kopřivová

Faculty of Arts, Charles University in Prague, Czech Republic

Abstract. This paper focuses on the automatic identification of idioms within the transcript of spoken discourse which are included in the spoken corpora ČNK (PMK, ORAL2006 and ORAL2008). In PMK, the idioms were manually searched for and identified, so it is possible to compare the efficiency of automatic and manual identification and describe the advantages and disadvantages of both approaches.

1 Source Corpora of Spoken Language

When searching idioms we use data from three spoken corpora, which are part of ČNK: Prague spoken corpus, ORAL2006 and ORAL2008.

Prague spoken corpus (the PMK) is the oldest part of corpus of spoken language within the Czech National Corpus. Recordings were taken in 1988–1996. Only the transcripts of recordings are available for the search. PMK was counterbalanced by four sociolinguistic variables, each of which can take two values. Three of them are related to the speaker: gender (male – female), age[2] (younger than 35 – older than 35), education (higher – lower), and the fourth to the type of speech (formal and informal). By formal we understand such situations in which the recording person was asking questions and the speaker replied to them in longer answers. As far as informal situations are concerned the topic here was not influenced by anything and it was a free conversation between the speakers. The transcription was aiming to capture the spoken language as accurately and clearly as possible and the concept is close to the folkloristic transcription. The transcribed text was also manually tagged and lemmatized. As a part of this lemmatisation is also marking of idioms which like other collocations are assigned with a multi-word lemma. The PMK is therefore a good reference corpus for assessing the success of an automatic idioms search.

The ORAL2006 corpus consists of transcriptions of recordings that were made in the years 2002–2006 in Bohemia (i.e. not in Moravia and Silesia). All were taken only in informal situations: They present conversations of speakers who knew each other and who had a friendly relationship (often they are family members, friends or acquaintances) and the recordings were taken in their natural environment. This corpus follows the manner of transcription and classification of the speakers of the PMK. Due to the extension of the collection to all Czech regions a new category was added: dialectal area in which the speaker spent his or her childhood, because at this time idiolects are being formed. The corpus is not balanced and only the transcript is available for the search.

[1] This article was created during the implementation of the project Czech National Corpus (LM2011023) funded by the Ministry of Education, Youth and Sports as a part of Major Infrastructure Projects for VaVal.

[2] The recording focused on the language of adults, so only people older than approximately twenty were recorded.

The ORAL2008 corpus follows the ORAL2006 in the manner of collection and transcription. It contains transcriptions of recordings from the years 2002–2007 and is counterbalanced by the characteristics of speakers: gender, age, education, dialect region of residence in childhood. Because these two corpora were greatly similar, we combined their data for the purpose of looking up the idioms and created a corpus to which we will further refer to as **ORAL**. **The ORAL corpus** does not contain all the data included in the reference corpora ORAL2006 and ORAL2008 and it is smaller (it consists of 1,691,474 word forms).

The ORAL and the PMK corpora are therefore different in the time of the recordings, in the inclusion of a formal type of speech in PMK and some little details in the manner of transcription. We believe that these differences do not preclude comparing the success rate in the idiom search.

2 Transcription of Spoken Language

The choice of spoken language transcription for the corpora is not easy. On the one hand, there is an effort to capture spoken language as accurately as possible; on the other hand, it is necessary to take into account the future users and allow them an easy way to search and orient themselves in the found transcript. The amount of data that must be processed manually must be also taken into consideration, as that can often lead to inconsistencies and errors. The easiest the way of transcription, the faster processing it enables. And last but not least, the way of transcribing is also affected by further processing such as morphological annotation of the corpus.

For the transcription in the PMK a relatively simple transcription was chosen. It retains elements of written language such as syntactic punctuation[3] and word boundaries. However, prosodic features, overlays (segments in which several speakers are talking at the same time) or special metalinguistic information about the situation are not recorded.

Other phenomena of spoken language are captured in the transcription of spoken corpora.

These are, for example, word repetitions (*že by to bylo spíš* ***na na*** *škodu*)[4], unfinished words (*aby* ***mo****, mohla být slušně živá*); indicating incomprehensible sections using dashes; and marking unfinished replicas (*a naši právě … a hele co von měl teda za školu*?), reduced forms of common words (prototype *prtože, prže, páč*), frequent use of demonstrative pronouns (*ten* ***to, to, to*** *se lilo do těch forem*), so-called "padding words" (*vono* ***vlastně*** *to prostředí*), hesitation filler words (*prostě přesně tak jako. to sou* ***eee****, jak sou pojistky a tak*), responsive sounds (***hm****, už jsem tamto využila*), idiolectal expressions, neologisms and garbled words (*a hrabě se tam* ***zaši*zasek***), changes in sentence perspective (*tam ten Stáňa ten co bylo, ten ta skříňka jako, no ten na to jídlo jo*), and correcting and restarting the speech (*ten plamen, tak ten ta teplota se roznáší i, protože litina je*).

[3] The transcription of the spoken language is commonly used pause punctuation, which is more suitable for its capture.

[4] Presented examples are from PMK or ORAL corpora.

These facts can break the connection within idioms and make their automatic search impossible (for example, the idiom is finished by the other speaker and therefore it appears in a different segment, separated by punctuation; for more information see 4.1).

3 Search Procedure FRANTA

FRANTA (Idiom ANnotation and Text Analysis) is an automatic search program that looks up collocations (idioms and collocations based on the Dictionary of Czech Phraseology and Idiomatics) in corpus data. It works with an unambiguously morphologically tagged text. The program for discontinuous idioms search allows you to specify morphological information or lemma of each word and a possible change in word order.

It can also be specified whether these are contiguous or non-contiguous combinations of words, i.e. positions within the idiom where any other words can be found are marked. Phrases that are automatically found are then identified and can be searched by using the corpus manager.

4 Comparison of Manual and Automatic Annotation Search on the PMK data

In manually marked collocations in the PMK, the idioms are marked with the so-called **collocation lemma** (hereafter CL); with the idioms marked by the FRANTA program, this lemma is presented as an attribute called **kolok**. In the automatic identification of idioms, all the components as word forms are assigned their lemma; when marking manually, only the collocation lemma attribute is specified.

4.1 Processing of Spoken Corpus

4.1.1 The Issue of Morphological Identification of Words

The use of syntactic punctuation in the transcriptions in the PMK and ORAL corpora helps to identify idioms automatically but it is complicated by the absence of capital letters at the beginnings of sentences. A further challenge is to capture also reduced variations of words, as especially the ORAL corpus tries to cover as many variations as possible. Some frequent words are characterized by a number of variants (e.g. the word *"nějaký*" can be pronounced and written as *něaký, náký, ňáky, ňákej, ňáké, nějakej*).

All the variations of words cannot be included into the dictionary of automatic morphological analysis; therefore an auxiliary automatic procedure has been created, in which the unknown forms and unusually written words are converted into literary variations (the forms remain, only the literary lemmas are added to unknown forms).

4.1.2 Morphological Tagging

In the automatic morphological disambiguation and automatic identification of constant collocations correctness of morphological tagging of word forms is crucial. The automatic search and marking of idioms and of constant collocations are based on morphological lemma and tag – associated with a specific use of the word in speech or text. The procedure for searching of constant collocations is however also a part of morphological disambiguation. The result of comparing manual and automatic marking of idioms is that any mistakes in automatically disambiguated idioms have been corrected and added to the automatic disambiguation.

4.1.3 Identification of Idioms and Constant Collocations

Discontinuity of spontaneous spoken language also makes the identification of idioms more difficult. Speakers often meander in their speech, interrupt their speech by various collocations, such as *abych tak řekl, já si myslím, jak se říká*, and hesitation filler words or responsive sounds, such as *hm, ehm, že jo, ty vole*; repetition of words or truncated words also often occur. Individual parts of an idiom can even be part of different sentences. Search of some idioms (verbal and nonverbal) is done only in one clause (the border is punctuation); other (clause expressions, proverbs, similes) are only searched within sentences.

4.2 Differences in Manual and Automatic Tagging of Idioms

4.2.1 The Quantitative Difference (Range of Manual Annotation)

Both annotations differ in range of marking idioms, multi-word names and phrases. In a manual annotation, also occurrence of such things as people's names (*Hruškan Hruškovič Hruškanovič*), municipalities (*Horní Mokropsy*), states (*Česká republika*), multi-word numerals (*dva tisíce pět set*), names of institutions (*Akademie věd*), movies (*Díky za každé nové ráno*), books (*Encyklopedie vědy a techniky*), songs (*Boleslav, Boleslav, překrásné město*), etc. were marked.

Unusual variations of proverbs, similes and quotations are also identified only in the manual annotation: *hlavně zdraví a ostatní si koupíme*; *žena nosí zástěru proto, aby zakrývala hanbu svého muže*; *hvězdná obloha nad námi a kategorický imperativ v nás*; *jako když slon žebrá o cukr*.

In the manual annotation also variations of known proverbs that you cannot search automatically can be found:

transcription: ale tak je to jenom ta výjimka co potvrzuje pravidlo
CL: to je výjimka co potvrzuje pravidlo

transcription: no ale, když chce někdo psa bít, viď, jak se říká, tak se dycky ta hůl na to najde, viď
CL: kdo chce psa bít hůl si vždy najde

Constant prepositional phrases (composed prepositions), such as *ve prospěch (něčeho), ve skrytu, ve směru, ve smyslu, ve snaze, ve spojení s, ve spojitosti s, ve spolupráci s*, are also marked manually in the PMK. One of the options in the search procedure FRANTA also allows an automatic search and marking of these collocations. However, this was not applied for the identification of idioms.

Noncontiguous multiword conjunctions (multipart conjunctions), such as *zda-nebo, buď-nebo, jak-tak, chvíli-chvíli, tak-jako, tak-jak, spíš-než, když-tak*, are in the manual annotation searched for and marked:

buď sem někde dělal, nebo sem učil
že se lišili jak vědomostma, tak vlasně i jednáním
že když je mladá, tak chce mít děti

To automatically determine where these sentence structures occur is not possible.

Nonverbal idioms are in manual marking usually marked with CL "To je…" ("This is ..."). The manually tagged part of the PMK identifies a total number of 397 different phrases of this type. Phrases such as *to je výjimka, to je volovina, to je vina, to je věc,* or *to je tragedie* are identified this way, although their phraseological meaning can be discussed.

In the automatically annotated corpus only 65 different phrases of this type are marked; most of the others are marked just as nonverbal idioms, or as part of a verbal idiom with the verb *být* 'to be': „. (*to je k nevydržení – nebýt k vydržení*). Not only because of these an option should be considered whether there should be something like a common phraseology lemma which would involve the use of an idiom in different types (nonverbal, verbal, sentence).

The following collocations which have no phraseological meaning are also manually annotated: *to je kůň, to je vidět, to znamená, to závisí, u mě, u nich, u něj, u nás, u vás, vidět do, za mě, za nás.*

From the above it is clear that there is no sense in a quantitative comparison of idioms found automatically, using the program FRANTA, and manually.

In the cases where manual and automatic annotations are used on clearly identified idioms and where the same lemma is applied, the automatic annotation is usually more successful. Unlike the manual annotation, it searches for and marks all occurrences of an idiom. Here are some examples of the largest differences in the number of annotated occurrences (automatic – manual): *něco takového* (141A-67M), *od té doby* (36A-4M), *to je všechno* (61A-4M), *nebo tak nějak* (31A-2M), *vykašlat se na to* (13A-6M), *nedá se říct, že* (46A-1M). Some idioms were not manually annotated at all but were searched for only automatically, for example: *jádro pudla, svatá trpělivost, že tě to baví, hluboká příčina, hluboký spánek, hluboký zájem, hned se roznést, hříšná myšlenka, hnát se za penězi, hodně už pamatovat, holý život, horká linka, horkou jehlou, z blbosti, živá váha, mrazení v zádech, špinavé peníze – praní špinavých peněz, sametová revoluce*.

On the other hand, some idiom occurrences are intentionally identified manually, as they may have their own specific meaning, and so/thus it is necessary to examine the wider context, intonation of speech or background of the situation to determine which use it is. These include phrases such as the following: *to je absurdní, to je úžasné, to je bedna, to je báječné, to je hlavní, ve velkém, v první řadě, nad hrobem*.

a na Dlouhym je to vidět,... prostě opravdu, že je to bedna
no, to je bedna, no, no, zvuková nějaká.

4.2.2 The Qualitative Difference in Tagging. Double Word Entries in the Transcription of Spoken Language

For some words (especially proverbial compound words and foreign words) there are doublets also in the written language. This group is further enlarged by misspellings, garbled words or mistakes in the transcription. These different ways of transcription make automatic identification of idioms more difficult. It is difficult to discover these inconsistencies even when the transcriptions are being checked, especially when multiple annotators are involved.

These include the following types: *popravdě řečeno – po pravdě řečeno, důvody jsou nasnadě – důvody jsou na snadě, jakpak by ne – jak pak by ne, notabene – nota bene, a priori – apriori, a tak dále – atakdále, brát to ze široka – vzít to zeširoka, vyzkoušet si nanečisto – vyzkoušet si na nečisto*. One example of such inconsistencies in manual annotation is the transcription *to je hais,* when a lemma is *to je hajs*.

When comparing the extent of both types of annotation, differences arise mainly in such cases when the program FRANTA or the manual annotation marks only a part of the idioms used. As a result, the same occurrence is assigned with a different collocation lemma. The advantage of automatic marking is that lemmas are more general and that marking is then more systematic and uniform. In the case of manual annotation, the degree of generalizations for a specific occurrence depends on the annotator's choice. Sometimes these may even differ from the actual transcriptions or the original expression of the speaker, and even annotations by the same annotator may not be consistent. On the other hand, annotators are also able to assign CL to updated or faulty variants of idioms.

transcription: *bič kerej si na sebe ušili*
CL: *bič který si na sebe upletl*

transcription: *dálnice a tím to zhasne jako*
CL: *a tím to hasne*

transcription: *udělá kravinku*
CL: *udělat kravinu*

transcription: *je to takový trošku na pláč*
CL: *to je k pláči*

transcription: *a zboří se svět*
CL: *svět se nezboří*

transcription: *si stanoví člověk za cíl*
CL: *stanovit si cíl*
colloc: *stanovit si za cíl*

transcription*: já sem spal spánkem spravedlivým*
CL: *spát spánkem spravedlivých*

In some cases it may be variations which are becoming more frequent (e.g. *a tím to zhasne*, appears in the ORAL corpus), in other cases the speaker cannot recall the correct expression (e.g. *bič kerej si na sebe ušili*).[5]

Other examples of where the lemmas of both annotations are significantly different (first is the CL, after the dash is kolok.):

pověsti kolujou – kolují pověsti, tak si trhni nohou – trhnout si nohou, ani mě to nehne – to mě ani nehne, to je houby platné/být houby platný – být houby platný, chlebem živ jest člověk – chlebem živ je člověk, nic si z toho nedělej/nic si z toho nedělá – nic si z toho nedělat, odtržený vod života – odtržený od života, zdá se mu že se znova narodil – znova se narodit

Another difference in the form of an idiom lemma is that the substitute pronoun *to* (it) is part of the lemma in manual tagging, whereas in automatic tagging it is either not part of the lemma at all, or it is replaced by a general substituting word such as *něco*, *někdo, nějaký* (something, someone, some), etc.

CL: *pouštět to druhým uchem ven*
kolok: *pouštět druhým uchem ven*

CL: *zažít to na své kůži*
kolok: *zažít na své kůži*

5 Conclusion

The following table shows the number of occurrences of idioms in each corpus Percentage-wise, there are more idioms in the PMK corpus than there are in ORAL. PMK contains older records than the ORAL corpus, which contains also records of informal situation. Further work is needed to explain the difference.

[5] [See: 7]

Corpus	ORAL	PMK
Number of positions	2,479,837	846,562
Number of idioms	30,795	16,357
Percent of occurence of idioms	3.00	4.77

This article discusses the first use of the automatic identification of idioms in spoken corpora using the FRANTA program. So far, this program has been used only with the written corpora.[6] Thanks to the manual idiom marking in the PMK corpus, it was possible to compare the advantages and disadvantages of automatic identification. New incentives to improve the disambiguating procedures came from this comparison and new idioms were found to be included in the program FRANTA. This comparison also brings about several questions related to idiomatic searches: What should a lemma look like for alternative idioms or for idioms which may take the form of a nonverbal idiom and also a sentence idiom? Are these lemmas to be divided as they are now or should the users be able to find all occurrences using some kind of "hyperlemma" and then decide what is of interest for them? We hope that this article will encourage more discussion on the form of phraseology processing within the corpora.

References

[1] Čermák, F. et al. (2007). *Frekvenční slovník mluvené češtiny*. Karolinum, Praha.

[2] Hnátková, M. (2006). Typy a povaha komponentů neslovesných frazémů z hlediska lexikálního obsazení. In *Studie z korpusové lingvistiky*, pages 142–167, Nakladatelství lidové noviny, Ústav Českého národního korpusu, Praha.

[3] Hnátková, M. (2011). Výsledky automatického vyhledávání frazémů v autorských korpusech. In *Korpusová lingvistika Praha 2011, sv. 3. Gramatika a značkování korpusů*, pages 171–185, Nakladatelství lidové noviny, Ústav Českého národního korpusu, Praha.

[4] Kopřivová, M. and Hnátková, M. From a Dictionary to a Corpus. In press.

[5] Kopřivová, M. (2008). Frazeologie v mluvených korpusech na základě PMK. In Kopřivová, M. and Waclawičová, M., editors, *Čeština v mluveném korpusu*, pages 149–160, Nakladatelství lidové noviny, Ústav Českého národního korpusu, Praha.

[6] Kopřivová, M. and Waclavičová, M. (2006). Representativeness of Spoken Corpora on the Example of the New Spoken Corpora of the Czech Language. In *Proceedings of the international conference "Corpus linguistics – 2006",* pages 174–181, St. Petersburg University Press, St. Petersburg.

[7] Schindler, F. (1993). *Das Sprichwort im heutigen Tschechischen: empirische Untersuchung und semantische Beschreibung*. Sagner, Munchen.

[8] Waclawičová, M., Křen, M., and Válková, L. (2009). Balanced corpus of informal spoken Czech: compilation, design and findings. In *Proceedings of the 10th Annual Conference of the International Speech Communication Association* (Interspeech 2009), ISCA, pages 1819–1822, Brighton.

[6] Cf. [2], [3], etc.

The Corpus CzeSL in the Service of Teaching Czech for Foreigners – Errors in the Use of the Pronoun *který*

Andrea Hudousková

Faculty of Arts, Charles University in Prague, Czech Republic

Abstract. The paper introduces the Czech learner corpus CzeSL as a source of relevant linguistic information about the errors of non-native speakers of Czech. It focuses on errors in the use of the relative pronoun *který* (who/which/that) and compares the scale of attested mistakes with the description of relative clauses in textbooks and grammars of Czech. It reaches the conclusion that an analysis of corpus data may considerably contribute to the description of Czech as a foreign language.

1 Introduction

Learner corpus is a representative set of texts of non-native speakers that illustrates the complex process of language acquisition. It is also a source of information about errors made by learners at different language levels and thus can be used for theoretical purposes, teaching and creating textbooks and handbooks of the target language.

This paper focuses on errors in the use of the Czech relative pronoun který (who/which/that) made by non-native speakers. The paper is organized as follows. Section 2 introduces the Czech learner corpus CzeSL. In section 3 the most frequent types of errors are distinguished on the basis of the corpus data. Sections 4 and 5 concentrate on how the pronoun který and relative clauses are explained in textbooks and grammars of Czech as a foreign language and whether their presentation is sufficient and adequate with respect to errors found in the CzeSL. Section 6 summarizes in what way the learner corpus may contribute to the description of Czech as a foreign language and indicates how it may help to improve the explanation of grammar and to create language textbooks.

2 CzeSL – Corpus of Non-native Czech Speakers

CzeSL is a non-reference corpus of non-native Czech speakers, the first version of which was published in 2012 in the size of 2 million of words.[1] This version of the corpus includes written texts, mostly students' essays produced in lessons and, to a lesser extent, texts of theses. The corpus contains texts of all language levels and of different first languages – Slavic, non-Slavic and non-Indo-European. Although in this respect the corpus is not absolutely balanced, it is still a source of representative linguistic data valuable for the study of Czech as a foreign language and consequently also for teaching purposes: "(...) generally, the sense of creating an emended and linguistically annotated

[1] Besides the pilot Slovene corpus PiKUST, CzeSL is the only existing corpus of a Slavic language. The size of CzeSL is comprable to that of German or French learner corpora. For the parameters of the corpus CzeSL see [22].

corpus is, inter alia, to allow the teachers of Czech as a foreigner language to find out quickly and effectively what types of mistakes and to what extent are made by learners of Czech whose mother tongue is not Czech, (...)" [18, p. 61][2].

3 Errors in the Use of the Pronoun *který*

The data from CzeSL based on the observation of the use of all forms of the relative pronoun *který*[3] indicate the following types of errors[4]:

1. *který* used instead of the possessive relative pronouns *jehož, jejíž, jejichž* (of which);
2. *který* used instead of the relative pronoun *jenž* (who/which/that)[5];
3. *který* used instead of the relative pronouns *kdo* (who)*, co* (what);
4. *který* not immediately following the modified noun phrase;
5. the form of *který* determined by the case of the modified noun and not by the valency of the verb in the relative clause;
6. *který* used instead of a more appropriate relative adverb;
7. other errors.

The next subsections deal with the mentioned types of errors in more detail.

3.1 *Za účelem kterého*[6] – *který* Instead of a Possessive Relative

The most frequent error (52 tokens), which is a consequence of negative transfer from other Slavic languages, is that the genitive form of the pronoun *který* is used instead of possessive relatives *jehož, jejíž, jejichž*, as demonstrated by the examples (1) and (2)[7].

1. tak ten člověk vubec ne bude chapat jiného človeka, **v jazyce kterého [v jehož jazyce]** ta barva je
2. že tajemnost dozoru/sledovaní nesmí být delší, než je za potřeby k ochraně zájmů, **za účelem kterých [za jejichž účelem]** tito opatření bylá zavedená

---

[2] „(...) smyslem tvroby emendovaného a lingvisticky anotovaného korpusu obecně je mj. umožnit učitelům češtiny jako cizího jazyka rychle a efektivně zjišťovat, jakých typů chyb a v jaké míře se dopouštějí studenti češtiny, pro něž čeština není mateřským jazykem. (...)"

[3] For the forms *který, kterého, kterému, kterém, kterou, kterých, kterým, kterými* all tokens in the corpus were checked, for the forms *kteří, které, která* random samples of 500 tokens, in sum 4 155 occurrences of the lemma *který*.

[4] Errors in the form of the pronoun *který,* i.e. errors in case, gender, number and animacy or in other respects incorrect forms were ignored.

[5] *Jenž* is claimed to be a bookish equivalent of *který*. However, this is not true of all its forms. Furthermore, as will be shown in subsection 3.2, the two pronouns cannot be interchanged in all cases.

[6] *For the purpose of which*

[7] Appropriate forms are given in square brackets.

The use of the genitive form of the pronoun *který* instead of the possessive relative is substandard in Czech and in written texts it occurs only marginally, as evidenced by the data from the Czech synchronic corpus SYN2010. The use of the pronoun *který* is there limited only to four collocations given in (3). However, even in these cases the possessive relative is more frequent.[8]

3. na základě kterého (42) / na jehož základě (161)
   v rámci kterého (49) / v jehož rámci (173)
   v důsledku kterého (3) / v jehož důsledku (31)
   v průběhu kterého (1) / v jehož průběhu (45)[9]

3.2 *Většina z kterých*[10] – *který* Instead of the Pronoun *jenž*

Another relatively numerous group of errors (13 tokens) includes collocations like *většina (z) kterých* (majority of which)*, každý z kterých* (each of which)*, spousta z kterých* (plenty of which)*, mnohé z kterých* (many of which)*, jeden z kterých* (one of which)*, hlavní z kterých* (the most essential of which) in which the relative pronoun *který* is used instead of the pronoun *jenž*, as demonstrated by the examples (4)–(9).

4. Česká republika láká turisty z celého světa svými vynikajícími historickými památkami, **většina kterých [z nichž většina]** se náchazí ve hlavním městě – v Praze
5. A není to lehké z dvou důvodu, **jeden z kterých [z nichž jeden]** je jazyk jako takový
6. a proto pálíme velký slaměný strašák (jako symbol) Maslenica se skládá ze sedmi dnů, **každý z kterých [z nichž každý]** má svůj název
7. Při detailnějším zkoumání určitě najdeme i víc shodných prvků, **hlavními z kterých [z nichž hlavními]** jsou ale zase opakující se motivy – láska, smrt, Praha.
8. Pozoruje, k jakým chybám dochází při běžné mluvě tehdejší společnosti, **spousta z kterých [z nichž spousta]** přetrvává do dneška.
9. Má obrovské plány, **mnohé z kterých [z nichž mnohé]** již jsou uskutečněny.

On the contrary, in the corpus SYN2010 no examples of such use of the pronoun *který* were found and only the genitive form of the pronoun *jenž* was attested in these cases.

3.3 *Někdo, který*[11] – *který* Instead of the Pronouns *kdo, co*

In five examples *který* was used instead of the pronouns *kdo* (who)*, co* (what), eventually *jenž* (who/which/that). The respective collocations *někdo, který* (someone that)*; každý,*

[8] The number of tokens attested in the corpus SYN2010 is given in parentheses.
[9] "on the basis of which; in the frame of which, in the consequence of which, in the process of which"
[10] *Majority of which*
[11] *Someone that*

který (everybody that)*; něco, které* (something that)*; všechno, které* (everything that) are demonstrated in the examples (10)–(13).

10. a dostat něco od někoho, **o kterém [o němž / o kom]** víme , že on myslel tehdy na nás
11. Můžu doporučit každému, **který [kdo]** má možnost dostat takový pobyt , aby tuto šanci využil
12. pokaždě jdu pěšky po Nové Městě, najdu něco **kterého [co]** jsem nikdy nepoznamenala
13. Lež je všechno, **které [co]** mění fakta a manipulace – všechno, které mění vědomí

3.4 Violation of Postposition

In nine instances the relative clause did not immediately follow the modified noun, as illustrated by the examples (14) and (15).

14. Ma ovalný obličej, **oči** ma šedozelené, **které** lemují delší černé řasy
15. **Socha** Jana Nepomuckého se dostá do středu zajemu, **kterou** si v praze hladí pro štěstí.

3.5 Wrong Case of the Pronoun *který*

In nine structures the relative *který* had wrong case, in agreement with the form of the modified noun, not according to the valency of the verb in the relative clause, as shown in examples (16) and (17).

16. jet do nějakého hradu **kterého [který]** má legendu nebo pohádku
17. Dost často můžeme vidět oficiálního oblečenýho muže s batohem nebo holku, **kterou [která]** si oblékla šaty a obula tenisky.

3.6 *Který* Instead of a Relative Adverb

In nine instances the pronoun *který* was used instead of more appropriate time or place relative adverbs *kde* (where)*, kdy* (when). It concerned collocations of general time and place expressions with a relative, such as *čas* (time) */ doba* (period) */ den* (day) */ víkend* (week-end)*, ve které(m)* (in which)*; místo, ve kterém* (place in which). Although the use of the pronoun *který* in these cases is mostly not excluded,[12] the use of relative adverbs *kdy* a *kde* is more common, as supported by the data from SYN2010. Examples from the corpus CzeSL are given in (18)–(20).

18. Byl to čas, **ve kterém (kdy)** se ukázalo ostřejší, kdo obstojí a kdo ne.
19. Co budeš dělat 12. května, **v kterém (kdy)** začíná Mezinárodní hudební festival Pražské jaro.

[12] However, in the example (19) with the ellipsis of the modified noun the use of the pronoun *který* is not possible.

20. vždicky dostane dobrou radu a je místo, **ve kterém (kde)** na něho čekají a myslí

3.7 Other Errors

Sporadically, the pronouns *který* and *jaký*[13] were confused, as illustrated by the example (21).

21. Příroda tam nádherná, nejhezčí **kterou [jakou]** já jsem videla ve své životě

Marginally, the pronoun *který* was used substantively, i.e. without a noun it would depend on, as illustrated by the example (22).

22. Když jsem našla **kterou** se mi líbí, nejdřív přečtu, a potom hledám slovník.

3.8 Summary

The data from the corpus CzeSL demonstrate that the majority of attested errors follows from using the pronoun *který* instead of possessive relatives on one hand and the pronoun *jenž* on the other hand. On the contrary, violation of postposition of the relative clause after the modified noun phrase, choosing wrong form of the relative or using the relative *který* instead of the pronouns *kdo* and *co* are by far less frequent.

4 The Pronoun *který* in Textbooks of Czech for Foreigners

The errors analyzed in section 3 occur in syntactic structures whose active knowledge is expected only at higher language levels. In the reference description of the level B2 for Czech [1] there is mentioned an active knowledge of relative clauses with the pronouns *kdo, co, jaký, který, čí* (who, what, which/that, whose). A user of Czech at this level should understand structures with the relative pronoun *jenž,* with the possessive relatives *jehož, jejíž, jejichž* and with the colloquial noninflected relative *co.* However, active use of these structures is not expected and it is required only at higher language levels.

Indeed, while explaining relative clauses to pre-intermediate and intermediate students of Czech at levels B1–B2, it is not feasible to provide exhaustive information about the use of the pronoun *který*. Hence, most of the textbooks limit themselves to an overview of the declension of this pronoun.[14] Surprisingly, even the textbook *Čeština pro pokročilé* (Czech for advanced students) [5] provides only the paradigms of relative pronouns *který* and *jenž.*

The textbook *Český krok za krokem–B1* (Czech Step by step – B1) points out that in common Czech the noninflected pronouns *co* (what) and *jak* (how) are used instead of the standard Czech pronoun *který*. It also notes the difference between the use of relative pronouns *kdo* (who) and *který* (that), as illustrated by the example (23).

23. Je tady někdo, **kdo** nepije kávu? – To je ten člověk, **který** nepije kávu? [12, p. 163]

[13] The pronoun *jaký* is used to speak about quality.
[14] For instance, [21], [2], [15], [17].

24. "Is there anybody who does not drink coffee? – Is this the person that does not drink coffee?"

The textbook *Čeština pro středně a více pokročilé* (Czech for intermediate and more advanced students) explicitly mentions that "the possessive relative pronoun cannot be replaced by the relative pronoun *který*"[15] [4, p. 193] and also notes the difference in the use of pronouns *kdo/co* and *který* with non-specific and specific reading, as shown in the example (24).

25. Ten, **kdo** se občas napije, není ještě alkoholik. – Ten student, **který** to udělal, nemohl být při smyslech. [4, p. 193]

    "Who drinks a glass from time to time is not an alcoholic. – The student that did it could not be conscious."

However, the analyzed textbooks[16] neither point out the postposition of the relative clause after the modified noun nor focus on the use of relative adverbs.

5 The Pronoun *který* in Grammars of Czech (for Foreigners)

Neither Czech grammars provide an exhaustive description of structures with relative pronouns. Apart from the respective paradigms they note only the existence of colloquial noninflected relatives *co* (what) and *jak* (how).[17] *Příruční mluvnice češtiny* (Handy grammar of Czech) and *Mluvnice češtiny 3* (Grammar of Czech 3) draw attention to the correct use of the possessive relatives *jehož, jejíž, jejichž*. *Mluvnice češtiny 3* also advices in detail how relatives are to be used in collocation with different pronouns (personal pronouns, *někdo* (someone), *něco* (something), *všechno* (everything), *všichni* (everybody), *každý* (each) etc.). *Příruční mluvnice češtiny* limits itself to the remark that "relative attributive clauses related to substantive pronouns *někdo* and *něco* are introduced by the relatives *kdo* and *co*"[18] [16, p. 494].

A comprehensive account of relative expressions is given in the grammar of Czech for foreigners *Čeština – jazyk cizí* (Czech – foreign language) [19]. This handbook points out the correct use of possessive relatives. It also focuses on the use of the relative pronoun *kdo* (who) with the pronouns *ten* (that), *každý* (each) a *všichni* (everybody) and the use of the relative *co* (what) with pronouns *to* (it) a *všechno* (everything).

However, none of the grammars mentions the structure *většina z nichž* (the majority of which) dealt with in subsection 3.2. Although this collocation is unproblematic for native speakers, the data from CzeSL indicate that it is an error made frequently by non-native speakers of Czech.

[15] „přivlastňovací vztažné zájmeno nemůže být nahrazeno vztažným zájmenem *který*"

[16] See References.

[17] Cf. [11], [10], [16], [7], [6].

[18] „vztažné přívlastkové věty vztahující se k substantivním zájmenům typu *někdo* a *něco* jsou uvozeny relativy *kdo* a co."

6 Language Errors and Teaching of Czech

While teaching methods in the past considered errors as a defective and undesirable phenomenon,[19] today's most widespread communicative method benefits from them: "An adequate analysis of an error (its typology and specific information value) becomes a valuable feedback both for the teacher and the non-native speaker and an outstanding source of relevant information. An error is not any more thought of negatively, on the contrary it is considered as a natural phenomenon, as an inevitable and integral part of a complex process of acquiring a foreign language code."[20] [14, p. 101]

By virtue of the existence of learner corpora the awareness of types and frequency of language errors is no more limited to personal experience of a teacher and it may be objectively checked in a balanced set of texts. The corpus makes possible to carry out practical researches focused on particular phenomena in a broader language context (cf. [23]). It thus offers great possibilities not only for teaching foreign languages, but especially for creating textbooks and handbooks for non-native speakers.

The explanation of grammar should be precise and correspond by its extent to the language level of students. The corpus allows taking into account problematic aspects of a particular language phenomenon with regard to the level and the first language of non-native speakers: "Just mapping the phenomena difficult for particular groups of learners would already be useful for Czech. Since in the process of material presentation it is very difficult to abstract from the description of the grammar system from the point of view of a native speaker, the learner corpus should be the basis that could help to radically change this situation."[21] [23, p. 134]

Learner corpus is a rich source of data that may help to recognize the errors occurring in the process of language acquisition. In the virtue of the metadata[22] contained within the corpus it will allow a number of studies aimed at particular language phenomena and different groups of non-native speakers.

References

[1] Adamović, A. and Holub, J. (2005). *Čeština jako cizí jazyk: jak se učíme jazykům, jak je vyučujeme a jak v jazycích hodnotíme*. Univerzita Karlova, Praha.
[2] Adamović, A., Ivanovová, D., and Hrdlička, M. (2007). *Basic Czech II*. Karolinum, Praha.
[3] Adamović, A. and Hrdlička, M. (2010). *Basic Czech III*. Karolinum, Praha.

[19] For an overview of teaching methods from the Middle Age until present see [13].

[20] „Adekvátní rozbor chyby (její typologie a specifická výpovědní hodnota) se tak stává pro vyučujícího i pro jinojazyčného mluvnčího cennou zpětnou vazbou a nenahraditelným zdrojem relevatních informací. K chybě se již nepřistupuje negativně, hodnotí se naopak jako přirozený jev, jako nevyhnutelná a integrální součást složitého procesu nabývání znalosti jinojazyčného kódů (…)."

[21] „Pro češtinu by však mělo být přínosné už samo zmapování obtížných jevů pro jednotlivé skupiny žáků. Protože při materiálové prezentaci je velmi obtížné odhlédnout od popisu gramatického systému z hlediska rodilého mluvčího, žákovský korpus by měl být právě tím podkladem, který by tuto situaci mohl pomoci zásadně změnit."

[22] Metadata are data about the text, its origin and the way it was gathered, about the learner, his first language, the length of study of the target language etc., cf. [22]. The publication of the metadata in CzeSL is envisaged in near future.

[4] Bischofová, J. (2011). *Čeština pro středně a více pokročilé*. Karolinum, Praha.
[5] Confortiová, H. and Turzíková, M. (2008). *Čeština pro pokročilé.* Karolinum, Praha.
[6] Cvrček, V. (2010). *Mluvnice současné češtiny*. Karolinum, Praha.
[7] Čechová, M. (2000). *Čeština – řeč a jazyk*. ISV nakladatelství, Praha.
[8] Czech National Corpus – SYN2010 (2010). Prague: Institute of the Czech National Corpus, Faculty of Arts, Charles University in Prague. Accessible at: `http://www.korpus.cz`.
[9] Czech National Corpus – CzeSL-plain (2013). Prague: Institute of the Czech National Corpus, Faculty of Arts, Charles University in Prague. Accessible at: `http://www.korpus.cz`., retrieved 9 May 2013.
[10] Daneš, F. (1987). *Mluvnice češtiny 3. Skladba*. Academia, Praha.
[11] Havránek, B. and Jedlička, A. (1981). *Česká mluvnice*. Státní pedagogické nakladatelství, Praha.
[12] Holá, L. and Bořilová, P. (2009). *Česky krok za krokem 2 (B1)*. Akropolis, Praha.
[13] Hrdlička, M. (2009). *Gramatika a výuka češtiny jako cizího jazyka: k prezentaci gramatiky českého jazyka v učebnicích češtiny pro cizince*. Karolinum, Praha.
[14] Hrdlička, M. (2012). *Jazyková chyba a práce s ní v jazykovém vyučování*. In *Čeština – cílový jazyk a korpusy*, pages 89–108, TUL, Liberec.
[15] Hronová, K. and Hron, J. (2008). *Čeština pro cizince: Czech for foreigners: A1-A2, B1: mini/medium*. Didakta, Praha.
[16] Karlík, P., Rusínová, Z., and Nekula, M. (1996): *Příruční mluvnice češtiny*. Nakladatelství Lidové noviny, Praha.
[17] Kestřánková, M., Šnaidaufová, G., and Kopicová, K. (2010). *Čeština pro cizince, úroveň B1*. Computer Press, Brno.
[18] Petkevič, V. et al. (2012). *Anotace chybových textů v českém žákovském korpusu*. In *Čeština – cílový jazyk a korpusy*, pages 61–88, TUL, Liberec.
[19] Poldauf, I. and Šprunk, K. (1968). *Čeština jazyk cizí: mluvnice češtiny pro cizince*. Státní pedagogické nakladatelství, Praha.
[20] Remediosová, H. and Čechová, E. (2005). *Chcete mluvit česky?* Harry Putz, Liberec.
[21] Rešková, I. and Pintarová, M. (1999). *Communicative Czech: Intermediate Czech.* Karolinum, Praha.
[22] Šebesta, K. (2012). Parametry žákovských korpusů a CzeSL. In *Čeština – cílový jazyk a korpusy*, pages 13–34, TUL, Liberec.
[23] Škodová, S. (2012). Nástin využití žákovských korpusů pro jazykové vyučování. In *Čeština – cílový jazyk a korpusy,* pages 125–138, TUL, Liberec.
[24] Štindlová, B. (2012). Chybové taxonomie a možnosti chybové anotace v žákovských korpusech. In *Čeština – cílový jazyk a korpusy,* pages 35–60, TUL, Liberec.

Delimitation of Participles in the Manual Morphological Annotation

Agáta Karčová

Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences, Bratislava, Slovakia

Abstract. This paper is concerned with annotation of participles in the manually annotated subcorpus r-mak (1.0–4.0) of the Slovak National Corpus. Passive and active participles are classified as special non-finite verbs by Slovak grammarians, but in the Slovak National Corpus these are classified as a separate part of speech. Boundaries between participles and other word classes are not always clear, thus they sometimes migrate to the categories of adjectives and nouns. Separating participles from adjectives and nouns was one of the most difficult problems in the manual annotation of the subcorpus r-mak. The paper also gives attention to borderline cases (homonyms, word forms created by analogy) and words where we had to modify formal criteria for their classification (words with notable semantic shift from the motivating word; substantivised participles).

1 Participles in General

Participles comprise a specific part of speech class in Slovak, characterised by its hybrid form and content. Active and passive participles are traditionally classified as non-finite verbs[1]. Participles, as well as verbs, exhibit grammatical categories of intention and aspect, but they also have to agree with the noun they modify in gender, number and case. Semantically, participles exhibit dynamic features; their meaning originates in the motivating verb. In many cases they can be shifted semantically, so the direct link between "motivating word" (verb) and "motivated word" (participle) weakens. Thereby, the semantic aspect of a participle may be reduced, or rather the participle may show static properties of entities so that the participle becomes an adjective. The form of participles is identical to the adjectives (corresponding with the adjective-like declension paradigms of *pekný* and *cudzí*) and the syntactic function of participles is identical to the adjectives (they can be used either as an agreeing attribute, predicate nominal or complement). This is the reason that the process of participle adjectivisation is frequent. Some participles may undergo the process of conversion into a substantive (substantivisation). Thus, substantivized participles often denote persons of specific functions or roles, e.g. *cestujúci* (traveler, an active participle of *to travel*), *vedúci* (leader, active participle of *to lead*), *účinkujúci* (acting, active participle of *to act*), *obžalovaný* (accused, a passive participle of *to accuse*), etc. Therefore, boundaries between adjectives and participles are rather unclear.

The classification of participles (most notably, *-n-/-t-*[2], i.e., passive participles) is a perennial problem in Slovak linguistics. The most precise description is given in [13], [8] and [14]. As stated in [13, p. 495], the most appropriate name would be *-n-/-t- príčastie* (-n-/-t-

[1] neurčitý slovesný tvar

[2] participle derived by -n-/-t- suffix

participle) since the meaning of formally passive participles is often not passive and often they are not used in passive constructions. Based on comprehensive research, J. Sejáková [14] agrees with the term and defines, in addition, the new term *n/t-ová jednotka* (-n-/-t-unit) to refer to the lexemes that are difficult to classify.

A detailed paradigmatic (including lexis, semantics, word-formation and grammar) and syntagmatic analysis is needed to classify the part of speech category of word ending with *-ný (-ený) / -tý, -iaci (-aci) / -úci*. J. Sejáková [14, pp. 31–34] uses the terms *pól adjektívnosti* (pole of adjectivity) and *pól slovesnosti* (pole of verbality), which demonstrated fuzziness of categorization. The authors of monograph [8, p. 209] give examples *ohnutý chrbát* (curved back) that illustrates that the word *ohnutý* may express either an inherent quality (only seemingly a consequence of an action) or an acquired feature (permanent or temporary quality that is caused by an action).

The Slovak vocabulary includes adjectives suffixed by -n-/-t-, which have a similar form as adjectival participles. In some cases, a transformational test has shown that adjectives can be easily differentiated from participles (as stated by Sejáková [14]), e.g. *novopostavený dom* (newly built house), *ukričaná žena* (rambunctious woman), *predpojatý človek* (prejudiced person), *sčítaný študent* (well-read student). Either their assumed motivating verb does not exist (**predpojat'*), or the word begins with a prefix or prefixoid that does not occur with its motivating verb (**novopostavit'*) or the word does not correlate with a motivating verb (no aspect congruence, no semantic congruence, etc.).

If it is possible to derive an adverb or an abstract noun from an participle, or the participle can undergo the formation of comparative and superlative, then this indicates its adjectivisation. E.g. from the adjectivised participle *unaven-ý* (tired) we can form an adverb *unaven-e* (tiredly), abstract substantive *unaven-ost'* (tiredness), comparative *unaven-ejší* (more tired), superlative *naj-unaven-ejší* (most tired) [see: 13, p. 502]. This is not a general rule (it affects only some adjectivised participles), therefore it cannot be used as a generally valid criterion for delimitation.

2 Participles in the Slovak National Corpus

2.1 Frequency of Participles

Participles are quite frequently used in Slovak written texts. The manually morphologically annotated subcorpus of the *Slovak National Corpus* (SNK) called *r-mak* was created at the Slovak National Corpus Department of the Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences. The current 4th version contains about 1.2 million tokens and was released in 2013.[3]

The number of participles in the corpus *r-mak-4.0* (1,199,326 tokens) is 16,332 hits (*Query:* `[tag="G.*"]`). They represent 1.36% of all the tokens (including the non-word ones, such as punctuation or numerals). The number of unique participles (word forms) is 8,796 (6.41% of all the unique word forms), which gives 3,675 unique lemmas (6.74% of all the unique lemmas). The subcorpus *r-mak-4.0* contains predominantly fiction and journalistic texts. The portion of professional texts is lower (19.0%), which may affect the

[3] Further information on the corpus size and text types can be found on the website of the Slovak National Corpus, Department of Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences (`http://korpus.juls.savba.sk/stats.html`).

statistical results (the greatest concentration of participles was expected in professional texts). The largest corpus of Slovak (at the time of writing) *prim-6.0* contains 1.6% participles in professional texts (subcorpus *prim-6.0-public-prf*), 1.07% in journalistic texts (prim-6.0-public-inf) and 1.05% in fiction (*prim-6.0-public-img*).[4]

In conclusion, when querying the subcorpus r-mak-4.0 we get 13,557 hits of passive participles with the suffixes *-ný/-tý* and only 2,775 hits of active participles with the suffixes *[-ú/-u/-ia]ci*. The number of words belonging to different parts of speech categories in the subcorpus *r-mak-4.0* is illustrated in the Figure 1.[5]

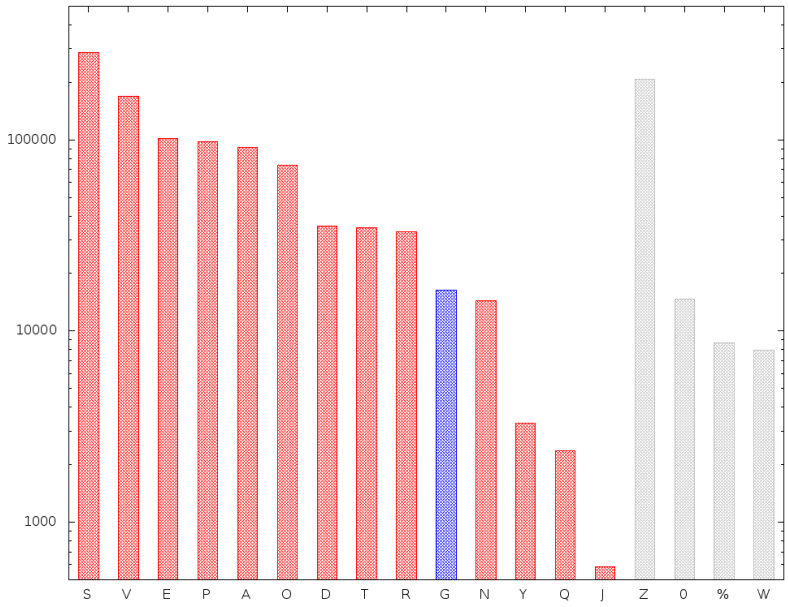


Fig. 1. Total number of tokens in the subcorpus r-mak-4.0 by POS

S – nouns, V – verbs, E – prepositions, P – pronouns, A – adjectives (91 577 hits), O – conjunctions, D – adverbs, T – particles, R – reflexive morphemes *sa*/*si*, **G – formal participles (16,332 hits)**, N – numerals, Y – conditional morpheme *by*, Q – undefinable part of speech, J – interjections, Z – punctuation, 0 – numbers, % – citation, W – abbreviations

2.2 Annotation of Participles in the SNK

Lemmatization and morphological tagging are an important part of a corpus. In undertaking the corpus research one needs to consider the reliability of a tagger [see: 9, p. 169]. The manually annotated corpus *r-mak* is assumed to be virtually error-free. Errors were kept to

[4] Corpus *prim-6.0* was annotated automatically, therefore the provided analysis is only approximate.

[5] Шимкова [17, p. 391] gives an overview of the frequency of parts of speech classes in the first three versions of the subcorpus *r-mak (1.0, 2.0, 3.0)*.

a minimum by 3-level control using the semi-automated tools [11]. The corpus has been manually disambiguated by two annotators. Their results were automatically compared and manually corrected [5, p. 61].

Naturally, the variability of participles and their unclear classification affected the way of their manual annotation. There were several possibilities for classifying participles, each of them presenting its advantages and disadvantages: 1. to establish an adjective-like class including adjectives as well as participles (chosen by, e.g. the Czech National Corpus; [7]), not distinguished in any way (with a few minor exceptions); 2. to consider paradigmatic and syntactic features for the *-n-/-t-* and *-iaci-/-úci* units. In this case, the annotation would have taken too long (given the nature and size of the annotation); 3. to establish a special formal class of participles considering their form and derivation only (synchronic approach). "We consider the participles to be a separate part of speech class, not a declined form of a verb – while definitely possible, this would lead up to some singular categorization, e.g. verbs with case" [6, p. 56] Morphological annotation in the SNK is based on formal morphology and the combination of attributive and positional systems of morphological tagging [see 17, pp. 387–388]. The selected criterion follows traditional classification of participles as non-finite verbs. This also allows easy searching for formal participles and gives reasonably precise information on the number of adjectives and formal participles.

During application of the formal approach, we observed that in the heterogeneous group of words suffixed by *-ný (-ený) / -tý, -iaci (-aci) / -úci* there are units that have to be disambiguated and require greater effort when being classified. Delimitation of participles from adjectives and nouns was one of the most difficult problems of the whole manual annotation. Our goal was to further specify the annotation of participles to get unified and more logical system. In the paper we discuss such borderline cases.

The user guide *Tokenizácia, lematizácia a morfologická anotácia Slovenského národného korpusu* ([4]; hereinafter User guide) was designed with a goal to build an annotated subcorpus (the User guide was used for all the subcorpus versions from 1.0 to 4.0). The User guide provides a description of tokenization, lemmatization and morphological annotation. "The tagset is highly functional and pragmatic, although some allowances had to be made to accommodate the traditional analysis of Slovak morphology and part of speech categories" [6, p. 41]. The tags are of various length, but the order of characters is obligatory. The tagset covers the traditional 10 part-of-speech categories and several non-word categories (19 categories in total). There were many borderline cases that had to be specified, e.g. verbal nouns, nouns with adjectival paradigms as well as participles (which are a separate part of speech category).

"The borderline cases are as follows: … 3. Active and passive participles as well as deverbalised adjectives. We classify these cases as separate part of speech category (`G`) – *písaný, otvorený, obutý, píšuci, hrajúci, stojaci*. Formal participles are distinguished from adjectives on the grounds of their form and origin" [4, p. 5]. This is explained as follows: "Participles (*prestretý, zívajúci* – `G`) are considered to be a separate part-of-speech category because they are dynamic and often unclear position in between both adjectives and verbs" [4, p. 7].

Following the guidelines, a given participle is marked `[k]` for active and `[t]` for passive. The categories for gender, number and case congruence use the same characters as for the adjectives. Participles can also exhibit a degree of comparison. Positive or irrelevant degree is marked `[x]`, whereas comparatives `[y]` and superlatives `[z]` occur rather rarely.

category	category type	tag character
part of speech	participle	G
type	active	k
	passive	t
gender congruence	masculine animate	m
	masculine inanimate	i
	feminine	f
	neutral	n
number congruence	singular	s
	plural	p
case congruence	nominative	1
	genitive	2
	dative	3
	accusative	4
	vocative	5
	locative	6
	instrumental	7
degree	positive	x
	comparative	y
	superlative	z

Table 1. Composition of a 'participle' tag

2.3 Cases Requiring Disambiguation

2.3.1 Homonymy of words and word forms

It was difficult to distinguish homonymous lexemes using the formal criterion. To classify the part of speech category of these homonymous lexemes, one should carefully consider the context and origin of a word, e.g. whether a word is derived synchronically from a verb – e.g. the word *rafinovaný*. The expression *rafinovaný cukor* (refined sugar; derived from the verb *rafinovať*) is an -n-/-t- participle, assuming that it was motivated by the verb *rafinovať* (refine). The expression *rafinovaný človek* (cunning man) is an adjective, because the verb *rafinovať* does not exist with corresponding meaning. The surrounding text together with the meaning of the lexeme determine the part of speech classification.

In analogical cases requiring disambiguation, the formal approach classifies lexemes corresponding to the usual conception of Slovak morphology, e.g. (1) *nesúci* človek (unsuitable person) and (2) človek *nesúci* drevo (person carrying wood, derived from the verb *niesť*); (1) *zvrátený* človek (immoral, scrofulous man) and (2) *zvrátený* beh udalostí (reversed action, derived from the verb *zvrátiť*); (1) *rezervovaný* človek (reserved, shy person) and (2) *rezervovaná* vstupenka (reserved ticket, derived from the verb *rezervovať*); (1) *sčítaný* študent (well read, educated student; from the verb *čítať*, the verb *sčítať* does not have this meaning) and (2) *sčítaná* suma (sum in total, derived from the verb *sčítať*). One advantage of the formal approach is that the existence of corresponding verbs can be easily verified, for some compounds and words with prefixoids we get a disparity, e.g.: *pracujúci* (worker) > *pracovať* (to work), *spolupracujúci* (collaborator) > *spolupracovať* (to collaborate), but: *cestujúci* (traveler) > *cestovať* (to travel), *spolucestujúci* (fellow-traveller) > **spolucestovať* (*to fellow-travel).

2.3.2 Delimitation of Active and Passive Participles from Adjectives

During the process of annotation several questions arose. We had to decide how to tag deverbalized words, but which have slightly different meanings. For instance, *vynikať* (be very good, be different from others) > *vynikajúci* (excellent, outstanding), *skúsiť* (to try) > *skúsený* (experienced), *poľahčiť* (to make easier) > *poľahčujúci* (mitigating), etc. Disambiguation also had to be performed on words that either share similar written forms as participles but are not derived from verbs or have either archaic or uncommon verb as their motivating word, e.g. *disciplinovaný* (disciplined) > *disciplinovať* (to discipline), *nadudraný* (sulky) > *nadudrať sa* (to sulk), *okrídlený* (winged) > *okrídliť* (to wing), *livrejovaný* (liveried) > **livrejovať* (to livery), *mrežovaný* (latticed) > **mrežovať* (to lattice), etc. This is a quite natural phenomenon in a language: "There is a group of words in between the motivated and completely demotivated lexemes. Their motivation is rather unclear at present, so there is an inconsistency between the genetic and synchronic motivation" [3, p. 25]. In this case, the diachronic aspect of the language is notably significant because the motivated word can still exist in the language even if its motivating word is an out-of-vocabulary or uncommon word, e.g. *slýchať* (to hear) > *neslýchaný* (outrageous). In certain cases, only a thorough study will show if a motivating word had ever occurred in a language and if a word form was created by analogy (to existing word forms), e.g. *melírovaný* (streaked), *premrštený* (exorbitant), *opodstatnený* (justified).

The formal approach is focused on the synchronic aspect of language, but the decision how to classify word forms with an unclear synchronic motivation had to be made. In ambiguous cases, annotators (including the author) have followed predominantly formal criteria. Generally, we tagged words as participles if there was a clear motivating verb, including any out-of-vocabulary, rarely used or semantically marked verbs which are nevertheless corresponding in meaning. When considering the correspondence between participle and verb, minor discrepancies were allowed. Semantic correspondence was considered crucial. Once the meaning of a lexeme markedly differs from the meaning of the motivating verb, e.g. *skúsiť* (to try) > *skúsený* (experienced) etc., the word was not considered to be a participle, strict formal criteria would have lead to distorted conclusions. But the non-congruence of aspect was admitted, for instance, the word *varený* (meaning *just cooked*, expressing finished action) was considered to be a participle, although the imperfective aspect of the source verb *variť* (to cook) expresses an unfinished action.

Despite many factors influencing the disambiguation, there was only a small number of disagreeing tags. The following active participles were tagged incorrectly: *horiaci* (burning), *raziaci* (punching), *školiaci* (training), *vládnuci* (ruling), *svetielkujúci* (luminous), *vzývajúci* (invoking), *lietajúci* (flying), *jasajúci* (exultant), *žiadajúci* (requesting). Passive participles were more often derived from verbs in perfective aspect: *neoverený* (untested), *nevyriešený* (unresolved), *obnosený* (worn), *prikovaný* (transfixed), *pokrčený* (crumpled), *roztvorený* (unfolded), which seems logical because words with resultative meaning tend to behave like adjectives which denote static features. The errors could have been made accidentally or by analogy. While correcting these cases, formal criteria have been applied. For all these words there exists a corresponding motivating verb with an identical meaning, e.g. *skúmať* (to examine) < *skúmaný* (examined).

2.3.3 Delimitation of Active and Passive Participles from Substantives with Adjective-like Paradigm

Disambiguation of participles and substantives is based on determining the syntactic function of a word in a certain syntagm. Some of the passive and active participles became nouns (in the process of substantivisation) so that they are classified as nouns in lexicographical works such as *Krátky slovník slovenského jazyka* [10], *Slovník súčasného slovenského jazyka* [1], [2], etc. There is an unclear distinction not only between participles and adjectives but also between participles and substantivized participles. Many lexemes have become nouns, e.g. *vedúci/vedúca katedry* (head of department), in both the masculine and feminine gender. There is a group of words classified as participles which are homonymous with substantivized participles most often in the role of an agreeing attributive, e.g. *muž vedúci vozidlo* (a man driving a vehicle). Apart from this, there are several words where the process of substantivization is still ongoing.

The subcorpus *r-mak-4.0* contains words which were sometimes tagged as active participles and at another time as nouns: *neveriaci* (doubting), *trpiaci* (suffering), *veriaci* (faithful), *vidiaci* (sighted), *vedúci* (leading), *kupujúci* (buying), *cestujúci* (traveler), *umierajúci* (dying), *pracujúci* (working), *účinkujúci* (performing), *vystavujúci* (exhibiting). Although it is not large, the overall number of occurrences in the corpus is non-negligible (*r-mak-4.0* subcorpus contains 79 occurrences of the word *vedúci* (head) and 26 for *pracujúci* (worker)). A significant amount of word forms was assigned to substantivized participles, such as: *veriaci* (believer; 40, 14), *vedúci* (head; 35, 44), *cestujúci* (traveler; 14, 4), *umierajúci* (dying; 3, 3) and *pracujúci* (worker; 3, 23).[6]

In most cases, the part-of-speech tagging conformed with the syntactic function of words. Examples include the word *cestujúci* (traveler), used either with a superordinate noun *cestujúca osoba* (traveling person) or as part of a predicate nominal or a complement. The word has been always tagged as a active participle; its usage in the role of subject or object led to its systematic tagging as a noun with an adjective-like paradigm:

`Gk`:

Osoba *cestujúca* / `Gkfs1x` rýchlosťou blízkou rýchlosti svetla by videla, že farba svetla vpredu …
'A person *traveling* at the speed of light would have noticed that the colour of light ahead...'

Ich kazatelia plnili funkciu *cestujúcich* / `Gkmp2x` spovedníkov a učiteľov
'Their preachers were in the function of *traveling* confessors and teachers)'

`SA`:

Povedali nám, že ďalej smú len *cestujúci* / `SAmp1`
'They told us that only *travelers* are permitted to go on'

Oslobodenie od dovozného cla v prípade alkoholu a tabakových výrobkov sa nepriznáva *cestujúcemu* / `SAms3` mladšiemu ako 18 rokov
'Purchase of duty-free alcohol and tobacco products is denied for *travelers* under 18'

There was a tendency to classify word forms in the role of subject or object as nouns, with the same word forms in other roles being classified as participles:

[6] First number in brackets indicates number of nouns with adjective-like paradigm, second number shows number of participles.

`Gk`:

veriaci moslimovia / ženy / kresťania / katolíci / prírodovedec
'*believing* Muslims / women / Christians / Catholics / biologist'

vedúci predstaviteľ / osobnosť / gól / postavenie
'*leading* representative / person / goal / status'

pracujúci osoba / médiá / mládež / otrok
'*working* person / media / youth / slave'

`SA`:

usmerňoval *veriacich* v zboroch
'he guided the *believers* in choirs'

vedúci katedry
'*head* of department'

pracujúci vyšli z tovární a úradov
'*workers* stepped out of the factories and offices'

The subcorpus *r-mak-4.0* contains 14 pairs of formally the same passive participles and nouns (substantivized participles) which can be classified as participles in a certain context. The words often tagged as nouns are as follows: *nezamestnaný* (unemployed; 10, 16), *obvinený* (accused; 5, 17), *poškodený* (damaged; 5, 11), *ranený* (hurt; 2, 5) and *unesený* (kidnapped; 3, 12)[7].

Both active participles and substantivized passive participles may change their part-of-speech category; derivationally therefore they are considered morphologically motivated lexemes [see 12, p. 20]. Some passive participles have been converted into nouns without any changes in their form. They differ semantically, participles convey state or action, substantivized participles refer to the entity related to the state or action. The words have taken on a new syntactic function. These word forms may adopt behaviour of either participles or substantivized participles.

Some of the morphologically motivated nouns derived from participles can be used only to a limited extent, e.g. *poškodený, obvinený, unesený*. The nouns are usually used in legal texts or historical legal texts. Otherwise, they fulfill the function of a participle.

Examples include:

`Gt`:

Máš *poškodenú* pamäť
'Your memory is *damaged*'

text je silne *poškodený*
'the text is very *damaged*'

`SA`:

Kým v predošlom právnom systéme sa *poškodený* / `[SAms1]`, resp. jeho príbuzenstvo snažili, …
'While in the previous legal system the *victim* or his relatives made efforts …'

Part-of-speech tagging was influenced by the syntactic function of words. This function reflects semantic shifts and is influential when participles are converted into nouns and vice versa. This retroactively affects the morphological nature of words.

[7] see footnote No. 6

3 Conclusions

This paper has briefly characterized the word class of participles which are usually classified as non-finite verbs. The class includes words which can adopt either adjective-like behaviour or can behave like nouns with an adjectival paradigm. This frequent POS conversion is assisted by the mutual semantic features of the words, as well as their forms being identical. After conversion, the lexemes also gain the possibilities to form comparative and superlative forms or to have adverbs or nouns derived from them.

We describe the rules of participle annotation in the subcorpus *r-mak*. We have shown several options for tagging participles. With hindsight (after comparing all versions of the morphologically annotated subcorpus *r-mak* 1.1-4.0), we can confirm that the formal approach to various groups of words requiring special treatment (polysemous words, words with unclear POS classification, words with unclear motivation, etc.) has been appropriately selected.

We have given a brief description of special cases, such as homonymous lexemes classified according to semantic features, e.g. lexeme *nesúci* (carrying), or analogically created words that are not participles because their motivating verb does not exist, e.g. *cestujúci* (traveler) > *cestovať* (travel), *spolucestujúci* (fellow-traveler) > **spolucestovať* (*fellow-travel). The formal approach was applied to a limited extend, therefore we separated such cases in which the semantic shift of participle from the meaning of the motivating verb was prominent, e.g. *skúsiť* (try) > *skúsený* (experienced, having an ability).

The aim of this research was to examine words sharing the same form but differing in their POS category. At the boundary between nouns and participles were the following words: *veriaci, vedúci, cestujúci, umierajúci, pracujúci, nezamestnaný, obvinený, ranený, unesený*; and at the boundary between adjectives and participles were: *pokrčený, roztvorený* and *skúmaný*. In delimitation of participles and substantivized participles it was necessary analyse wider context because the syntactic function of words is decisive.

The morphological annotation provides concise information about the morphological features of each words. A word can be classified as a participle if it shares a similar or identical meaning with its motivating verb. The formal approach enabled a logical and precise annotation of this variable part-of-speech category of participles. The obtained results might be used in the further synchronic or diachronic research concerning all functions of words (semantic, word-formation, syntactic, etc.).

References

[1] Buzássyová, K. and Jarošová, A., editors (2006). *Slovník súčasného slovenského jazyka. A – G*. VEDA, Bratislava. 1134 p.

[2] Buzássyová, K. and Jarošová, A., editors (2011). *Slovník súčasného slovenského jazyka. H – L*. VEDA, Bratislava. 1087 p.

[3] Furdík, J. (2004). *Slovenská slovotvorba (teória, opis, cvičenia)*. Náuka, Prešov. 200 p.

[4] Garabík, R., Gianitsová, L., Horák, A., and Šimková, M. (2004). Tokenizácia, lematizácia a morfologická anotácia Slovenského národného korpusu. Retrieved from `http://korpus.juls.savba.sk/publications.html` on 1 September 2013.

[5] Garabík, R. and Gianitsová-Ološtiaková, L. (2005). Manual Morphological Annotation of Slovak Translation of Orwell's Novel 1984 – Methods and Findings. In *Computer Treatment of Slavic and East European Languages*, pages 59–66, VEDA, Bratislava.

[6] Garabík, R. and Šimková, M. (2012). Slovak Morphosyntactic Tagset. *Journal of Language Modelling*, 0(1):41–63.

[7] Hajič, J. (2000). Popis morfologických značek – poziční systém. Retrieved from `http://www.korpus.cz/bonito/znacky.php` on 1 September 2013.

[8] Horecký, J., Buzássyová, K., Bosák, J., et al. (1989). *Dynamika slovnej zásoby súčasnej slovenčiny*. Vydavateľstvo Slovenskej akadémie vied, Bratislava. 436 p.

[9] Jelínek, T. (2008). Morfologické značkování a lemmatizace v korpusech ČNK. In *Gramatika a korpus 2007*, pages 169–179, Academia, Praha.

[10] Kačala, J., Pisárčiková, M., and Považaj, M., editors (2003). *Krátky slovník slovenského jazyka*. VEDA, Bratislava. 985 p.

[11] Карчова, A. and Шимкова, M. (2006). Морфологічна анотація текстів словацького національного корпусу. In *Лексикографічний бюлетень 13.*, pages 71–76, Інститут української мови Національної академії наук України, Київ.

[12] Ološtiak, M. (2009). Spolupráca slovotvornej motivácie s inými typmi lexikálnej motivácie. *Jazykovedný časopis*, 60(1):13–34.

[13] Ružička, J., editor (1966). *Morfológia slovenského jazyka*. Vydavateľstvo SAV, Bratislava. 895 p.

[14] Sejáková, J. (1995). *Adjektivizácia n/t-ových príčastí v súčasnej slovenčine*. PhD thesis, JÚĽŠ SAV, Bratislava. 248 p.

[15] Slovak National Corpus – prim-6.0-public-all. (2013). Bratislava: Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences. Accessible at: `http://korpus.juls.savba.sk/`.

[16] Slovak National Corpus – r-mak-4.0. (2009). Bratislava: Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences. Accesible at: `http://korpus.juls.savba.sk/`.

[17] Шимкова, M. (2008). Морфологическая разметка частиц речи в Словацком национальном корпусе и возможности её использования в процессе создания толкового словаря. In *Труды международной конференции «Корпусная лингвистика – 2008»*, pages 387–395, Издательство Санкт-Петербургского университета, Санкт-Петербург.

Corpus Based Identification of Czech Light Verbs

Václava Kettnerová[1], Markéta Lopatková[1], Eduard Bejček[1], Anna Vernerová[1], and Marie Podobová[2]

[1] Faculty of Mathematics and Physics, Charles University in Prague, Czech Republic

[2] Institute of Slavonic Studies, The Academy of Sciences of the Czech Republic, Prague, Czech Republic

Abstract. In this paper, we describe a corpus based experiment focused on the possibility of identifying Czech light verbs. The experiment had two main aims: (i) to establish the inventory of Czech light verbs entering into combinations with predicative nouns, and (ii) to verify the adopted criteria for distinguishing light usages from full usages of the given verbs. As for the inventory of light verbs, we propose and verify the hypothesis that the possibility of light usages of verbs is related to their semantic class membership rather than to their high frequency. In the second part of the experiment, we exploited the compiled inventory of Czech verbs inclined to occur as light verbs. The criteria adopted for distinguishing light usages of a verb from its full ones were applied to selected corpus sentences with these verbs. Three annotators were asked to determine whether a verb occurrence in an extracted corpus sentence corresponds to the full or to the light usage of the given verb. The feasibility of this task has been proven by the achieved κ_w statistics 0.686 and by the inter-annotator agreement 85.3%. As a result of this experiment, we obtained 893 verbonominal combinations of Czech light verbs and predicative nouns. These combinations will be further utilized for the lexicographic representation of these phenomena.

1 Introduction

As a verb represents the most important syntactic unit in a language, the description of its syntactic structure belongs to the primary tasks of both theoretical and computational linguistics. At present, the basic information on syntactic behavior of verbs is provided in many lexical resources. However, the description of advanced syntactic properties of verbs is still missing. Light verbs belong to such advanced linguistic phenomena. In this case, the syntactic structure of a sentence is not solely determined by a verb alone but also by a predicative noun with which the verb combines. Predicative nouns exhibit two characteristic properties: (i) they denote an event meaning, e.g. a process, a state, or a property, not a resultative meaning, and (ii) they are characterized by a set of valency complementations. See the following examples with the verb *nést* 'to carry' (1)–(2). Unlike the syntactic structure with the full usage of the verb in (1), the syntactic structure with the light usage of the verb (2) is affected also by the predicative noun *odpovědnost* 'responsibility'. This predicative noun expresses 'the state of being responsible' and contributes its valency complementation *za bezpečnost* 'for security' to the resulting syntactic structure.

(1) *Učitel nese sešity.*
'The teacher is carrying exercise books.'

(2) *Učitel nese odpovědnost za bezpečnost žáků.*
the teacher – carries – responsibility – for the security of pupils
'The teacher is responsible for the security of (his) pupils.'

While easily mastered by native speakers, light verbs pose a serious challenge for foreign speakers as well as for automated language processing (esp. for machine translation, information extraction, information retrieval, question answering, etc.) [11]. It has been recognized for a long time that both language learning and NLP tasks would be facilitated by a lexical resource providing an explicit and systematic representation of these phenomena. Such a representation should be based on a thorough theoretical analysis of light verbs.

Despite being subject to many analyses, many aspects of light verbs are still not clear. Even a clearcut definition of light verbs as a specific group of verbs is lacking [6]. In accordance with [2], we consider light verbs as a specific usage of a verb that loses individual semantic properties and retains only some of semantic facets of its full verb counterpart. To acquire semantic capacity, a light verb combines with a predicative noun which contributes its individual semantic features to a resulting complex predicate. From this point of view, each verb can potentially be used as a full or a light verb.

In this paper, we report on a corpus based analysis of Czech light verbs that enter into combinations with predicative nouns. Considering the wide range of issues related to light verbs, this analysis focuses on the possibility to identify them. Such an identification requires operational criteria for distinguishing light usages of a verb from its full usages, see examples (1)–(2). The criteria adopted here were verified in the parallel annotation of a large amount of corpus data obtained from the Czech National Corpus.[1] The annotation process and the criteria used in the annotation are described in detail in Section 3.

At the beginning of the annotation, lemmas of the verbs that are prone to combine with predicative nouns were necessary to select. Compiling the inventory of such verb lemmas is described in Section 2. In order to draw up this inventory, the valency lexicons VALLEX[2] and PDT-VALLEX[3] were explored. Let us briefly introduce these two valency lexicons.

1.1 Lexical Resources

Both VALLEX and PDT-VALLEX take the Functional Generative Description (henceforth FGD, [12]) as their theoretical background, and both are human as well as machine readable.

VALLEX is a valency lexicon of Czech verbs (see esp. [14], [8]) providing rich syntactic information on roughly 2,730 lexemes containing 6,460 lexical units ('senses'). Unlike traditional dictionaries, VALLEX treats a pair of perfective and imperfective aspectual counterparts as a single lexeme – if perfective and imperfective verbs were counted separately, the size of the lexicon would virtually grow to 4,250 verb entries. Almost one half

[1] http://ucnk.ff.cuni.cz/
[2] http://ufal.mff.cuni.cz/2.5/
[3] http://ufal.mff.cuni.cz/lindat/PDT-Vallex.html/

of the lexical units are sorted into 22 rough semantic groups (e.g. verbs of communication, motion, transport, exchange, mental action). At present, this lexicon inconsistently describes several randomly selected light verb usages: most light verb usages are subsumed under valency frames corresponding to full verb usages, less of them are represented by separate valency frames. In both cases, an explicit indication of the light verb usage is missing. A preliminary theoretical analysis of an interaction between verbal and nominal valency structures within light verb constructions and the possibility of its adequate description in the VALLEX lexicon is provided in [7].

PDT-VALLEX is a valency lexicon which was built during the annotation of the Prague Dependency Treebank[4] (henceforth PDT [4]) as a supporting annotation tool designed for preserving data consistency in the annotated corpus, see esp. [5], [13]. It describes valency behavior of Czech verbs, nouns, adjectives and adverbs in a fully formalized way (7,500 verbs, 3,800 nouns, 800 adjectives, and a few adverbs). Only those senses of words that occurr in the annotated data of PDT or some other treebanks in the PDT family (Prague Czech-English Dependency Treebank[5] and Prague Dependency Treebank of Spoken Language)[6] are recorded. The information on syntactic structures with light verbs is simply listed in pairs of unlinked separate valency frames: (a) in valency frames of light verbs, in which the valency complementation occupied by a predicative noun is labeled with the CPHR functor (for CompoundPHRaseme, see [4]), and (b) in valency frames of predicative nouns.

2 Inventory of Czech Light Verbs

Although light verbs and their full verb counterparts always have identical forms, they only agree in some of their semantic aspects [2], [9]. For instance, the light verb *nést* 'to carry' in (2) loses the individual meaning of its full counterpart in (1): the meaning of the full verb can be characterized as 'the shared movement of a physical object and a person who controls the shared path', but in the light usage in (2), the lexical properties expressing the physical movement are suppressed; instead, the lexical features of 'permanent presence of a certain sense experienced by a person wherever they move' are foregrounded. The individual lexical semantic property 'a sense of moral commitment or obligation to somebody or something' is supplied by the predicative noun *odpovědnost* 'responsibility' with which the verb combines.

The question arises whether the verbs that are inclined to combine with predicative nouns share some common features on the basis of which they can be characterized. According to [3], high frequency verbs, occurring in various semantic and syntactic contexts, have a strong tendency to lose their individual semantic properties and to combine with predicative nouns. The hypothesis was formulated on Swedish verbs.

However, Czech verbs do not confirm this hypothesis: we sorted Czech verbs according to their frequency in the Czech National Corpus[7] (henceforth CNC); modal verbs and the verb *být* 'to be' (with primarily auxiliary function) were excluded. We compared the obtained list of the first 50 Czech verbs with the verbs with the CPHR functor in PDT-VALLEX, i.e., those verbs that are classified as used as light verbs in at least one of their

[4] http://ufal.mff.cuni.cz/pdt2.0/

[5] http://ufal.mff.cuni.cz/pcedt2.0/

[6] http://ufal.mff.cuni.cz/pdtsl/cz/

[7] The balanced subcorpus of contemporary Czech texts SYN2000 was used.

occurrences in PDT. Only 32% of the high frequency verbs contain at least one valency frame with the CPHR functor in PDT-VALLEX.

Thus we tried to establish the criteria for the identification of Czech verbs with possible light usages on another basis. As a starting point, we carried out a tentative survey of the verbs with the CPHR functor in PDT-VALLEX.[8] This study revealed an interesting fact: verbs with valency frame(s) describing light usages fall into just a few semantic groups. They express exchange (e.g. *vzít* 'to take', *dát* 'to give'), location (e.g. *pokládat, položit* 'to put down'), motion (e.g. *přicházet, přijít* 'to come'), transport (e.g. *vést* 'to lead'), or they refer to an action in a generic way (e.g. *dělat* 'to do').

In the next step, we further explored the idea that the capacity of a verb to be used as light is related to their semantic class membership. We sorted all verbs belonging to one of the above mentioned semantic groups in VALLEX according to their frequency in CNC.[9] The verbs designating actions in a generic way are not grouped together in a specific semantic class in this lexicon. However, as they represent a significant group of verbs allowing for light usages, we have included all 17 verb lemmas with generic meaning obtained from PDT-VALLEX directly into our experiment. The most frequent verbs of exchange, location, motion, and transport, plus the verbs with generic meaning were chosen as candidates for light verbs. From the list of candidates, verbs with less than six valency frames in VALLEX were removed. In order to achieve a satisfactory coverage of verbs in CNC, we selected first 59 most frequent verb lemmas – this number covers 20.0% of total verb occurrences in CNC.

The resulting inventory of selected verbs was exploited in the second part of the experiment focusing on the possibility of distinguishing light usages from full usages. This strategy to identify Czech lemmas allowing for light usages gave more satisfactory results than frequency – 48 from the overall 59 selected verbs (81.4%) were used as light verbs, based on the findings of our experiment (Section 3).

Alternatively, there was an option to directly use the verbs with the CPHR functor from PDT-VALLEX as candidates for light verbs. This method would eliminate verb lemmas predicating only as full verbs; however, many verbs that can form a light usage would be missed: 27 out of 48 verbs with a light usage that occurred in the annotation do not have the CPHR functor in PDT-VALLEX.

3 Annotation of Light Verbs

In this section, we describe in detail the annotation of corpus sentences based on 59 selected Czech verb lemmas. For each of these selected verb lemmas included in the experiment, 100 random sample sentences were extracted from the CNC. Thus the annotated data size is 5,900 sentences per each annotation.

Three human annotators in parallel were asked to determine whether a verb occurrence in an extracted sentence corresponds to a full or a light usage of the given verb (thus the

[8] PDT-VALLEX contains 148 Czech verb lemmas with at least one valency frame containg the CPHR functor.

[9] Although phase verbs are usually considered to represent light verbs, they are not indicated by the CPHR functor in PDT-VALLEX. As a result, these verbs were included in our experiment only if they fall into some of the above mentioned semantic groups in VALLEX. For instance, the verb *skončit* 'to finish' was included in the experiment as it is classified both as a phase verb and as a verb of location according to VALLEX.

overall number of annotated sentences is 17,700). The main aim of this annotation was to examine the native speakers' agreement on the interpretation of light verb usages. To facilitate the interpretation, the annotators could take a context of one preceding sentence into consideration. When the annotators indicated that a given occurrence of a verb is a light usage, they had to determine the whole verbonominal combination of the given light verb and a predicative noun. Also, an uncertainty flag indicating that the annotators are not quite sure could be attached to a positive answer.

3.1 Criteria for Distinguishing Light Verb Usages from Full Ones

At the beginning of the annotation, it was necessary to single out criteria for distinguishing light verb usages from their full usages. For this purpose, we have adopted two criteria mentioned in the rich bibliography on light verbs – the reduction test and the test of coreference of nominal and verbal complementations.

Reduction Test. The reduction test, proposed in [10], is based on the assumption that it is the predicative noun (not the light verb) that represents the semantic core of the entire verbonominal combination. As a result, it is the predicative noun that stands for the whole verbonominal combination and it cannot be omitted from the combination – in contrast to the light verb. This test consists of the sequence of two syntactic operations: (i) relativization and (ii) omission of the light verb. When these operations are applied on a particular syntactic structure – e.g. on the sentence structure in (2) repeated here as (3), (i) the relativization results in (4) and (ii) the omission results in (5) – the semantic invariant between (4) and (5) is clearly preserved. However, when this test is applied to a full verb usage – e.g. on (1) repeated here as (6), (i) the relativization results in (7) and (ii) the omission results in (8), the semantic invariant is not preserved: (8) is obviously not equivalent to (7).

(3) *Učitel nese odpovědnost za bezpečnost žáků.*
the teacher – carries – responsibility – for the security of pupils
'The teacher is responsible for the security of pupils.'

(4) *Odpovědnost, kterou za bezpečnost žáků nese učitel.*
the responsibility – which – for the security of pupils – carries – the teacher

(5) *Odpovědnost učitele / Učitelova odpovědnost za bezpečnost žáků.*
'The teacher's responsibility for the security of pupils.'

(6) *Učitel* nese sešity.
'The teacher carries exercise books.'

(7) *Sešity, které nese učitel.*
'Exercise books, which the teacher carries.

(8) *Sešity učitele / Učitelovy sešity.*
'Teacher's exercise books.'

Coreference Test. The second criterion stipulates that some of valency complementations of a light verb and a predicative noun within the resulting complex predication must be referentially identical. This condition is imposed esp. on the ACTor of the event

expressed by a predicative noun. For example, in the verbonominal combination of the light verb *udělat* 'to make' and the predicative noun *dojem* 'impression', resulting in the verbonominal combination *udělat dojem* 'to make an impression', the ACTor of the predicative noun *dojem* corefers with the ACTor of the verb *udělat*, see Figure 1. Apart from the ACTor, other valency complementations of a predicative noun can be coreferentially related to verbal complementations of a light verb. For example, in the combination *dostat příkaz* 'to get an order', two valency complementations from the nominal valency frame ACTor and ADDResse corefer with the verbal valency complementations ORIGin and ACTor, respectively, see Figure 2.

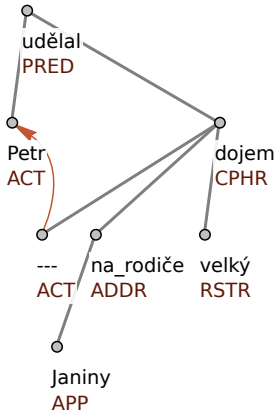


Fig. 1. The (simplified) dependency tree for the sentence *Petr udělal na Janiny rodiče velký dojem* 'Peter made a great impression on Jane's parents'

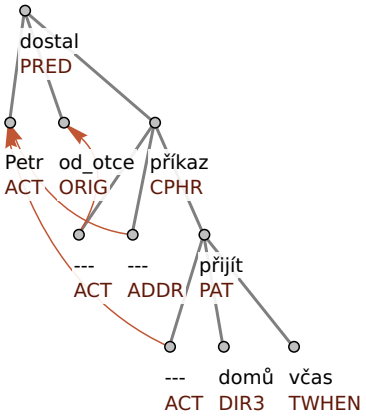


Fig. 2. The (simplified) dependency tree for the sentence *Petr dostal od otce příkaz přijít domů včas* 'Peter was given the order from his father to come home on time'

Supporting Criteria. In addition to the reduction test and the coreference test, the annotators could rely on other auxiliary criteria. These criteria are based on the observation that the possibility for predicative nouns to be pronominalized (9)-(10) or to be asked for by wh-questions (9)-(11) is highly restricted in verbonominal constructions with a light verb.

(9) *Petr upadl do rozpaků.*
Peter – fell – in embarrassment

(10) **Petr upadl do toho.*
Peter – fell – in this

(11) **Do čeho Petr upadl?*
in what – Peter – fell?

3.2 Annotation Task

In case that the annotators concluded that a light usage of a verb is present in a given sentence (also a special flag indicating an uncertainty of light usage could be used), they had two tasks. First, they had to indicate the whole combination of the light verb and the predicative noun. In case that a sentence contained coordinated predicative nouns combined with a single light verb, the annotators had to determine all of the predicative nouns combined with the given light verb. Second, after identifying a light usage of a verb, the annotator had to determine with which verbal valency complementation the valency complemenation 'ACTor' of the given predicative noun is coreferential.

The annotated data size and overall statistics on the annotations are summarized in Tables 1 and 2.

Annotated verbs	59
Annotated sentences for each verb	100
Parallel annotations	3
Total annotated sentences	17,700

Table 1. Annotated data size

Annotator:	A	B	C
Verb lemmas with light usages	40	35	49
Verb lemmas with uncertain light usages	6	13	4
Light verb usages	713	522	699
Uncertain light verb usages	110	256	287
Full verb usages	5,077	5,122	4,914
Total verb usages	5,900	5,900	5,900
Found verbonominal combinations	843	796	1,002
Coreference between nominal 'ACTor' and some of verbal complementation(s)	615	649	755

Table 2. Overall statistics on the annotations

3.3 Inter-Annotator Agreement

Table 3 provides inter-annotator agreement (IAA), Cohen's κ and Artstein and Poesio's α_κ statistics on the annotated 17,700 sentences.

First two columns represent IAA, i.e., the percentage of sentences with an agreement out of all annotated sentences (the first column); if also the combinations "yes-maybe" is considered to be an agreement (the second column), the pairwise inter-annotator agreement naturally rises.

The first row introduces the average pairwise agreement. In our case of three parallel annotations, it is the sum of agreements of three possible annotation pairs divided by three. The second row shows an IAA for all three annotations together, which is a more rigid

measure than the pairwise average: e.g. combinations "yes-yes-maybe" and "yes-maybe-no" are considered as disagreements for an exact match (left column), thus obtaining 0 and 0, respectively (whereas corresponding pairwise average values in the first row would be $\frac{1+0+0}{3} = \frac{1}{3}$ and $\frac{0+0+0}{3} = 0$, respectively). The latter combination "yes-maybe-no" is a disagreement (=0) even with uncertainty tolerance (but would be rated $\frac{1+0+0}{3} = \frac{1}{3}$ in the average pairwise match).

The third column represents Cohen's κ (generalized for multiple annotators in the last row), i.e., the inter-annotator agreement above the chance counted from individual annotators' preferences for their answers (roughly speaking, from their 'average answers'). The last column is a weighted variant of κ (called κ_w for two annotators and α_κ for multiple annotators, see an extended version of [1]). Weights for disagreement were set as follows: "yes-no" is not an agreement at all ($a_{\text{yes,no}} = 0$), but "yes-maybe" and "no-maybe" is counted as a partial agreement ($a_{\text{yes,maybe}} = \frac{2}{3}$ and $a_{\text{no,maybe}} = \frac{1}{3}$). The third row shows generalizations of κ coefficients for multiple annotators, which is claimed to be "a better practise" than an average of pairs by [1].

	IAA exact match	IAA uncertainty tolerance	κ unweighted	κ_w, α_κ weighted
Average pairwise match	89.7%	92.6%	0.602	0.688
Match of all three annotators	85.3%	88.9%		
Agreement above chance for three annotators			0.600	0.686

Table 3. Inter-annotator agreement and κ statistics of three parallel annotations

3.4 Golden Data

The sentences with exact agreement across three involved annotations form the so called golden data. The sentences with disagreement were manually re-annotated in order to resolve disagreement and to unify the annotations. On the basis of the golden data, we obtained verbonominal combinations which can be further applied in the lexicographic description of Czech light verbs. The overall statistics on the golden data is provided in Table 4. The numbers of light usages of each verb involved in the experiment are provided in the Appendix.

Annotated sentences	5,900
Sentences with light verb usages	855
Sentences with uncertain light verb usages	18
Sentences with full verb usages	5,027
Verbonominal combinations	893
Verb lemmas annotatated	59
Verb lemmas with light usages	48

Table 4. Overall statistics on the golden data

4 Conclusion

We have described an experiment with the identification of Czech light verb usages. This experiment consisted of two parts: in the first part, we have explored the possibility to identify the inventory of Czech verbs allowing for light usages; in the second part, we have examined the criteria adopted for distinguishing light usages of a verb from its full ones.

As a result of the first part, we have suggested the hypothesis that the possibility of Czech verbs to be used as light verbs is connected to their semantic class membership (exchange, location, motion, transport verbs and verbs with generic meaning) rather than to their high frequency. However, the hypothesis has to be further examined esp. on low frequency verbs belonging to the selected semantic groups.

In the second part of the experiment, the reliability of distinguishing light usages of a verb from its full ones on the basis of the adopted criteria – the reduction test and the coreference test – has been examined. The achieved inter-annotator agreement (IAA 85.3% and κ_w 0.686) appears to be promising.

Further, as a result of the annotation process, the golden data that consists of the sentences with exact agreement and the sentences with a resolved disagreement has been obtained. The golden data contains 893 instances of combinations of light verbs and predicative nouns. These data will be further exploited as the lexical stock in the lexicographic representation of Czech light verbs in the valency lexicon of Czech verbs VALLEX.

Acknowledgments

The research reported in this paper has been supported by the Czech Science Foundation GA ČR, grant No. GA P406/12/0557 and partially grant No. GA P406/10/0875. This work has been using language resources stored and/or distributed by the LINDAT-Clarin project of MŠMT (project LM2010013).

References

[1] Artstein, R. and Poesio, M. (2008). Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555–596. Extended version is available at: `http://cswww.essex.ac.uk/Research/nle/arrau/`.

[2] Butt, M. (2010). The Light Verb Jungle: Still Hacking Away. In Mengistu Amberber, B. B. and Harvey, M., editors, *Complex Predicates in Cross-Linguistic Perspective*, pages 48–78. Cambridge University Press, Cambridge.

[3] Cinková, S. (2009). *Words that Matter: Towards a Swedish-Czech Colligational Dictionary of Basic Verbs*, volume 2 of *Studies in Computational and Theoretical Linguistics*. Institute of Formal and Applied Linguistics, Prague.

[4] Hajič, J., Hajičová, E., Panevová, J., Sgall, P., Pajas, P., Štěpánek, J., Havelka, J., and Mikulová, M. (2006). *Prague Dependency Treebank 2.0*. Linguistic Data Consortium, Philadelphia, PA, USA.

[5] Hajič, J., Panevová, J., Urešová, Z., Bémová, A., Kolářová, V., and Pajas, P. (2003). PDT-VALLEX: Creating a Large-Coverage Valency Lexicon for Treebank Annotation. In *Proceedings of The Second Workshop on Treebanks and Linguistic Theories*, volume 9, pages 57–68.

[6] Hanks, P., Urbschat, A., and Gehweiler, E. (2006). German Light Verb Constructions in Corpora and Dictionaries. *International Journal of Lexicography*, 19(4):439–457.

[7] Kettnerová, V. and Lopatková, M. (2013). The Representation of Czech Light Verb Constructions in a Valency Lexicon. In *Proceedings of the Dependency Linguistics Conference, DepLing 2013*. (accepted).

[8] Lopatková, M., Žabokrtský, Z., and Kettnerová, V. (2008). *Valenční slovník českých sloves*. Nakladatelství Karolinum, Praha.

[9] Macháčková, E. (1979). *Analytická spojení typu sloveso + abstraktní substantivum (analytické vyjadřování predikátů)*. Ústav pro jazyk český ČSAV, Praha.

[10] Radimský, J. (2010). *Verbo-nominální predikát s kategoriálním slovesem*. Editio Universitatis Bohemiae Meridionalis, České Budějovice.

[11] Sag, I. A., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. (2002). Multiword Expressions: A Pain in the Neck for NLP. In *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2002)*, pages 1–15, Mexico City, Mexico.

[12] Sgall, P., Hajičová, E., and Panevová, J. (1986). *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. Reidel, Dordrecht.

[13] Urešová, Z. (2011). *Valence sloves v Pražském závislostním korpusu*, volume 8 of *Studies in Computational and Theoretical Linguistics*. Institute of Formal and Applied Linguistics, Prague.

[14] Žabokrtský, Z. and Lopatková, M. (2007). Valency Information in VALLEX 2.0: Logical Structure of the Lexicon. *The Prague Bulletin of Mathematical Linguistics*, (87):41–60.

Appendix: List of verbs annotated within the experiment

This appendix lists the verbs involved in the annotation process sorted with decreasing frequency in CNC. It provides the numbers of their light usages stored in the golden data. Some of these verbs were not found capable of creating light usages (at least in the examined data) – these have a dash in the second column.

verb	light usages	*uncertain* light usages
mít	36	1
jít	7	
stát	3	1
dát	23	1
dostat	30	2
uvést	12	
přijít	13	
dělat	14	
získat	43	
patřit	–	
vést	38	
udělat	26	1
najít	3	
zůstat	3	
dojít	35	
vrátit	2	
nechat	4	
platit	–	
vzít	6	
dosáhnout	52	1
skončit	64	
připravit	9	1
ukázat	1	
přijmout	26	
vydat	15	
počítat	–	
vypadat	–	
vyjít	2	1
chodit	–	
postavit	4	1

verb	light usages	*uncertain* light usages
sedět	–	
držet	4	3
ztratit	43	1
uzavřít	34	1
ležet	2	
měnit	–	
vybrat	–	
zastavit	9	
podat	19	
pohybovat	–	
jezdit	–	
založit	3	
nést	28	
končit	50	1
přidat	4	
projít	7	
obrátit	4	
učinit	30	
vystoupit	2	
stavět	4	1
sejít	2	
udržet	22	1
položit	15	
padnout	27	
převzít	37	
pustit	20	
hodit	–	
přejít	17	
zvednout	1	

Agents Expressed by Prepositionless Instrumental Modifying Czech Nouns Derived from Intransitive Verbs[1]

Veronika Kolářová

Faculty of Mathematics and Physics, Charles University in Prague,
Czech Republic

Abstract. The present paper aims to provide corpus-based description of Czech deverbal nouns that allow for modification by Agent expressed by prepositionless instrumental, A_1(Ins). As the starting point we give frequency data of selected nouns derived from transitive verbs. Then we focus on nouns derived from intransitive verbs and show that modification by A_1(Ins) is possible not only with nouns derived from verbs that can be passivized, but also with nouns the source verbs of which cannot be changed to passive. The latter issue represents the most contributive finding of the paper; it concerns especially nouns derived from reflexive verbs, both transitive and intransitive. We also improve the up-to-now description by taking into consideration not only Czech nouns derived from verbs by productive means (e.g. *domlouvání* 'talking') but also the non-productively derived ones (e.g. *domluva* 'caution'), mostly left aside. Finally, the corpus material also gives an evidence for usage of theoretically ungrammatical constructions in which the second complementation (A_2) is omitted on the surface and only A_1(Ins) is expressed, e.g. *vyhrožování zaměstnavatelem* 'threatening by the employer', *domluva strážníky* 'caution by police officers'; the corpus-based examples lead to the revision of the statement about ungrammaticality of such constructions.

1 Introduction

Agents expressed by prepositionless instrumental (Ins) modifying Czech deverbal nouns have been studied mainly in connection with two topics: (i) relation of a nominalized structure with Agent in the form of Ins, i.e. A_1(Ins), to the corresponding passive verbal construction, and (ii) an action meaning (reading) of the noun modified by A_1(Ins); in the present paper, we focus on the topic (i), while the topic (ii) is only marginally discussed[2]. Both topics have been primarily studied on the material of nouns derived from transitive verbs; nouns derived from intransitive verbs have been marginally dealt with, on the basis of only few examples, and thus they deserve to be studied in detail. Traditionally, Czech nouns derived from verbs by productive means are in the centre of attention (e.g. "transitive" nouns *ošetření* 'treating', *přednášení* 'lecturing', and "intransitive" nouns *domlouvání* 'talking-IPFV' / *domluvení* 'talking-PFV', *dotýkání se* 'touching-IPFV' / *dotknutí se* 'touching-PFV'), while nouns derived from verbs by non-productive means or by the zero suffix are often left aside (e.g. the "transitive" noun *výuka* 'teaching / instruction' and "intransitive" nouns *domluva* 'caution', *dotyk* 'touch').

[1] The research reported in the paper was supported by the Czech Science Foundation under the project P406/12/P190.

[2] The possibility to be modified by A_1(Ins) serves as one of criteria for identifying an action meaning (reading) of the noun [7], [1], [21, p. 22], [20]. In real communication, A_1(Ins) is very rare [14, p. 123], [16, p. 80], [12, p. 59].

Our approach to issues of valency of Czech deverbal nouns is based on the theory of valency (especially valency of verbs) as developed in the framework of Functional Generative Description by [17], [18]. In accordance with this approach we consider the complementation expressed by prepositionless instrumental to be Agent (Actor, ACT).

As for the relation of nominalized structures with A_1(Ins) to verbal passive constructions, the structures given in (2) and (4) are considered to be parallel to the structures given in (1) and (3), cf. [11], [9], [18], among others.

(1) *lékař ošetří pacienta*
'the doctor will treat the patient.'

(2) *lékařovo ošetření pacienta*
'doctor's treating of the patient'

(3) *pacient byl ošetřen lékařem*
'the patient was treated by a doctor.'

(4) *pacientovo ošetření lékařem*
patient-ADJ.POSS treating-NOM.SG doctor-INS.SG
'patient's treating by a doctor / treating of the patient by a doctor'

However, passivization is not limited to syntactically transitive verbs. According to [10] Czech verbs can be passivized when they have minimally two actants (A_1 and A_2), one of which (A_1) affects the second one (A_2). This situation covers not only syntactically transitive verbs, but also some intransitive ones, cf. (5). As for reflexive verbs, they are considered not to allow to be changed to passive, however, some exceptions exist, cf. (6) to (9).

(5) *Blondýnkám je nadržováno / pomáháno / lichoceno*
blond-DAT.PL be-3.SG.PRES favour-PASS.PART / help-PASS.PART / flatter-PASS.PART
'Blonds are favoured / helped / flattered'

(6) *Soudce se paní kuchařky tázal / dotázal, zda…*
judge-NOM.SG REFL lady-GEN.SG cook-GEN.SG ask-3.SG.PRT if…
'The judge asked the lady cook if…'

(7) *Paní kuchařka byla soudcem tázána / dotázána, zda…*
lady-NOM.SG cook-NOM.SG be-3.SG.PRT judge-INS.SG ask-PASS.PART if…
'The lady cook was asked by the judge if…'

(8) *Petra se dotklo jednání toho člověka*
Peter-GEN.SG REFL offend-3.SG.PRT action-NOM.SG that-GEN.SG man-GEN.SG
'action of that man offended Peter'

(9) *Petr byl dotčen jednáním toho člověka*
'Peter was offended by the action of that man'

According to [9, pp. 40-41] and [20, pp. 43-44], Czech deverbal nouns derived from intransitive verbs allow for modification by A_1(Ins) when a noun is derived from non-ergative verbs, cf. (10) to (13), while with nouns derived from non-accusative verbs modification by A_1(Ins) is ungrammatical, cf. (14) and (15).

(10) *holkám je nadržováno učitelem*
girl-DAT.PL be-3.SG.PRES favour-PASS.PART teacher-INS.SG
'girls are favoured by the teacher'

(11) *nadržování holkám učitelem*
favouring girl-DAT.PL teacher-INS.SG
'favouring girls by the teacher'

(12) *synovi je domlouváno starostlivou matkou*
son-DAT.SG be-3.SG.PRES talk-PASS.PART worried-INS.SG mother-INS.SG
'son is talked to by his worried mother'

(13) *domlouvání synovi starostlivou matkou*
talking.NOM.SG son-DAT.SG worried-INS.SG mother-INS.SG
'talking to the son by his worried mother'

(14) **propadnutí obci majetkem*
passing village-DAT.SG property-INS.SG
'passing to the village by the property'

(15) **unikání strážníkům šťastným vězněm*
escaping.NOM.SG policeman-DAT.PL happy-INS.SG prisoner-INS.SG
'the escaping to the policemen by the happy prisoner'

Considering both nouns derived from transitive and intransitive verbs, modification by A_1(Ins) is expected only when A_2 is present [9, p. 40], [20, p. 41], cf. (16) to (18).

(16) **přednášení Evou*
lecturing Eve-INS.SG
'lecturing by Eve'

(17) **vyprávění ovčí babičkou*
telling sheepish-INS.SG grandma-INS.SG
'the telling by the sheepish grandma'

(18) **nadržování učitelem*
favouring teacher-INS.SG
'favouring by the teacher'

Karlík [9] studied in detail the relationship between internal structure of Czech nouns derived from verbs by productive means and their syntactic behaviour and claims that the structures given in (2) and (4) do not show structural differences corresponding with the active – passive voice distinction, therefore A_1(Ins) cannot be licensed through passivization (see also [21, p. 22]).

2 Nouns Modified by A_1(Ins): Corpus-based Observations

In the present paper, we focus on nouns derived from intransitive verbs (Section 2.2), because they have been only marginally dealt with and so they are less theoretically described than nouns derived from transitive verbs. "Intransitive" nouns do not represent the typical examples of nouns modified by A_1(Ins) and even the linguistic intuition of native speakers sometimes fails when language correctness of such constructions is discussed. In such a case we need sufficient source of data to prove or disprove our expectation or hypothesis; representative and balanced corpus data are irreplaceable then.

Non-typical and also rare examples of "intransitive" nouns modified by A_1(Ins) can be better evaluated when compared with typical examples of nouns modified by A_1(Ins), i.e. with nouns derived from transitive verbs. Thus we also probe into valency behaviour of selected "transitive" nouns (Section 2.1) in order to see how often these nouns occur with the modification by A_1(Ins). With most of "transitive" nouns we do not have doubts about the language correctness of the A_1(Ins) modification, however frequency of the modification based on corpus material has not been examined yet.

We do not use the terms non-accusative and non-ergative verbs, but try to identify particular semantic classes of the nouns (e.g. nouns of communication, nouns of mental action, nouns of motion)[3].

In the paper, we present results of searching for "intransitive" as well as "transitive" nouns modified by A_1(Ins) in morphologically annotated subcorpora of the Czech National Corpus (CNC; Český národní korpus); the following five CNC subcorpora were used: SYN2000, SYN2005, SYN2006PUB, SYN2009PUB and SYN2010. The nouns with the A_1(Ins) modification were mostly searched for by the following queries: ([lemma="…"] [!(tag="[Z|V|J].*")]{0,5} [tag="N...7.*"]) or ([lemma="…"] [!(tag="[Z|R|V|J].*")]{0,5} [tag="N...7.*"]). All found examples were manually checked and absolute frequencies of the examples that really match the required structure are summarised in Tables 1-4; we

[3] For identification of the appropriate semantic class of the nouns, we use semantic classification of source verbs of the nouns, captured in the valency lexicon of Czech verbs, VALLEX [15]. However, VALLEX provides the information on the semantic classes only for selected verbs, thus some source verbs of nouns that we study in the paper are not semantically classified, e.g. *dožít se* 'to live to'*, dovolávat se* 'to call for'*, ujmout se* 'to take care'*, vzdát se* 'to surrender'.

separate examples of nouns modified by both A_1(Ins) and another participant A_2 (category A in the Tables) from constructions with A_1(Ins) modification only (category B in the Tables; for more details see Section 3).

The manual checking of all found examples includes syntactic as well as semantic analysis of found strings. For example, we excluded all strings in which the form of instrumental could be interpreted as another participant or a free modification, especially the complementations with the semantico-syntactic function of means (marked in the paper by the functor MEANS, e.g. *lokální ošetření kortikoidem*.MEANS 'local treatment by a corticoid', *připojení telefonem*.MEANS 'connection using the phone', *dotazování telefonem*.MEANS 'questioning using the phone') and direction "which way" (marked in the paper by the functor DIR2, e.g. *průnik obranou*.DIR2 'penetration through the defence'). We had also to exclude numerous ambiguous constructions, cf. examples with the nouns *dotýkání se* 'touching-IPFV' and *dotknutí se* 'touching-PFV', exhibiting ambiguity of ACT and MEANS, cf. (19) and (20), and the example with the noun *výpomoc* 'help', illustrating ambiguity of ACT and Patient (PAT), cf. (21).

(19) *dotýkání se oblaků*.PAT *konečky prstů*.ACT/MEANS
touching REFL cloud-GEN.PL fingertip-INS.PL
'touching of clouds by fingertips'

(20) *dotknutí se země*.PAT … *dolní končetinou*.ACT/MEANS (SYN2010)
touching REFL ground-GEN.SG lower-INS.SG extremity-INS.SG'
'touching of the ground by the lower extremity'

(21) *výpomoc přímou pečovatelskou službou*.ACT/PAT (SYN2009PUB)
help direct-INS.SG nursing-INS.SG service-INS.SG
'help by direct nursing service / help with direct nursing service'

2.1 Nouns Derived from Transitive Verbs

Nouns derived from transitive verbs represent very large group of nouns. For the present corpus-based study we selected 15 productively derived nouns (13 non-reflexive nouns, see Table 1, and 2 reflexive nouns, see Table 2) and 3 non-productively derived nouns (Table 3).

First we focused on nouns given as examples in the literature, i.e. *ošetření* 'treating', *vyprávění* 'telling' and *přednášení* 'lecturing'. Then we selected polyvalent (bivalent and trivalent) nouns representing particular semantic classes, e.g. nouns of communication (e.g. *oznamování* 'announcing', *výuka* 'teaching'), nouns of exchange (e.g. *odebrání* 'taking away', *vrácení* 'returning', *dodávka* 'delivery'), nouns of mental action (e.g. *uvědomování si* 'being aware'), nouns of ingestion (e.g. *požití* 'ingestion', *konzumace* 'consumption'). As for productively derived nouns, both aspectual counterparts were searched for (e.g. *odebírání* 'taking-IPFV away' / *odebrání* 'taking-PFV away'), provided they exist.

According to absolute frequencies[4] of productively derived nouns, given in Table 1 and Table 2, we can see that perfective nouns more often occur with A_1(Ins) modification than imperfective ones.

As for the semantic classification, productively derived nouns of communication and nouns of mental action modified by A_1(Ins) are the least frequent semantic classes. However, although some of them represent really isolated examples, we can see that they are unquestionably grammatically correct constructions. The second participant (mostly PAT) is expressed by prepositionless genitive, cf. (22), (23) and (25), or a possessive pronoun, cf. (26), or it is omitted on the surface, cf. (24).

(22) *důvody* *neoznamování* *trestného činu*.PAT *občany*.ACT (SYN2000)
reason-NOM.PL non-announcing-GEN.SG criminal-GEN.SG offence-GEN.SG citizen-INS.PL
'reasons for non-announcing of a criminal offence by citizens'

(23) *přednesení* *revizních zpráv*.PAT *jednotlivými členy*.ACT (SYN2000)
presentation-NOM.SG audit-GEN.PL report-GEN.PL particular-INS.PL member-INS.PL
'presentation of audit reports by particular members'

(24) *rozlišení základních vypravěčských typů (vyprávění postavou*.ACT *účastnou v ději – vyprávění vypravěčem*.ACT *stojícím mimo děj)* (SYN2005)
'distinguishing basic narrative types (telling by a character taking part in an action – telling by a narrator being outside an action)'

(25) *způsob uvědomování si okolního světa*.PAT *danou postavou*.ACT (SYN2010)
way being_aware-GEN.SG REFL outside-GEN.SG world-GEN.SG given-INS.SG character-INS.SG
'the way of being aware of the outside world by the given character'

(26) *u těchto statků můžeme předpokládat větší míru jejich*.PAT *uvědomění jednotlivcem.*ACT *než v případě...* (SYN2005)
'concerning this property we can suppose a larger degree of their realization by an individual than in case…'

[4] Absolute frequencies given in Tables 1–4 would be better interpreted when supplemented with relative frequencies (i.e. the ratio of the absolute frequencies to the frequency of lemmas of the nouns as such).

Noun	SYN 2000		SYN 2005		SYN 2006 PUB		SYN 2009 PUB		SYN 2010		Total	
	A	B	A	B	A	B	A	B	A	B	A	B
odebírání 'taking-IPFV away'	1	0	0	0	1	0	3	0	0	0	5	0
odebrání 'taking-PFV away'	3	0	2	0	8	1	19	1	1	0	33	2
ošetřování 'treating-IPFV'	0	0	3	2	1	3	3	8	0	1	7	14
ošetření 'treating-PFV'	0	4	3	6	6	33	14	129	0	5	23	177
oznamování 'announcing-IPFV'	0	0	0	0	1	0	0	0	0	0	1	0
oznámení 'announcing-PFV'	2	0	0	1	14	2	18	9	0	1	34	13
požívání 'consuming-IPFV'	3	0	0	0	1	1	18	4	0	1	22	6
požití 'ingestion-PFV'	3	0	13	1	4	0	44	1	2	0	66	2
přednášení 'lecturing-IPFV'	0	0	0	0	0	0	0	0	0	0	0	0
přednesení 'presentation-PFV'	1	0	0	0	1	0	3	0	0	0	5	0
vracení 'returning-IPFV'	1	0	1	0	3	0	6	0	1	0	12	0
vrácení 'returning-PFV'	3	1	2	2	17	4	27	3	1	3	50	13
vyprávění 'telling-IPFV'	0	0	0	3	0	1	0	0	0	0	0	4
Total	17	5	24	15	57	45	155	155	5	11	**258**	**231**

Table 1. Productively derived, non-reflexive, "transitive" nouns modified by A_1(Ins): Ratio of presence of the second participant to its absence

Noun	SYN 2000		SYN 2005		SYN 2006 PUB		SYN 2009 PUB		SYN 2010		Total	
	A	B	A	B	A	B	A	B	A	B	A	B
uvědomování si 'being-IPFV aware REFL'	0	0	0	0	0	0	0	1	1	0	1	1
uvědomění si 'being-PFV aware REFL'	2	0	2	0	0	0	0	0	1	0	5	0
Total	2	0	2	0	0	0	0	1	2	0	**6**	**1**

Table 2. Productively derived, reflexive, "transitive" nouns modified by A_1(Ins): Ratio of presence of the second participant to its absence

For the present study, we selected three non-productively derived "transitive" nouns, representing three semantic classes, i.e. nouns of exchange (*dodávka* 'delivery'), nouns of ingestion (*konzumace* 'consumption') and nouns of communication (*výuka* 'teaching'). According to absolute frequencies given in Table 3, the nouns are comparably frequent when they are modified by A_1(Ins). The second participant (mostly PAT) is expressed by prepositionless genitive, cf. (26a), or a possessive pronoun, cf. (27), or it is omitted on the surface, cf. (54) to (56) below.

Noun	SYN 2000		SYN 2005		SYN 2006 PUB		SYN 2009 PUB		SYN 2010		Total	
	A	B	A	B	A	B	A	B	A	B	A	B
dodávka 'delivery'	5	1	4	1	4	0	10	2	1	0	24	4
konzumace 'consumption'	1	0	5	1	8	4	23	8	2	3	39	16
výuka 'teaching'	0	1	1	1	3	0	17	8	3	3	24	13
Total	6	2	10	3	15	4	50	18	6	6	**87**	**33**

Table 3. Non-productively derived, "transitive" nouns modified by A_1(Ins): Ratio of presence of the second participant to its absence

(26a) *přehodnotili <u>dodávku</u> tepla*.PAT *firmou*.ACT *Thermo DDK* (SYN2000)
re-evaluate-PRT delivery-ACC.SG heat-GEN.SG company-INS.SG Thermo DDK
'(they) re-evaluated delivery of heat by the company Thermo DDK'

(27) ... *alkohol. Myslím tím jeho*.PAT <u>*konzumaci*</u> *špičkovými hráči*.ACT.
(SYN2009PUB)
... alcohol. I mean it-PRON.POSS consumption-ACC.SG top-INS.PL player-INS.PL
'... alcohol. I mean its consumption by top players.'

2.2 Nouns Derived from Intransitive Verbs

Concerning nouns derived from intransitive verbs, our method is to predict particular nouns that, according to our linguistic intuition, could allow for the modification by A_1(Ins) and then to verify whether the nouns occur with the modification in the selected CNC subcorpora, mentioned above. We elaborated lists of both productively and non-productively derived "intransitive" nouns. We applied the same procedure as with the "transitive" nouns (described in Section 2), including manual checking of all found examples in all five CNC subcorpora used. However, the "intransitive" nouns modified by A_1(Ins) are considerably less frequent than the "transitive" ones (on the average, we found 2 examples of each "intransitive" noun in some of the five CNC subcorpora). Thus we cite the absolute frequencies of the respective constructions only in one summarizing table (Table 4). Again, examples of nouns modified by both A_1(Ins) and another participant or complementation (A_2; category A in the Table) are separated from constructions with A_1(Ins) modification only (category B in the Table; for more details see Section 3).

"Intransitive" nouns and their modifications	Nouns derived from non-reflexive verbs		Nouns derived from reflexive verbs	
	Productively derived nouns (5 lemmas)	Non-productively derived nouns (2 lemmas)	Productively derived nouns (8 lemmas)	Non-productively derived nouns (2 lemmas)
A_1(Ins) + A_2 (category A)	9	1	14	2
A_1(Ins) only (category B)	3	5	3	0
Total	**12**	**6**	**17**	**2**

Table 4. "Intransitive" nouns modified by A_1(Ins): Ratio of presence of the second complementation to its absence (on data of five CNC subcorpora)

Although the examples of "intransitive" nouns modified by A_1(Ins) are rather rare we consider the constructions to be grammatically correct[5]. In the following sections, we classify the nouns according to the form of the second complementation, distinguishing two basic groups of the "intransitive" nouns, i.e. nouns derived from verbs that can be passivized (Section 2.2.1) and nouns derived from reflexive verbs (Section 2.2.2).

[5] The situation is similar to that of productively derived, "transitive" nouns of communication, discussed in Section 2.1.

2.2.1 Nouns Derived from Verbs that Can be Passivized

Considering nouns derived from verbs that can be passivized, we started with the two nouns mentioned in [9, p. 41] and [20, p. 43], i.e. *domlouvání* 'talking-IPFV' and *nadržování* 'favouring-IPFV', and then extended the list by semantically or syntactically similar nouns, especially by nouns of communication and nouns of mental action with a participant, i.e. Patient or Addressee (ADDR), in the dative form. We searched for 17 productively derived nouns[6] and for 5 non-productively derived nouns[7]. A_1(Ins) was found with 3 productively derived nouns, i.e. *vyhrožování* 'threatening-IPFV', *napomáhání* 'helping-IPFV / aiding-IPFV', *porozumění* 'understanding-PFV', cf. (27a) to (29), and with 2 non-productively derived ones, i.e. *výpomoc* 'help', cf. (30), and *domluva* 'caution'; the noun *domluva* 'caution' occurred with A_1(Ins) only, cf. (57) in Section 3.

(27a) *napomáhání tomuto trestnému činu.*PAT *státními orgány*.ACT (SYN2009PUB)
aiding this-DAT.SG criminal-DAT.SG offence-DAT.SG state-INS.PL body-INS.PL
'aiding and abetting by state (power) bodies'

(28) *vyhrožování rozhodčím.*ADDR *trenérem*.ACT (SYN2006PUB)
threatening referee-DAT.PL coach-INS.SG
'threatening to the referees by the coach'

(29) *porozumění věci.*PAT *širší veřejností*.ACT (SYN2006PUB)
understanding issue-DAT.SG general-INS.SG public-INS.SG
'understanding the issue by the general public'

(30) … *okomentoval výpomoc domácímu týmu.*ADDR *sudími*.ACT … *trenér* (SYN2009PUB)
comment-PRT help-ACC.SG home-DAT.SG team-DAT.SG referee-INS.PL coach-NOM.SG
'the coach commented on the help to the home team by the referees'

After that, we searched the CNC subcorpora for nouns the source verbs of which can be passivized and, at the same time, they can be modified by a participant (mostly PAT), or an obligatory free modification (direction "where", marked by the functor DIR3) which is expressed by a prepositional group. We searched for 9 productively derived nouns[8] and

[6] Namely *domlouvání* 'talking-IPFV' / *domluvení* 'talking-PFV', *důvěřování* 'trusting-IPFV', *křivdění* 'wronging-IPFV' / *ukřivdění* 'wronging-PFV', *lichocení* 'flattering-IPFV', *nadávání* 'scolding-IPFV', *nadržování* 'favouring-IPFV', *napomáhání* 'helping-IPFV / aiding-IPFV', *podlézání* 'bootlicking-IPFV', *pomáhání* 'helping-IPFV', *porozumění* 'understanding-PFV', *spílání* 'berating-IPFV', *uvěření* 'coming to believe', *vynadání* 'dressing down', *vyhrožování* 'threatening', *zabránění* 'preventing-PFV / prevention'.

[7] Namely *domluva* 'caution', *lichotka* 'flattery', *nadávka* 'insult', *výhrůžka* 'threat', *výpomoc* 'help'.

[8] Namely *pronikání* 'penetrating-IPFV' / *proniknutí* 'penetrating-PFV', *přihlížení* 'taking into account-IPFV' / *přihlédnutí* 'taking into account-PFV', *přispívání* 'contributing-IPFV' / *přispění* 'contributing-PFV', *přistoupení* 'joining-PFV / accession', *vniknutí* 'penetrating-PFV / entry', *vzpomínání* 'remembering-IPFV'.

for 1 non-productively derived noun, i.e. *průnik* 'penetration'. A_1(Ins) was found only with 2 productively derived nouns, i.e. *přistoupení* 'joining-PFV / accession' and *vniknutí* 'penetrating-PFV / entry', cf. (31) and (32).

(31) *možnost <u>přistoupení</u> k dluhu.*PAT *rodinnými příslušníky.*ACT (SYN2009PUB)
possibility accession-GEN.SG to debt-DAT.SG family-INS.PL member-INS.SG
'possibility of accession to the debt by the family members'

(32) *při neoprávněných <u>vniknutích</u> do krypty.*DIR3 *samozvanými správci.*ACT *bývalého koncentračního tábora* (SYN2006PUB)
'during unjustified entries to the crypt by self-proclaimed administrators of the former concentration camp'

2.2.2 Nouns Derived from Reflexive Verbs

Reflexive verbs are considered not to allow to be changed to passive. However, our corpus-based material shows that some nouns derived from reflexive verbs can be modified by A_1(Ins). It concerns especially productively derived nouns (see examples below). Contrary to expectations, A_1(Ins) was found also with one non-productively derived noun (derived from verbs with a participant expressed by prepositionless genitive), i.e. *dotyk* 'touch', cf. (40). As for the reflexive particle *se / si* accompanying nouns derived by productive means (the particle is labeled by REFL in following examples), according to occurrences found in CNC subcorpora used, the particle is often kept but it can also be omitted. Non-productively derived nouns do not keep it at all [8, p. 188].

The most numerous subgroup of the nouns derived from reflexive intransitive verbs is represented by the nouns derived from verbs with a participant expressed by prepositionless objective genitive, e.g. productively derived nouns *dotazování se* 'questioning-IPFV', *dotknutí se* 'touching-PFV', *dovolání se* 'calling-PFV for', *dožití se* 'living-PFV to', *ujímání se* 'taking-IPFV care', *vzdání se* 'waiving-PFV', *zmocnění se* 'seizing-PFV / seizure', see (33) to (39), and one non-productively derived noun, *dotyk* 'touch', cf. (40). Although some nouns derived from verbs with a participant (PAT or ADDR) expressed by prepositionless genitive allow for modification by PAT or ADDR expressed by a possessive pronoun (e.g. *jejich.*ADDR *dotazování* 'their questioning', cf. Kolářová, to appear), there is no occurrence of combination of PAT or ADDR expressed by a possessive pronoun and Actor expressed by prepositionless Ins, but only occurrences of combination of PAT or ADDR in prepositionless genitive and Actor in instrumental.

(33) *<u>dotazování</u> 31 analytiků.*ADDR *agenturou.*ACT *Bloomgerg* (SYN2000)
questioning-NOM.SG 31 analyst-GEN.PL agency-INS.SG *Bloomgerg*
'questioning of 31 analysts by the agency Bloomberg'

(34) *<u>dotknutí</u> míče.*PAT *předchozím hráčem.*ACT (SYN2009PUB)
touching ball-GEN.SG preceding-INS.SG player-INS.SG
'touching of the ball by the preceding player'

(35) *dovolání se neplatnosti.*PAT *smlouvy tím.*ACT*, kdo neplatnost sám způsobil* (SYN2009PUB)
calling_for REFL invalidity-GEN.SG contract-GEN.SG that-INS.SG
'calling for the invalidity of the contract by that who caused the invalidity himself'

(36) *dožití se konce.*PAT *pojištění pojištěným.*ACT (SYN2009PUB)
living_to REFL end-GEN.SG insurance-GEN.SG insured-INS.SG
'living to the end of the insurance by the insured'

(37) *ujímání se zvířátek.*PAT *hodnými lidmi.*ACT (SYN2006PUB)
taking_charge REFL (small_)animal-GEN.PL good-INS.PL people-INS
'taking charge of small animals by good people'

(38) *vzdání se tohoto práva.*PAT *zaměstnavatelem.*ACT (SYN2010)
waiving REFL this-GEN.SG right-GEN.SG employer-INS.SG
'waiving of this right by the employer'

(39) *zmocnění se televize.*PAT *teroristy.*ACT (SYN2009PUB)
seizure REFL television-GEN.SG terrorist-INS.PL
'seizure of the television by terrorists'

(40) *Dotyk sítě.*PAT *hráčem.*ACT *není chybou* (SYN2006PUB)
touch net-GEN.SG player-INS.SG is not a mistake
'Touch of the net by a player is not a mistake.'

Modification by A_1(Ins) could also be possible with nouns derived from reflexive intransitive verbs with a participant expressed by prepositionless dative. We searched for the following productively derived nouns, i.e. *posmívání se* 'laughing-IPFV', *vysmívání se* 'mocking-IPFV'*, vyhýbání se* 'avoiding-IPFV' / *vyhnutí se* 'avoiding-PFV', and for one non-productively derived noun, i.e. *výsměch* 'mockery'. However, A_1(Ins) was found only with the non-productively derived noun *výsměch* 'mockery', cf. (41).

(41) *výsměch právu.*PAT *zástupcem.*ACT *státní moci* (SYN2009PUB)
mockery law-DAT.SG representative-INS.SG state power-GEN.SG
'mockery of law by a state power representative'

Agent expressed by the form of instrumental is possible also with some nouns derived from reflexive intransitive verbs with an obligatory free modification expressed by a prepositional group (or an adverb), e.g. productively derived noun *vloupání se* 'breaking-PFV in / break-in', cf. (42).

(42) *vloupání neznámým pachatelem.*ACT *do kiosku.*DIR3 *se spotřebním zbožím* (SYN2000)
'break-in by an unknown perpetrator into the kiosk with consumer goods'

3 Constructions with A_1(Ins) only

It has been already mentioned that both nouns derived from transitive and intransitive verbs are expected to allow modification by A_1(Ins) when A_2 is present; in other words, constructions in which only A_1(Ins) is expressed are considered to be ungrammatical, cf. Karlík [9, p. 40], Procházková [20, p. 41] and examples (16) to (18) above.

[1] In the present paper, on the basis of studied corpus material, we would like to point out that various nouns occur with A_1(Ins) not only when A_2 is present, but also when A_2 is omitted on the surface.[9] Thus the theoretical statement about ungrammaticality of such constructions should be specified.

A classification of deletion types is closely related to the type of coreference between the deleted word and its antecedent; the coreference may be grammatical[10] or textual; for types of deletions in nominalized structures see [12, pp. 83-86], among others.

We have found numerous occurrences of deletion of A_2 based on textual coreference. The antecedent of the deleted A_2 can be easily determined from the previous or following context, however the coreference relation cannot be explained by grammatical properties of the constructions. The deletions of A_2 based on textual coreference apply to both nouns derived from transitive verbs and nouns derived from intransitive verbs. Again, it concerns nouns derived by productive means as well as nouns derived by non-productive means. As for productively derived nouns, constructions in which A_2 is omitted on the surface are represented by both perfective and imperfective nouns. However, according to absolute frequencies of "transitive" nouns, given in Table 1 and Table 2, the constructions are more frequent with the perfective nouns; with some imperfective nouns they do not occur at all.

---

[9] Even one of the nouns listed in the constructions that are in the literature considered to be ungrammatical, i.e. the noun *vyprávění* 'telling', occurs in CNC subcorpora with A_1(Ins) only, cf. (17) and (24) above, and Table 1.

[10] We have found few occurrences in which the antecedent of the omitted A_2 could be identified on the basis of grammatical coreference; it concerns several verbs exhibiting the property of Control and their derivates. For example, we think it is the case of the verbs *předurčit / určit koho / co k čemu* 'to predetermine sb to do sth / for sth', and their adjectival derivates, i.e. *předurčený / určený k čemu* 'predetermined for sth'; as for the verbs (not typical representatives of verbs of Control), the grammatical coreference relation can be identified between the Controller (i.e. PAT(Acc) of the verbs *předurčit / určit* 'to predetermine') and the unexpressed Controllee, which is ACT within the active embedded objective clause or its nominalization modifying the verbs *předurčit / určit* 'to predetermine' (e.g. *předurčit koho, aby vykonal něco / k vykonání čeho* 'to predetermine sb to do sth / for doing sth') or PAT within the passive embedded objective clause or its nominalization modifying the verbs *předurčit / určit* 'to predetermine' (e.g. *předurčit výrobek k tomu, aby byl konzumován / ke konzumaci* 'to predetermine the product to be consumed / for consumption'). Typically, Controllee is an unexpressed "subject" of an infinitival construction modifying a verb of Control, e.g. *odhodlat se odejít / k odchodu* 'to resolve to leave / for leaving', however, also some verbs of Control without possibility to express the respective complementation by an infinitival construction exist, see [19]. We assume that in constructions of the verbs *předurčit / určit* 'to predetermine' and their adjectival derivates, the coreference relation between omitted A_2 and its antecedent can be interpreted on the basis of grammatical coreference, cf. *výrobek (je) předurčený / určený ke konzumaci lidmi.*ACT 'the product (is) predetermined for consumption by people', i.e. *konzumace výrobku lidmi* 'consumption of the product by people' is concerned.

"Transitive" productively derived nouns with A_1(Ins) only (e.g. perfective *odebrání* 'taking away', *vrácení* 'returning', *oznámení* 'announcing', and imperfective *ošetřování* 'treating', *požívání* 'consuming', *uvědomování si* 'being-IPFV aware / realization') are illustrated in (43) to (48).

(43) *čtyři [psi] jsou ... volní k odebrání novým chovatelem*.ACT (SYN2009PUB)
'four [dogs] are ... free for taking away by a new breeder'

(44) *průkazka bude, po vrácení poštou*.ACT*, uložena u nich* (SYN2009PUB)
'the identity card will be, after returning by the post office, deposited at their place'

(45) *Po oznámení rodiči*.ACT *policisté začali po neznámém muži pátrat.* (SYN20009PUB)
'After announcing by parents, policemen began to search for an unknown man.'

(46) *jeho zdravotní stav vyžaduje nezbytně ošetřování jinou osobou*.ACT (SYN2006PUB)
'his health condition requires indispensably treating by another person'

(47) *Maso nakažených zvířat je nevhodné pro požívání lidmi*.ACT. (SYN2006PUB)
'Meat of infected animals is not fitting for consuming by people.'

(48) *Soubor práv… byl budován po staletí uvědomováním si lidskou inteligencí.*ACT. (SYN2009PUB)
'Legal code was created during centuries by being_aware REFL by human intelligence.'

"Intransitive" productively derived nouns with A_1(Ins) only (e.g. *napomáhání* 'helping-IPFV / aiding-IPFV'*, vyhrožování* 'threatening-IPFV'*, vloupání se* 'breaking-PFV in / break-in'*, dotazování se* 'questioning-IPFV', *vzdání se* 'surrendering-PFV') are exemplified in (49) to (53).

(49) *jakékoli napomáhání sestřičkou*.ACT *je … vyloučeno.* (SYN2009PUB)
'any helping by the nurse is … excluded'

(50) *horníci mluvili především o vyhrožování zaměstnavatelem*.ACT (SYN2005)
'miners talked mainly about threatening by the employer'

(51) *klasické vloupání neznámým pachatelem*.ACT. (SYN2000)
'classic break-in by an unknown perpetrator'

(52) *při běžném dotazování pracovníkem*.ACT (SYN2009PUB)
'during common questioning by the worker'

(53) *zánik platnosti zaregistrované ochranné známky např. vzdáním se jejím majitelem.*ACT (SYN2006PUB)
'expiration of the registered trademark e.g. by surrendering by its owner'

As for non-productively derived nouns, the "transitive" nouns with A_1(Ins) only, i.e. *dodávka* 'delivery', *konzumace* 'consumption', *výuka* 'teaching / instruction', are illustrated in (54) to (56).

(54) *v oblasti finálních dodávek velkou specializovanou firmou*.ACT (SYN2000)
'in the field of final deliveries by a big specialized company'

(55) *někteří lidé volí možnost výuky soukromým lektorem.*ACT. (SYN2009PUB)
'some people choose the possibility of teaching by a private language assistant'

(56) *rostliny, které... nejsou vhodné ke konzumaci člověkem.*ACT (SYN2006PUB)
'plants which ... are not fitting for consumption by a man'

The "intransitive" non-productively derived nouns with A_1(Ins) only, i.e. *domluva* 'caution', *výpomoc* 'help', are exemplified in (57) and (58). We find it interesting that there is even no occurrence of the noun *domluva* 'caution' modified by A_1(Ins) and A_2 (hypothetical example *domluva dětem strážníky* 'caution / talking to children by police officers'); there are three occurrences of the noun *domluva* 'caution' modified by A_1(Ins) only, i.e. without any other participant expressed.

(57) *Po domluvě strážníky*.ACT *děti z místa odešly.* (SYN2009PUB)
'After caution by police officers children leaved the place.'

(58) (in the context of the 30. anniversary of the occupation of former Czechoslovakia in the year 1968)
*30. výročí přátelské výpomoci spojeneckými armádami.*ACT (SYN2006PUB)
'30. anniversary of the friendly help by allied armies'

As for "transitive" nouns, the ratio of presence of A_2 (category A) to its absence on the surface (category B), in case a noun is modified by A_1(Ins), is captured in Tables 1-3; concerning "intransitive" nouns, the ratio is given in Table 4.

However, regardless the numerous examples of constructions with A_1(Ins) only, given above, there are nouns that probably really do not allow for modification by A_1(Ins) without expression of A_2 on the surface, and thus constructions with these nouns are hypothesized to be ungrammatical, cf. (59).

(59) ??*ujímání se hodnými lidmi.*ACT (introspective example)
taking_charge-NOM.SG REFL good-INS.PL people-INS
'taking charge by good people'

4 "Intransitive" Nouns: Discussion of the Results

In this section we summarize and discuss main observations concerning "intransitive" nouns that, according to the corpus material, allow for A_1(Ins). The total number of the "intransitive" nouns (lemmas) that occurred with A_1(Ins) is 17 (i.e. 13 productively derived nouns and 4 non-productively derived nouns) and the total number of occurrences of A_1(Ins) modifying the nouns is given in Table 4. However, despite the enlarged and corpus material, we do not answer the question why some nouns do not occur with A_1(Ins).

A_1(Ins) occurs with nouns derived from verbs that can be passivized as well as with nouns derived from reflexive verbs that do not allow to be changed to passive. These observations seem to correspond to Karlík's claim that the structures given above in (2) and (4) do not show structural differences corresponding with the active – passive voice distinction and thus A_1(Ins) is considered not to be licensed through passivization.

Modification by A_1(Ins) is possible with "intransitive" nouns representing various semantic classes: e.g. nouns of communication (e.g. *dotazování se* 'questioning', *vyhrožování* 'threatening'; *domluva* 'caution'), nouns of mental action (e.g. *porozumění* 'understanding'), nouns of motion (e.g. *vniknutí* 'penetrating'), nouns of location (e.g. *vloupání se* 'break-in'), nouns of contact (e.g. *dotknutí se* 'touching', *zmocnění se* 'seizing / seizure'; *dotyk* 'touch'), nouns of combining (e.g. *přistoupení* 'joining / accession').

Considering forms of the second complementation, the "intransitive" nouns can be modified by A_1(Ins) and at the same time by the second complementation (A_2) expressed by a prepositionless case (not only prepositionless dative, mentioned in the literature, but also prepositonless genitive)[11] as well as by a prepositional group (or an adverb).

As for nouns derived from verbs with a participant expressed by prepositionless genitive, an analogy to constructions corresponding to verbal transitive constructions is possible, cf. (60) and (61). The original adverbal case (i.e. Acc vs. Gen) does not seem to be so important. A_1(Ins) serves as one of possible forms for expression of Agent, used especially in cases when a noun denotes an action and other forms of Agent (possessives or genitive) are not possible or they are not proper from another reason; for example, it is well-known that a possessive adjective can be derived only under certain conditions; as for the genitive form of the Agent, in case A_2 is expressed, it would lead to constructions with double post-nominal genitives, cf. (62). One of the reasons for usage of A_1(Ins) instead of A_1(Gen) in case A_2 is omitted on the surface, is probably the fact that the genitive form may be syntactically ambiguous, thus the form of instrumental is used to avoid the ambiguity; we have in mind especially the case of the syntactic ambiguity of ACT and PAT, cf. (63), or the syntactic ambiguity of ACT and ADDR, cf. (64). Usage of A_1(Ins) instead of A_1(Gen) in order to avoid syntactic ambiguity of the genitive form holds also for nouns derived from transitive verbs, cf. (65) and (66).

(60) *<u>přepadení</u> televize.*PAT *teroristy.*ACT (introspective example)
attacking television-GEN.SG terrorist-INS.PL
'attacking of the television by terrorists'

[11] Prepositionless instrumental is probably impossible as such a hypothetical construction would consist of two participants expressed by Ins, e.g. ??*<u>pohrdání</u> kým*.PAT *kým*.ACT 'contempt of sb by sb', ??*<u>nákaza</u> čím*.PAT *kým*.ACT 'getting infected with sth by sb'.

(61) *zmocnění se televize*.PAT *teroristy*.ACT (SYN2009PUB)
seizure REFL television-GEN.SG terrorist-INS.PL
'seizure of the television by terrorists'

(62) *dožití pojištěné osoby*.ACT *sjednaného konce*.PAT *pojištění* (SYN2009PUB)
living_to insured-GEN.SG person-GEN.SG agreed-GEN.SG end-GEN.SG insurance-GEN.SG
'living of the insured person to the agreed end of the insurance'

(63) *sudí pískají každý dotyk hráče*.ACT/PAT *jako faul* (SYN2006PUB)
'Referees signal by a whistle every touch of the player as a foul.'
'the player touches / the player is touched'

(64) *dotazování pracovníka*.ACT/ADDR (introspective example)
questioning-NOM.SG worker-GEN.SG
'questioning of the worker, i.e. the worker asks / the worker is asked'

(65) *poskytovat informace pouze na základě zmocnění rektora*.ACT/PAT (introspective example)
to give information only on the basis of authorization rector-GEN.SG
'to give information only on the basis of authorization by the rector / of the rector'

(66) *poskytovat informace pouze na základě zmocnění rektorem*.ACT (SYN2006)
to give information only on the basis of authorization rector-INS.SG
'to give information only on the basis of authorization by the rector'

Other "intransitive" nouns seem to use the instrumental form of Agent analogically as well, although the second complementation is expressed by the form different from prepositionless genitive.

Considering nouns derived from verbs by productive means, we suppose the nouns modified by A_1(Ins) exemplified in the present paper denote an action. Also several nouns derived from verbs by non-productive means occurred with A_1(Ins), e.g. *domluva* 'caution'. It would be interesting to study in detail whether the non-productively derived nouns denote an action as well. However, this issue goes beyond the major topic of this paper, and so we leave it for further research.

5 Conclusion

Czech nouns derived from intransitive verbs, both productively and non-productively derived nouns, allow for modification by A_1(Ins) to a higher extent than it has been expected. However, in comparison with "transitive" nouns they are less frequent. On the basis of corpus material, we considerably increased the list of "intransitive" nouns that

allow for A_1(Ins) modification and provided more detailed classification of the nouns according to the form of the second complementation and the semantic class the noun belongs to. It has turned out that modification by A_1(Ins) is possible not only with nouns derived from verbs that can be passivized, but also with nouns the source verbs of which cannot be changed to passive (it concerns especially nouns derived from reflexive verbs, both transitive and intransitive). Modification by A_1(Ins) is possible even when the second complementation A_2 is omitted on the surface, which should lead to the revision of the non-specific statement about ungrammaticality of such constructions.

References

[1] Alexiadou, A. (2001). *Functional Structure in Nominals. Nominalization and ergativity*. John Benjamins Publishing Company, Amsterdam/Philadelphia.

[2] Czech National Corpus – SYN2000 (2000). Prague: Institute of the Czech National Corpus, Faculty of Arts, Charles University in Prague. Accessible at: `http://www.korpus.cz`.

[3] Czech National Corpus – SYN2005 (2005). Prague: Institute of the Czech National Corpus, Faculty of Arts, Charles University in Prague. Accessible at: `http://www.korpus.cz`.

[4] Czech National Corpus – SYN2006PUB (2006). Prague: Institute of the Czech National Corpus, Faculty of Arts, Charles University in Prague. Accessible at: `http://www.korpus.cz`.

[5] Czech National Corpus – SYN2009PUB (2010). Prague: Institute of the Czech National Corpus, Faculty of Arts, Charles University in Prague. Accessible at: `http://www.korpus.cz`.

[6] Czech National Corpus – SYN2010 (2010). Prague: Institute of the Czech National Corpus, Faculty of Arts, Charles University in Prague. Accessible at: `http://www.korpus.cz`.

[7] Grimshaw, J. (1991). *Argument Structure*. The MIT Press, Cambridge, Mass.

[8] Karlík, P. (2000). Valence substantiv v modifikované valenční teorii (Valency of nouns in a modified valency theory). In Hladká, Z. and Karlík, P., editors, *Čeština – univerzália a specifika 2*, pages 181–192, Masarykova univerzita, Brno.

[9] Karlík, P. (2004a). Mají dějová substantiva slovesný rod? (Are derived action nominals sensitive to the active – passive voice distinction?) In Hladká, Z. and Karlík, P., editors, *Čeština – univerzália a specifika* 5, pages 33–46, Nakladatelství Lidové noviny, Praha.

[10] Karlík, P. (2004b). Pasivum v češtině (The passive voice in Czech). *Slovo a slovesnost*, 65:83–112.

[11] Karlík, P. and Nübler, N. (1998). Poznámky k nominalizaci v češtině (Notes on nominalization in Czech). *Slovo a slovesnost*, 59:105–112.

[12] Kolářová, V. (2010). Valence deverbativních substantiv v češtině (na materiálu substantiv s dativní valencí) (Valency of deverbal nouns in Czech: With a special regard to nouns with dative valency). Karolinum, Praha.

[13] Kolářová, V. Adverbální předmětový genitiv a jeho protějšky v nominálních konstrukcích: Případ posesiva (Adverbal objective genitive and its counterparts in nominal constructions: The case of possessives). In *Sborník z konference Slovo a tvar v štruktúre a v komunikácii*, Bratislava, 2012. In press.

[14] Křížková, H. (1968). Substantiva s dějovým významem v ruštině a v češtině (Nouns with action meaning in Russian and Czech). In *Kapitoly ze srovnávací mluvnice ruské a české III. O ruském slovese*, pages 81–152, Academia, Praha.

[15] Lopatková, M., Žabokrtský, Z., and Kettnerová, V. et al. (2008). *Valenční slovník českých sloves* (Valency dictionary of Czech verbs). Univerzita Karlova v Praze, Nakladatelství Karolinum, Praha.

[16] Novotný, J. (1980). Valence dějových substantiv v češtině (Valency of non-productively derived nouns in Czech). Sborník pedagogické fakulty v Ústí nad Labem. SPN, Praha.

[17] Panevová, J. (1980). *Formy a funkce ve stavbě české věty* (Forms and functions in the structure of Czech sentences). Academia. Praha.

[18] Panevová, J. (2000). Poznámky k valenci podstatných jmen (Notes on valency of nouns). In Hladká, Z. and Karlík, P., editors, *Čeština – univerzália a specifika* 2, pages 173–180, Masarykova univerzita, Brno.

[19] Panevová, J., Řezníčková, V., and Urešová, Z. (2002). The theory of control Applied to the Prague Dependency Treebank (PDT). In *Proceedings of the 6th International Workshop on Tree Adjoining Grammars and Related Frameworks*, pages 175–180, Università di Venezia.

[20] Procházková, V. (2006). *Argument structure of Czech event nominals*. Master Thesis, Univesity of Tromsø.

[21] Veselovská, L. (2001). K analýze českých deverbálních substantiv (On the analysis of Czech deverbal nouns). In Hladká, Z. and Karlík, P., editors, *Čeština – univerzália a specifika 3*, pages 11–27, Masarykova univerzita, Brno.

Corpus-based Online Word Formation Exercises for Advanced Learners of English – Challenges and Solutions

Grzegorz Krynicki

Faculty of English, Adam Mickiewicz University, Poznań, Poland

Abstract. The paper presents the design and operation of an online platform for word formation practice. The system is based on a pre-defined list of pairs of base and derived forms and usage examples drawn automatically from the British National Corpus. A procedure for the extraction of example sentences is outlined. Results of 372 users' interacting with the system for over 4.5 month are reviewed. The question about what factors influence users' evaluation of specific exercises as more difficult is addressed. The results may be relevant in the area of language testing, preparation of examination materials, student-teacher online interaction and teaching English word formation.

1 Word Formation in Learning English as a Foreign Language

Advanced learners of English willing to expand their vocabulary appreciate the study and practice of word formation. Knowing how to combine a small set of particles (prefixes like *non-*, *im-*, *de-* or suffixes like *-able*, *-ish*, *-ly*) with a few base words may increase learner's vocabulary significantly and with minimum effort. For example, knowing what these particles mean and the meaning of the simple root morpheme *port* ('to send' or 'carry') most advanced learners would also probably guess the meanings of *export*, *import*, *deport*, *portable* and *transport*. To look at it from another perspective – in the British National Corpus (BNC) the prefix over- begins 2013 different word types [4, p. 13]. Knowing how over- influences the meaning of words it is attached to allows to know close to half of what each of these 2013 words means.

The set of corpus-based gap fill exercises in word formation described in this paper is based on two major sources, a word formation list of 1,929 tokens [11] and the British National Corpus of 100 million tokens. The word formation list was prepared for 1–3BA and 1MA students of the Faculty of English, Adam Mickiewicz University in Poznań, Poland by teachers of practical English. However, they can be useful to all advanced learners of English (from B2 to C2 in CEFR). The list of word forms was compiled based on articles and course-books used by students of English philology. They illustrate most word-formation mechanisms (prefixation, suffixation, compounding, clipping etc.) and cover a wide range of general topics. All example sentences were drawn from the BNC (BNC 2001) to ensure that the language used in these activities is authentic and varied.

These exercises are aimed at advanced students of English who want to:

- improve their receptive and productive command of English vocabulary,
- inductively learn English word formation rules,
- overcome the interference from their native tongue morphology,
- master vocabulary in authentic sentence context,
- learn Polish equivalents of English complex words (although the knowledge of Polish is not necessary to benefit from all other aspects of the exercises),
- practice for advanced English grammar tests and examinations,

- improve their skills in dictionary word lookup – dictionaries often provide definitions for simpler words and leave the creation of complex words with relatively intuitive meanings to the user.

The system may also be used by EFL teachers and test designers. Although new corpora and other lists of base form – derived form pairs can easily be added by the administrator to create new exercises, the online interface does not allow manipulating these resources. Free unrestricted online access to the exercises is possible at the following address: `http://wa.amu.edu.pl/~krynicki/wf`.

2 How to Use the System

When the user logs into the system, he will see a table of 6 columns (Fig. 10, last page of this paper). In the 2nd column of the table, 10 example sentences are listed, each with a gap that needs to be filled with a word form derived from the base word given in the 3rd column. If the example sentence is too ambiguous, the user may click "More" to see additional example sentences. If are ready to see the answer, click "Answer" in the 4th column. The user compares his answer with the answer that appears in the 5th column and mark check-box in the 6th column if the user's answer differed in any way from the answer provided by the system. Once the user has done all the 1,929 exercises, he will have the possibility to export the difficult items to a tab-separated text file so that he can drill them in spaced memory software, e.g. [1] or [10].

All examples were drawn from the corpus automatically so it may happen that even top students will have problems guessing the missing word form on the basis of a single ambiguous example sentence. For this reason, the option of viewing two additional example sentences has been provided. If the user clicks "More" – a new sentence will drop down below the already visible example. If the exercises are used to practice for a written examination, it should be kept in mind that in most exams where word form gap fill exercises appear the user will not have the possibility to see more than 1 example sentence. Moreover, the user will have to write his answers not just think about them as is the case with this system. For these reasons, before providing the answer, the user should try to mentally spell the word and mark it as difficult if he makes the slightest mistake.

If the user does not know what the English word form means in Polish, a list of equivalents will appear in a balloon tip when the user hovers his mouse pointer over most word forms. If the word form is clicked, the user will be redirected to a form where he can edit the Polish equivalents of the word form and English example sentences. The editions will be visible to others after they have been accepted by the administrator. In the system, Polish equivalents were drawn automatically from various electronic English-Polish dictionaries without any regard to their part of speech (POS), order in which they originally appeared or phrases they may be used in.

Each student had a different order of sentences submitted. The order was generated in a pseudo-random fashion during his first visit. Randomization was adopted as a precaution against students who would like to solve the exercises simultaneously on different computers and help each other. Every time the student logs into the system he can continue his work without having to repeat the exercises he has already done.

At the bottom of the screen the user sees the progress bar so that he can monitor how many exercises out of 1,929 he has done.

3 Selection of Example Sentences

Automatic selection of example sentences was conducted taking into consideration the length of the candidate sentences and the number of proper names they contained. Roughly, the more the example sentence approached the "ideal" length and the fewer proper names it had, the more chances it had of being selected.

1. The BNC corpus was split into approx. 6 million sentences.
2. Corpus entities were converted to Windows-1252 encoded text to make their tokenization and display easier, e.g. the entity *&bquo;* used in BNC to denote a double quotation mark is not a standard HTML entity and was converted to *"*.
3. Tokenization and down-casing, e.g. *She can't stand her mom's "complaints".* was converted to *she can not stand her mom's "complaints".*
4. By the rule of the thumb
   - Sentences of 80 characters or fewer were excluded as they were considered to provide not enough context to guess the gapped word. Although excessively long sentences often contain material irrelevant for the guessing of the gapped word, the upper limit for the sentence length was not set. The ideal sentence length was set at 160 characters;
   - Sentences containing a capital letter anywhere else than at the beginning of the sentence were excluded in the first stage to minimise the number of proper names and abbreviations in the example sentences.
5. For each of the remaining sentences:
   - Base form of each word in the sentence was obtained by consulting lemmatized word frequency lists [5];
   - If the base form was present in the WA list (among lower-case derived words), the sentence was considered a potential example of the usage of this base form;
   - Potential examples were ordered from the ones closest to the ideal length to the ones farthest from the ideal length. In this order example sentences were submitted to the student.
6. If 3 sentences meeting the above criteria were not found for a word form from the WA list, in the second stage, the missing sentences were filled in from those containing capital letters elsewhere than at the beginning of the sentence in the increasing order of the number of capital characters they contained.

As an effect of this procedure, in 92.3% of exercises the word form was illustrated by 3 example sentences, in 4.5% of exercises 2 example sentences were used and 1 sentence was used to illustrate the usage of the remaining 3.2% of word forms.

No of sentences	Frequency	Relative frequency
1	62	0.0321
2	86	0.0446
3	1,781	0.9233

Table 1. Frequency of exercises with 1, 2 or 3 example sentences

The author is aware of many imperfections the above algorithm has, especially in the view of solutions proposed by e.g. [6] or [3]. In future stages of the project, parameters that characterize the readability, complexity and stylistic properties of the examples will be considered.

4 Students' Judgements about the Difficulty of the Exercises

Two freshman groups attending classes in FCE General English and CAE English Grammar were suggested to use the platform to prepare for their final practical English examination. A notice about the exercises was also published on a Moodle site devoted to practical English examination that all and only Faculty members had access to. The notice additionally informed that example sentences used in the exercises would not appear in the final exam. Students were also reminded that the word formation component of the exam will include only the words from the WA list.

The Faculty members included over 1,500 BA and MA students from B2 to C2 CEFR levels. Over the period of 4 months and 20 days (Apr 9th – Aug 29th), 389 students logged into the system at least once. For 17 of them, there is no evidence of them doing any exercises as batches of more than 10 completed exercises were evaluated. The remaining 372 students did 417.3 out of of 1,929 exercises on average (21.6%). 45 students completed all 1,929 exercises. Each exercise was solved by at least one student. A unique exercise was solved by 23.8 student on average.

In order to improve the interface and the content of the exercises as well as to aid the preparation of tasks for the final practical English examination, an analysis of the students' responses was conducted. Students' responses included information about which exercises they found difficult. The difficulty judgements were then related to properties of the prompt base word, expected word form and the example sentence from which it was extracted.

4.1 Statistics on Students' Judgements

The user of the system was encouraged to mark as difficult the exercise in which any mistake was made. He was informed that once all the exercises have been completed, difficult items could be exported for drilling in spaced memory software. 254 out of 372 students (68.3%) marked at least one exercise as difficult. Among these students, the average number of items marked as difficult was 109.0 i.e. 16.2% (standard deviation of 231.2). This constituted 26.1% of all the exercises they tried to solve on average. The maximum percentage of exercises a single student marked as difficult 68.2% (225 of 330). 1,919 exercises out of 1,929 were marked as difficult at least once. Fig. 2 illustrates the distribution of exercises with different levels of difficulty according to students.

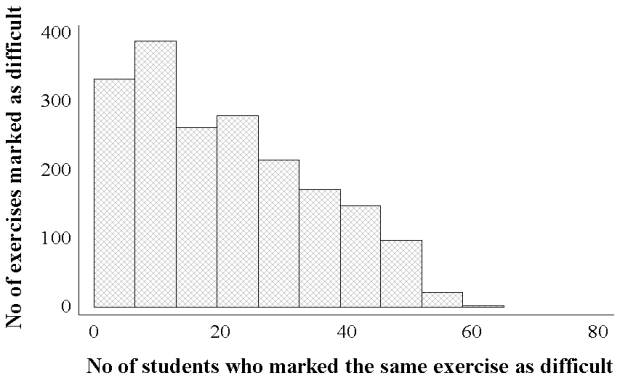


Fig. 1. Histogram of 1,919 exercises marked difficult at least once by 254 students with the numbers of times one exercise was marked difficult and observations pooled into 10 classes.

Students found the exercises rather challenging. One exercise was marked difficult by 61 students. Half of the exercises were marked difficult 19 or more times. 36 exercises were marked difficult exactly once.

4.2 Factors Potentially Influencing Students' Judgements

In this study, the influence of the following factors on users' evaluation of exercise difficulty was considered: student's language competence (for a sample of students), properties of base form hint and the gapped word form, number of example sentences illustrating the use of the word form and their length as well as availability of Polish translation of the word form. The potentially significant factors that were not considered include student's language aptitude and the properties of the context in which word form appeared were not considered.

Properties of the base word and derived form included their frequencies in BNC frequency list [5], similarity of POS codes listed for them in the BNC list as well as their graphemic similarity.

To obtain these properties for base and derived forms, the BNC list was preprocessed in the following way:

- Complex tags were "rounded" to general categories of nouns, adjectives, adverbs, verbs (e.g. NN0 common noun and NN1 singular common noun were pooled into NN category). All other POS were discarded;
- From portmanteau tags, used in CLAWS to indicate where the system was uncertain between two possible analyses, only the first one was chosen;
- Frequencies of identical words within the same rounded and simplified POS were added.

The resulting list contained 365,040 nouns, 68,460 verbs, 190,086 adjectives and 9,739 adverbs.

Graphemic similarity between base form and word form was another characteristic whose influence on exercise difficulty was considered. Graphemic similarity was expressed by two parameters. First, by the longest common prefix between the two

forms. It was calculated as the maximum number of characters that the two words shared at their beginnings. Second, by the longest common subsequence ratio (LCSR) determined by dividing the length of their longest common subsequence by the length of the longer word [9]. It was hypothesized that the greater the similarity between the two forms, the easier is will be to guess the derived from given the base form and the less frequently an exercise including them will be marked as difficult. Table 2 includes an extract from the list of 1929 word forms annotated for similarity and frequency information (for complete list refer to [8]).

4.3 Significance Tests of Factors Potentially Influencing Students' Judgements

Statistical significance tests were used to identify factors that had a significant influence over users' judgements about exercise difficulty. The dependent variable in all tests was the number of times a given exercise was judged difficult (ranging from 0 to 61). The independent variables of word frequency were grouped under 4 or 5 variables: 0 if 0 frequency was observed, 1–3 for data points up to 25th, 50th and 75th percentile respectively and 4 otherwise (Table 3). Longest common subsequence ratio, longest common prefix lengths ("prefix" in a pattern-matching rather than linguistic sense) and length of the first sentence were grouped as presented in Table 4.

Base	*POS*	*Base freq*	*Word form*	*WF POS*	*WF freq*	*Intersection*	*Long. Prefix*	*LCSR*	*Pol. equ.*	*No sent.*
abandon	nv	1316	abandonment	n	496	n	7	0.64	1	3
able	j	30410	ability	n	9135		2	0.43	1	3
normal	jd	12452	abnormal	j	810	j	0	0.75	1	3
normal	jd	12452	abnormality	n	287		0	0.55	1	3
normal	jd	12452	abnormally	d	151	d	0	0.6	1	3
cite	v	282	above-cited		151		0	0.6	0	3
abstain	jnv	129	abstainer	n	3	n	7	0.78	1	3
abstain	jnv	129	abstention	n	99	n	4	0.6	1	3
abstain	jnv	129	abstinence	n	150	n	4	0.6	1	3
abstain	jnv	129	abstinent	j	10	j	4	0.67	1	2

Table 2. First 10 word forms annotated for base form, parts of speech, frequencies in BNC, intersection of the sets of POS tags for each form, length of the longest common prefix, longest common subsequence ratio (LCSR), information about whether the option of displaying Polish translation of the word form was available, number of example sentences that illustrated the use of the word form. POS abbreviations: j – adjective, d – adverb, v – verb, n – noun. The whole list is available at `http://wa.amu.edu.pl/~krynicki/wf/table2.csv`.

Grouping variable	*Range*	*Freq of base form (x)*	*Frequency of derived form (x)*
1	0 <x≤ 25%	5-1,658	1-68
2	25% <x≤ 50%	1,659-4,372	69-273
3	50% <x≤ 75%	4,373-11,650	274-1,062
4	75% <x≤ max	11,651-129,547	1,063-48,374

Table 3. Transformation of Frequency of base form in BNC and Frequency of derived form into 4 grouping variables

Grouping variable	*LCSR*	*Longest common prefix*	*Length of the first sentence*
0	0-0.15	0-2	52-110
1	0.16-0.50	3-4	111-159
2	0.51-0.60	5	160
3	0.61-0.67	6	161-200
4	0.68-1	7-11	201-391

Table 4. Transformation of Longest common subsequence ratio, Longest common prefix and Length of the first sentence into 5 grouping variables

In all tests at least one of ANOVA assumptions was violated – the standardized skewness and/or kurtosis was outside the range of -2 to +2 for at least one of the factors and/or the difference between the smallest standard deviation and the largest was greater than 3 to 1. Therefore, Kruskal-Wallis Test (KWT) was used to test significance of most factors.

Language competence

General English written test results were known for 21 of 254 students who marked at least one exercise as difficult. A relatively weak positive correlation was found between student's results and the number of exercises he marked difficult (Spearman rank correlation coefficient=0.22, p=0.023). This indicates that marking exercises was not directly related to language competence but it may have rather reflected student's willingness to review items in the future to remember them better in the practical English exam, student's diligence in general or student's preference for reviewing items in spaced memory software rather than using web interface.

Base form frequency

KWT was used to test the null hypothesis that the medians of grouping variable of *Times judged difficult* (i.e. how many times an exercise was marked difficult by all students) within each of the 4 levels of the grouping variable of *Frequency of base word* are the same. The test statistic K=10.15 and p=0.0173, which is a significant result at the 0.05 level. Fig. 2 presents a Box-and-Whisker plot of the dependent variables against the factor. Boxes extend to 1st and 3rd quartile, whiskers extend to the maximum observations. Notches that do not overlap indicate medians that are significantly different. Therefore, contrasts between the levels 1:2, 1:3, 2:4 and 3:4 are statistically significant.

In other words, exercises using most and least frequent base forms as hints are judged significantly more difficult than those using hints of frequency between 25th and 75th percentile.

Possible reasons may be related to higher derivational productivity of most frequent base forms and their higher ambiguity. Low-frequency base forms may be of less help as a hint because of their lower familiarity to students.

Word form frequency

The aim of the second test was to test the null hypothesis that the medians of Times judged difficult within each of the 4 levels of Frequency of derived form are the same. The test statistic K=15.1029, p=0.0017, which is significant at 0.05 level.

Derived forms of low frequency were difficult to guess if gapped from an exercise probably because of their low familiarity to students.

Scalar of intersection of POS tag sets for base and derived forms

Consider two sets, one containing POS tags listed in BNC for base word used as a hint in our word formation exercise and the other containing POS tags for the form derived from the hint but gapped in the example sentences. Intersection of these two sets is the set of POS tags base and derived forms have in common. The scalar (or cardinal) of the intersection is the number of POS tags shared by both forms. Fig. 4 illustrates the result of KWT of *Times judged difficult* against the *POS intersection scalar*.

This effect to some extent may be explained by the fact that students assume derivation usually changes morphosyntactic category of the base form. Therefore, the greater the overlap between POS tags of the two forms, the more problematic such derivation may appear. This effect is also reinforced by the fact that the greater the intersection, the greater the POS set of each form and the greater their ambiguity.

Longest common prefix and LCSR

The first level of *Longest common prefix* (0 indicating prefixes of 0–2 characters) differs significantly from all the other levels (Fig. 5) with respect to the difficulty of exercises containing forms that begin with this prefix.

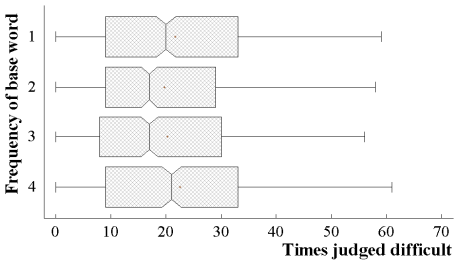


Fig. 2. Times judged difficult vs. Frequency of base word. K=10.15, p=0.0173

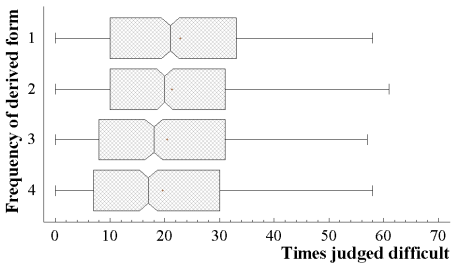


Fig. 3. Times judged difficult vs. Frequency of derived form. K=15.1029, p=0.0017

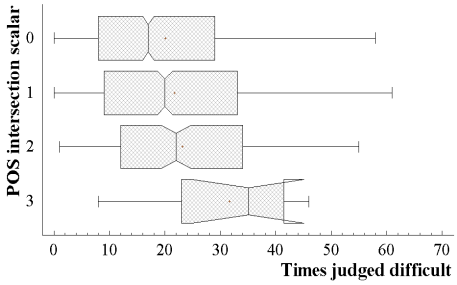


Fig. 4. Times judged difficult vs. POS intersection scalar. K=15.3427, p=0.0015

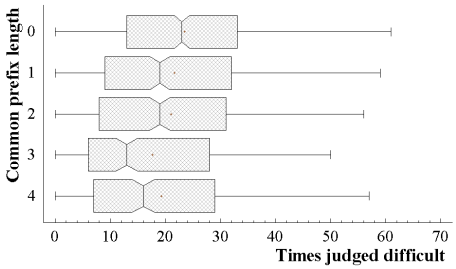


Fig. 5. Times judged difficult vs. Common prefix length. K=49.625, p=0.0000

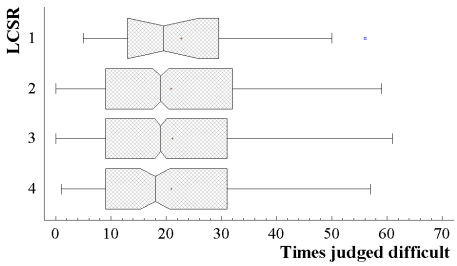


Fig. 6. Times judged difficult vs. LCSR. K=0.3798, p=0.9443

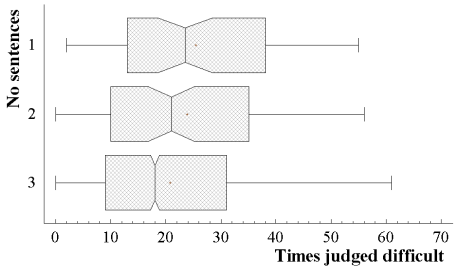


Fig. 7. Times judged difficult vs. Number of example sentences. K=9.20819, p=0.0100

Short common prefix or lack of it may make exercise more difficult. Similarity anywhere within the words as expressed by LCSR (Fig. 6) does not significantly affect how exercises containing them are evaluated (p=0.9443).

Number of example sentences and sentence length

Students' judgements indicate that having 3 example sentences made their task significantly easier than when they have just 1 example (Fig. 7). It is also possible however that the greater difficulty of exercises with 1 example sentence may follow from the fact that if only 1 usage example meeting criteria described in 3 was found in BNC for the given word form it must be rare and therefore difficult no matter how many examples it would be illustrated with.

Sentences shorter than 111 characters as well as those longer than 200 increase the chances that the student will find the exercise difficult (Fig. 8). This last result may have been reinforced by time pressure before the exams – reading lengthy sentences may have been considered by students a waste of time.

Polish equivalents

After trying to guess the English derived form, the user could look up the correct answer in English and make sure he knew its Polish equivalents. Learning new Polish meanings could influence his decision about whether to mark the exercise as difficult. KWT revealed a significant relationship between the presence of Polish equivalent and the difficulty of the exercise (Fig. 9).

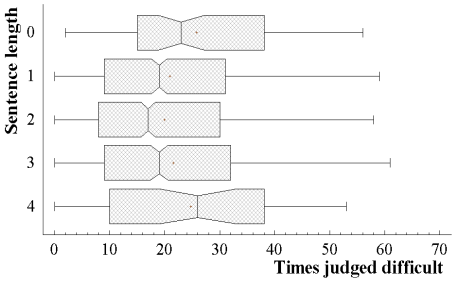


Fig. 8. Times judged difficult vs. Sentence length. K=15.1209, p=0.0045

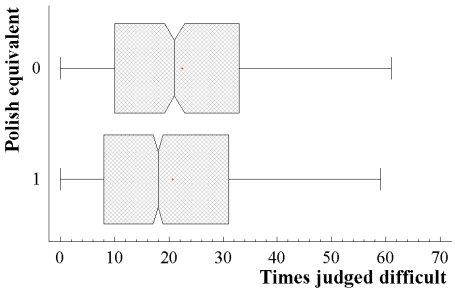


Fig. 9. Times judged difficult vs. Polish equivalent. K=3.8961, p=0.0484

5 Conclusions

The experiment described in this paper indicates that even a simple method of automatic selection of example sentences from a sufficiently large corpus may result in useful and engaging word formation exercises. The benefit of exercises of this form is not limited to acquiring correct English word-formation rules. Due to authentic sentence context, they develop the learner's grammar skills, teach meaning and meaning relationships and collocations. Moreover, a well designed word formation exercise with appropriate context is not only more effective but also more interesting than isolated word lists (c.f. [2, p. 28]).

Practical conclusions that follow from the above study may include:

- Word forms derived from base forms by other processes than prefixation are considered more difficult and should probably be paid greater attention to by learners and teachers;
- Word formation exercises using most and least frequent base forms as hints are more challenging than hints of average frequency;
- Students should be aware that derivation does not always change morphosyntactic category of the base form;
- With automatically extracted examples, it is important that they have alternatives;
- The absence of L1 equivalents of gapped word forms increases the perception of the exercise as a difficult.

Future version of the system will incorporate methods of example sentence extraction so that the context of the gapped word form is balanced for frequency and so that important collocations of the word form are represented. Other forms of word formation exercises will also be introduced.

References

[1] Anki. (2013). Spaced repetition software. URL: http://ankisrs.net.

[2] Balteiro, I. (2011). Awareness of L1 and L2 Word-formation Mechanisms for the Development of a More Autonomous L2 Learner. *Porta Linguarum*, 15:25–34.

[3] Didakowski, J., Geyken, A., and Lemnitzer, L. (2012). Automatic example sentence extraction for a contemporary German dictionary. In *Proceedings EURALEX 2012*, pages 343–349, Oslo.

[4] Joandi, L. (2012). Productivity Measurements Applied to Ten English Prefixes. URL: `http://urn.kb.se/resolve?urn=urn:nbn:se:su:diva-81966`.

[5] Kilgarriff, A. (1995). BNC database and word frequency lists. URL: `http://www.kilgarriff.co.uk/bnc-readme.html`.

[6] 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 XIII EURALEX International Congress*, pages 425–432.

[7] Krynicki, G. (2013a). Corpus-based word form gap fill exercises for advanced learners of English. URL: `http://wa.amu.edu.pl/~krynicki/wf/`.

[8] Krynicki, G. (2013b). WA list of ~1929 word forms annotated for frequency and similarity information. URL :`http://wa.amu.edu.pl/~krynicki/wf/table2.csv`.

[9] Melamed, I. D. (1999). Bitext maps and alignment via pattern recognition. *Computational Linguistics*, 25(1):107–130.

[10] Supermemo. (2013). Spaced repetition software. URL: `http://www.supermemo.pl/`.

[11] Szczegóła, T. (2012). WA list of ~1929 word forms was compiled by and is available on WA UAM PNJA. Moodle site: `http://wa.amu.edu.pl/moodle/mod/resource/view.php?id=32196`.

[12] The British National Corpus, version 2 – BNC World. (2001). Oxford University Computing Services on behalf of the BNC Consortium. Accessible at: `http://www.natcorp.ox.ac.uk/`.

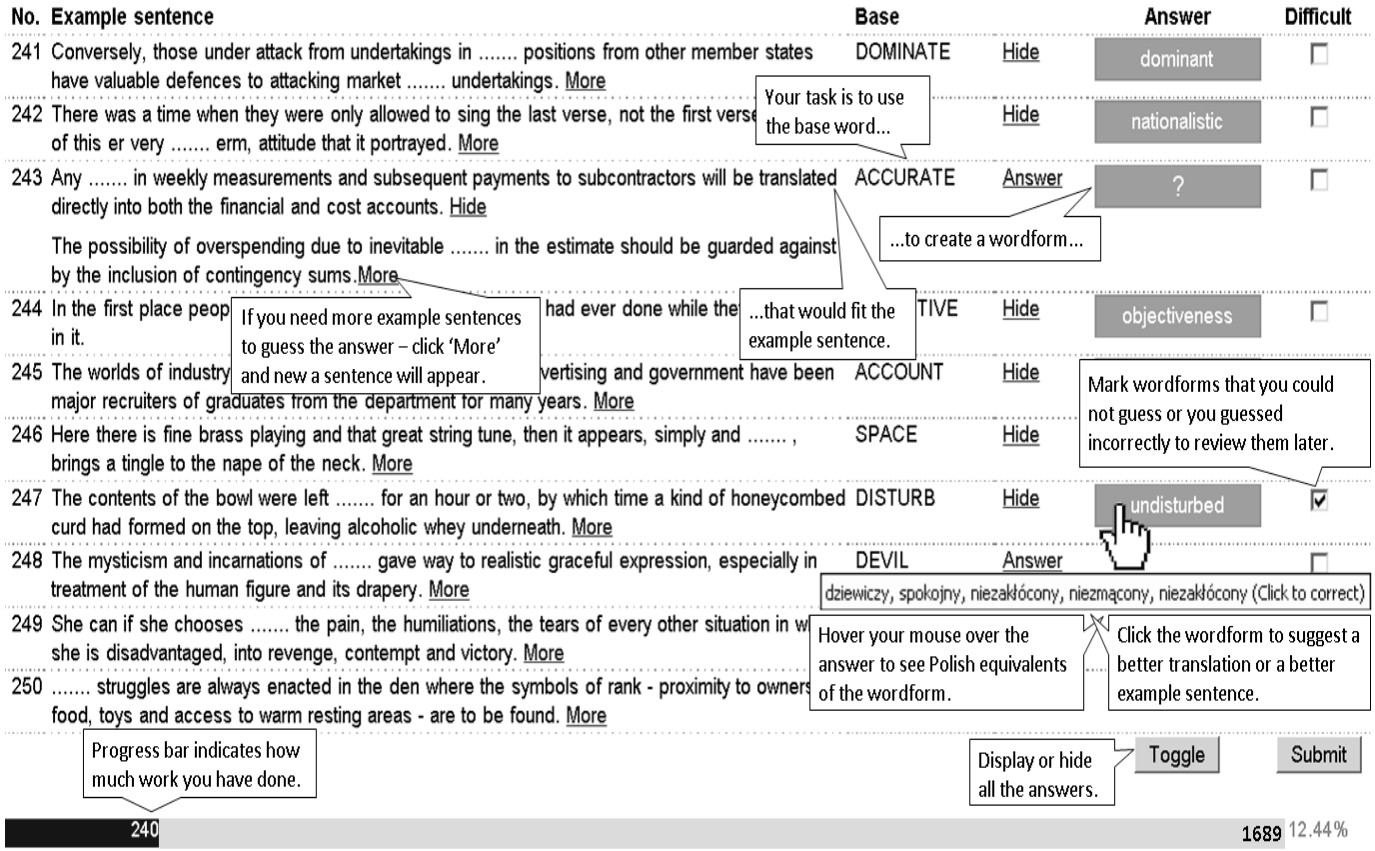


Fig. 10. A screenshot of the exercise panel with callouts explaining all functionalities.

Experimenting with Slovak Wikipedia as a Source for Language Technologies

Michal Laclavík[1], Štefan Dlugolinský[1], and Michal Blanárik[2]

[1] Institute of Informatics, Slovak Academy of Sciences, Bratislava, Slovakia
[2] Faculty of Informatics and Information Technologies, Slovak University of Technology, Bratislava, Slovakia

Abstract. In this paper we discuss how Slovak Wikipedia can be used for Natural Language Processing tasks on Slovak texts. We briefly discuss similar work done in past and also provide several experiments on Slovak Wikipedia related to Entity Search or Named Entity Recognition. In the paper we also try to motivate future research on Slovak Wikipedia, since valuable data can be gathered for building and improving Language Technologies.

1 Introduction

Wikipedia is a well known source of human knowledge created and maintained by crowd. It contains a variety of human maintained information on many topics of knowledge as well as facts, relations on entities such as people, organizations or locations. English Wikipedia was used in many ways [1] to create NLP tools including "wikifiers" such as Wikipedia Miner[1], Illinois Wikifier [2] or DBPedia Spotlight[2] [3]. Tools like these identify Wikipedia entities in a text. They are based on Wikipedia downloadable archives[3] or DBPedia[4], which contains structured information in a from of RDF graphs. Slovak Wikipedia was not explored much so far for NLP tasks, however some early work exists on this topic.

In this paper we discuss two experiments focused on Slovak Wikipedia parsing, entity search, as well as named entity recognition. Experiments are provided to show the potential of the Wikipedia as a text corpus with additional information rather then showing results of ready to use NLP tools.

Wikipedia is not composed only of articles but it also includes links representing relations among entities mentioned in the articles. Links contain anchor texts representing alternative names (inflected forms, abbreviations) or properties of addressed entities, which can be used in NLP tasks. In addition, DBPedia includes structured information now available in 111 languages including Slovak. Tools such as DBPedia Spotlight can be built [4] for multiple languages including Slovak with limited number of NLP tools.

2 Wikipedia Parsing, Indexing and Search

In this chapter, we describe our first experiment, which was aimed on extraction of Wikipedia links and their anchor texts. We have used a dump of the Slovak Wikipedia from the 30th April 2013. It contained 310,571 articles (including redirects to other articles).

[1] http://wikipedia-miner.cms.waikato.ac.nz
[2] https://github.com/dbpedia-spotlight/dbpedia-spotlight/wiki
[3] http://en.wikipedia.org/wiki/Wikipedia:Database_download
[4] http://dbpedia.org

The processing of the articles was performed by a MapReduce application on a Hadoop cluster in two stages: parsing and indexing. The size of Slovak Wikipedia is manageable on single machine, but we have used MapReduce approach since ready made tools for text processing are available. We have used and customized such tools for Wikipedia parsing and indexing.

The parsing stage has been performed within a map phase of the MapReduce application. Articles from the dump file have been processed one by one by multiple mappers. There have been links pointing to other articles (i.e. outlinks) extracted from each parsed article together with anchor texts and additional outlink data. Additional outlink data included:

- title of an article on which the outlink was pointing (extracted from URL)
- title of an article which contained the outlink
- title of the outlink
- path in article structure, where the out-link was found (built up of titles and subtitles of article sections).

If there was a redirect processed instead of a standard Wikipedia article, we have treated it as an outlink and its title became the outlink's anchor text. There have been totally 4,212,467 outlinks extracted in the parsing stage and emitted by mappers to one reducer in the second (indexing) stage.

Emitted outlinks have been combined by their URLs (URLs of Wikipedia articles) and converted to inlinks, so we got for each referenced Wikipedia article all its inlinks. We have further processed only those articles, which belong to the main Wikipedia namespace (all encyclopedia articles, lists, disambiguation pages, and encyclopedia redirects) and indexed them using Solr[5].

There have been 696,874 articles with 3,977,843 inlinks indexed. The number of indexed articles compared to the number of parsed articles was higher because there were many non-existing articles referenced in the Slovak Wikipedia, which is about 55%. The average number of inlinks per article was 5.71. If we consider only those articles, which were referenced more than one time, then there have been in average 13.60 inlinks per article and median was 3 inlinks per article.

We have created a searchable corpus of Wikipedia entities (articles) represented by their titles, links from the other pages and anchor texts. User can query the corpus using Solr web interface[6] or REST web services.

In Figure 1, we can see a screenshot of results for "Ľudovít Štúr" article. We can see anchor texts as well as other pages (inlinks) referencing this article. In addition, other metadata is present in the corpus, which can serve as a resource for creating training sets for NLP methods such as lemmatization, stemming or named entity recognition. Similar possibilities are discussed in chapter 3.

[5] http://lucene.apache.org/solr/

[6] http://147.213.75.180:8080/stevo/skwikislovco/browse

Fig. 1. Screenshot from Solr search interface

3 Named Entity Recognition

The second experiment was focused on named entity recognition. We have taken an XML dump of the Slovak Wikipedia from 21st February 2013 and focused on two types of named entities: person names and locations. We have exploited anchor texts of links, which referred to these kinds of entities and collected their inflected forms. More detailed information about links can be found in Table 1.

Links	9,935,074
Links with inflected forms	549,740

Table 1. Basic information about Slovak version of Wikipedia

3.1 Person and Location Extraction

We have used two extraction methods for person names in a basic form. The first one relied on a specific markup pattern typical only for mentioning person: "[[*person_name*]] (*[[*date_of_birth*]])", e.g. "[[Ľudovít Štúr]] (* [[1815]])". The second method was based on infobox information fields related to person only: date of birth and name. There have been 16,454 and 11,404 person names in a basic form extracted with the first method and the second method respectively. The total number of unique person names in a basic form was 22,511. The list of names was complemented by their inflected forms discovered in anchor texts of links. The final list of person names contained 42,500 names.

Extraction of location entities was similar to the second method used for person names. The only difference was that for locations, we have used information fields related to geographic coordinates instead of date of birth and name fields. There have been 37,121 location names in a basic form extracted and complemented by their inflected forms discovered in anchor texts of links. All together we have discovered 37,603 different location names.

3.2 Experiment and Evaluation

We have trained Named Entity Recognition using Apache OpenNLP toolkit. The model was trained on data obtained from Wikipedia. Person names and locations have been tagged for algorithm that performs training of the model.

Training file for recognition of person's names	
Number of sentences	184,602
Number of tagged names	91,915
Training file for recognition of location	
Number of sentences	40,579
Number of tagged locations	38,538

Table 2. Training set data

The models for recognition of person names and locations were trained with 500 iterations on training files made in advance. In Table 3 we summarize achieved results applied on training data. The models perform pretty well on these files, but also it has been shown on that parameter cutoff, which determines how many times a specific feature must occur to be added to the model, should be set on higher value for the person recognition than for location recognition.

	Cutoff	Precision	Recall	F-Measure
Evaluation of trained model for persons recognition	3	0.901	0.617	0.733
	5	0.896	0.839	0.867
Evaluation of trained model for locations recognition	3	0.998	0.995	0.997

Table 3. Testing NER on trained data

In order to evaluate trained models in real use, we had manually tagged person names in 10 sample articles and locations in other 3 sample articles, both from sme.sk web site. Person names recognition was evaluated on model trained with parameter cutoff set on 5, because this model performed better in previous test. In Table 4 we show the results that have been achieved by applying trained models on sample articles. The models recognized entities with high precision but with poor recall. This behavior is similar that can be achieved with gazetteer, because trained models mainly recognized entities, which were present in training data set, but recognition of new entities was quite poor.

	Precision	Recall	F-Measure
Evaluation of trained model for persons recognition	0.891	0.372	0.517
Evaluation of trained model for locations recognition	1.0	0.292	0.433

Table 4. NER performance

NER on Slovak text have several big challenges. The one of big challenges is having several inflected forms of same named entity. In order to identify entities correctly, we need to group identified inflected forms of named entities. To cope with this challenge we have analyzed suffixes of extracted named entities. This list is the outcome of analysis

of difference between lemma (base form) and inflected form of names previously obtained from Wikipedia. In this list is only that suffix which occurred more than 0.2 percent of all the suffixes were detected.

Suffix	Number of occurrences	Suffix	Number of occurrences	Suffix	Number of occurrences
a	5,543	**ov**	130	**ea**	55
om	4,406	**ová**	127	**s**	54
ovi	1,323	**ova**	103	**e**	53
m	779	**ovho**	88	**ových**	52
ho	589	**mu**	76	**ovom**	44
ou	541	**ove**	71	**ému**	41
ej	325	**ovou**	62	**o**	41
ovej	192	**vi**	59	**ovo**	41
ého	189	**eho**	56	**i**	40
us	143	**ovu**	56	Other:	1,345

Table 5. List of most occurred suffixes in people names and locations

Merging inflected forms of named entities was based on Levenshtein distance between two words with addition of obtained list of suffixes. This algorithm performed pretty well in case of merging word with same lemma but it failed in case of change in lemma of word during process of inflection.

4 Conclusions

In this article, we have discussed possibilities of using Slovak Wikipedia for Natural Language Processing. We have conducted several experiments analysing Wikipedia for the task of named entity recognition or statistics on word suffixes of selected named entity types. Experiments do not provide ready to use solutions for named entity recognition, lemmatization, stemming or other NLP tasks, but show the use pattern of growing Wikipedia resource. Our intent was to show that Slovak Wikipedia can serve as decent source for Language Technology evaluation and training supporting various NLP tasks.

Acknowledgements

This work is supported by projects: VEGA 2/0185/13 and CLAN APVV-0809-11. This publication is also the result of the project implementation ITMS: 26240220072 supported by Operational Programme Research & Development funded by the ERDF.

References

[1] Medelyan, O., Milne, D., Legg, C., and Witten, I. H. (2009). Mining meaning from Wikipedia. *International Journal of Human-Computer Studies*, 67(9):716-754, DOI=10.1016/j.ijhcs.2009.05.004.

[2] Ratinov, L., Roth, D., Downey, D., and Anderson, M. (2011). Local and global algorithms for disambiguation to wikipedia. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 1375–1384, Association for Computational Linguistics, Portland, Oregon, URL: http://dl.acm.org/citation.cfm?id=2002472.2002642.

[3] Mendes, P. N., Jakob, M., García-Silva, A., Bizer, Ch. (2011). DBpedia spotlight: shedding light on the web of documents. In Ghidini, Ch., Ngonga Ngomo, A.-C., Lindstaedt, S. and Pellegrini, T., editors, *Proceedings of the 7th International Conference on Semantic Systems (I-Semantics '11)*, pages 1–8, ACM, New York, Accessible at: DOI=10.1145/2063518.2063519, http://doi.acm.org/10.1145/2063518.2063519.

[4] DBPedia, DBPedia Spotlight Internationalization. (2013). URL: https://github.com/dbpedia-spotlight/dbpedia-spotlight/wiki/Internationalization.

Query Interface for Diverse Corpus Types

Tomáš Machálek and Michal Křen

Faculty of Arts, Charles University in Prague, Czech Republic

Abstract. The paper describes a version of NoSketch Engine that is being developed at the Czech National Corpus. The focus is on its technical description, as well as newly implemented functionality from user's point of view. The paper also explains our motivation for adaptations of the software and outlines the plans we have for the near future. Although it is being developed primarily to suit our needs, we believe it can be useful for other research groups as well.

1 Introduction

Within the framework of Large Research, Development and Innovation Infrastructures, the Czech National Corpus (CNC) strives mainly for continuous mapping of Czech by the means of building language corpora and providing access to them free of charge for experts as well as for general public. The data aim to be large, balanced and carefully processed to ensure high reliability of linguistic research based on them.

Needless to say, possibilities of how regular users can exploit the data are necessarily limited by the functionality implemented in the tools available. As most of the users primarily need a modern, fast and user-friendly corpus query engine, the paper describes the work that has been done in the CNC in this respect.

2 Overview of Recent Situation

For a very long time, the CNC corpora can be queried using Bonito. It is an open-source stand-alone client application that connects to Manatee [6] server via non-standard TCP port using a specific communication protocol. From users' point of view, it is an application that requires installation and the non-standard port may be blocked. Bonito certainly has its merits, quite rich functionality and it is thus still popular among the CNC users. However, it is also outdated and its development was stopped a long time ago.

In 2011, a modern web-based query interface called NoSketch Engine (NoSkE, [7]) was publicly released at `http://nlp.fi.muni.cz/trac/noske/`. Being based also on Manatee as its server part, it overcomes the technical limitations of Bonito mentioned above while retaining and enriching its functionality (with a few exceptions, e.g. limited possibility to mark and categorize concordance lines). Therefore, NoSkE was a natural choice as future substitute of Bonito in the CNC, and it was thus made available as a parallel option for querying corpora at `http://korpus.cz/corpora/`. NoSkE is an open-source version of commercial corpus management system Sketch Engine (SkE, [3]) that is developed, maintained and run by Lexical Computing Ltd at `http://www.sketchengine.co.uk`. While NoSkE is a regular, yet powerful corpus query interface, SkE features a number of additional components not present in NoSkE,

namely the word sketches and corpus-building functionality. However, a license needs to be paid in order to use SkE, which is not a realistic option for several thousand CNC users.

In addition to monolingual corpora, we also needed a query system able to handle – with adequate functionality – the architecture and size of InterCorp [10], [5], a large parallel corpus for 30+ languages. Until recently, it was not possible to find such a system, and also the support for parallel corpora in NoSkE was insufficient. In particular, 1:1 alignment was required and further work with language parallel with the primary one was substantially restricted (e.g. no statistics available). This is why the CNC started development of special parallel query interface based on Manatee and called Park [4]. It is in service since 2008 at `http://korpus.cz/Park/`. Its development was, regrettably, rather slow due to many reasons; the most serious obstacle was found to be the lack of proper support for parallel corpora in Manatee API. Therefore, some features had to be implemented in Park on the interface level, i.e. very inefficiently, and some others are still missing, namely frequency distribution and collocations.

Missing features in Park were also the main reason why it was decided to make InterCorp available also as a set of separate monolingual corpora via the standard monolingual query interface. This was just a temporary solution that enabled to use functions missing in Park when working with the individual languages, although the alignment could not be utilized this way.

3 Solution

As a consequence, there were 4 possibilities available to access the CNC in 2012. All of them used Manatee as a server, which means that the CQL query language was the same across all the interfaces. However, users had to make their choice before they could start querying the corpora which interface to use: retiring monolingual Bonito, web interface Park tailored to parallel corpus InterCorp (with limited functionality), modern NoSkE for full-fledged access to monolingual corpora (as a substitute of Bonito) and also to the individual language versions of InterCorp (as a substitute of Park). As this situation was obviously far from being satisfactory, our efforts aimed at unification and centralization of the 4 access points into a single one.

Another important issue we considered was customization, as our past experience shows that it is desirable to be able to add new functionality to the existing tools and, more generally, to adapt them according to our needs, as well as to the user feedback. In other words, the CNC as a research infrastructure should be independent of ready-made tools and their authors' priorities and/or opinions about the particular features that naturally differ. On the other hand, we certainly did not want to reinvent the wheel, so the obvious way to go was adaptation and further maintenance of a well-established open-source system.

As a solution, we decided to continue the use of Manatee-based user interfaces and to choose NoSkE. It is an open-source corpus query system licensed under GNU GPL 2 that includes also Manatee and finlib, fast indexing library. The whole system features adequate functionality, it supports large corpora and, last but not least, it is being continuously developed. However, it proved desirable to modify the interface (NoSkE) in many ways to make it more suitable to the individual needs of CNC. Since summer 2012, we have thus been working on a CNC version (fork) of NoSkE. The major drawback of NoSkE, i.e. the insufficient support for parallel corpora, has been solved after negotiations with

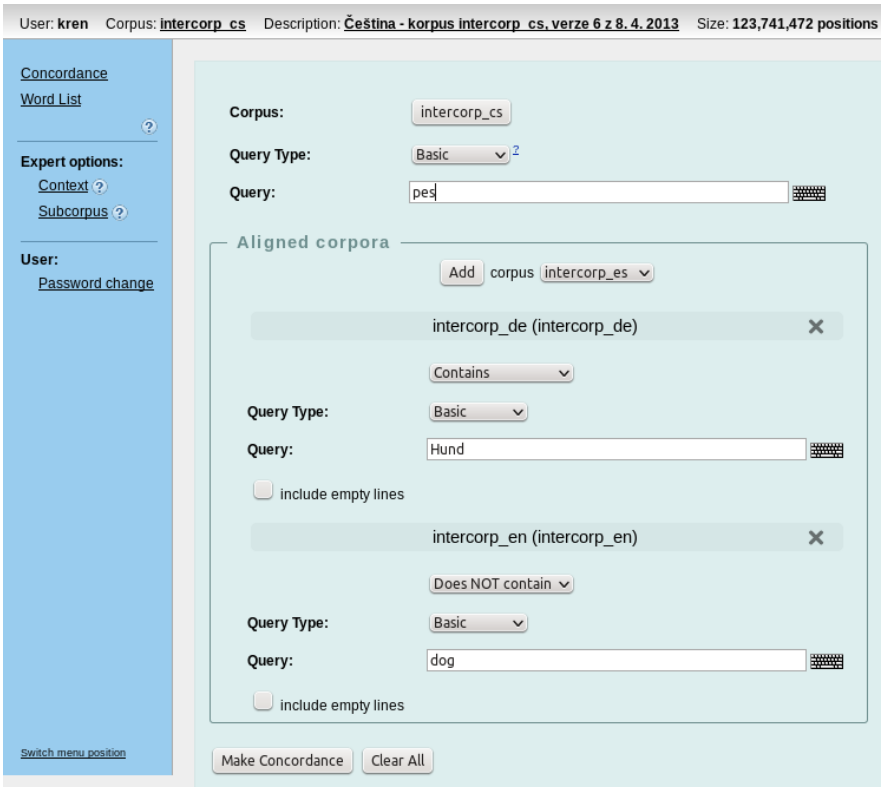


Fig. 1. Building a parallel query with the CNC version of NoSkE. Please note that the form differs from the official release in many ways.

Lexical Computing that implemented the features we requested into the core of the system by the end of 2012. The new functionality has been released also under GNU GPL 2. As a result, it has been made possible to search any number of parallel corpora at the same time (Fig. 1), display the translations next to each other and to use complete set of post-processing functions (filter, sort, statistics, etc.) individually on any of them. This was the most important milestone to reach our goal: to have at our disposal a fine-tuned, customizable interface for querying diverse corpus types, including parallel and spoken ones (cf. below).

4 Technical Description

Server side of the original NoSkE interface is written in Python as a common CGI web application running within Apache HTTP server. Client side is based on HTML templates accompanied by JavaScript (mostly based on jQuery library) to make the interface more user-friendly.

The CNC extensions and modifications of the application have been carried out while keeping in mind the following requirements:

- merging future official releases of NoSkE with the CNC version should be as smooth as possible;
- new functionality should be preferably implemented as a reusable component;
- legacy or unnecessary code should be changed/removed.

Obviously, these criteria also restricted our possibilities in terms of modifications of the original code. As the most important consequence, we had to avoid fundamental changes in application's design which made some more complex enhancements harder to implement because the original application was not designed with regard to them (e.g. audio support for spoken corpora).

These restrictions affected especially the server side of the application. For this reason, current changes and enhancements of server-side code are rather scarce and limited to:

- improved logging: all warning or error-like events are processed, no unnecessary technical information is provided to a user while internally as much as possible useful information is written to the log;
- installation process: while the original application uses `make`, our version can be installed simply by editing the; `config.xml` configuration file and putting application's files to a proper destination directory (where Apache web server expects them to be located);
- unnecessary (e.g. SkE-related) code and deprecated Python language constructs (e.g. so called "old style" objects) were removed.

As for the client-side code, we decided to make more fundamental architectural changes, especially towards better modularization. For this purpose we used JavaScript library RequireJS which allows keeping track of module dependencies and provides a simple way how to keep the scope of imported functionality as local as possible (i.e. each module should be code-visible only where it is actually needed). Also the organization and interdependencies of client-side code in general has been simplified (e.g. JavaScript code scattered across HTML files).

New functions were generally easier to implement than the changes, especially if they relied on existing server-side services. One exception includes audio support for spoken corpora which required rewriting of concordance fetching and displaying routines to be able to work with both user selected and internally required structures and positional attributes. Client-side audio player has been implemented using SoundManager 2 library which allows building custom user interfaces, as we required a minimalist solution with essential controls only.

The development process itself relies on use of a distributed versioning system (Mercurial in our case) which allows not just to keep track of realized changes, but it also provides a simple way how to simultaneously develop multiple versions, independently test new features, merge them to other versions etc. Such a system also makes potential contributions from other developers much easier.

The new code is open-source, it is available under GNU GPL 2 (i.e. the same license as the original one) and anyone is welcome to make a copy of our official repository `https://bitbucket.org/ucnk/bonito-2/` with its full history, make changes and ask us to accept them back (so called "pull request"). The version labeling is based on the original one, i.e. the standard Manatee and NoSkE version numbers are supplemented by our independent version numbering prefixed with the `ucnk` string.

Despite all the changes and fixes there is still a work to be done within the application internals. Unit testing provides a way how to programmatically test individual components which would make future changes in the application less error-prone. The deployment process (i.e. process of installation or updating of the application) should be improved and more automated. Last but not least, a better code development documentation needs to be written, as the original code contains almost no comments.

5 New Functionality

The new features that have already been implemented in the CNC version of NoSkE can be divided into three groups:

1. Adaptation for deployment in the CNC
   - CNC-specific localization (mostly adjustment of Czech translations);
   - password change option;
   - advanced user query logging;
   - extended configuration options for system administrators centralized in file `config.xml`.
2. Minor general improvements
   - showing ARF [8] and relative frequency (i.p.m.) for every query result (the latter is also available in the frequency distribution);
   - possibility to address positions inside KWIC in frequency distribution and sort;
   - line numbering in frequency distribution and collocation candidate list;
   - lists of pre-set attribute values displayed when creating a subcorpus are supplemented with scrollbars (useful for long lists);
   - possibility to create subcorpora also using a user-defined `within` condition;
   - context size in positions;
   - possibility to set shuffled display of concordances as a default option.
3. Major general enhancements
   - hierarchical arrangement of available corpora together with displaying detailed information about them (Fig. 2);
   - tag builder: interactive selection of morphological categories as an option for CQL query type (Fig. 3);
   - support for spoken copora: possibility to play audio segments (Fig. 4).

We shall perhaps comment on the individual items of the last group and explain their importance. With the growing number of publicly available corpora, we found it necessary to offer their clear arrangement so that related corpora are displayed next to each other. In our current implementation, the global hierarchy of all available corpora is defined in `config.xml`. Right after individual users log in, the actual set of corpora accessible to them is retrieved from the database and arranged in the pre-defined order. As users are typically not granted access to all available corpora, the resulting hierarchy is selected as a subset of the global one. Apart from the hierarchy, a number of additional information about the individual corpus is displayed: brief description, web link (both can be defined in `config.xml`, too) and totals for individual positional attributes and structures. We also plan to add a handy citation information that would facilitate proper citing the corpus data.

The interactive and user-friendly selection of morphological categories is a feature that has been requested many times by the CNC users. The tag builder has been implemented

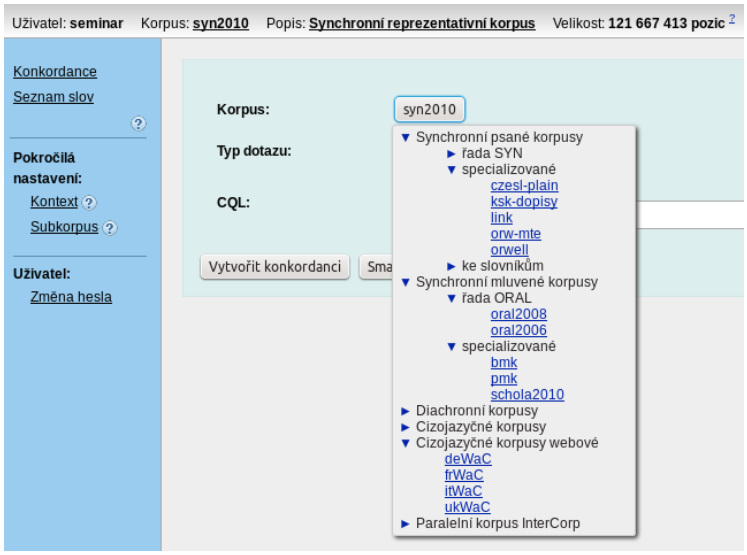


Fig. 2. Hierarchy of available corpora

as a JavaScript component for the Prague positional tagset [1], but it can be applicable to some other tagsets as well. It is data-driven, i.e. only combinations of categories that actually exist in the corpus data are displayed and can be selected. The tag builder can be configured also in `config.xml`, including the descriptions of the individual categories and their translations into other languages (currently Czech and English).

The support for spoken corpora enables to use the interface also for an additional modality – the sound. In particular, corpus ORAL2013 [9] is planned to be released by the CNC by the end of 2013. ORAL2013 is a new, sociolinguistically balanced corpus of informal spoken Czech that features manual alignment with the sound on the level of segments (sequences of typically 5–10 words that constitute a natural unit). For ORAL2013, it will thus be possible to hear actual realization of every expression by means of a single click in the interface.

6 Conclusion and Future Plans

Although the development of the CNC version of NoSkE was motivated primarily to suit our needs, we believe it can be useful for other institutions as well. We are ready not only to share the code and provide help with deployment of the software, but also to cooperate with other research groups on its further development. Currently, we plan to enhance its functionality in the following main directions:

- implementation of word collocation profiles [2];
- user-friendly graphical module for displaying the contents of selected (sub)corpus;
- facility that would enable selection and/or categorization of the concordance lines.

The latter two enhancements are needed to surpass the functionality of Park and Bonito, respectively. After making the enhanced NoSkE fully operational, it will be pos-

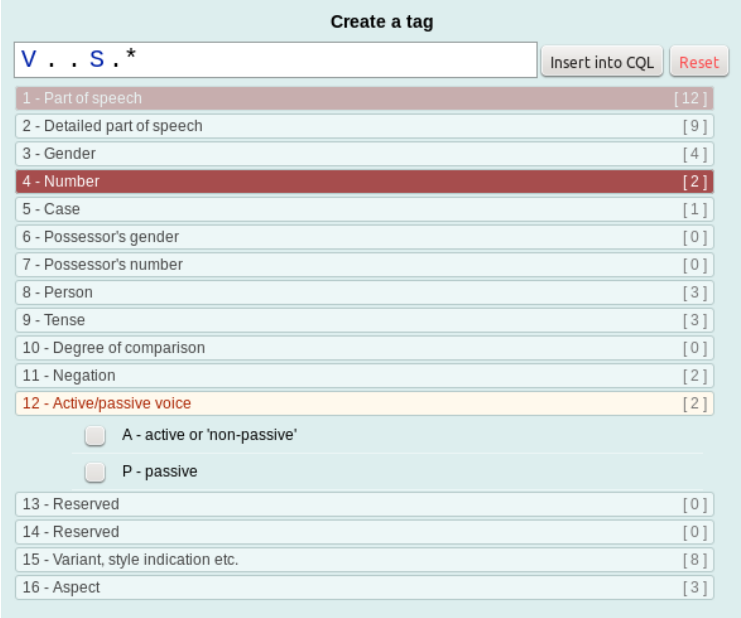


Fig. 3. Tag builder

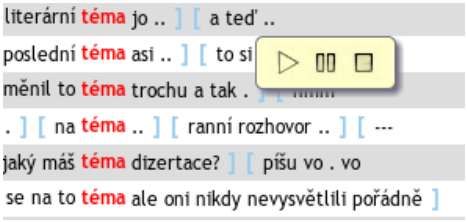


Fig. 4. Playing one of the audio segments (enclosed in brackets)

sible to discontinue the service of both retiring interfaces, and thus to eventually establish a single access point served by the CNC version of NoSkE. We also plan its integration into a new CNC web portal that is currently being prepared, which will cause graphical redesign of the software.

Acknowledgement

This paper resulted from the implementation of the project Czech National Corpus (LM2011023) funded by the Ministry of Education, Youth and Sports within the framework of Large Research, Development and Innovation Infrastructures.

References

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

[2] Keibel, H. and Belica, C. (2007). CCDB: A corpus-linguistic research & development workbench. In *Proceedings of the Corpus Linguistics Conference*, Birmingham.

[3] Kilgarriff, A., Rychlý, P., Smrž, P., and Tugwell, D. (2004). The Sketch Engine. In Williams, G. and Vessier, S., editors, *Proceedings of the 11th EURALEX International Congress*, pages 105–116, Lorient.

[4] Křen, M., Rosen, A., Štourač, M., Vavřín, M., and Vondřička, P. (2011). Paralelní korpus InterCorp po sedmi letech. In Čermák, F., editor, *Korpusová lingvistika Praha 2011. 2 Výzkum a výstavba korpusů*, pages 105–115. NLN, Praha.

[5] Rosen, A. and Vavřín, M. (2012). Building a multilingual parallel corpus for human users. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 2447–2452, ELRA, Istanbul, Turkey.

[6] Rychlý, P. (2000). *Korpusové manažery a jejich efektivní implementace*. FI MU, Brno.

[7] Rychlý, P. (2007). Manatee/Bonito – a modular corpus manager. In *1st Workshop on Recent Advances in Slavonic Natural Language Processing*, pages 65–70, Brno.

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

[9] Válková, L., Waclawičová, M., and Křen, M. (2012). Balanced data repository of spontaneous spoken Czech. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 3345–3349, ELRA, Istanbul, Turkey.

[10] Čermák, F. and Rosen, A. (2012). The case of InterCorp, a multilingual parallel corpus. *International Journal of Corpus Linguistics*, 13(3):411–427.

The Effect of Stop Words Elimination on Sequence Patterns Extraction in Comparable Corpora

Daša Munková[1], Michal Munk[2], and Michal Vozár[2]

[1] Faculty of Education, Constantine the Philosopher University, Nitra, Slovakia
[2] Faculty of Natural Sciences, Constantine the Philosopher University, Nitra, Slovakia

Abstract. Currently, the advertisements play an important role in our society. To improve the quality of textual data (advertising content) it is necessary to filter out noise textual data from important data. The aim of this work is to determine to what extent it is necessary to carry out the time consuming data preparation in the process of discovering sequential patterns in sets of English and Slovak advertisements. For this purpose, an experiment was conducted focusing on data preparation in these two comparable sets. We try to find out whether the use of a pre-existing stop words list or standardized stop words list e.g. Snowball Stop is suitable (effective) in context of textual sources such as advertisements. Furthermore, we try to find out to what extent the stop words elimination efects the quantity and quality of extracted rules. The results of analysis showed that only language has a significant impact on the quantity and quality of extracted rules in comparable English and Slovak advertisement sets, and that the Snowball stop words list is ineffective in analysis of short texts such as advertisements.

1 Introduction

The present era is characterised by enormous amount of available advertising content on one hand, but often a lack of knowledge on the other hand. A huge amount of advertasing data has a weak predictive value.

Data preparation is the most time consuming phase in the whole process of knowledge discovery. It converts a document transformation from an original textual data source into a form which is suitable for applying various methods of extraction, in order to transform unstructured form into structured representation, i.e. to create a new collection of texts fully represented by concepts [8]. According to Feldman and Sanger two steps of textual data preparation are inevitable. First, it is an identification of significant features in a way that is most computationally efficient and practical for pattern discovery. Second, it is an accurate capture of meaning of the content (on the semantic level) [8]. More detail about textual representation and data preparation for the purpose of data analysis is dealt with in books on the subject of text mining, content analysis etc.

Based on the huge amount of short texts collected and the nature and assumptions of the techniques, textual data has to be of a very good quality in order to be effective [6], [20], [27], [12]. To improve the quality of textual data, many authors have proposed different techniques to extract an effective stop word list for a particular corpus [6], [23], [22]. Stop words lists have not been examined in great detail, which has resulted in the use of pre-existing stop word lists. These might not be suitable in each context of the textual sources as evidenced in our experiment. Research in the area has identified the weaknesses of standardized stop words list [6], [4], [5], [24].

Short texts like advertisements are characterised by considerable expressive variability. The expressive variability consists of morpho-syntactic means of expressions, lexical means of expressions and stylistic means of expressions [10].

The aim of this paper is to determine to what extent it is necessary to carry out the time consuming data preparation in the process of discovering sequential patterns in English and Slovak advertisement sets. For this purpose, an experiment was conducted focusing on data preparation in these two comparable sets (corpora). Due to influence of works [3], [15] during the realisation of an experimental plan we used our own model for text representation which is similar to bag-of-words model [16], [17].

This paper is further divided into several chapters which are as follows: in chapter 2 we characterize content/function words and also stop words. We summarize related works dealing with stop words issues. We summarize the transaction/sequence model in chapter 3. Subsequently, we particularize research methodology in chapter 4. This chapter describes how we prepared texts in different levels of data preparation. Chapter 5 provides a summary of the experiment results in detail. Finally, the discussion of the results and a conclusion follows in the last chapter.

2 Words and Stop Words

Words differ in the role they perform. We can divide words into two groups. One group referring to objects, actions and properties is content words. The second group telling us how the words from the first group are mutually related is function words [14]. Linguists define two categories of words: open-class words and closed-class words. Open-class words represent content words and closed-class words represent function words.

In terms of parts of speech content words include nouns (objects), verbs (actions) and adjectives and adverbs (properties that quantify nouns and verbs). On the other hand, function words consist of determiners, pronouns, prepositions, conjunctions, numbers etc. [14].

Text is made up of a sequence of words, which are separated by a tokenization process [14]. Some frequent words make up most of the text. Stop words carry less important meaning than other words occurring in the text. Stop words are functional, general, and common words of the language that usually do not contribute to the semantics of the documents and have no read added value [3]. Myerson [18] stated two conditions for a stop word. It should have a high document frequency (DF) and the statistical correlations with all classification categories should be small. Zou et al. [7, 30] define a stop word as a word with stable and high frequency in documents. According to M. Khosrow [13] stop words are words having no significant semantic relation to the context in which they exist.

For example, in English language, articles ”the, a, an”, prepositions ”on, up etc.” conjunctions ”and, or etc.”, pronouns ”it, us etc.” are usually defined as stop words. Stop words may also be document-collection specific words [13], [3], e.g. the word ”to help” would probably be a stop word in a collection of advertisements but certainly not in a collection of News articles. Several authors [25], [7], [29] have argued for the elimination of stop words which make the selection of the useful words more efficient and reduce the complexity of the structure of the document.

The most common approach to creating a stop words list is to manually assemble it from a list of words or terms having no natural useful information [6]. This approach is used by several authors [9], [29] and others.

3 Transaction/sequence Model

We used a transaction/sequence model for text representation, similar to a bag-of-words model, which allows us to investigate the relationships among examined features and to search for associations among identified words in our set (corpus). The structure and data character predetermine the use of specific methods for analysis - data modelling. In the case of the use of a transaction/sequence model for text representation, it is mainly association rule analysis and sequence rule analysis.

Examined variables: Language, Text ID, Sentence ID, Transaction/Sequence ID - it consists of the previous two/three variables, Sequence - an order of words in a text/sentence, Word-individual words, Part of speech – morphological words classification (nouns, verbs, adjectives, adverbs, articles, pronouns, prepositions, conjunctions and others), and Stop words – Snowball list.

4 Research Methodology

Short texts like advertisements are characterised by a number of slogans, phrases, words, symbols etc. To improve the quality of textual data it is necessary to filter out noise textual data from important data. Noise textual data are data (text) not relevant to the task at hand [3], [21], [19]. Stop words are good examples of noise data. In our experiment we used a pre-existing list of stop words for English (Snowball stop words list for English) and similar for Slovak [26]. We tried to find out whether using a pre-existing stop word list or a standardized stop words list like Snowball Stop words list is suitable in context of textual sources like advertisements.

We aimed at specifying the inevitable steps to improve the quality of textual data (advertising content) represented by transaction/sequence model. We focused on sequence identification and stop words elimination. We tried to find out to what extent the elimination of stop words has an effect on the quantity and quality of extracted rules.

In particular we assessed the impact of these techniques on the quantity and quality of the extracted rules representing sequential patterns in comparable advertisement sets (EN, SK). Experiment was conducted using the following steps:

1. Text collection (Data collection-comparable advertisement sets).
2. Format removal.
3. Data preparation on different levels:
   (a) a sentence sequence identification without stop words removal for English corpus (File EN1),
   (b) a sentence sequence identification with stop words removal for English corpus (File EN2),
   (c) a sentence sequence identification without stop words removal for Slovak corpus (File SK1),
   (d) a sentence sequence identification with stop words removal for Slovak corpus (File SK2).
4. Data analysis - searching for sequential patterns in individual files. We used STATISTICA Sequence, Association and Link Analysis for sequence rules extraction. It is an implementation of algorithm using the powerful a-priori algorithm [2], [1], [11], [28] together with a tree structured procedure that only requires one pass through data.

5. Understanding of the output data – a production of data matrices from the analysis outcomes, defining assumptions.
6. Comparison of results of data analysis elaborated on various levels of data preparation from the point of view of quantity and quality of the found rules – sequential patterns.

We articulated the following two assumptions:

1. we expect that stop words elimination will have a significant impact on the quantity of extracted rules, and
2. we expect that stop words elimination will have a significant impact on the quality of extracted rules, in terms of their basic measures of quality in examined comparable advertisement corpora.

5 Results

5.1 Data Understanding

Text is rarely translated sentence by sentence or word by word. Long sentences may be split into short sentences or vice versa. Therefore our analysed texts represent collections of short texts - advertisements from the comparable sets (Slovak advertisement corpus and English advertisement corpus) i.e. we created two different sets according to language (Slovak and English) with the same subject matter. They write about the same topics (products), but they are not translations of each other (no direct translations).

The experiment used two different corpora. A corpus of English written advertisements contains over 31,390 words. The second, Slovak corpus of written advertisements consists of 28,070 words. We used our own analyser for determining the parts of speech. Among the most frequent parts of speech in English advertisement corpus are nouns with portion higher than 26%, verbs and adjectives with portion higher than 14%, then others and pronouns, each with approximately 10% of the total number of words. For Slovak advertisement corpus, there is a difference: nouns with portion higher than 36%, adjectives with portion higher than 18%, verbs with portion higher than 14% and then conjunctions with approximately 10% of the total number of words.

Body	$\Longrightarrow$	Head	SK1	SK2	EN1	EN2
(verb)	$\Longrightarrow$	(preposition), (noun)	1	0	0	0
...		...				
(adjective)	$\Longrightarrow$	(verb)	1	1	1	1
Count of derived sequence rules			65	73	45	50
Percent 1's			57.02	64.04	39.47	56.14
Percent 0's			42.98	35.96	60.53	56.14
Cochran Q Test	Q=20.20266;df=3;p<.000154					

Table 1. Incidence of discovered sequence rules in particular files

Based on Snowball list of stop words, 42.59% of stop words were determined in English advertisement corpus. Pronouns and prepositions are the parts of speech most frequently used as stop words, with portion higher than 21%, followed by others and verbs with portion higher than 15% of stop words. A similar stop words list was used for the Slovak advertisement corpus where 26.75% of stop words were identified. From the point of view of parts of speech, pronouns and prepositions are the parts of speech most frequently used as stop words, with portion higher than 33%, then verbs and pronouns with higher than 11% of the total number of words used. In English advertisement corpus nouns, verbs, adjectives, pronouns and others (articles, interjection and symbol) belong to the most frequently occurring parts of speech. On the contrary, in Slovak advertisement corpus nouns, adjectives, adverbs and conjunctions belong to the most frequently used. The differences are mainly in the verb incidence, pronoun, conjunction and others. Based on the cross-tabulation analysis there is a low dependency between the incidence of parts of speech and language in case of Slovak vs. English advertisement corpus, the contingency coefficient (V=0.27) is statistically significant (Chi-square=416.7343; df=8; p=0.0000), i.e. the incidence (use) of parts of speech depends only on the language of corpus (SK or EN).

Furthermore, we investigated whether there is also a difference in the incidence of parts of speech in stop words in Slovak and English advertisement corpus.

The results of cross-tabulation analysis showed that there is a medium dependency between the incidence of parts of speech in stop words and language in case of Slovak vs. English advertisement corpus, the contingency coefficient (V=0.37) is statistically significant (Chi-square = 280.1117; df = 8; p = 0.0000), i.e. the incidence of parts of speech in stop words depends on the language (Slovak or English).

5.2 Comparison of the Quantity of Extracted Rules in Examined Files

The analysis (Table 1) resulted in sequence rules, which we obtained from frequented sequences fulfilling their minimum support (in our case min s = 0.1). Frequented sequences were obtained from identified sequences based on the length of sentence.

Most rules were extracted from files with sentence sequence identification without stop words in Slovak corpus; exactly 73 were extracted from the file (File SK2), which represents over 64% of the total number of found rules. Based on the results of Q test (Table 1), the zero hypothesis, which reasons that the incidence of rules does not depend on individual levels of text preparation or language is rejected at the 1% significance level.

Statistically significant differences were proven only in language in terms of the average incidence of found rules. That means that only language has an important impact on the quantity of extracted rules. Naturally, the Slovak language belongs to a inflected (morphologically richer) language family than the English language, so the morphological differences between them are axiomatic.

On the contrary, elimination of the stop words has no significant impact on the quantity of extracted rules in particular languages.

Results of cross-tabulation analysis showed, that the differences among files with a without stop words consist of 23 rules extracted from the File SK1 which represents 20% (Table 1) and also 31 extracted rules from the File SK2 representing 27%. After stop words elimination, a number of rules has increased by 7% in case of Slovak corpus. The similar results were obtained for English corpus (Table 2). A number of extracted rules has

SK1/SK2	0	1	∑
0	18	31	49
	15.79%	27.19%	42.98%
1	23	42	65
	20.18%	36.84%	57.02%
2	41	73	114
	35.96%	64.04%	100.00%
Mc Nemar (B/C)			
Chi^2=0.9; df=1; p=0.3408			

EN1/EN2	0	1	∑
0	38	31	69
	33.33%	27.19%	60.53%
1	26	19	45
	22.81%	16.67%	39.47%
2	64	50	114
	56.14%	43.86%	100.00%
Mc Nemar (B/C)			
Chi^2=0.3;df=1;p=0.0000			

SK1/EN1	0	1	∑
0	34	15	49
	29.82%	13.16%	42.98%
1	35	30	65
	30.70%	26.32%	57.02%
2	69	45	114
	60.53%	39.47%	100.00%
Mc Nemar (B/C)			
Chi^2=7.2; df=1; p=0.0072			

SK2/EN2	0	1	∑
0	38	3	41
	33.33%	2.63%	35.96%
1	26	47	73
	22.81%	41.23%	64.04%
2	64	50	114
	56.14%	43.86%	100.00%
Mc Nemar (B/C)			
Chi^2=16.7; df=1; p=0.0000			

Table 2. Cross-tabulations - Incidence of rules: (a) SK1 x SK2, (b) EN1 x EN2, (c) SK1 x EN1, (d) SK2 x EN2

increased by 4% after stop words removing. Elimination of stop words has influence on an increase of a number of extracted rules from Slovak as well as from English advertisement corpus but this increase is not statistically significant.

By contrast, when comparing the results of comparable Slovak and English sets, where these sets were pre-processed at the same level, these differences are statistically significant. In case of files with stop words (Table 2), a number of extracted rules has increased by almost 18% and in case of files without stop words (Table 2) it has increased by more than 20% in favour of Slovak corpus.

5.3 Comparison of the Quality of Extracted Rules in Examined Files

Quality of sequence rules is assessed by means of two indicators [2]: support and confidence. Results of the sequence rule analysis showed differences not only in the quantity of the found rules, but also in the quality. Kendall's coefficient of concordance represents the

degree of concordance in the support of the found rules among examined files. The value of coefficient (Table 3a) is 0.21, while 1 means a perfect concordance and 0 represents discordancy.

Support	Mean	1	2
EN1	30.8393		****
EN2	34.2194	****	****
SK1	35.9736	****	
SK2	36.8069	****	
Kendall Coeff. of Concordance 0.2100			

Confidence	Mean	1	2
File EN1	47.2998	****	
File EN2	50.8743	****	****
File SK1	52.3577	****	****
File SK2	52.9699		****
Kendall Coeff. of Concordance 0.2500			

Table 3. Homogeneous groups for (a) support of derived rules; (b) confidence of derived rules

From the multiple comparison (Tukey HSD test) two homogenous groups (Table 3a), one consisting of files File EN2, File SK1 and File SK2; and other consisting of files File EN1 and File EN2 were identified in terms of the average support of found rules. Statistically significant differences on the level of significance 0.05 in the average support of found rules were only proved among File EN1 and files File SK1, File SK2, i.e. again only between languages.

There were demonstrated differences in the quality in terms of confidence characteristics values of the discovered rules among individual files. The coefficient of concordance values (Table 3b) is 0.25, while 1 means a perfect concordance and 0 represents discordancy. From the multiple comparison (Tukey HSD test) two homogenous groups (Table 3b), the first consisting of files File EN1, File EN2 and File SK1, the second consisting of files File EN2, File SK1 and File SK2 were identified in terms of the average confidence of found rules. Statistically significant difference on the level of significance 0.05 in the average confidence of found rules was proved between File EN1 and File SK2.

Results (Table 3a, Table 3b) show that the largest degree of concordance in the support and confidence is among the rules found in the files without stop words removal and files with stop words removal. On the contrary, discordance is between languages. Again it was proven that stop words elimination has no impact on the quality of extracted rules and only language has a significant impact on the quality of extracted rules.

6 Discussion and Conclusion

Data preparation represents the most time consuming phase and an important task in the whole process of knowledge discovery. Based on the nature of the processed textual data many authors have proposed different techniques to improve the quality of textual data. This has resulted in an extraction of an effective stop words list and in the formulation of general or standard stop words lists such as Snowball stop word list (it consists of 72 words for English and of 85 words for Slovak, due to morphological richness of Slovak language).

Two comparable advertisement corpora (English and Slovak), each of 600 advertisements were used in this experiment. Both corpora consist of different sizes (advertisement consists of 7 words up to 32 words) and categories of advertisement texts (e.g. products, services etc.). Stop words are the most frequently occurring words. They are function words and carry less important meaning than other words occurring in the advertisement. The elimination of such words should lead to higher efficiency (it reduces the corpus size approximately by 20–30%).

The first assumption, removing stop words has no significant impact on the quantity of extracted rules in both comparable corpora (SK, EN), was not proved. Only language has a statistically significant impact on the quantity of extracted rules. Elimination of stop words has influence on an increase in the number of extracted rules from Slovak as well as from English advertisement corpus. However this increase is not statistically significant.

The second assumption was also not proved. Stop words elimination has no significant impact on the quality of extracted rules in both examined corpora (SK, EN). It was again showed that only language has a significant impact on the quality of extracted rules.

On the basis of our experiments, stop words elimination, defined in Snowball stop words list, has no significant impact on the quantity and quality of extracted rules. Only language has a significant impact on the quality and quantity of extracted rules. Results showed that the quality and quality of extracted rules depend on language of corpus (whether it is English or Slovak advertisement corpus) and not on Snowball stop words list elimination from the comparable corpora.

It is important what kind of stop words list is used (standard or generated). In this study a Snowball list of English stop words (and its equivalent for Slovak) was proved more likely ineffective for advertisement corpus. The question remains whether the removing of stop words really impacts the quantity and quality of extracted rules. Therefore, in further research we will attempt to propose effective stop words list for advertisement corpora (EN, SK) and focus on identifying the impact of proposed list of stop words in knowledge extraction.

References

[1] Agrawal, R. and Srikant, R. (1994). Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases*.

[2] Agrawal, R., Imielinski, T., Swami, A. N. (1993). Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, pages 207–216.

[3] Alajmi, A., Saad, E. M., and Darwish, R. R. (2012). Toward an arabic stop-words list generation. *International Journal of Computer Applications*, 46(8):8–13.

[4] Chakrabarti, S., Dom, B., Agrawal, R., and Raghavan, P. (1997). Using taxonomy, discriminants, and signatures for navigating in text databases. In *Proceedings of the 23rd International Conference on Very Large Databases*, pages 446–455.

[5] Chakrabarti, S., Dom, B., Agrawal, R., and Raghavan, P. (1998). Scalabe feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies. *The VLDB Journal*, 7:163–178.

[6] Choy, M. (2012). Effective listings of function stop words for twitter. *Knowledge Information System*, 3(6):8–11.

[7] El-Khair, I. A. (2006). Effect of stop words elimination for arabic information retrieval: A comparative study. *International journal of Computing and Information Sciences*, 4(3):119–133.

[8] Feldman, R. and Sanger, J. (2007). The text mining handbook. *Cambridge University Press*.
[9] Fox, C. (1992). Lexical analysis and stoplists. *Information Retrieval – Data Structures and Algorithms*, pages 102–130.
[10] Gromová, E. (2003). *Teória a didaktika prekladu*. Univerzita Konštantína Filozofa Nitra, 190 p.
[11] Han, J., Lakshmanan, L. V. S., Pei, J. (2001). Scalable frequent-pattern mining methods: an overview. *Tutorial notes of the seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.
[12] Jung, W. (2004). An investigation of the impact of data quality on decision performance. In *Proceedings of the 2004 International Symposium on Information and Communication Technology (ISICT '04)*, pages 166–171.
[13] Khosrow, M. (2009). *Encyclopedia of Information Science and Technology*. Information Science, Second Edition.
[14] Koehn, P. (2010). Statistical machine translation. *Cambridge University Press*.
[15] Munk, M., Kapusta, J., and Švec, P. (2010). Data pre-processing evaluation for web log mining: Re-construction of activities of a web visitor. *International Conference on Computational Science, ICCS 2010, Procedia Computer Science*, 1(1):2273–2280.
[16] Munková, D. and et al. (2012). Analysis of social and expressive factors of requests by methods of text mining. In *Pacific Asia Conference on Language, Information and Computation, PACLIC 26*, pages 515–524.
[17] Munková, D., Munk, M., and Vozár, M. (2013). Data pre-processing evaluation for text mining: Transaction/sequence model. *International Conference on Computational Science, ICCS 2013 Procedia Computer Science*.
[18] Myerson, R. B. (1996). Fundamentals of social choice theory. *Discussion Paper*, (1162).
[19] Nisbet, R., Elder, J., and Miner, G. (2009). *Handbook of statistical analysis and data mining applications*. Academic Press, Elsevier.
[20] R. Cooley, B. M. and Srivastava, J. (1999). Data preparation for mining world wide web browsing patterns. *International Jurnal of Advanced Computer Science and Application*, pages 1–27.
[21] R. Nisbet, J. E. and Miner, G. (2009). Handbook of statistical analysis and data mining applications. *Academic Press, Elsevier*.
[22] Rose, S., Engel, D., Cramer, N., and Cowley, W. (2010). Automatic keyword extraction from individual documents. In *Text Mining: Applications and Theory, John Wiley and Sons, Ltd.*
[23] Salton, G. (1971). *The SMART Retrieval System-Experiments in Automatic Document Processing*. Upper Saddle River, NJ. USA.
[24] Silva, C. and Ribeiro, B. (2003). The importance of stop word removal on recall values in text categorization. In *Proceedings of the International Joint Conference on Neural Networks*, volume 3, pages 1661–1666.
[25] Sinka, M. P. and Come, D. W. (2003). Evolving better stoplists for document clustering andweb intelligence. In *Proceedings of the 3rd Hybrid Intelligent Systems Conference*, Australia. IOS Press.
[26] Snowball (2013). http://snowball.tartarus.org/algorithms/english/stop.txt.
[27] Tayi, G. K. and Ballou, D. P. (1998). Examining data quality. *Communications of the ACM*, 41(2):54–57.
[28] Witten, I. H. and Frank, E. (2000). Data mining: Practical machine learning tools and techniques. Morgan Kaufmann, New York.
[29] Yao, Z. and Ze-wen, C. (2011). Research on the construction and filter method of stopword list in text preprocessing. In *Proceeding of the Intelligent Computation Technology and Automation (ICICTA)*, pages 217–221. IEEE, Los Alamitos.
[30] Zou F., Wang F. L., Deng X., Han S., and Wang L. S. (2006). Automatic construction of chinese stopword list. In *Proceedings of the 5th WSEAS International Conference on Applied Computer Science*, pages 1010–1015, Hangzhou, China.

Valency of Selected Primary Adjectives in the SYN2010 Corpus

Kateřina Najbrtová

Faculty of Arts, Masaryk University, Brno, Czech Republic

Abstract. The study focuses on the valency of primary adjectives on the basis of the linguistic material provided by the SYN2010 corpus. The adjective valency constitutes a rather neglected part in the extensive literature dealing with the question of valency, only a few authors address it [11], [12]; adjective valency appears in the form of fragments in some articles [9], [10] etc.; a monograph treatment is provided by [6]. Using contemporary linguistic material, this study attempts to clarify if the valency described in Prouzová's works, and it does so by focusing on the comparison of the conclusions reached by her and Kopřivová; a short valency lexicon of primary adjectives is built as a result of this aim. Another aim of the study is to obtain a model of the primary adjective valency for future comparison between primary and derived adjectives valency.

1 Valency in General

It is an impossible task to determine a single definition of valency, individual conceptions mutually differ (more or less). The individual definitions agree only in the name of the founder of the valency concept in linguistics – L. Tesnière. If an attempt to abstract from the individual definitions is made, it can be stated that valency refers to an ability of a verb (but also of other action expressions) to bind other expressions with predefined (and required) forms to itself. As it has been already pointed out, it is not only the verbs which have a valency, but also the nouns, adjectives and adverbs – for more detail see e.g. [9, p.1].

Generally, it is possible to distinguish two (or rather three) different approaches in the Czech linguistics: a semantic approach (F. Daneš, M. Grepl and P. Karlík), a syntactic approach (P. Sgall), and possibly also an approach inferring the valency of units from their syntactic and semantic properties (Panevová, a modified valency theory of P. Karlík) ([17, pp. 8–10], [10, pp. 29–30]).

The positions which a unit binds to itself can be divided into obligatory ones which are closely associated with the semantics of the unit whose meaning is not complete without this complementation, and potential ones (in the terminology of Daneš) or facultative ones (in the terminology of FGP) which are not significantly associated with the semantics of the unit; in some cases it is difficult to determine the boundary between the different complementations.

A dialogue/question test (a test of grammatical completeness) based on the un/acceptability of the answer: "I do not know" can be used as a measure. A speaker who used a unit cannot respond "I do not know" to a question asking about the obligatory member ([10, p. 35], [17, p. 22]). E.g. Je náchylný. – K čemu? – *Nevím. X Je velký. – Jako kdo? / Jak moc? – Nevím. (lit. "He is prone". – "To what?" – "*I do not know." X "He is high." – "Like who? / How high?" – "I do not know.")

In this study valency is understood to refer to an expression of the semantic properties of a word, therefore it is used similarly as in two printed valency dictionaries [14] and [15].

1.1 Adjective Valency

Primary and secondary adjectives without a doubt possess a valency. They can be developed in their valency by a simple grammatical constituent, phrase or clause, the type of the complementation does not depend on the grammatical constituent function of the given adjective. If an adjective has the ability to require a valency complementation, it can be a part of the sentence predicate, a predicate or otherwise; adjectives which do not have a valency cannot be a predicate (this statement is not true if reversed) ([8, pp. 66–68], [12, p. 265]). The repertory of valency complementations is more limited than in verbs, and this applies also in the case of deverbal adjectives. (One of the valency complementations, usually the agent, is included in the unit for which the adjective serves as a determinant, and it is therefore never possible to find an adjective with more than two valency positions [9, p. 9].

In determining the obligatory/potential character of adjective valency positions, the dialogue test is never absolutely reliable. Panevová [9, p. 7] notes: "The complementations ascertained by this test are semantically obligatory, but on the surface they are usually omissible ..."[1] Therefore, this study relies mainly on the corpus data and it uses the indisputable data of frequency instead of the potential/obligatory character.

1.2 The Semantics of the Complementation

Although the detailed semantic determination of the type of complementation is not used any further, it is important to be mentioned. In deverbal adjectives it is better to work with the same group of complementations which are used in the case of verbs (e.g. [10]), in non-deverbal adjectives (understand primary and deverbal adjectives with obscured motivation), the obligatory character of the complementation is established first. In a simplified way it is possible to say that we usually assign the value Pat (grammatically used case, the main feature is binding capacity, e.g. *plný* (lit. *full*), *rovnocenný* (lit. *equal*)) to adjectives with a single valency complementation (if complementations traditionally used for verbs cannot be used). We assign the value Addr (e.g. *věrný komu* (lit. *loyal to whom*)), Regard (*hrdý na koho/co* (lit. *proud of who/what*)), Purpose (e.g. *důležitý k čemu / pro co* (lit. *important for what*)), in isolated cases Direction (e.g. *kolmý na co* (lit. *perpendicular to what*)) or Cause/Origin (e.g. *nešťastný z koho/čeho* (lit. *unhappy because of who/what*)) to single valency adjectives in whose case complementations from the verbal repertory can be used. Double valency adjectives are accompanied by complementations of the type Pat and Addr (e.g. *dlužný komu co* (lit. *owing to whom what*)) – in terms of primary adjectives this is nevertheless an isolated case, other double valency adjectives represent already obscured derivates (e.g. *vděčný za co komu/čemu* (lit. *grateful to whom for what*)) [9, pp. 9–12].

[1] „Doplnění zjištěná tímto testem jsou sémanticky obligatorní, i když na povrchu jsou zpravidla vypustitelná…"

2 Primary and Secondary Adjectives, the Situation in Available Literature

Two types of adjectives are distinguished on the basis of the morphemic composition: primary underived adjectives and secondary derived adjectives (this is only one of the possible divisions, division according to the meaning, possibly according to the specific content (e.g. [7, p. 70]) is not relevant for the purpose of this study). The primary adjectives, e.g. *velký* (lit. *big*), *hezký* (lit. *nice*), represent qualitative (qualifying) adjectives in an overwhelming majority [7, p. 70], [4, p. 281].

For the purposes of this study, the term primary adjective is used to refer to such an adjective which is not derived directly from a verb, or possibly to an adjective whose motivation is no longer apparent, therefore it cannot be assumed that its valency field is derived from an original verb.

A list of primary adjectives which is included in [12] was chosen as the starting material. As the author herself states, she gained the material by making excerpts from [16] and supplemented it from [13] and also from her own linguistic intuition.

Different groups of primary adjectives with a valency can be found in several studies, articles and grammar books e.g. [11], [8, pp. 72–75], [6], but the list put together by Prouzová is the most complete one in the available literature.

Prouzová divides adjectives according to their obligatory valency complementation into groups with complementation in the genitive, dative, instrumental and the summary group with complementation in the prepositional case. Complementations in the form of an infinitive and a subordinate clause are included in the individual groups designated by the name of the case.

The monograph *Valence českých adjektiv* (lit. *The Valency of Czech Adjectives*) [6] based on the data of the SYN2000 corpus, divides adjectives in a similar way – the basic groups are adjectives with a preposition-free case, adjectives with a preposition, adjectives with an infinitive and adjectives with a subordinate clause. The author does not distinguish between primary and derived adjectives, she relies mainly on the frequency value.

There are several different valency dictionaries for the Czech language, both in the electronic and in the printed form; their origin can be traced back to the 1990's. (A detailed overview is given for example by [17, pp. 10–17]). *Slovník slovesných, substantivních a adjektivních spojení* (lit. *Dictionary of Verbal, Substantive and Adjective Connections*), [15] from 2005 represents the latest endeavour. All dictionaries are primarily focused on the valency of verbs, valency of other parts of speech is dealt with only marginally or is missing altogether – e.g. [15] records nouns and adjectives only selectively, with a focus on deverbal nouns and adjectives. Therefore, the above mentioned dictionaries are not used in this study.

Although a valency dictionary of 200 adjectives is included in [6], it does not record all the adjectives described in [12]. It is one of the aims of this study to record this difference.

The main aim of this study is to verify if the stated valency works in the most frequented adjectives from the given groups, and to do so on the material provided by the synchronic SYN2010 corpus (therefore with the distance of almost 30, or rather 10 years). At the same time a short valency lexicon of primary adjectives which can be further expanded is built.

The SYN2010 corpus (100 m of verbal forms) is a synchronic, referential, lemmatized and morphologically tagged corpus; it includes texts from 2005–2009 (genre composition: 40% fiction, 27% professional literature, 33% journalism). The lemmatization and morphological tagging in the SYN2010 corpus is the most reliable one in comparison with all the SYN corpora (e.g. [2]).

3 A List and Frequency of Adjectives

The following tables provide a list of primary adjectives adopted from [12]. Her division of adjectives into those with a genitive, dative and instrumental valency and a valency with prepositional cases is preserved. Only more accurate division into separate cases was added to the set of adjectives with prepositional cases valency. The number next to each adjective refers to the total frequency in the SYN2010 corpus (regardless of concrete cases; some adjectives appear more than one). The input enquiry was modified so that the found concordance lines would include only adjectives without the negative prefix "ne-"[2] (Affirmative and negated units frequently have a common lemma, e.g. *nápadný* (lit. *conspicuous*) – 3,693 occurrences out of which 1,500 occurrences are that of *nenápadný* (lit. *inconspicuous*); the actual number of occurrences is therefore 2,193.). Units which are negative in their basic form constitute an exception. In this case the procedure was the opposite (if it was necessary).

[2] For example [lemma="nápadný" (lit. *conspicuous*) & word!="nenápadn.*" (lit. *inconspicuous*)], in certain cases the mobile vocal *e* has to be taken into account.

GEN.	freq.	DAT.	freq.	INSTR.	freq.
plný/full	29,646	podobný/similar	30,013	?známý/known	32,636
?schopný/able	19,358	jasný/clear	24,564	jistý/sure	29,297
blízký/close	18,039	svatý/holy	13,079	slavný/famous	12,334
?vzdálený/remote	8,030	drahý/dear	11,616	nebezpečný/dangerous	11,509
prostý/void	5,781	cizí/strange	10,925	povinný/obliged	7,061
hodný/worthy	5,184	?příjemný/pleasant	10,457	vinný/guilty	2,422
mocný/powerful	4,528	milý/kind	7,616	nápadný/conspicuous	2,193
důstojný/dignified	1,645	rovný/straight	5,984	pověstný/proverbial	1,248
daleký/distant	1,498	příbuzný/related	5,979		
sytý/full up	674	přítomný/present	5,856		
účastný/sympathetic	178	?zřejmý/obvious	5,814		
		vzácný/valuable	5,794		
		věrný/loyal	3,655		
		příznivý/favourable	3,496		
		pohodlný/comfortable	2,935		
		nepřátelský/hostile	2,316		
		?zjevný/evident	1,883		
		?přiměřený/adequate	1,611		
		?povědomý/familiar	756		
		úměrný/proportional	713		
		adekvátní/adequate	690		
		rovnocenný/equal	612		
		libý/likeable	235		
		konformní/conformist	131		
		kongeniální/ of the same genius	30		

PREPOSITIONAL CASES					
GEN. (+prep.)	freq.	DAT. (+prep.)	freq.	ACC. (+prep.)	freq.
smutný/sad (z)	6,007	dobrý/good (k)	113,116	dobrý/good (pro)	113,116
nešťastný/ unhappy (z)	5,279	?nutný/necessary (k)	20,997	důležitý/ important (pro)	37,694
laskavý/ amiable (k)	2,737	?vhodný/suitable (k)	18,256	?nutný/ necessary (pro)	20,997
nervózní/ nervous (z)	2,577	náročný/demanding (k)	8,800	?vhodný/ suitable (pro)	18,256
veselý/ cheerful (z)	2,252	zlý/bad (k)	7,613	bohatý/rich (na)	11,969
?žhavý/ burning (do)	1,676	hotový/ready (k)	7,339	slabý/weak (na)	9,151
		?nezbytný/essential (k)	6,528	náročný/demanding (na)	8,800
		ochotný/willing (k)	6,455	zlý/bad (na)	7,613

		hrubý/rough (k)	5,883	?nezbytný/ essential (pro)	6,528
		spravedlivý/fair (k)	2,545	*chudý/poor* (na)	5,930
		hodný/good (k)	5,184	hrubý/rough (na)	5,883
		pozorný/attentive (k)	982	hodný/good (na)	5,184
		něžný/tender (k)	1,823	užitečný/useful (pro)	4,086
		?zralý/ripe (k)	1,693	opatrný/careful (na)	2,616
		lhostejný/indifferent (k)	1,549	?nadšený/ enthusiastic (pro)	2,558
		hluchý/deaf (k)	852	hrdý/proud (na)	2,282
		drzý/cheeky (k)	767	?zvědavý/ curious (na)	2,280
		bezohledný/ thoughtless (k)	653	pyšný/proud (na)	1,891
		bezbranný/ defenceless (vůči)	571	?zralý/ripe (na)	1,693
		?náchylný/prone (k)	509	pozorný/ attentive (na)	982
		?zaujatý/ interested in (vůči)	479	líný/lazy (na)	921
		kompetentní/qualified (k)	457	choulostivý/ sensitive (na)	605
		milostivý/merciful (k)	400	?náchylný/prone (na)	509
		ohleduplný/considerate (k)	327	bezcenný/worthless (pro)	495
		bezcitný/heartless (k)	251	lenivý/indolent (na)	146
		imunní/immune (vůči)	228		
		benevolentní/ benevolent (k)	132		
		povolný/compliant (k)	86		
		svolný/compliant (k)	75		
		beztaktní/tactless (k)	4		

LOC. (+prep.)	freq.	INSTR. (+prep.)	freq.
jistý/sure (v)	29,297	hotový/finished (s)	7,339
?zkušený/experienced (v)	5,294	svolný/compliant (s)	75
kompetentní/qualified (v)	457		
zručný/skilful (v)	336		
povolný/compliant (v)	86		

Notes: *Another type of complementation is often possible for the majority of given adjectives (typically relative clauses and infinitives) – for more details see lists in [12] and [6].*

Some frequent valency adjectives are not included (like velký (lit. big), malý (lit. small) etc.), because they occur neither in Prouzová's list nor in Kopřivová's list. The study could be expanded by addition of mentioned adjectives and adjectives similar to them to the original list.

Adjectives printed in italics can be considered as non-requiring complementations [9, p. 8].

Adjectives marked with a question mark are rather deverbative (this problem is especially connected with adjectives ending in -[áeě]n- and -[iyeéa]t-, its satisfactory solving is difficult and depends on the depth of used etymology, besides other things). Nevertheless correction of Prouzová's list is not an aim of this study.

3.1 Valency Frame of Selected Adjectives

The scope of this work does not allow the recording of valency frames of all adjectives, and the criterion of frequency was therefore used to choose the five most frequent ones in the group of non-prepositional and prepositional cases.

A simple method of recording which is used also by [6] was adopted for the purposes of this study. In the case of a valency with a non-prepositional case, the case of the relevant word is also stated; in the case of a prepositional case, the case of the word and the preposition are both stated, similarly the conjunction which introduces a subordinate clause is also given. Only the right valency is recorded, and no attempt is made to determine if the positions are potential/obligatory.

The valency complementation given by [12] is printed in *italics*, the valency complementation given by [6] is in the **bold** typeface, the ***combination*** of both methods shows that the given valency was recorded by both sources. Newly added complementations are indicated by underlining.

The valency complementations with the highest frequency are given in parentheses (in the infinitive the possible complementations proceed from the first position on the right from the KWIC). In the cases of complementations by a subordinate clause only one example is given for illustration. Valency complementations are ranked according to their frequency in the SYN2010 corpus, from the most frequent complementation to the least frequent one.

PLNÝ/FULL

A – Sgen:	7,464 (*lidí, života, slz*) (lit. *of: people, life, tears*)
A – toho, co SENT:	2 (*p. toho, co bude muset udělat, p. toho, co naplňovalo moje srdce*) (lit. *f. of he will have to do, f. of what has been filling my heart*)
A – toho, že SENT:	1 (*noviny p. toho, že jsme jim udělali ostudu*) (lit. *newspaper f. of that we have embarrassed them*)

SCHOPNÝ/ABLE

A – INF:	13,568 (*pochopit, vyrovnat, soustředit*) (lit. *understand, balance, concentrate*)
A – Sgen:	529 (*slova, pohybu, provozu*) (lit. *of: word, movement, working*)
A – pro ACC:	27 (*zdravotní, mě, své, … výkupné, Brňany*) (lit. *for: health, me, my, … ransom, citizens of Brno*)

A – k DAT:	12 (*tomu, ní, takovém, … obývání, životu*) (lit. *to: that, her, this, … inhabitation, life*)
A – toho, aby SENT:	1 (*s. toho, že by se každá země*) + 1 (*neschopni toho, aby se rozšířili*) (lit. *a. that every country*) + (lit. *una. to expand*)
BLÍZKÝ/CLOSE	
A – Sdat:	337 (*rychlosti, nule, věku*) (lit. *to: speed, zero, age*)
A – Sgen:	32 (*druhu, rychlosti, přátel*) (lit. *to: kind, speed, friends*)
A – Sinstr:	22 (*jednáním, citoslovci, rodem*) (lit. *by: behaviour, interjections, familyline / genus*)
A – k DAT:	3 (*blízkých k ránu a večeru*) (lit. *to: the morning and evening*)
A – toho, aby SENT:	1 (*blízek toho, aby*) (lit. *close to*)
A – tomu, aby SENT:	1 (*blízká tomu, aby*) (lit. *close to*)
VZDÁLENÝ/REMOTE	
A – od GEN:	255 (*sebe, nás, místa*) (lit. *from: oneself, us, place*)
A – Sdat:	59 (*horizontu, světu, životu*) (lit. *to: horizon, world, life*)
A – Sgen:	17 (*ideálu, krajnosti, lyzolu (chodeb)*) (lit. *of: ideal, limit, lysol (of corridors)*)
A – toho, aby SENT:	0
PROSTÝ/VOID	
A – Sgen:	95 (*problémů, vody, touhy*) (lit. *of: problems, water, desire*)
A – toho, aby SENT:	0 (jen typ *tak prosté to ovšem nebývá* – odpovídá témuž dotazu) (only type *"it does not used to be so easy"* – it corresponds to the same query)

PODOBNÝ/SIMILAR	
A – Sdat:	2,378 (*listům, člověku, lidem*) (lit. *to: leafs, human, people)*
A – Sinstr:	?40 (*vzhledem, (Slunci), tvarem;* často chybné značkování) (lit. *by: appearance, (sun), form)*; often wrong tagged cases)
A – v LOC:	0 (spojení s předložkou *v* v 82 případech) (connection with preposition *v* in 82 cases)
A – SinstrSdat:	0
JASNÝ/CLEAR	
A – že SENT:	6,231 (*bylo naprosto j., že jediným možným smyslem*) (lit. *it was absolutely c., that the only possible sense)*
A – zda SENT:	3,759 (*není j., zda ministerstvo války*) (lit. *it is not c., if the Ministry of War)*
A – jestli SENT:	145 (*j., jestli jsi při smyslech*) (lit. *c. if you are in your mind)*
A – Sdat:	18 (*kapitánovi, stratégovi, Petru*) (lit. *to: captain, strategist, Peter)*
A – pro ACC:	8 (*celý, případ, všechny*) (lit. *for: whole, example, all)*

SVATÝ/HOLY

A – Sdat:	?19, resp. 1 (*zběsilost je s. potomkům*; jinak nevalenční nebo jiné*; o svatém Janě, ve Svatém Janě nad Malší*) (lit. *furiousness is h. to offspring*; otherwise non-valency complementation, see above)

DRAHÝ/DEAR

A – Sdat:	?109, resp. 6 (*bude d. mystikům;* jinak chybné značkování – 103 případů) (lit. *will be d. to mystics;* otherwise 103 wrong tagged cases)

CIZÍ/STRANGE

A – Sdat:	8 (*lidstvu, světu, krasoduchům*) (lit. *to: mankind, world, "nicespirits"*)

ZNÁMÝ/KNOWN

A – Sinstr:	121 (*koncem, výrokem, výskytem*) (lit. *by: end, statement, occurrence*)
A – tím, že SENT:	114 (*nechvalně z. tím, že je prolezlý chybami*) (lit. *ingloriously known by being full of mistakes*)

JISTÝ/SURE

A – Sinstr:	169 (*úspěchem, vítězstvím, životem*) (lit. *about: success, victory, life*)
A – v LOC:	29 (*kramflecích, tom, posteli*) (lit. *in: "heels", that, bed*)
A – tím, že SENT:	84 (*j. tím, že jel v souladu s předpisy*) (lit. *s. that he has driven in accordance with rules*)
A – tím, jestli SENT:	5 (*j. tím, jestli by si potom nemysleli*) (lit. *s. if they do not think after*)
A – tím, zda/zdali SENT:	3 (*j. tím, zdali to, co se posléze přihodilo*) (lit. *s. if that, what has happened after*)

SLAVNÝ/FAMOUS

A – Sinstr:	?8, resp. 4 (*s. likérem benediktinka, s. chovem ryb*) (lit. *for: liqueur "benediktinka", farming of fish*)
A – tím, že SENT:	2 (*s. tím, že fotí bezdomovce, s. tím, že jeho hrdinové měli*) (lit. *f. for taking photos of homeless people, f. that his heroes had have*)

NEBEZPEČNÝ/DANGEROUS

A – pro ACC:	87 (*společnost, člověka, chodce*) (lit. *to: society, human, pedestrian*)

A – tím, že SENT:	11 (*n. tím, že v sobě obsahuje tři formy*) (lit. *d. that he contains three forms in himself*)
A – Sinstr:	3 (*n. směrem dopředu, n. rychlostí*) (lit. *by forward direction, speed*)

POVINNÝ/OBLIGED

A – INF:	2421 (*splnit, zachovávat, zajistit*) (lit. *fulfil, maintain, guarantee*)
A – k DAT:	63 (*dani, odběru, registraci*) (lit. *to: tax, offtake, registration*)
A – pro ACC:	57 (*všechny, každého, orgány*) (lit. *for: all, every, organs*)
A – Sinstr:	25 (*výkonem, řezníky, úctou*) (lit. *by: performance, butchers, respect*)
A – tím, že SENT:	0

DOBRÝ/GOOD

A – INF:	3,385 (*vědět, konzultovat, připomenout*) (lit. *know, consult, remind*)
A – k DAT:	775 (*jídlu, tomu, snědku*) (lit. *to: meal, that, eating*)
A – že SENT:	255 (*je d., že to říkáš*) (lit. *it is g. that you are telling it*)
A – pro ACC:	253 (*nás, všechny, lidi*) (lit. *for: us, all, people*)
A jako:	203 (*ty, každá, každý, oni*) (lit. *as: you, she-every, he-every, they*)
A – na ACC:	174 (*to, tom, zub, nohy*) (lit. *for: it, that, tooth, feet*)
A – aby SENT:	142 (*nebylo by d., aby se o něm vědělo*) (lit. *it would not be g. to have known about him*)
A – do GEN:	20 (*té, toho, výuky, něj, polévky*) (lit. *for: it, that, teaching, him, soup*)
A – při LOC:	14 (*tom, tomhle, 18 stupních, obléhání*) (lit. *during: that, that, 18 degrees, siege*)
A – to, že SENT:	3 (*dobré tím, že zkouškové začíná*) (lit. *g. that exam period is starting*)

DŮLEŽITÝ/IMPORTANT

A – INF:	2,550 (*mít, vědět, znát*) (lit. *have, know, know*)
A – pro ACC:	862 (*nás, mě, život*) (lit. *for: us, me, life*)
A – aby SENT:	712 (*strašně d., abyste ho nezvedal*) (lit. *very i. not to lift him up*)
A – že SENT:	411 (*je d., že při práci pro zpravodajskou službu*) (lit. *i. that during the working for intelligence service*)
A jako:	?271 (*to, pro, jeho, ty*) (lit. *as: it, for, his, you*)
A – v LOC:	102 (*této, tom, případě, situaci*) (lit. *in: this, that, case, situation*)
A – zda SENT:	60 (*d., zda si tu skutečnost přiznám*) (lit. *i. if I admit the reality to myself*)

A – při DAT: 58 (*stříhání, rozhodování, řešení*) (lit. *during: cutting, decision making, solving of*)
A – na ACC: 35 (*práci, srdci, tom*) (lit. *to: work, heart, that*)
A – k DAT: 22 (*udržení, tomu, pochopení*) (lit. *to: sustaining, that, understanding*)
A – u GEN: 17 (*dětí, utrechtského, ohrožených*) (lit. *children, Utrecht's, endangered*)
A – pro to, aby SENT: 0

JISTÝ/SURE – see above

NUTNÝ/NECESSARY
A – INF: 11,838 (*dodat, podotknout, říci*) (lit. *add, remark, say*)
A – aby SENT: 608 (*opravdu n., abychom se dozvěděli všechno*) (lit. *really n. to get to know everything*)
A – k DAT: 443 (*tomu, dosažení, životu*) (lit. *to: that, achievement, life*)
A – pro ACC: 424 (*vstup, zachování, jeho*) (lit. *for: entry, maintaining, his*)
A – v LOC: 148 (*zájmu, rámci, tomto*) (lit. *in: interest, framework, this*)
A – při LOC: 89 (*každé, jejich, výpočtu*) (lit. *during: every, their, calculation*)
A – na ACC: 82 (*tuto, něj, ně, základu, začátku*) (lit. *this, he, they, base, start*)
A – z GEN: 61 (*důvodu, hlediska, nich*) (lit. *from: reason, point of view, they*)
A – do GEN: 47 (*budoucna, ceny, konce*) (lit. *for: future, price, end*)
A – s INSTR: 42 (*nimi, tím, ním, velkým, vývojem*) (lit. *with: they, that, he, big, development*)
A – po LOC: 31 (*každém, době, výměně*) (lit. *after: every, period, exchange*)
A – u GEN: 25 (*některých, dna, většiny*) (lit. *by:some, bottom, majority*)
A – před INSTR: 25 (*jejich, počátkem, samotným*) (lit. *before: their, beginning, itself*)
A – za GEN: ?20 (žádné spojení předložky *za* s genitivem) (no connection of preposition *za* with genitive)

VHODNÝ/SUITABLE
A – INF: 2,155 (*použít, mít, používat*) (lit. *use, have, use*)
A – pro ACC: 1468 (*všechny, děti, použití*) (lit. *for: all, children, using*)
A – k DAT: 404 (*tomu, použití, výrobě*) (lit. *for: that, using, production*)
A – aby SENT: 176 (*v., aby se choval jako slon*) (lit. *s. not to behave like an elephant*)
A – na ACC: 120 (*přípravu, to, zimu*) (lit. *for: preparation, it, winter*)
A – do GEN: 107 (*slunného, vašeho, skalek*) (lit. *for: sunny, yours, rock gardens*)

A jako:	62 (*příloha, součást, prezence*) (lit. *as: supplement, component, attendance*)
A – v LOC:	61 (*době, případě, situacích*) (lit. *in: period, case, situations*)
A – při LOC:	47 (*omezení, žaludečních, zažívacích*) (lit. *during: restriction, stomachal, digestive*)
A – po LOC:	23 (*skončení, celou, několika*) (lit. *after: ending, whole, some*)
A – u GEN:	12 (*zdravotníků, nejistých, plotů*) (lit. *medics, uncertain, fences*)
A – z GEN:	11 (*hlediska, líce, oken*) (lit. *from: point of view, face, windows*)

3.2 Specifications of Searching

Only adjectives in the affirmative form in the positive were used in the search for all phenomena. Moreover, some procedures and findings which are given by [6, pp. 18–68] were used for the purposes of this study.

The complementation in the non-prepositional case was found by removing adjectives in every given case (the possibility of finding a modifier which was not the subject of the study was thus decreased) and by subsequent filtering of the first right position from the KWIC (where the occurrence of the valence complementation is most likely). A similar approach was used in the search for the complementation with a preposition. In examining the infinitive complementation, the focus was only on the interval of positions (1,3) on the right from the KWIC (the probability that infinitives are valency complementations of an adjective decreases with distance; frequency lists nevertheless proceed from the frequency distribution of the position (1,1)). Possible complementations in the form of a subordinate clause were sought by means of a lemma of specific conjunctions and the assumption that it is a single clause. For examples see the notes[3].

[3] Searching of valency complementation in dative (after removing KWIC in dative) – P-filtr (1,1) [tag="N...3.*" & tag!="A...3.*"].

Searching of valency complementation in infinitive – P-filtr (1,3) [Vf.*].

Searching of the connection *slavný tím, že* (lit. *famous for*) [lemma="slavný"] ?[lemma="ten"] [word="\,"] ?[tag="J.*"] within <s/>.

Searching of the connection *vhodný, aby* (lit. *suitable to*) – with specific conjunction [lemma="vhodný" & tag="A.*"] [word="\,"] ?[word="aby"] within <s/>.

Searching of the connection *nebezpečný tím, že* (lit. *dangerous)* [word="nebezpečn.*"] [lemma="ten"] [word="\,"] ?[tag="J.*"] within <s/> – adjective *nebezpečný* (lit. *dangerous*) is lemmatized as *bezpečný* (lit. *safe*).

4 Conclusion

Using the material of the SYN2010 corpus, the study attempted to describe the valency frames of several primary adjectives, and it used the lists of adjectives given in [12] and [6] as a basis.

The differences between the valency frames given in both studies were described, and the given frames of selected adjectives were tested on the contemporary material of the SYN2010 corpus. The result is an extensive group of primary valency adjectives accompanied by frequency statistics and a detailed description of valency in five most frequent adjectives from each group – valency complementations are ranked according to their frequency in the SYN2010 corpus.

In the case of two adjectives (*plný* (lit. *full*), *blízký* (lit. *close*)) new possible complementations were found (*plný toho, že SENT* (lit. *full of*); *blízký tomu, aby SENT* (lit. *close to*)). On the other hand, some of the given complementations were not found at all (*vzdálený toho, aby SENT* (lit. *remote of*); *prostý toho, aby SENT* (lit. *void of*); *podobný v LOC* (lit. *similar in*); *podobný Sinstr Sdat* (lit. *similar*); *povinný tím, že SENT* (lit. *obliged in*); *důležitý pro to, aby SENT* (lit. *important for*)). On the basis of these findings the two existing lists can be revised.

The adjectives included in the list established by [6] are accompanied with data on frequency (in the SYN2000 corpus). By comparing these data with the results of this study, the conclusion was reached that the frequency distribution of valency complementations established on the basis of the SYN2010 corpus did not fundamentally differ from the results reached by [6] (despite the higher level of lemmatization of the SYN2010 corpus). Minor confusions in the order of the valency complementations in some adjectives and a single prominent difference in the adjective *vhodný* (lit. *suitable*) are insignificant.

(In the case of the adjective *známý* (lit. *known*), [6] gives the order of complementations: *že SENT – Sinstr* with a difference in the order of hundreds of occurrences; in the material of this study the order is the opposite with the difference in the number in the order of single cases; similarly in the case of *povinný* (lit. *obliged*), the order of complementations given is: *pro ACC – k DAT*; in the material of this study it is the opposite; possibly in the case of *důležitý* (lit. *important*) the order is: *v LOC – jako*; in the material of this study the order is again the opposite. Differences of the same type appear also in the case of adjectives *vzdálený* (lit. *remote*), *jistý* (lit. *certain*), *dobrý* (lit. *good*), *nutný* (lit. *necessary*). In the case of the adjective *vhodný* (lit. *suitable*) the given order of complementations is: *pro ACC – INF* with a difference in the number of occurrences in the order of single cases, this study, however, found the opposite order of complementations with the difference in the order of hundreds of occurrences.)

It remains to be added that in cases where there is a possibility of an infinitive complementation, the infinitive is always more frequent than other complementations (this concerns adjectives in the function of predicates), see for example *dobrý* (lit. *good*), *vhodný* (lit. *suitable*), *nutný* (lit. *necessary*). Complementations with a subordinate clause, on the other hand, show the least frequency (see *plný* (lit. *full*), *schopný* (lit. *able*), *blízký* (lit. *close*), *známý* (lit. *known*),etc.).

Isolated mistakes in the morphological tagging and the need for a correct interpretation of originally acquired data (see adjectives *podobný* (lit. *similar*), *svatý* (lit.

holy), *drahý* (lit. *dear*), *slavný* (lit. *famous*), *důležitý* (lit. *important*), *nutný* (lit. *necessary*)) were pointed out. Ambiguity in the lemmatization – when a negated form is assigned an affirmative lemma but not all negated forms are included in this lemma – causes also certain difficulties in the searching.

The study could be expanded both quantitatively (examination of valency in a higher number of, or in all, adjectives) and qualitatively – due to the limited scope the obligatory/potential character of valency complementations and also the semantic roles of complementations were set aside.

References

[1] Czech National Corpus – *SYN2010* (2010). Prague: Institute of the Czech National Corpus, Faculty of Arts, Charles University in Prague. Accessible at: `http://www.korpus.cz`, retrieved 2 March 2013.

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

[3] Karlík, P., Nekula, M., and Pleskalová, J., editors. (2002). *Encyklopedický slovník češtiny (The Encyclopedic Lexicon of Czech Language).* Nakladatelství Lidové noviny, Praha.

[4] Karlík, P., Nekula, M., and Rusínová, Z., editors. (2003). *Příruční mluvnice češtiny (The Hand Grammar of Czech Language)*. Nakladatelství Lidové noviny, Praha.

[5] Kocek, J. (2013). *Korpus SYN2010 (Corpus SYN2010).* Accessible at: `http://ucnk.ff.cuni.cz/syn2010.php`, retrieved 2 March 2013.

[6] Kopřivová, M. (2006). *Valence českých adjektiv (The Valency of Czech Adjectives)*. Nakladatelství Lidové noviny, Praha.

[7] *Mluvnice češtiny 2. Tvarosloví (Grammar of Czech Language 2. Morphology).* (1986). Academia. Praha.

[8] *Mluvnice češtiny 3. Skladba (Grammar of Czech Language 3. Syntax).* (1987). Academia. Praha.

[9] Panevová, J. (1998). Ještě k teorii valence (Once more to the Theory of Valency). *Slovo a slovesnost*, 59:1–14.

[10] Panevová, J. (1999). Valence a její univerzální a specifické projevy (Valency and its Universal and Specific Manifestations). In Hladká, Z. and Karlík, P., editors, *Čeština – univerzália a specifika*, pages 29–37, Masarykova univerzita, Brno.

[11] Piťha, P. (1982). K otázce valence u adjektiv (Towards the Valency of Adjectives). *Slovo a slovesnost*, 40:113–118.

[12] Prouzová, H. (1983). K valenčním vlastnostem primárních adjektiv v češtině (On the Valency of Underived Adjectives in Czech). *Slovo a slovesnost*, 44:265–274.

[13] Slovník spisovného jazyka českého I-IV (Dictionary of the Standard Czech Language I-IV). (1989). Academia, Praha.

[14] Svozilová, N., Prouzová, H., and Jirsová, A. (1997). Slovesa pro praxi: valenční slovník nejčastějších českých sloves (Verbs for Practice: Valency Lexicon of the most Frequent Czech Verbs). Academia, Praha.

[15] Svozilová, N., Prouzová, H., and Jirsová, A. (2005). Slovník slovesných, substantivních a adjektivních vazeb a spojení (Dictionary of Verbal, Substantive and Adjective Relations and Connections). Academia, Praha.

[16] Šmilauer, V. (1966). Novočeská skladba (Modern Czech Syntax). SPN, Praha.

[17] Urešová, Z. (2011). *Valence sloves v Pražském závislostním korpusu (The Valency of Verbs in Prague Dependency Treebank)*. ÚFAL, Praha.

Event Extractor: Email Events Detection and Calendar Integration

Filip Ogurčák[1] and Michal Laclavík[2]

[1] Faculty of Informatics and Information Technologies, Slovak University of Technology, Bratislava, Slovakia

[2] Institute of Informatics, Slovak Academy of Sciences, Bratislava, Slovakia

Abstract. In this paper, we discuss the event detection from Slovak emails. We have gathered small corpus of Slovak emails with events. Then we analyze this corpus in order to understand the structure of email events and its main parts such as time, place and title of the event. We also introduce the approach for detection of time and name of event and evaluate it on the corpus. In addition we describe Thunderbird plug-in, which supports the event detection functionality and its integration with Google calendar as well as its server part which is used for event corpus creation.

1 Introduction

Email as a medium of communication becomes main technology for everyday exchange of information. Everyday usage of email has become inseparable part of both work and personal live.

Information exchanged by email is mostly written in natural language, therefore not understandable for computers. We can use specified software to include tasks or meeting requests in email message or other types of events, which can be described. Time consumption while using these software solutions can be great and it is definitely more comfortable to write our message in natural language.

Email events are being sent in different form and languages, therefore language dependent natural language processing is needed. We focused on Slovak language and domain containing events with time and activity.

Event detection is supported by some extend in email clients, but only if events are shared in standardized way, using hCalendar[1] or iCalendar[2] microformat standards. Gmail recently introduced limited text based event detection functionality on English. We have decided to integrate our approach for event extraction with Thunderbird email client and Google Calendar, however main contribution of the paper lies in created and analyzed corpus of Slovak email events. The corpus experiment shows statistics of email event parts (time, title, place) in email event messages. On these results, appropriate event detection algorithms can be build.

2 Event Corpus

In order to extract events from emails we need to understand how event entities (event time, title and place) are represented in email and where they are located in the email. In this chapter we discuss occurrence of these entities in emails in order to find useful patterns for their extraction. For this purpose we had collected 83 emails which contain 100 events. This corpus is then used for statistical evaluation of event entities dependency.

[1] http://microformats.org/wiki/hcalendar

[2] http://microformats.org/wiki/icalendar-implementations

2.1 Corpus Creation

To create corpus we have used Thunderbird plug-in described in chapter 4. Plugin had limited functionality of event time and title detection, which could be then corrected by users manually. So the users set date, title and place of event. This event was then added to user calendar and it was also added to our database together with text of the email. Because event was added to user calendar, precondition is that is relevant for him.

All collected data were then manually checked and corrected in case of user's spelling mistake. We have removed emails in language other than Slovak. In this way we have collected 100 events contained in 83 emails. All these emails are written in Slovak language and can be grouped to personal, work and automatic generated emails. Length of these emails varies from 24 to 6,435 words (average is 750 and median 386 words), so our corpus contains short and long emails. And because users varies in diacritics usage, our corpus contains emails with diacritics (23%) and without diacritics (77%) as well.

We can say the corpus is made up by the heterogeneous emails and covers all types of email usage.

2.2 Experiment and Corpus Statistics

The main purpose of the experiment is to understand where in the message date and time of beginning, date and time of ending, title and place of event is located. We focus especially to understand where date and time of event is located in text, because it is the most important entity which unambiguously identify event. Then we focus on the dependence of date and the event title, and in last step we focus on the dependence of place, date and title of event. This experiment was conducted over the data from corpus and all data were processed and evaluated manually. The results of the experiment are presented in three tables for date and time (Table 1), title (Table 2) and location (Table 3) of the event.

In the first table (Table 1) we identified dates and times localized in subject, text and history of message. We have also identified dates and times defined by numbers and words, where some dates defined by words depends on the date of sending and some not.

	Start date	**Start time**	**End date**	**End time**
Defined by numbers	53%	76%	20%	25%
Defined by words	67%	6%	11%	1%
Independent from the date of sending	12%	100%	27%	100%
Depends on the date of sending	88%	0%	73%	0%
Is located in the subject	3%	0%	0%	0%
Is located in the text	78%	77%	26%	26%
Is located in the history	7%	5%	0%	0%
Not present in email	12%	18%	74%	74%
It can be deduced from the beginning of the event			92%	0%

Table 1. Location and representation of the date and time in the email

In Table 1 we can see summary of the experiment regarding date and time of the event. Following findings are interesting: 67% of events have date of start defined by words and 88% of them depends of the date of sending message; 76% of events have time of start defined by numbers and only 6% by words; 74% of messages do not define the end of the event in the text of the message, but in up to 94% percent of them the date of ending is the same as date of beginning; about 77% of messages include the date and time of beginning in the text. Based on these facts, we can say that we need to focus mainly on the dates defined by words and times defined numerically in date and time extraction, focusing mainly on the text of message.

In second table (Table 2) we identified titles of events localized in subject, text and history of message. In events localized in text we then focused on the dependence of location of the date to find out how these event titles can be discovered in text.

	Event name
Is email subject	49%
Is part of email subject	23%
Is located in the text	40%
In the same sentence as the time	65%
In the sentence before the time	5%
Somewhere before sentence with time (different as previous)	12%
In the sentence after the time	5%
Somewhere after sentence with time (different as next)	13%
Is located in the history	0%
Not present in the email	5%

Table 2. Event title in the email and its dependence on the date of event

In Table 2 we can see summary of the experiment regarding title of the event. Following findings are interesting: 72% of events have title localized in the subject and 26% in the same sentence as date of event. Based on these facts, we can say that we need to focus mainly on the subject of message or sentence with the date of event.

In the last table (Table 3) we identified places of events localized in subject, text and history of message. In events localized in text we then focused on the dependence of location of the date and title to find out how these event places can be discovered in text.

		Event place
Is located in the subject		2%
Is located in the text		49%
	In the same sentence as the time	71%
	In the sentence before the time	9%
	Somewhere before sentence with time (different as previous)	0%
	In the sentence after the time	14%
	Somewhere after sentence with time (different as next)	6%
	In the same sentence as the title	19%
	In the sentence before the title	0%
	Somewhere before sentence with title (different as previous)	0%
	In the sentence after the title	1%
	Somewhere after sentence with title (different as next)	14%
Is located in the history		4%
Not present in the email		7%
Is not defined		38%

Table 3. Event place in the email and its dependence on the date and title of event

In Table 3 we can see summary of the experiment regarding place of the event. Following findings are interesting: 38% of events have not defined place of event; 34% of events have title localized in the same sentence as date of event. Based on these facts, we can say that we need to focus mainly on the sentence with the date of event. Because location of the event may not be defined at all, we do not need to pay much attention for him.

Based on the results of experiment, we are able to identify which of the extracted data is relevant and therefore important for the user. We are also being able to separate several events, which may be present in the same email. The results of this experiment would be applied in the method implementation and at the stage of increasing precision and recall of this extraction method.

3 Event Extraction Method

We can use two different approaches to extract events from text [2]: knowledge engineering or machine learning. Both approaches use defined rules to extract entities, but knowledge engineering use rules defined by domain expert, and machine learning use extensive training set to automatically extract this rules. Another simple approach is to

use gazetteer (list) based extraction [6], which we use for days for month names extraction when they are defined by words. It can be also used for event place detection using list of geographical locations or user defined list of locations such as room numbers.

In this paper we focus on knowledge engineering approach in which we manually create rules to extract required entities. For this purpose we use GATE, an open source tool for text processing which use defined JAPE rules [5] to process text.

The disadvantage of this approach is too many irrelevant results which are need to be detected and removed. In further work, we would like to use machine learning approach to filter these irrelevant results and extract only relevant ones. For this purpose we would like to use machine learning OpenNLP projects [4] trained with data in our corpus. Using this approach, we should be able to reach the required precision and recall of our extraction method.

3.1 Date and Time Extraction

In the extraction of date and time it is necessary to recognize two ways in which the data can be defined: defined numerically or defined by words.

In date and time defined numerically we start with standardized format [3]:

[DAY].[MONTH].[YEAR] [HOUR]:[MINUTE]

We can define individual data ranges (day is in range 1-31, month 1-12, hour 0-23, ...) so it is easy to find a pattern for extraction. However, we must take care of format variations in user usage, so it is possible to find multiple patterns.

This patterns can be easily transform to regular expression patterns with which we extract date and time of event.

Complicated situation occurs when date or time is defined by words. We must use gazetteers which contains all this words, and because every word has affected the date and time otherwise, we must define formula for every word. On the other hand, there is a finite set of these words so gazetteer creation is possible.

3.2 Title Extraction

In the extraction of title we use experiment results (Table 2) and focus only to title localized in subject or sentence with date of event.

To extract title from the subject we easily remove special marks defining that message is replied (Re:) or forwarded (Fwd:) and we get whole subject which theoretically represent title of event.

To extract title from the sentence with date of event we decided to split the sentence according to commas, dates and places of events, so our sentence will be in next format:

[POSSIBLE TITLE 1], [POSSIBLE TITLE 2] [DATE] [POSSIBLE TITLE 3] [PLACE] [POSSIBLE TITLE 4].

As you can see we have multiple possibilities where title of event can be located. Because title must contains noun in its title, we use POS tagger to reduce this possibilities. The result of this reduction is only 1 or 2 phrases which with 26% of probability represent title of event.

3.3 Place Extraction

In the extraction of place we use experiment results (Table 3) and focus only to place localized in sentence with date of event or in sentence after this date.

To extract place from this sentences we use the fact, that place of event contains noun and usually contains preposition (in Slovak: v, na, do, pred, ...). We can also define set of locations (cities, states, firms, cafes…) and use gazetteers to hold them. At the end we define rules which contains results from this gazetteers and POS tagger and use them for extraction.

3.4 Evaluation

In previous paragraphs we describe hypotheses to extract date, time, title and place of event. To evaluate these hypotheses we use relevant events from corpus and manually apply hypotheses on them. The results of this evaluation are described in Table 4. Because hypotheses were applied only on relevant data, expected precision is expressed by interval only.

	Precision	**Recall**
Date and time extraction	40–93%	91%
Title extraction	20–45%	71%
Place extraction	10–40%	83%

Table 4. Expected extraction precision and recall

To evaluate extraction of date and time we also implement our hypothesis to GATE using JAPE rules and Gazetteers, and apply these rules to another corpus. The result of this evaluation is described in Table 5 and it is almost same as expected.

	Precision	**Recall**
Date and time extraction	66%	89%

Table 5. Precision and recall of implemented extraction method

As you can see the results (Table 4 and Table 5), described extraction method reaches good recall for all extracted entities but not as well precision. It is due to many irrelevant results returned from this method. In next work we would like to use created corpus to train machine learning method witch reduce this irrelevant results and increase precision as well.

4 Thunderbird Event Extractor Plug-in

Thunderbird plug-in represents the client part of the developed application. Its main function is in connecting Thunderbird (email message), the extraction server and Google calendar [1]. In this section we describe functionality of the plug-in.

Plug-in first sends text of email to extraction server. The server performs data analysis and after processing send extracted data back to the plug-in. Next step is user authorization on Google Oauth 2.0[3] authorization server, where unique access tokens are generated with help of user interaction. These tokens are necessary for communication with Google services. Finally request is send to Google Calendar API[4], which will create event in the user's calendar.

Communication is based on REST[5] requests with strictly specified data structure collected in JSON object. Example of this object may look like this:

```
{"MessageID":692,
"Name":["Druhe kolo osobneho pohovoru"],
"Place":[""],
"DateFrom":["2013-05-09","2013-05-13"],
"TimeFrom":["16:42","10:00"],
"DateTo":["2013-05-09","2013-05-13"],
"TimeTo":["17:42","11:00"],
"Description":""}
```

Plug-in can be activated from pop-up menu or by pressing the "G" on keyboard. When started for the first time, two GUI windows will appear – first for authorization and second for event creating and editing. When authorization tokens are valid, the authorization window will not appear next time.

The event window (see Fig.1) is the main window and it is displayed every time user runs the plug-in. This window contains only necessary components to create new event: event title, duration, place and description. Text boxes are filled in with information detected by the extraction server. User can select from multiple values or edit the information. Because the extraction server provides multiple extraction methods, user can change used extraction method in options menu (see Fig.1).

[3] http://code.google.com/intl/sk-SK/apis/accounts/docs/OAuth2Login.html

[4] http://code.google.com/intl/sk-SK/apis/calendar/v3/reference.html

[5] http://www.xfront.com/REST-Web-Services.html

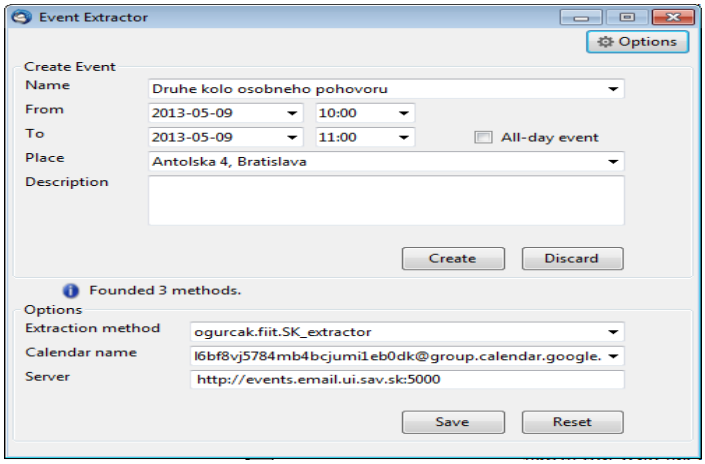


Fig.1. Thunderbird Event Extractor plugin with open options menu

Web Service of the extraction server is composed of two parts: one for extraction and one for capturing requests. The whole architecture is suitable to add new extraction method or new client. This way, it can be used with another email client or calendar.

5 Conclusion

In this paper we have presented our approach for event detection in Slovak emails based on knowledge engineering approach with defined rules for extraction. For this purpose we have prepared corpus with annotated events, where we have searched for dependencies between the date, time, title and place of the event. Based on these results we deduced hypotheses for the extraction of individual entities and evaluate them on corpus.

Because we used knowledge engineering approach with defined rules, we focused on providing more data by cost of precision and let user to choose relevant extracted information. In the future we would like to use created corpus and machine learning to increase precision as well.

Acknowledgement

This work is supported by project VEGA 2/0185/13. This publication is also the result of the project implementation ITMS: 26240220072 supported by Operational Programme Research & Development funded by the ERDF.

References

[1] Ogurčák, F. (2012). *Extrakcia udalostí z emailovej komunikácie a integrácia s kalendárom*. FIIT STU, Bratislava.
[2] Paralič, J. et al. (2010). *Dolovanie znalostí z textov*. Equilibria, s.r.o., Košice.
[3] Pravidlá písania a úpravy písomností (2011). Technická norma, STN 01 6910. 1. 4. 2011. Slovenský ústav technickej normalizácie, Slovenská republika.
[4] Baldridge, J. (2005). *The openNLP project*. URL: `http://opennlp.apache.org/`.
[5] Thakker, D., Osman, T., and Lakin, P. (2009). *Gate jape grammar tutorial*. Nottingham Trent University, UK; Phil Lakin, UK, Version, 1.
[6] Laclavík, M. and Šeleng, M. (2012). *Vyhľadávanie informácií*. Nakladateľstvo STU, Bratislava. 141 p.

Formal (Morpho)Syntax Properties of Reflexive Particles *se*, *si* as Free Morphemes in Contemporary Czech

Vladimír Petkevič

Faculty of Arts, Charles University in Prague, Czech Republic

Abstract. On the basis of the SYN2010 corpus of contemporary Czech comprising 100 million tokens, the paper investigates main formal (morpho)syntactic properties of Czech reflexive particles *se* and *si*. It is mainly the relation of these particles to their "base words" (which can be verbs, deverbal adjectives or deverbal nouns) that is discussed here. This relation primarily concerns different mutual word order positions of a base word and "its" particle with respect to the part-of-speech of the base word, context-free character of word order positions of the baseword-reflexive pairs, and also kinds of haplology of reflexives. The specification of properties of reflexives being described here helps (i) to improve rule-based automatic morphological disambiguation of contemporary Czech, and (ii) to improve automatic parsing of Czech. The paper can, moreover, be regarded as a contribution to the theory of syntax of Czech which has not paid due attention to the formal behaviour of reflexives yet.

1 Introduction

Czech reflexive particles *se* and *si* belong to the most frequent word forms in Czech. In the SYN2010 corpus of contemporary Czech comprising 100 million word forms, there are:

- 2348948 occurrences of reflexive *se* (the 2nd most frequent word form in the SYN2010 corpus, ca 2.34% of all tokens)
- 594709 occurrences of reflexive *si* (the 12th most frequent word form in the SYN2010 corpus, ca 0.6% of all tokens)

in various functions. In this study I will specify formal conditions concerning their relation to their "base words" (= verbs, deverbal adjectives, deverbal nouns) without which they cannot exist in a grammatical clause. By associating reflexives with their base words, their functions are partly determined, too. The conditions, if implemented as formal rules, help to improve automatic morphological disambiguation and syntactic analysis of Czech.

The paper deals with the following topics:

(a) part-of-speech ambiguity of the word forms *se* and *si* (Sect. 2);
(b) formal (morpho)syntactic properties of reflexive particles *se* and *si* (Sect. 3).

2 Part-of-speech Ambiguity of the Word Forms *se* and *si*

Both word forms, i.e. *se* and *si*, are part-of-speech (POS) ambiguous, and first of all, it is necessary to identify the forms *se* and *si* as reflexives (reflexive particles/pronouns) rather than as a preposition or a verbal form. As a rule-based POS disambiguation of the form *se* was thoroughly studied by Oliva in his seminal article [2] (see also [3]), I will present

only basic facts accompanied by examples; and in Section 3 I will study other (morpho)syntactic properties of these reflexives, already assuming that the reflexive forms *se* and *si* are identified as such by the Czech tagging system in the most reliable way as reflected in the tagging of the SYN2010 corpus.

2.1 Word Form *se*

The word form *se* is:

- either a reflexive particle/pronoun (referred to as *reflexive* in the sequel) that is polyfunctional (it expresses a direct reflexive object, it is a particle of truly reflexive verbs/adjective/nouns, it also co-forms non-periphrastic passive etc.)
- or the vocalized form of the preposition *s* (E. *with*) requiring the instrumental or genitive case; the vocalized form stands immediately in front of a word form beginning with a sibilant (*s*, *z*) or with a specific type of consonant cluster.

Examples:

(1) *Stali jsme* ***se***(Refl) *služebníky Božími a přestali jsme být služebníky lidí.*
E. lit. *We became servants of God and ceased to be servants of people.*

Here *se* is the reflexive particle obligatorily associated with the truly reflexive verb form *stali* (E. *became*).

(2) *To my, když jdeme* ***se***(Prep) *starým na večeři, tak to vypadá trochu jinak.*
E. lit. *We, when we go* ***with*** *the old man to have a dinner, it looks slightly different.*

Here *se* is the vocalized form of the preposition *s* requiring the following word form *starým* (E. *old man*) to be in the instrumental case.

Vocalization criteria are specified in [1], a detailed description of POS disambiguation rules of the word form *se* is presented in [2]. These and also some other rules (their number is up to 30!) implemented in the system of a rule-based morphological disambiguation of Czech (cf. [3]) have been used for an automatic morphological disambiguation of corpora of contemporary Czech of the SYN series (SYN2005 and subsequent ones) with 99.7% accuracy. In the sequel, I will assume that *se* has already been correctly identified as a reflexive with the given accuracy. Moreover, the frequency of *se* as the reflexive and as the preposition radically differs in the SYN2010 corpus:

- frequency of *se* as a reflexive: 94–95%;
- frequency of *se* as a preposition: 5–6%.

2.2 Word Form *si*

The word form *si* is:

- either reflexive particle/pronoun (referred to as *reflexive* in the sequel) expressing primarily the **medium voice** of the corresponding base word, i.e. a subject performs an action **for** himself/herself

- or a verb form *jsi* (= 2nd pers. sg. present tense of the verb *být* (E. *be*)) written in a non-standard orthography.

Examples:

(3) *Dovedete **si***(Refl) *představit, jak příjemně je osamocenému člověku*?
E. lit. *Can you **for yourself** imagine, how pleasant it is for a lonely person?*

(4) *Co **si***(Refl | Verb) *četl?*
E. lit. *What **for himself** did he read?* or also
E. *What did you read?*

Sentence (4) is ambiguous, having the following interpretations:

(4a) *Co **si***(Refl) *četl?*
E. *What did he read (**for himself**)?*

Here the subject reads "for himself", i.e. *si* is used in the medium meaning, whereas in:

(4b) *Co **si***(Verb) *četl?* (= *Co **jsi***(Verb) *četl?*)
E. *What did you read?*

si is to be interpreted as a non-standard orthography of the word *jsi*, i.e. the 2nd pers. sg. present tense of the verb *být* (*E. to be*), conforming – with the past participle *četl* (E. *read*) – the preterite tense of the 2nd pers. sg.

Unlike the rule-based POS disambiguation of the form *se* (analyzed in [2]), the rule-based POS disambiguation of the form *si* has not yet been described. In general, it is extremely difficult to correctly disambiguate the word form *si* due to the kind of ambiguity shown above. There are, however, many contexts in which it is possible to correctly disambiguate occurrences of *si* as a reflexive, e.g. in sentence (5) and (6)[1]:

(5) *Kuřáci samoty na slamníku zítřka, cpaném pilinami, **si***(Refl) *rozestlali*(**Plural**).
E. lit. *The smokers of loneliness on a straw mattress of tomorrow, stuffed with sawdust, **for themselves** made their bed.*

(6) *Jak **si***(Refl) *přeju, aby to už skončilo!*
E. lit. *How much I **for myself** wish, so that it already ended!*

In (5) it is impossible to interpret the word form *si* as a singular verbal auxiliary form *jsi* because of the presence of the plural past participle *rozestlali* (E. *they made their beds*): in Czech the auxiliary and the past participle must agree in number. In (6), *si* cannot be interpreted as a form of the verb *být* (E. *to be*), since there is another verbal form in the present tense (*přeju*, E. *I wish*) in the same clause.

[1] The disambiguation team (Tomáš Jelínek, Pavel Květoň, Karel Oliva, Vladimír Petkevič) at the Institute of Theoretical and Computational Linguistics, Faculty of Arts, Charles University, implemented the disambiguation rules that correctly identify the occurrences of the word form *se* and *si* in the corpora of Czech with more than 99.5% accuracy.

Moreover, in written corpora there is an insignificant number of *si* having the function of the auxiliary *jsi*; so I assume that the word form *si* has been disambiguated as a reflexive with almost 100% accuracy.

Thus in the sequel, I assume that the occurrences of the forms *se* and *si* have already been disambiguated as reflexives and now I will investigate their formal (morpho)syntactic properties.

3 Formal (Morpho)Syntactic Properties of Reflexives *se* and *si*

At the beginning it necessary to say that some of the properties specified below have already been used for automatic POS rule-based disambiguation of the forms *se* and *si* briefly discussed above (cf. [2]).

3.1 Relation of the Free Morpheme *se* and *si* to its Base Word

Reflexives *se* and *si* are free morphemes, but in a well-formed sentence they cannot stand alone since every occurrence of the reflexive is always bound to its **base word** in a clause. These base words can be: non-passive verbal forms, deverbal adjectives or deverbal nouns. If an occurrence of the reflexive *se* or *si* has no base word counterpart in a clause, the sentence is not (morpho)syntactically well-formed, most often due to the aposiopesis:

(7) *Já* ***se*** *na to...*
E. lit. *I will* ***myself*** *on it...*

and the hearer will supply the missing word(s) based on the situational context.

The basic problem consists in the automatic identification of an appropriate base word for a given occurrence of the reflexive *se*/*si*. For instance, in sentence

(8) *Myslím, že* ***se*****$_1$** *dobře* ***vyzná*****$_1$** *v anatomii.*
E. *I think that* ***s/he knows a lot*** *about anatomy.*

the human hearer will have no problems with the identification of the verbal base word *vyzná* (E. *s/he knows a lot*) as associated with the reflexive *se* since both words form a multiword unit with an unambiguous meaning. However, for any automatic analysis it is a big problem: an occurrence of the reflexive can be associated not only with a verbal form, but also with an appropriate deverbal adjective or deverbal noun in the same sentence/clause as its base word. The problem is even more complicated if there are more base word candidates for a given occurrence of *se* or *si* in a sentence as in:

(9) ... *Říká jedna z brněnských hereček,* ***přející*****$_1$** ***si*****$_1$** *zůstat v anonymitě.*
E. lit. ... *Says one of the Brno actresses,* ***wishing*****$_1$** ***for herself*****$_1$** *to remain in anonymity.*

In (9), reflexive *si* is associated with the adjective *přející* (E. *wishing*) rather than with the verbal form *zůstat* (E. *remain*) or with the noun *anonymitě* (E. *anonymity*) or even with the verbal form *říká* (E. *says*).

Furthermore, the situation is complicated if there are more occurrences of reflexives in a sentence, as in:

(10) ***Přál*$_1$ *si*$_1$ *podívat*$_2$ *se*$_2$** *jí do krku.*
E. lit. *He* ***wished*$_1$ *(for himself*$_1$*)*** *to* ***look*$_2$ *(himself*$_2$*)*** *her into the throat.*

(11) *Když* ***si*$_1$ *třesoucími*$_2$ *se*$_2$** *prsty* ***uvazoval*$_1$** *šátek kolem krku...*
E. lit. *When he with* **Refl**$_1$ ***trembling*$_2$** **Refl**$_2$ *fingers* ***bound*$_1$** *a scarf round his neck...*

The indices indicate the appropriate association. Specially, in (11) the deverbal adjective *třesoucími* (E. *trembling*) is a truly reflexive adjective immediately followed by its *se*.

First, generally valid statements will be specified and then I will analyze the properties of individual parts of speech of the base words associated with their reflexives.

Statement 1. A base word as well as its reflexive belong to the same clause.

Statement 2. Every base word can be associated with one occurrence of a reflexive at most.[2]

Both statements are obvious and they can be advantageously exploited for disambiguation and parsing if a clause consists of discontinuous parts where an occurrence of *se/si* and an associated base word belong to different parts of a non-contiguous clause.

Statement 3. Every occurrence of a reflexive can be (morpho)syntactically associated with two base words at most.

In standard cases, there is only one base word with which an occurrence of reflexive is associated; in the case of haplology (see Section 3.2 below) such an occurrence is associated with two words in a clause.

Statement 4. Base words and their associated reflexives conceived of as bracketed pairs form a context-free (= properly nested) structure.

This statement has the following meaning: With respect to the word order, every two pairs are either disjoint or one pair is fully embedded in another one (pairs are said to be properly nested). Thus, the following structures are well-formed, i.e. context-free (the indices identify the pairs):

(12a) ... *se*/*si*$_1$ ... baseword$_1$ ... *se*/*si*$_2$ ... baseword$_2$ ...
(12b) ... *se*/*si*$_1$ ... baseword$_1$ ... baseword$_2$ ... *se*/*si*$_2$ ...
(12c) ... baseword$_1$ ... *se*/*si*$_1$ ... *se*/*si*$_2$ ... baseword$_2$ ...
(12d) ... baseword$_1$ ... *se*/*si*$_1$ ... baseword$_2$ ... *se*/*si*$_2$

(12e) *se*/*si*$_1$ ... *se*/*si*$_2$ ... baseword$_2$ ... baseword$_1$
(12f) *se*/*si*$_1$ ... baseword$_2$ ... *se*/*si*$_2$ ... baseword[3]

[2] Cf. also [2, p. 310]
[3] The structures: (12g) baseword$_1$ ... *se*/*si*$_2$ ... baseword$_2$ ... *se*/*si*$_1$ and
(12h) baseword$_1$ … baseword$_2$ ... *se*/*si*$_2$ ... *se*/*si*$_1$ are impossible in Czech as well, see Sect. 3.1.1.

whereas the following "clinched" (context-sensitive) schemes are impossible in Czech sentences:

(13a) ... *se/si*$_1$ ... baseword$_2$ ... baseword$_1$ ... *se/si*$_2$ ...
(13b) ... *se/si*$_1$ ... *se/si*$_2$ ... baseword$_1$ ... baseword$_2$ ...
(13c) ... baseword$_1$ ... *se/si*$_2$ ... *se/si*$_1$ ... baseword$_2$ ...
(13d) ... baseword$_1$ ... baseword$_2$ ... *se/si*$_1$ ... *se/si*$_2$ ...

For instance, sentences (10) and (11) above comply with the following scheme, where the pairwise nesting is well-formed, respectively:

(10') ... baseword$_1$ *si*$_1$ baseword$_2$ *se*$_2$
(11') ... *si*$_1$ baseword$_2$ *se*$_2$ ... baseword$_1$

Intuitively it seems that the property of sentence structure dealt with in Statement 4 holds also for other languages that contain reflexives as free morphemes.

In a way, Statement 4 is related to Statement 5 below which seems to capture a universal feature valid for all languages of the world that have compound sentences with discontinuous clauses.

Statement 5. The parts of a discontinuous clause in a compound sentence form a context-free (= properly nested) structure.

This statement claims that parts of different clauses are – with respect to the word order – properly embedded, i.e. if a clause C1 is split into two or more parts and at least one part of another clause C2 is placed between some two parts C1a and C1b of C1, then **all the parts** of C2 are placed between C1a and C1b. For instance, the following structure is correctly nested:

(14) *Main1*$_a$ *Sub1*$_a$ *Sub2* *Sub1*$_b$ *Main1*$_b$

Here *Main1*$_a$ and *Main1*$_b$ denote two discontinuous parts of the main clause *Main1*, and both parts of the *Sub1* clause, i.e. *Sub1*$_a$ and *Sub1*$_b$, are embedded in the *Main1* clause. Moreover, there is another clause, *Sub2*, entirely embedded in *Sub1* and thus splitting *Sub1* into two discontinuous parts *Sub1*$_a$ and *Sub1*$_b$.

Let us now study the mutual position of reflexives *se/si* and their base words according to the parts of speech of the base word.

3.1.1 Verb as a Base Word

A reflexive *se/si* can either follow, or precede its base verb (the verb form cannot be that of a passive participle). The word order position of *se/si* is primarily determined by the fact that both *se* and *si* are enclitics.

A. A verb form preceding "its" *se/si* precedes it:
(a) either immediately,
(b) or there can stand between the base word and its *se/si* the following words only:

- subordinate conjunction *-li* (E. *if*)
- conditional particles of the 1st, 2nd and 3rd person sg. and pl. (*bych*, *bysem*, *bys*, *bychom*, *bysme*, *byste*, *by*; E. *would*)
- the following present tense forms of the verb *být* of the 1st and 2nd per. sg. and pl. (*jsem*, *sem*, *jsi*, *si*, *jsme*, *sme*, *jste*, *ste*; E. *I_am*, *you_are*, *we_are*, *you_are*)
- particle *už*, *již* (E. *already*), *prý* (E. *allegedly*).

B. If a verb follows "its" *se/si*, it can stand in an arbitrary distance from the reflexive in the same clause, even if the clause consists of two discontinous parts where the reflexive *se/si* is placed in the left-hand part and the base word in the right-hand part of the clause. It can even be stated – as our corpora show – that the closer the reflexive *se/si* is to the verb, the more numerous such structures are. There is even a certain regularity in the decrease in frequency with an increasing distance between the two words as shown in Table 1.

Distance in words	# of occurrences
0	702,451
1	299,612
2	187,238
3	97,026
4	44,141
5	20,194
6	9,585
7	4,406
8	2,084

Table 1. Distance between an occurrence of the preceding reflexive *se/si* and its baseword

It is clear that with the wider distance, the number of occurrences sharply decreases (roughly with the power of 2).[4]

As the reflexive *se/si* is a clitic, it stands (in both word order cases A and B) in a clitic cluster which occupies the second syntactic position in a clause. Inside the clitic cluster, reflexive *se* takes up the following word order position:

- conjunction *-li* (E. *if*)
- conditional particles of the 1st, 2nd and 3nd person sg. and pl. (*bych*, *bysem*, *bys*, *bychom*, *bysme*, *byste*, *by*; E. *would*)
- the following present tense forms of the verb *být* of the 1st and 2nd per. sg. and pl. (*jsem*, *sem*, *jsi*, *si*, *jsme*, *sme*, *jste*, *ste*; E. *I_am*, *you_are*, *we_are*, *you_are*)
- **reflexive *se***
- dative pronominal clitics
- accusative pronominal clitics
- other clitics.

[4] I did not study compound structures in which the reflexive *se/si* and its base word are separated by an embedded clause. This case is extremely rare, but fully grammatical.

Moreover, short adverbs or particles *už*, *již* (E. *already*), *prý* (E. *allegedly*) can stand anywhere after the conjunction *-li* in a cluster:

(15) *Nevím také, nebylo-**li by se**$_1$ v tomto případě možno **dovolávat**$_1$ fondu...*
E. lit. *I also do not know, whether it were not **Refl**$_1$ in this case possible to **appeal**$_1$ to the fund...*

Here the clitics are ordered in accordance with the order specified above.

3.1.2 Deverbal Adjective as a Base Word

The reflexive *se/si* can be associated with an adjective as its base word only if the adjective is a deverbal one derived from a present transgressive (its lemma ending in *-ící*, *-oucí*), or from a past transgressive (its lemma ending in *-vší*). The reflexive can either follow, or precede "its" adjective.

A. An adjective Adj preceding "its" *se/si* precedes it:
- either immediately
- or there can stand between the base word and its *se/si* the following words only: particles *už*, *již* (E. *already*), *prý* (E. *allegedly*).

In this word order, Adj heading the adjectival group occupies unmarkedly the first position in the adjectival group and reflexive *se/si* occupies – in the trivial clitic cluster – the second position in this group (similarly as in a clause headed by a finite verb, cf. Sect. 3.1.1 above).

B. If an adjective Adj follows "its" reflexive *se/si*, it stands:
- either immediately behind *se/si*
- or there stand, between *se/si* and Adj, only the words belonging to the adjectival group headed by Adj, i.e. mainly adverbs or (very short) adverbials. Thus, adjectives tend to stand much closer to their reflexives than verbs. We see that in this word order, reflexive *se/si* occupies the second syntactic position in the corresponding adjectival group headed by Adj. This means that the reflexive must be preceded by one syntactic element, usually an adverbial (or adverbial group), thus taking up the first position in the adjectival group.

Example:

(16a) *Všechny ty rozesmáté tváře, radostně **si**$_1$ **hrající**$_1$ děti, přátelští dospělí, všechno mu to připadalo přehnané.*
E. lit. *All the laughing faces, cheerfully **Refl**$_1$ **playing**$_1$ kids, friendly adults, all of it* seemed to him as exaggerated.

(16b)* *Všechny ty rozesmáté tváře, **si**$_1$ radostně **hrající**$_1$ děti, přátelští dospělí, všechno mu to připadalo přehnané.*
E. lit. *All the laughing faces, **Refl**$_1$ cheerfully **playing**$_1$ kids, friendly adults, all of it seemed to him as exaggerated..*

The sentences (16a) and (16b) differ only in the mutual word order of the words *si*$_1$ and *radostně* (E. *cheerfully*). In the grammatical sentence (16a), the reflexive *si*$_1$ correctly takes up the second position in the adjectival group *radostně si hrající* (E. *cheerfully playing*), whereas in the ungrammatical sentence (16b) the reflexive *si*$_1$ occupies the inadmissible first position in the adjectival phrase *si radostně hrající* (E. *cheerfully playing*). Thus we see that the Wackernagel's rule about the second syntactic position of clitics holds not only for a clause headed by a finite verb, but also for an adjectival group headed by an adjective.

The reflexive seems to be always adjacent to the deverbal adjective, as claimed in [2, p. 308]. However, the following grammatically well-formed structures can also be encountered in the SYN2010 (or 1 billion SYN) corpus:

(17) *Uviděla tři věci: jachtu,* ***právě se*$_1$ *majestátně obracející*$_1$** *v zátočině...*
E. lit. *She saw three things: a yacht,* ***just Refl*$_1$ *magnificently turning*$_1$** *in the bay...*

(18) *Ostatní měli snad podle složitých a* ***stále se*$_1$ *samovolně měnících*$_1$** *pravidel hry sraz v nějakém jiném podniku.*
E. lit. *The others had probably according to complicated and* ***constantly Refl*$_1$ *arbitrarily changing*$_1$** *rules of the game a meeting in another establishment.*

(19) *... samozřejmě každá jiná a i jiného,* ***mně se*$_1$ *naprosto nehodícího*$_1$** *rozměru.*
E. lit. *... certainly every other and also another,* ***to me Refl*$_1$ *totally inappropriate*$_1$** *dimension.*

(20) *Čeští politici, a hlavně ti nastupující (****či se*$_1$ *o to snažící*$_1$****), si pořád jen těžce zvykají na to...*
E. lit. *Czech politicians, and mainly those starting (****or Refl*$_1$ *at it attempting*$_1$****), si pořád jen těžce zvykají na to...*

In sentences (17)–(20) the adjectival structures written in bold are grammatically well-formed, *se* taking up the second position in them but being separated from the adjective by an adverb (*majestátně* (E. *magnificently*) and *samovolně* (E. *arbitrarily*) in (17) and (18), respectively), or particle (*naprosto* (E. *totally*) in (19)), or even a prepositional group (*o to* (E. *at it*) in (20)).

3.1.3 Deverbal Noun as a Base Word

A deverbal noun is most closely associated with its reflexive *se/si*: *se/si* always immediately follows its deverbal noun, as, for instance, in:

(21) *Všimli si, že má hlavní snaha se soustřeďuje na* ***osvojení*$_1$ *si*$_1$** *určitých prvků.*
E. lit. *They noticed that my main endeavour concentrates on* ***mastering*$_1$ *for oneself*$_1$** *some elements.*

(22) *Taková úmluva není rozhodnutí řízené rozumem, nýbrž* ***přizpůsobení*$_1$ *se*$_1$** *reakcím druhého na vlastní chování.*
E. lit. *Such an agreement is not a decision governed by reason, but an* ***accommodation*$_1$ *of oneself*$_1$** *to the reactions of the other to one's own behaviour.*

Let us summarize: The reflexive *se/si* stands in the syntactically second position in a clause / adjectival group / deverbal noun group. Moreover, the "more nominal" the base word, the closer "its" *se/si* stands.

3.2 Haplology

In a grammatically well-formed sentence no base word can be associated with more than just one reflexive (cf. Statement 2 above), but an occurrence of a reflexive can be associated with two base word occurrences (cf. Statement 3 above). In this case, two adjacent occurrences of reflexives *se/si* having their different respective base words merge into one and the resulting reflexive *se* or *si* "serves" two base words. There are four theoretically possible combinations:

(i) *se* and *se* merging into *se*,
(ii) *se* and *si* merging into *se*,
(iii) *se* and *si* merging into *si*,
(iv) *si* and *si* merging into *si*.

This phenomenon of haplology may complicate morphological disambiguation, but mainly parsing.

It is typically *se* and *se* that are merged:

(23) *Snažil*$_1$ ***se***$_1$ *usmát*$_2$ *(se*$_2$*).*
E. lit. *He tried*$_1$ ***Refl***$_1$ *to smile*$_2$ (*Refl*$_2$)
E. *He tried to smile.*

but also *si* and *si*:

(24) ***Troufám***$_1$ ***si***$_1$ ***tipnout***$_2$ (***si***$_2$)*, že se ČSSD rozhodne tolerovat vládu.*
E. lit. *I* ***dare***$_1$ ***Refl***$_1$ *to* ***tip***$_2$ (***Refl***$_2$) *that ČSSD will decide to tolerate the government.*
E. *I dare to tip that ČSSD will decide to tolerate the government.*

and even *se* and *si* can merge. In this "hybrid" case, the unmarked (at least less frequent) *si* usually overrides *se* as in:

(25) *Ve 4 ráno jsem* ***si***$_{12}$ ***snažil***$_2$ ***vsugerovat***$_1$*, že sedím v sauně.*
E. lit. *At 4 in the morning* ***I Refl***$_{12}$ ***tried***$_2$ ***to put–into–my–mind***$_1$*, that I sit in the sauna.*

(26) *Tesař vyznávající pohodlí víc než společenskou upjatost a* ***nerozpakující***$_2$ ***si***$_{12}$ ***sundat***$_1$ *košili...*
E. *The carpenter adhering to comfort more than social prudery and* ***not hesitating***$_2$ ***Refl***$_{12}$ ***to take***$_1$ ***off*** *his shirt...*

In (25), the base word for the reflexive *si* is the verbal form *vsugerovat* (E. *put into one's mind*), the obligatory reflexive *se* for the truly reflexive verb *snažit* is overridden here by the "marked" form *si*. In (26), the base word for the reflexive *si* is the form

sundat and the obligatory form *se* associated with the truly reflexive verb *rozpakovat* was overridden by the "marked" reflexive *si*. However, the reversed overriding is also possible: *si* may be overridden by *se* as in:

(27) *Dovey* ***se*****$_1$** *ho* ***netroufala*****$_{12}$** ***zeptat*****$_1$***, kolik je mu let.*
E. lit. *Dovey* ***Refl*****$_1$** *him did not* ***dare*****$_{12}$** ***to ask*****$_1$***, how old he is.*
E. *Dovey did not dare to ask him, how old he is.*

In (27), the base word for *se* is the *se*-reflexive verb *zeptat* (E. *ask*); the obligatory form *si* for the verbal form *netroufala*, which is a *si*-reflexive verb, has been overridden by the reflexive *se* associated with *zeptat*.

Typically, an occurrence of the haplologized reflexive *se/si* is associated with the words belonging to the same part of speech, but we see in (26) that the haplologized reflexive *si* (merged from *se* and *si*) can be associated with two base words belonging to different parts of speech: a verb and an adjective. Another example of this type follows:

(28) ***Snažil*****$_1$** ***se*****$_{12}$** ***smějící*****$_2$** *dívku odvézt.*
E. lit. ***He tried*****$_1$** ***Refl*****$_{12}$** ***the laughing*****$_2$** *girl to carry away.*

Here *snažil* (E. *tried*) is a form of the truly reflexive verb *snažit* and the form *smějící* (E. *laughing*) is a form of the truly reflexive deverbal adjective, so here *se* is associated with two base words belonging to different parts of speech.

4 Conclusion

In this article, the main (morpho)syntactic properties of reflexives *se* and *si* in contemporary Czech were studied, especially their relation to their base words. It is primarily shown that – notwithstanding the conditions and criteria formally described in the paper – an automatic association of reflexives with their base words is a generally difficult task, especially for automatic parsing, and it should to be paid due attention.

References

[1] Kučera, K. (1984). K vokalizaci neslabičných předložek v současné češtině. *Naše řeč*, 67:225–237.

[2] Oliva, K. (2003). Linguistics-based PoS-tagging of Czech: disambiguation of *se* as a test. In Kosta, P., Błaszczak, J., Frasek, J., Geist, L., and Żygis, M., editors, *Investigations into Formal Slavic Linguistics (FDSL 4). Part 1. Contributions of the Fourth European Conference on Formal Description of Slavic Languages held at Potsdam University, November 28–30, 2001*, pages 299–314, Peter Lang, Frankfurt am Main.

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

Introduction to Online Learning

Katarína Pišútová

Centre for Information Technology, Comenius University, Bratislava, Slovakia

Abstract. One of the goals of this text is to provide introductory information about online learning. There are basic historical facts, different definitions, list of differences between face-to-face and online learning. The second part of this article focuses on development of online learning in Slovakia, issues that are specific to Slovak situation and suggestions on how to overcome some of these issues.

1 Introduction

The influence of the Internet on education has been the center of many professional debates in recent years. Growth in the number of courses, virtual institutions, and articles dealing with online learning is increasing at a rapid pace. When describing this development as "madness" and a "gold-rush", Boshier [7] relates to getting 71,600 hits from the Google search engine for the words "Virtual University" in March, 2001. When I repeated the same search in May 2013, the number of hits had grown to an overwhelming 200,000,000.

2 What is Online Learning

This chapter provides a short overview on history, characteristics, and definitions of online learning. It summarizes reasons for the growth of online learning and describes typical beliefs and perceptions spread among educators.

2.1 History

Online learning can be traced as a result of the convergence of two historic educational trends: distance education and use of technology in classrooms [3].

The history of using technology in the classroom is quite long. It started centuries ago with chalk and a blackboard and evolved rapidly in the twentieth century with more and more varying technologies such as overhead projectors, slide shows, film and videotapes, and computers. All these technologies served to help students to see better examples and illustrations - to be able to understand the explanations of their teacher better. The Internet came into education as another tool in a row of classroom technologies. In many classrooms computers, Internet and web-based materials are used as teaching aids in exactly the same way as any of the previously used technologies. And the production and publication of web-based materials can be very cheap and easy to use. The use of the computer alone has prompted many changes in the very approach of teaching, but now with prepared instructional programs using compact disc or the Internet, the teacher does not even need to remain present for instruction - the control over learning is shifting from the teachers to the learners.

Distance education has a long history as well. In a modern sense, distance education began in the nineteenth century when the first commercial correspondence colleges were established following the development of reliable and speedy postal service. People living in remote areas got the occasion to study at universities in London or other cities. Long before the Second World War, universities in North America were offering extension services to remote adults [2]. However, these courses were famous for high drop out rates and poor results on examinations. Then, in 1969, the British Open University was established and it has been marked as turning point in the development of distance education. This was the first institution designed solely and specifically for distance education and it started to combine and integrate all accessible technologies (print, broadcasting and face-to-face instruction). The institutional structure of the Open University was developed to be different than that of conventional universities and it succeeded in providing cost-effective courses of high quality. The structural model of the Open University was successfully copied in many countries around the world. As well, with the addition of distance learning options for their students, the Open University structure has influenced many institutions more oriented to "traditional" teaching and learning.

The importance of the British Open University is that it created a new way of teaching. For the first time there were teams of instructors, instructional designers and media specialists developing high quality teaching materials ahead of time with built-in interaction mechanisms and for repeated use with students taught by part-time tutors.

With the introduction of the Internet, student-teacher interaction gains more emphasis and can be developed in various ways. Suddenly, communication is required not just from teacher to student, but as well from student to teacher and also among individual students. It is possible now, for students to interact with each other without being at the same place at the same time.

Teachers soon realized that electronic materials developed for on-campus students could be used for distance students and vice-versa. So is it still a distance course? And does it matter? With ever-improving Internet technologies – and access – classroom teaching and distance learning are starting to converge and the interaction and control of the learners are rapidly becoming more important than the distance.

2.2 How to Define Online Learning

Online learning cannot be defined by distance between students and a teacher or by only using computers and the Internet. The main features and changes that online learning has brought into education are in interaction, sharing, cooperation and new control of the students over their learning process [11].

As in many new fields, the terminology around online learning is far from universal. Terms like Web-based learning, e-learning, networked learning, distributed learning, flexible learning, distance learning, computer-assisted learning, etc., are being used simultaneously and interchangeably by different authors and in different countries. The recent trend of mixing different technologies and supplementing face-to-face courses

with the Internet, or Internet-based courses with face-to-face teaching, according to students' needs and circumstances makes the terminology even more confusing[1].

For instance, Harasim [12] defines what she calls learning networks through cooperation: "Learning networks are groups of people who use computer mediated networks to learn together at the same time, place, and at the pace that best suits them and is appropriate to the task".

The Institute for Academic Technology at the University of North Carolina has provided in 1995 a more detailed definition of what they call distributed learning (as quoted by [3]): "A distributed learning environment is a learner centered approach to education, which integrates a number of technologies to enable opportunities for activities and interaction in both asynchronous and real-time modes. The model is based on blending a choice of appropriate technologies with aspects of campus-based delivery, open learning systems, and distance education. The approach gives instructors the flexibility to customize learning environments to meet the needs of diverse student populations, while providing both high-quality and cost-effective learning".

For the purpose of this text I will stick to the term "online learning" in a wider sense which includes fully online as well as appropriately blended courses.

3 What is Different?

This chapter summarizes changes brought into education by online learning. It provides descriptions of changing roles of teacher and student, changing of teacher-student-content relationships, and summarizes the ongoing debate over the quality of online learning.

3.1 Changing Roles for Teacher and Learner

As mentioned before, online learning brings some new options and features into the teaching and learning process.

Harasim [12], quoting a survey on 240 teachers and learners, summarizes positive changes as follows:

- The role of the teacher changes to that of the facilitator and mentor
- Students become active participants; discussions become deeper and more detailed
- Access to resources is expanded significantly
- Learners become more independent
- Access to teachers becomes equal and direct
- Interaction among teachers are increased significantly
- Education becomes learner centered; learning becomes self-paced
- Learning opportunities for all students are more equal; learner-learner group interactions are significantly increased
- Personal communication among participants is increased

[1] One of the Czech visionaries of online learning likes to say: "Both whisky and education are best when blended" [6].

- Teaching and learning is collaborative
- There is more time to reflect on ideas; students can explore on the networks; exchange of thoughts and ideas is expanded; the classroom becomes global
- The teacher-learner hierarchy is broken down. Teachers become learners and learners become teachers.

The negative changes observed in the same survey included:

- Longer preparation time for teachers
- Matching educational background with expectations. Each individual bases his/her expectations for online courses on previous educational experience. Novelty and the differences of studying online can bring positive or negative surprises. Many students expect that studying online is much easier than an online course. Relevant and detailed information has to be provided for the students at the beginning of the course
- Quality of learning. The online learning experience can be very different from the experience of face-to-face learning. Some students note improvements in the learning experience; some feel the experience is of a lower quality
- Flexibility of time schedule. One of the main features of distance and online learning is the fact that students can decide when and where to study. Enhanced flexibility and independence goes together with responsibility and self-discipline in competently fulfilling all the tasks necessary for the course. Encouragement and support from the tutor can play a significant role here
- Saving time by studying online. By studying online the learner can save the time otherwise spent by traveling to school. But as well, communicating using technology can take more time than face-to-face communication, so the time saving may not be as significant for some as others
- Miscommunication. Since most of the communication between learner and tutor is made via technology, misunderstandings are more likely to happen. A tutor should be aware of these risks and spend time and effort to lessen frustrations arising from miscommunication
- Loneliness, isolation. Communicating with tutors and peers using only technology can cause strong feelings of isolation and loneliness. Again, tutor's feedback, communication and support can help
- Lack of technical experience. Students who do not have previous knowledge and experience with online technology can view it as a very strong handicap. Every online course should have fast and reliable instructional and technical support – an easily reachable person ready to help to resolve such problems and to overcome frustrations.

4 Online Learning in Slovak Academia

Due to the fact that universities can charge only limited fees for their part-time courses, there is a lack of motivation to invest into moving these programs online to make them

accessible to a larger volume of students. University administrators (being overwhelmed by transition and lack of funding for basic needs) do not invest into building support structures for teachers or online students.

From 1995–1999, there was a European Union funded program called "Multi Country Cooperation in Distance Education", which was trying to promote cooperation in distance education development between EU member states and accession countries [17]. This project's goal was to build completely new distance education programs in cooperation with EU partners, based mostly on the UK Open University model. In Slovakia, the interest among universities in this project was not very high. It was felt that Universities had enough problems already without changing their form of delivery. Only a few universities, all technically oriented, applied for participation in this project. Within the Multi Country Project five distance learning centres were created at four different Slovak technical universities – Slovak Technical University in Bratislava, Technical University in Zvolen, University of Žilina and Technical University Košice.

These centres were provided with technical equipment, and employees were trained in designing, developing and teaching distance learning courses. From 1995–1999, Slovak centres participated in creating and conducting a number of distance learning courses. However, these courses were all in English and hence did not attract many Slovak students, and only three of them had any online component. In 1999, when the Multi Country Project had officially finished, a network of distance education centres remained in place, each as a part of their host university's infrastructure. But universities did not have funding and also not much interest to move into developing their own online courses.

On the other hand, this program created the first population of university employees trained in online instructional design who later supported the first Slovak online courses and training of people at other universities, non-governmental organizations and business companies. Since then, online courses at Slovak universities and non profit organizations have been developing based on the efforts of enthusiasts or funding opportunities either from independent donors - such as the Open Society Foundation, or EU funds e.g. the European Social Fund, ever since Slovakia became an EU member in 2004.

The problem with outside grant funding is that very often the university decides to create a particular course with a particular focus based on the grant call, rather than the mission of the institution. After the grant money runs out, the university often closes the course because they have no real interest in the focus area and hence no motivation in finding funds to continue teaching it. In this way, many courses based on EU and other grant money, simply cease to exist after the end of the grant.

In recent years, commercial companies and non-governmental institutions in Slovakia began offering some online courses. Online courses are mostly offered by technically oriented business companies and in most cases are automated lectures with self-testing, where students work individually with occasional contact with the instructor, and absolutely no contact with co-learners. A wide spread online learning commercial program is the international Cisco Networking Academies Program (CNAP), which again does not use collaboration (CNAP website, accessed on August 10, 2011).

Online learning initiatives of non-governmental institutions are even more dependent on different grants and external funding than universities, so survival of their online projects beyond the initial grant to create them is also quite rare.

4.1 Support Structures for Teachers in Online Learning at Slovak Universities

A number of universities that tried to set policies and establish systems for online learning seem to focus solely on technical aspects. The University of Žilina offers the Moodle open source environment for all their teachers. They created a department for running and maintaining servers, and provide technical support for teachers willing to try to transfer their courses. The support is also provided for students taking online courses [1]. However, the university does not provide any incentives for teachers to transfer their courses into online form, nor does it provide instructional design support. A similar situation exists at the Slovak Technical University, which also uses Moodle as a Learning Management System [13], and the Technical University of Košice, which developed a Learning Management System of their own [15].

Distance learning centres created within the Multi Country Project were mostly asked to "earn" their running costs, so designers at these centres focus on developing and creating commercially profitable courses in areas like law and management (mostly as automated self-test courses), and do not have the time or the motivation to provide instructional design help or services to their colleagues at the University.

One exception is the private College of Management, which was created by the City University of Seattle in the U.S. The College of Management offers online and blended courses, and provides teachers with training and support to the same standards as City University of Seattle. The College of Management also seems to be the only higher education institution in Slovakia that emphasizes use of collaborative activities in their coursework [14].

5 Where to Start with Online Learning

Canadian professor Tony Bates published recently a series of blog articles called Nine Steps to Quality Online Learning [5]. These steps present practical instructions for teachers who are considering starting to create their first online course. This section contains a summary of the 9 steps with some comments on application and context in the Slovak environment.

Step 1: Decide how you want to teach online

Because of all the differences of online teaching listed above, it is necessary to re-think pedagogically and methodologically your course when shifting online. Professor Bates uses analogy of comparing teaching in classroom and online to driving a car and piloting a small airplane. Very basic things are the same. You need to drive and move forward to get where you are going. But the traffic rules are slightly different, plane moves not only forward, back, right and left, but also up and down and when something goes wrong, then a typical car driving reaction to hit the break and stop would be smart while the plane is in the air.

When teaching online, there are more aspects and issues to consider and a direct transfer of classroom course into online form (such as filming a lecture and upload it on the website) is not usually the most effective way.

An online course needs to be *designed* to match student's needs. One of the most prominent need among online students is usually flexibility, so use of asynchronous tools, which will enable students to keep their flexibility is a key.

In any case, at first you need to think about needs of your students, your course content and figure out the best ways how to present the content and introduce student activities that your course in an online form.

Step 2: Decide what kind of online course it should be

Online learning comes in many forms. It can be seen as a continuum, from no use of technology (very rare these days) to classroom teaching enhanced by online learning (with students doing some online "homework" on top of their regular classroom meetings, to hybrid courses (where some of the meetings in classroom are replaced by online work) all the way to fully online course.

There are three factors influencing such decisions:

- Profile and needs of the students (generally older and more experienced students benefit more from fully online courses, also it is important to know how much flexibility in their schedule the students have or whether and how often they would be able to physically attend classroom meetings)
- Nature of the subject matter (some subjects or particular parts of course content are more suitable for online learning then others)
- Resources available (this includes your time as an instructor, what help is available to you, the experience of other instructors, software and availability of other online materials).

Step 3: Work in a team

Generally, if it is your first online course you should not need to create and teach your online course by yourself. You should primarily seek help in following areas:

- Learning management. Most universities (also in Slovakia) provide a learning management system (LMS) for their teachers and students. In most cases the LMS used is Moodle – for instance at Comenius University – `http://moodle.uniba.sk` Learning management system is a software system installed at the university server, where your course is created and taught from. Generally an LMS enable you to create and teach online course, to present the course content and to use different tools provided (like discussion fora, wikis, blogs, video conferences and chat tools) without needing to learn html code or any other form of programming language
- Technical support. Most universities provide technical support within their IT departments and if you need something specific for your course that the LMS

does not provide, you should be able to ask for help
- Methodological support. An institution providing online learning should have at least some experts in online methodology where faculty can get help and advice with planning their online courses. Most Slovak universities don't provide this kind of support. Comenius university hired their first methodological expert in 2012 and are planning to expand on this kind of support.

Step 4: Build on existing resources

It doesn't make sense to try to "reinvent the wheel" just because you teach an online course. So it makes sense to first check for materials already available before spending time and energy creating your own. The best sources to search would probably be Open Educational Resources (OER) and then of course your colleagues who teach similar subjects.

In particular, look for teaching modules (small pieces of online material, sometimes with student questions or activities), videos, animations and simulations already developed. The UK's Open University's Openlearn, Merlot, MIT's Open Courseware, iTunes U and the Khan Academy are well known resources. You can also get useful information on how to search for OER from Open Educational Resources Handbook.

Few open learning initiatives and resources have appeared also in Slovakia:

- Univerzita pre moderné Slovensko - `http://www.upms.sk/` (video lectures & tests, over 6,000 students)
- Textbook on Social Policy by Miroslav Beblavý `http://www.socialnapolitika.eu/`.

Step 5: Master the technology

As the first step, you need to know which learning management system your institution is using. Most institutions also provide training for these teachers in LMS use. If such training is available it is a good idea to take one. Knowing what options your LMS provides will give you more idea on what could be done within your course.

Only you know what your LMS can and cannot do and gaining some initial experience it will be time to try to master and use some cool external tools. LMS is usually capable to cover up to 90% of needs for an online course.

Step 6: Setting appropriate learning outcomes and goals for online learning

When transferring a course into online form, it makes sense to carry over also the original goals and outcomes for the course. However, online learning provides additional tools and opportunities so it makes sense to try to consider whether it would be smart to create some additional goals. The point for consideration would be:

- Online learning can easily be designed to develop 21st century skills such as independent learning, critical thinking, team work, initiative, collaborative learning, communication skills and information management by embedding them in the course.
- Students will develop information and communication technology skills needed to access the site of the course and to work online, also they can easily learn how to: find, evaluate, analyze and apply information appropriately within the course subject.
- You can bring experts or practitioners from the outside world as resources into the within the course or program.

Step 7: Create a well-designed curriculum and structure for the course

Teaching structure would include two critical and related elements:

- the choice, breakdown and sequencing of the curriculum (content)
- the deliberate organization of student activities by teacher or instructor (skills development; and assessment).

This means that in a strong teaching structure, students know exactly what they need to learn, what they are supposed to do to learn this, and when and where they are supposed to do it. In a loose structure, student activity is more open and less controlled by the teacher. In terms of the definition, 'strong' teaching structure is not inherently better than a 'loose' structure, nor inherently associated with either face-to-face or online teaching.

The three main determinants of teaching structure are:

- the organizational requirements of the institution
- the preferred philosophy of teaching of the instructor
- the instructor's perception of the needs of the students.

Another very important factor is the time necessary for student to succeed in an online course. In North America a rough estimate for a student studying a classroom based 3 credit course is to spend about 8-9 hours of study every week. Hence online courses are created trying to more-or-less match this estimate and counting on students needing to work 8-9 hours a week to complete a 3 credit course online.

Step 8: Communicate, communicate, communicate

The more students work online, the more isolated they feel. There is a lot of research that indicates the importance of instructor "presence"in an online course (for instance [16]). The challenge is to balance the "presence"and involvement online with the workload. It is easy to be overwhelmed by all online activity in an online class.

Step 9: Innovate and Evaluate

Even when you build your online course based on all the research and practical advice gained from literature, the way of dynamic of the group and reactions will be heavily dependent on your particular course structure and content, background and needs of students and your teaching style and involvement.

It is generally valid, that each and every online course could still be improved. Of course you might have more improvements to do after you teach your very first online course for the first time then later, but even experienced instructors are still finding ways of improvement.

Online form of teaching leaves electronic records (transcripts from discussions, chat exchanges, records from video conferences) – materials that should be reviewed after each teaching of a course and used for improvements.

References

[1] Bachratý, H. and Bachratá, K. (2008). Mathematical Distance Education and E-Learning. In *Proceedings of the ICETA Conference*, *Slovakia*, pages 197–200, Elfa, s.r.o., Košice.

[2] Bates, T. (1995). *Technology, Open Learning and Distance Education*. Routlege, London, UK.

[3] Bates, T. (2000). *Managing Technological Change. Strategies for College and University Leaders*. Jossey-Bass, San Francisco, California.

[4] Bates, T. (2004). The Promise and the Myths of E-Learning in Post-Secondary Education. In M. Castells, editor, *The Network Society: A Cross-Cultural Perspective*, Edward Elgar, Cheltenham, UK.

[5] Bates, T. (2012). 9 Steps to Quality Online Learning. URL: `http://www.tonybates.ca/2012/05/02/nine-steps-to-quality-online-learning-introduction/`, retrieved 11 September 2013.

[6] Bauerová, D. (2003). Vzdělávání v informační společnosti – Education in the Information Society. Presentation at ELEARN Conference, February 4–5, Žilina, Slovakia.

[7] Boshier, R. W. (2001). Virtual University Madness in the World Wide Goldrush. In *Virtual University – Proceedings of the 2nd International Conference*, pages 5–11, Slovak Technical University, Bratislava.

[8] Cisco Networking Academies Program Website. URL: `http://www.cisco.com/web/learning/netacad/index.html`, retrieved 10 August 2011.

[9] Clark, R. (1994). Media Will Never Influence Learning. *Educational Technology Research and Development*, 42(2):21–29.

[10] Conrad, R. M. and Donaldson, J. A. (2004). *Engaging Online Learner*. Jossey-Bass, San Francisco.

[11] Frydenberg, J. (2002). Quality Standards in eLearning: A Matrix of Analysis. International Review of Research in Open and Distance Learning, October 2002. URL: `http://www.irrodl.org/content/v3.2/frydenberg.html`, retrieved 29 May 2003.

[12] Harasim, L., Hiltz, S. R., Teles, L., and Turoff, M. (1995). *Learning Networks: A Field Guide to Teaching and Learning Online*. MIT Press, Cambridge, MA.

[13] Huba, M. (2008). Steps to Quality E-Learning. In *Proceedings of the ICETA Conference*, pages 211–214, Elfa, s.r.o, Košice.

[14] Hvorecký, J., Burík, V., and Žáry, I. (2007). MBA in Marketing via a World-Wide Virtual University. In *Proceedings of the ICETA Conference*, pages 253–258, Elfa, s.r.o., Košice.

[15] Kocur, D. and Košč, P. (2009). Recommednations of Institutional Implementation of E-learning Technologies. In *Proceedings of the ICETA Conference*, Slovakia, pages 49–52, Elfa, s.r.o., Košice.

[16] Mandernach, B., Gonzales, R., and Garrett, A. (2006). An Examination of Online Instructor Presence via Threaded Discussion Participation. *Journal of Online Learning and Teaching*, 2(4):248–260.

[17] Phare Multi-CountryProgramme in Distance Education – General Guidelines (1997). European Training Foundation, Torino, Italy.

[18] Phipps, R. and Merisotis, J. (1999). What's the difference? A review of contemporary research on the effectiveness of distance learning in higher education. Institute for Higher Education Policy, Washington, DC, URL: `http://www.ihep.com`, retrieved 24 February 2000.

[19] Pišútová, K. (2012) *Collaboration in Online Learning in Slovakia*. Dissertation. Open University, Milton Keynes, UK.

[20] Russell, T. (2001) The No Significant Difference Phenomenon: A Comparative Research Annotated Bibliography on Technology for Distance Education (IDECC, fifth edition).

[21] Russel, T. (2013). The No Significant Difference Phenomenon. URL: `http://teleeducation.nb.ca/nosignificantdifference/`, retrieved 25 and 28 May 2013.

Automatic Extraction of Multiword Units from Slovak Text Corpora

Ján Staš[1], Daniel Hládek[1], Jozef Juhár[1], and Martin Ološtiak[2]

[1] Faculty of Electrical Engineering and Informatics, Technical University of Košice, Slovakia
[2] Faculty of Arts, University of Prešov, Slovakia

Abstract. The paper describes the process of automatic extraction of multiword units from the Slovak text corpora gathered from the Internet. We propose a morphologically motivated and statistical approach for extraction of relevant multiwords from four specific areas: fiction, justice, broadcast news and web. We have ensured that the extracted multiwords represent the most suitable candidates for the given domain by filtering out out-of-domain multiword units. The proposed extraction scheme may be useful not only for many natural language processing and speech recognition tasks, such as topic detection, text categorization or statistical language modeling, but also for lexicographic, lexicological and comparative research in linguistics and Slovak language sciences. By analysing of the extracted multiword units we have also obtained basic knowledge about the possible errors encountered in the process of text normalization and morphological annotation of the used text resources.

1 Introduction

Multiword units can be characterized as a lexical structures made of a sequence of at least two or more lexemes that can be decomposed into simplex words and display *lexical*, *syntactic*, *semantic*, *pragmatic* or *statistical meaning* or *idiosyncrasy*, that occur together frequently in a given language, and are usually associated with a fixed set of situations or particular context [11]. They usually correspond to *compound words* ("*hardware*"), *nominal compounds* ("*peanut butter*"), *non-decomposable* or *decomposable idioms* ("*kick the bucket*"), *verb-particle constructions* ("*look up*"), *light verbs* ("*make a mistake*"), *phrasal verbs* ("*break down*"), *institutionalized phrases* ("*traffic lights*"), *proper names* ("*New York Rangers*"), *sentence fragments* ("*and so on*") and some other *collocations* that co-occur more frequent than would be expected by the chance.

Multiword expressions are widely used in many areas of *natural language processing and understanding*, *computational linguistics*, or *speech recognition* and in many tasks oriented on the field of *information retrieval*, *question answering*, *topic detection*, *text categorization*, *machine translation*, *word sense disambiguation*, *spelling correction*, or *statistical language modeling* [6].

Two algorithms for automatic extraction of multiword units from the Slovak text corpora based on a morphological and statistical analysis are described in this paper, primarily proposed for the lexicographic, lexicological and comparative research in linguistics. The acquired knowledge and obtained results have been used in several natural language processing and speech recognition tasks in Slovak [9] for better text categorization based on the most frequent keyphrases or domain-specific language modeling [3]. Similarly, we

have found that the extracted multiwords may detect a number of errors that were introduced by preprocessing and morphological annotation of the Slovak text data [2].

As it was mentioned earlier, the proposed algorithms for automatic extraction of multiwords from the Slovak text corpora is based on two different principles. *Morphologically motivated approach* uses morphosyntactic analysis and part-of-speech tags for extracting relevant multiwords with using predefined patterns. This approach comes from the research on automatic extraction semantically significant collocations from Czech text corpora, described and published in [7]. *Statistical approach* uses standard evaluation measures for extraction of multiwords from the text corpora based on the frequency of co-occurrence of two or more words. This approach continues in our previous research oriented on decreasing errors in recognition of short words in the speech recognition task using multiwords [11]. Combining both approaches, we can also improve the efficiency of the extraction process and focus our research only on the specific type of multiword units like idioms or phrases.

This article is organized as follows. Section 2 serves a brief review about the utilization of multiwords in our research. The text corpora used for automatic extraction of relevant multiword units is mentioned in Section 3. Section 4 presents the proposed morphologically motivated and statistical approaches. The most relevant examples of the extracted multiword units from the Slovak text corpora for both approaches and results of its combination are briefly described in Section 5. Section 6 summarizes knowledge obtained in the analysis of multiword units before extraction and concludes this paper.

2 Motivation

Previous work, described in [11], was oriented on the extracting multiwords for the purpose of eliminating errors in the recognition of short monosyllabic words like prepositions, conjunctions or participles in the connection with long polysyllabic words in the statistical language modeling as a part of the Slovak automatic transcription and dictation system for the judicial domain [9]. In this case, the statistical measures such as *absolute* and *relative frequency* of words in the context and *pointwise mutual information* conditioned by selected *linguistic constraints* for extracting appropriate multiword units are used. In the Table 1, we can see the result of this extracting procedure. The misrecognition errors of the short monosyllabic words in speech recogniton, at the beginning of the speech or after long pause, were eliminated about by third.

Absolute Frequency		**Relative Frequency from the Left**		**Pointwise Mutual Information**	
assimilation of voicing	similarity in letters	assimilation of voicing	similarity in letters	assimilation of voicing	similarity in letters
v právnej	*mu ukladá*	*v súčasnosti*	*po odtrhnutí*	*žuť žuvačku*	*húf fanúšičiek*
ked'sa	*len na*	*k revidovaniu*	*do ohnivej*	*očistných kúr*	*skrutkového oja*
z toho	*od dlžníka*	*s difrakciou*	*bez zapnutých*	*rys ostrovid*	*popadalo ono*
v paragrafe	*do omeškania*	*k sebazáchove*	*súd dedukoval*	*hus zagágala*	*poškriabal lak*
ak by	*som mal*	*pod Tatrami*	*ju uviazať*	*krčných žíl*	*potopil loď*

Table 1. Example of extracted the most relevant multiwords conditioned by linguistic constraints in previous research oriented to the area of speech recognition

It has been shown that a similar approach may be useful in lingustics for lexicographic, lexicological and comparative research in the project of building dictionary of multiword naming units, in which we collaborate with the *Institute of Slovak Studies, Media Studies and Library Studies* at the *Faculty of Arts, University of Prešov*. In our case, the multiword units in the role of keyphrases are well proven in the topic detection, text categorization or document clustering tasks, better than the isolated keywords. Therefore, the research in this area is more than necessary. Studying the knowledge of a related languages [6], [7], we adapt the proposed algorithm for the automatic extraction of multiword units based on a statistical methods and extend it using a set of morphological rules with predefined patterns.

By appropriate selection of morpho-syntactic patterns, it is possible to use the result from extraction process in the other tasks oriented on the area of natural language processing and speech recognition in the Slovak language. The acquired knowledge can be possibly used for improving algorithms for normalization [10] and morphological annotation [2] of the Slovak text data, because extraction process usually detects common mistakes in the preprocessing, so as it will be discussed in Section 6.

3 Text Corpora

A large amount of the text data used in the process of automatic extraction of multiword units was collected using an automatic system for text gathering designed in our laboratory, called *webAgent* [3], [10]. This system retrieves the text data from various web pages and electronic resources that are written in the Slovak language. In the next step, the text data are filtered from a large amount of grammatically incorrect words, symbols or numerals and normalized into their pronounced form. Finally, the processed text corpora are divided into smaller domain-oriented subcorpora ready, for example, for the training language models. Statistics of the number of words and sentences for particular text subcorpus is summarized in the Table 2.

It is important to note that the text corpus from the judicial domain was obtained from the *Ministry of Justice of the Slovak Republic*, in order to develop the automatic transcription and dictation system for their internal purpose [9]. The corpus of fiction was created from a number of electronic books freely available on the Internet.

Text Corpus	Words	Sentences	Average
Corpus of Fiction	101,234,475	8,039,739	12.5918
Judicial Corpus	565,140,401	18,524,094	30.5084
Broadcast News	554,593,113	36,326,920	15.2667
Web Corpus	748,854,697	50,694,708	14.7719
Other Text	55,711,674	4,071,165	13.6845
Total	2,025,534,360	117,656,626	17.2156

Table 2. Statistics on the text corpora

In the case of the automatic extraction of multiword units using morphology, it is necessary to have a morphologically annotated corpus. For morphological analysis we have used *Dagger* [2], the Slovak morphological classifier based on a hidden Markov model

for solving the problem of part-of-speech tagging and suffix-based word clustering function restricted by *manually morphologically annotated lexicon of words* [1] for solving the word sense disambiguation problem. The hidden Markov model was trained on a trigram statistics generated from *the manually morphologically annotated corpus* [1] together with the lexicon delivered by the *Slovak National Corpus Department* at the *Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences in Bratislava*, in 2008 year. As we can see in the Table 2, the processed and morphologically annotated corpora were divided into five domain-specific subcorpora, oriented on the domain of *fiction*, *justice*, *broadcast news*, remaining *web* and *other heterogeneous text*.

4 Automatic Extraction of Multiword Units

As has been shown in the Introduction of this paper, we propose two approaches for automatic extraction of the multiword units from the Slovak text corpora. Morphologically motivated approach that uses predefined patterns derived from the sequence of a morphological tags given a basic information about the part-of-speech of each word in the text corpora. The second one is the statistical approach using three statistical functions based on absolute co-occurrencies of two words in observed text corpora, relative frequency in the context of the left and the right word in a bigram and pointwise mutual information between two words, restricted by the list of the Slovak stopwords.

4.1 Morphologically Motivated Approach

Referring to the linguistics and lexicological research, we propose morphologically motivated approach for extracting relevant multiword units from text corpora written in Slovak, where only the first symbol from a morphological tag is considered to give essential part-of-speech information. Whereas the Slovak language belongs to the group of highly inflective languages, in the case of a noun that is preceded by an adjective, grammatical categories of gender, number and case have to be the same. The same rules are true for the most part-of-speech classes and their grammatical categories. Therefore, it is not necessary to consider remaining information about the other grammatical categories of given part-of-speech, contained in tags.

In the Table 3, the proposed morpho-syntactic patterns or schemes for automatic extracting multiword units for bigrams, trigrams and quadrigrams from the Slovak text corpora can be seen. We have focused on extracting of multiwords that carry a certain meaning and are useful for both linguistics and statistical research, applicable in computational linguistics, natural language processing and understanding, statistical language modeling and speech recognition.

By reason of more accurately extraction of multiwords, every word included in it passed through spellcheck lexicon the total size of about 6.5 million of unique words. Spellcheck lexicon consists from manually checked and corrected Slovak words contained in *dictionary of the Slovak automatic transcription and dictation system* [9], [3], *manually morphologically annotated lexicon of words* obtained from Slovak National Corpus [1] and morphologically annotated lexicon of the all inflected forms of all Slovak words, called *Tvaroslovník* [5] that was created at the *Institute of Informatics, Faculty of Science, Pavol Jozef Šafárik University in Košice*.

Type	Characterization	Scheme	Example
2-gram	adjective + noun	AS	*akademický rok*
	numeral + noun	NS	*prvá pomoc*
	noun + noun	SS	*delba práce*
	abbreviation + noun	WS	*USB kľúč*
3-gram	adjective + adjective + noun	AAS	*malý pohraničný styk*
	adjective + noun + noun	ASS	*národný úrad práce*
	adverb + adjective + noun	DAS	*výškovo nastaviteľný volant*
	numeral + adjective + noun	NAS	*druhá svetová vojna*
	noun + adjective + noun	SAS	*deň pracovného pokoja*
	noun + preposition + noun	SES	*beh cez prekážky*
	noun + numeral + noun	SNS	*univerzita tretieho veku*
4-gram	noun + preposition + adjective + noun	SEAS	*obchod s bielym mäsom*
	noun + preposition + noun + noun	SESS	*dohoda o vykonaní práce*
	noun + noun + conjunction + noun	SSOS	*kniha sťažností a prianí*

Table 3. Proposed part-of-speech scheme

4.2 Statistical Approach

In the statistical approach, three common measures based on an absolute co-occurence, relative frequency in the left or the right context of the word and pointwise mutual information for bigrams have been used. As it is noted in [11]:

- **Absolute frequency** (f_A) expresses the total number of occurrences of the item in the whole corpus.
- **Relative frequency** of the multiword unit **in the context** (f_R) indicates the percentage of all occurrences of the word y in the corpus in the context of the word x. This measure is known as the left/right Fisher coefficient.

$$f_R(x, y) = \frac{f_A(x, y)}{f_A(x)} \times 100\%. \tag{1}$$

- **Pointwise mutual information** (PMI) is a measure of how much of the actual probability of a particular co-occurrence of events $p(x, y)$ differs from what we would expect on the basis of the probabilities of the individual events and the assumption of independence $p(x)p(y)$.

$$PMI(x, y) = log_2 \frac{p(x, y)}{p(x)p(y)}. \tag{2}$$

  In general, PMI reflects events such as collocations that do not occur in language frequently, but usually have strong relation and plain meaning.

Compared to the previous morphologically motivated approach, these statistical measures express only syntactic dependences between two words within extracted multiword. Moreover, statistics of co-occurrences could be obtained only from bigram counts generated from particular text subcorpora, because with higher order n-grams the time and computational requirements usually rapidly arise. As it was in previous approach, all words, before the process of extraction of relevant multiword units, passed through spellcheck lexicon. Multiwords that contain at least one stopword, punctuation character, two numerals, or another non-significant or irrelevant words were also omitted. In the Table 4, we can see the list of the Slovak stopwords retrieved from manually morphologically annotated lexicon obtained from the Slovak National Corpus [1].

a, abo, aby, aj, ak, ako, akoby, akokoľvek, akonáhle, akože, akožeby, akurát, aký, akže, ale, alebo, aleže, ani, asi, asiže, asnáď, aspoň, aspoňže, avšak, azda, azdaže, ač, ačkoľvek, ačpráve, až, ba, behom, bez, bezmála, bezo, bezpochyby, beztak, blízko, bodaj, bol, bola, boli, bolo, bude, budem, budeme, budete, budeš, budú, buď, by, bár, bárs, bársby, celkom, cestou, cez, cezo, chvalabohu, cieľom, dnes, do, dobre, dokiaľ, dokonca, dokopy, dokým, dolu, doprostred, doslova, dozaista, dľa, ešte, eštože, fakticky, figu, hej, hlavne, hneď, ho, hoci, hore, hádam, hľadiac, i, iba, ibaže, ich, inakšej, inakšia, inakšie, inakšieho, inakšiemu, inakšom, inakšou, inakší, inokedy, iná,ináč, iné, iného, inému, iní, inú, iný, iných, iným, inými, inšej, inšia, inšie, inšieho, inšiemu, inšiu, inšom, inšou, iste, isteže, ja, je, jedine, jednako, jedno, jednoducho, jeho, jej, jemu, jest, jestli, ju, k, kam, každej, každom, každou, každá, každé, každého, každému, každí, každú, každý, každých, každým, každými, kde, kdeby, kdeže, kdežeby, kdežto, keby, kedykoľvek, keď, keďže, kiež, kiežby, konečne, končа, kto, ktohovie, ktokoľvek, ktorej, ktorejkoľvek, ktorejsi, ktorom, ktoromsi, ktoromkoľvek, ktorou, ktoroukoľvek, ktorousi, ktorá, ktorákoľvek, ktorási, ktoráže, ktoré, ktorého, ktoréhokoľvek, ktoréhosi, ktorékoľvek, ktorému, ktorémukoľvek, ktorémusi, ktorési, ktorí, ktorím, ktoríkoľvek, ktorísi, ktorú, ktorúkoľvek, ktorúsi, ktorý, ktorých, ktorýchkoľvek, ktorýchsi, ktorým, ktorýkoľvek, ktorými, ktorýmisi, ktorýmikoľvek, ktorýmkoľvek, ktorýmsi, ktorýsi, ktosi, ktovie, ktovieakej, ktovieako, ktovieakom, ktovieakou, ktovieaká, ktovieaké, ktovieakého, ktovieakému, ktovieakí, ktovieakú, ktovieaký, ktovieakých, ktovieakým, ktovieakými, ktoviekoľko, ktoviečo, ktoviečoho, ktoviečom, ktoviečomu, ktoviečím, ktože, ktožehovie, ktoževie, ku, kvôli, kým, kýmže, kýmkoľvek, lebo, len, lenže, lež, ma, maximálne, mať, medzi, menej, mi, miesto, mimo, mne, minimálne, mnohá, mnohé, mnohého, mnohému, mnohí, mnohích, mnohím, mnohú, mnohý, mnohých, mnohýkrát, mnohým, mnohými, mnohých, mnohýkrát, mnohým, mnohými, mnou, moja, moje, mojej, moji, mojich, mojim, mojimi, mojom, mojou, moju, možno, mu, musieť, musí, musím, musíme, musíte, musíš, my, myslím, má, máme, máte, môcť, môj, môže, môžebyť, môžem, môžeme, môžete, môžeš, môžme, môžte, môžu, môžuc, mňa, na, nad, nado, nadol, najmä, najprv, najskôr, nakoniec, nakoľko, nami, namiesto, nanajvýš, naopak, naozaj, napokon, napred, napriek, naprieč, naproti, naskrz, naveľa, navzdory, nazad, naši, našich, našu, nech, nechby, nehľadiac, neodhliadnuc, nepozerajúc, nepočínajúc, nepočítajúc, nerátajúc, nesúdiac, nevynímajúc, nezahŕňajúc, nezačínajúc, neďaleko, než, nie, niekto, niektorej, niektorom, niektorou, niektorá, niektoré, niektorého, niektorému, niektorí, niektorú, niektorý, niektorých, niektorým, niektorými, niekym, niekým, nielen, nielenže, nieto, niežeby, nijako, nič, niže, no, nolen, nové, nového, novému, noví, novú, nový, nových, novým, novými, novšej, novšia, novšie, novšieho, novšiemu, novšiu, novšom, novšou, novší, novších, novším, novšími, nuž, nám, nás, náš, o, obďaleč, od, odhliadnuc, odo, ohľadne, ohľadom, okolo, okrem, on, ona, oni, ono, ony, oproti, ovšem, ozaj, ozajže, paroma, po, pod, podo, podľa, pokiaľ, pokým, pomaly, pomerne, pomimo, ponad, ponevráč, popod, popravde, popred, popri, postupom, potom, povedzme, povedľa, povyše, poza, pozdĺž, pozerajúc, počas, počnúc, počínajúc, počítajúc, pravda, pravdaže, pre, pred, predo, predsa, presne, preto, pretože, prečo, pri, priam, približne, priebehu, priemerne, prirodzene, pritom, pričom, prospech, prosto, prostredníctvom, prosím, proti, práve, prípadne, raz, reku, respektíve, respektívne, rovno, rozhodne, rátajúc, s, sa, samozrejme, si, skoro, skraja, skrz, skutočne, skôr, slovom, sme, snáď, so, som, sotva, sotvaže, spod, spomedzi, sponad, sponiže, spopod, spopred, spopri, spovyše, spoza, sprava, spred, spredu, správy, späť, ste, stred, svoj, svoje, svojho, svoji, svojich, svojim, svojimi, svojmu, svojom, svojou, svojích, svojím, svojími, síce, sú, súdiac, sčasti, sťa, stiaby, ta, tak, takmer, takto, takže, tam, teda, tejto, temer, ten, tento, ti, tieto, tiež, to, tobôž, toho, tohto, tohoto, tom, tomto, tomuto, totiž, toto, trebárs, tu, tuším, tvoj, tvojími, ty, tá, táto, tú, túto, tým, týmito, týmto, u, uprostred, určite, už, v, vari, vaše, vcelku, veru, veruže, veď, vedže, viac, vinou, vlastne, vnútri, vo, von, vopred, voči, vplyvom, vpred, vprostred, vraj, vrátane, vskutku, vy, vyjmúc, vynímajúc, vyše, vzdory, vzhľadom, vám, vás, váš, vôbec, včítane, vďaka, však, všetok, z, za, zahŕňajúc, zaručene, začínajúc, zdola, zhora, zhruba, zmedzi, znad, zo, zrejme, zreteľom, zvlášť, zásluhou, zľava, áno, údajne, čerta, či, čili, čisto, čiže, čo, čoby, čopriam, čože, čožeby, čím, ďalej, ďaleko, ďalšej, ďalšia, ďalšie, ďalšiu, ďalšieho, ďalšiemu, ďalší, ďalších, ďalším, ďalšími, ťa, ťažko, že, žeby, žepriam, žiaľ, žiaľbohu

Table 4. List of the Slovak stopwords

4.3 Filtration of Out-of-Domain Multiwords

Using filtration of out-of-domain multiwords from every domain-specific text corpus ensures that the extraction result will contain only those units that are typical for the observed domain. Therefore, all multiword units that exist in each text corpora are omitted from the observation. For example, if we want to extract multiword units from judicial domain, all multiwords that occur in the corpus of fiction, broadcast news corpus, web corpus and the other texts, were excluded from this list. This step ensures that only strongly domain-dependent multiword units remain within one domain.

At the end of this section, it should be noted that the efficiency of the extraction process can be significantly enhanced by combination of morphologically motivated and statistical approaches together with filtration out-of-domain multiword units as well.

Scheme	Fiction	Justice
AS	*čierny jazdec*	*právny predchodca*
NS	*trom pátračom*	*oboch rodičov*
SS	*ušiak jimmy*	*zástupca navrhovateľa*
WS	*Al Parker*	*GSM číslo*
AAS	*vysoký štíhly muž*	*naliehavý právny záujem*
ASS	*veľkomožný pán kráľ*	*federálneho ministerstva financií*
DAS	*smrteľne bledú tvár*	*zvlášť hrubým spôsobom*
NAS	*dve veľké izby*	*tretia nápravnovýchovná skupina*
SAS	*pán hlavný inšpektor*	*vydania platobného rozkazu*
SES	*prstami do vlasov*	*úhradu vo výške*
SNS	*vodca troch pátračov*	*odsek päť ústavy*
SEAS	*klobúk so širokou strechou*	*zákon o kolektívnom investovaní*
SESS	*zážitky na prahu smrti*	*zástupca z radov advokátov*
SSOS	*výbor lásky a starostlivosti*	*ministerstvo poľnohospodárstva a výživy*

Scheme	Broadcast news	Web
AS	*finálová zostava*	*tichá lokalita*
NS	*ôsmeho kola*	*sedem nocí*
SS	*hovorkyňa magistrátu*	*číslo aukcie*
WS	*MHK Košice*	*PC komponenty*
AAS	*najvyšej futbalovej súťaže*	*celková podlahová plocha*
ASS	*legislatívna rada vlády*	*dostupný počet kusov*
DAS	*mierne lavínové nebezpečenstvo*	*mierne svahovitý terén*
NAS	*päť percentný nárast*	*päť izbový byt*
SAS	*poškodzovania cudzej veci*	*servis výpočtovej techniky*
SES	*boj o záchranu*	*prípojka na pozemku*
SNS	*rozdiel jedného gólu*	*výstraha prvého stupňa*
SEAS	*obvinenie z trestného činu*	*pozemok pre rodinné domy*
SESS	*násilia proti skupine obyvateľov*	*služby v oblasti ubytovania*
SSOS	*výrobcov piva a sladu*	*registrácia domén a webhosting*

Table 5. Example of the most relevant multiwords based on morphological analysis

5 Results

In this section a brief review of experimental results of automatic extraction of multiword units from the Slovak text corpora will be summarized.

In the Table 5, we can see some of the most relevant multiword units extracted using morphologically motivated approach after filtration described above. The number of extracted multiwords using morpho-syntactic patterns for each domain-specific text corpus are summarized in the Table 6.

As we can see from experimental results, extracted multiword units seem to be related with investigation area to a high degree. An interesting result is that the number of extracted items from the fiction domain is very high, which is probably caused by a high variability of sentence word-order. On the contrary, the judicial domain is characterized by a high number of common expressions and the number of extracted multiword units was lower than in the fiction domain.

The Table 7 shows example of the results of extraction of the most relevant multiword units using statistical methods for each domain-specific corpora after filtration process. This approach appears to be appropriate, if we would not have any morphological analyser.

	Fiction	Justice	BN	Web
2-grams	310,819	2,226	155,619	22,436
3-grams	225,217	1,426	66,858	9,226
4-grams	51,168	208	7,219	1,490
Total	587,204	3,860	229,696	33,152

Table 6. Number of extracted multiword units using morphologically motivated approach after filtering

Scheme	Fiction	Justice
Absolute Frequency	*čierny diviak*	*zastúpený opatrovníčkou*
Relative Freq. from the Left	*dieťa neuškrtila*	*právnych úkonov*
Relative Freq. from the Right	*usychajúcej kukurici*	*nepoleptaných častí*
Pointwise Mutual Information	*vychladlých horách*	*protinábožensk´eho zúrenia*

Scheme	Broadcast News	Web
Absolute Frequency	*zimný múčnik*	*udelený komentár*
Relative Freq. from the Left	*dvoma trafostanicami*	*noha rabujúceho*
Relative Freq. from the Right	*podstúpiac smrť*	*kosiacom obilie*
Pointwise Mutual Information	*meninovými blahopraniami*	*doznievajúcimi verdiktmi*

Table 7. Example of the most relevant multiwords based on statistical analysis

By combining morpho-syntactic patterns with statistical measures, only those multiword units that are frequent in the given language and have certain meaning can be effectively extracted and are usable in selected tasks in statistical or linguistics research. An example of combining the morphological with statistical approach in case of judicial domain is shown in the Table 8.

6 Discussion

The results of the process of automatic extraction of multiword units using morphologically motivated and statistical approaches and their combination were described in this paper. All words that enter to the process of extraction, passed through spellcheck lexicon at first. Then multiword units were extracted from the text corpora. Finally, out-of-domain events were filtered to obtain the strongly domain-dependent multiword units for each of domain.

The analysis of extracted multiword units from Slovak the text corpora has brought us new insights:

- Efficiency of the extraction procedure depends on the used spellcheck lexicon. For efficient domain-oriented tasks it is better to restrict the process of extraction of multiwords with the list of the most frequented wordforms from the examined domain;
- Efficiency can be also enhanced by introducing stemming or lemmatization into to process of extraction of multiword units;
- Analysis of multiwords reveals hidden errors in the process of text normalization, namely in numeral transcription or abbreviation expanding steps. This knowledge can help to us eliminate this errors in further research;

Scheme	Absolute Frequency	Pointwisse Mutual Information
AS	*denným úrokom*	*patetickým bľabotom*
NS	*obom odporcom*	*mnohonásobnými napadnutiami*
SS	*faktúrou čislo*	*korelátmi anxiety*
WS	*GSM číslo*	*RTG ožarovač*

Scheme	Relative Frequency from the Left	Relative Frequency from the Right
AS	*prvotným vyrozumeniam*	*smerovací dokument*
NS	*dvoch respondentov*	*nultí zamestnanci*
SS	*spôsobe vstávania*	*pootvorenie poklopu*
WS	*CD rádiomagnetofón*	*CMOS kamier*

Table 8. Example of the most relevant multiwords based on combining morphological and statistical approaches for the judicial domain

- Analysis also reveals errors in word sense disambiguation during morphological analysis and part-of-speech tagging. In many cases, morphological tags were changed, mainly in case of new out-of-vocabulary words that were assigned into the category of abbreviations or numerals;
- It would be interesting to find the intersection between multiword units extracted only using morphological analysis and without any knowledge about part-of-speech tags. Which statistical measure gives the similar results compared to the morphologically motivated approach?;
- In the future research it would be appropiate to focus on the higher-order n-grams, especially in the case of the statistical extraction of multiword units;
- Verb-particle constructions, phrasal verbs and idioms were not included into the process of extraction of multiword units based on the morphologically motivated approach;
- The list of stopwords should be extended with other frequent but less meaningful words too;
- The future research should be focused on the introduction of a complex tool for extracting collocations from text corpora based on the number of statistical methods, called the *Ngram Statistics Package* (`Text::NSP`) [4], [8].

It should be noted that the extraction of the multiword units from the judicial domain and using them as keyphrases in the task of categorization text for in-domain and out-of-domain data for statistical language modeling and adaptation caused increased accuracy of our speech recognition system about 5.54%, relatively. Therefore, this research also plays an important role in the field of a natural language processing and speech recognition.

Acknowledgments

The research presented in this paper was supported by the Ministry of Education, Science, Research and Sport of the Slovak Republic under the research project APVV-0342-11 (25%) and Research and Development Operational Program funded by the ERDF under the project ITMS-26220220141 (75%).

References

[1] Garabík, R. (2006). Slovak morphology analyzer based on Levenshtein edit operations. In *Proceedings of the 1st Workshop on Intelligent and Knowledge oriented Technologies, WIKT 2006*, pages 2–5, Bratislava, Slovakia.

[2] Hládek, D., Staš, J., and Juhár, J. (2012). Dagger: The Slovak morphological classifier. In *Proc. of the 54th International Symposium ELMAR 2012*, pages 195–198, Zadar, Croatia.

[3] Juhár, J., Staš, J., and Hládek, D. (2012). Recent progress in development of language model for Slovak large vocabulary continuous speech recognition. In Volosencu, C., editor, *New Technologies - Trends, Innovations and Research*, pages 261–276, InTech, Rijeka, Croatia.

[4] Kohli, S. (2006). Introducing an object oriented design to the Ngram Statistics Package. Technical report, Department of Computer Science, University of Minnesota, Duluth, USA.

[5] Krajči, S. and Novotný, R. (2012). Projekt Tvaroslovník - slovník všetkých tvarov všetkých slovenských slov. In *Sborník příspěvkú 11. ročníku konference ZNALOSTI 2012*, pages 109–112, Mikulov, Česká republika.

[6] Manning, C. D. and Schütze, H. (1999). *Foundations of statistical language processing*. The MIT Press, Cambridge, Massachusetts, USA.

[7] Pecina, P. and Holub, M. (2002). Sémanticky signifikantní kolokace: automatická detekce kolokací v českém textovém korpusu. Technical Report ÚFAL/CKL TR–2002–13, Ústav formální a aplikované lingvistiky, Matematicko-fyzikální fakulta, Univerzita Karlova, Praha, Česká republika.

[8] Petersen, T., Banerjee, S., McInnes, B. T., Kohli, S., Joshi, M., and Liu, Y. (2011). The Ngram Statistics Package (Text::NSP): A flexible tool for identifying Ngrams collocations and word associations. In *Proceedings of the Workshop on Multiword Expressions: From Parsing and Generation to the Real World*, pages 131–133, Portland, Oregon, USA.

[9] Rusko, M., Juhár, J., Trnka, M., Staš, J., Darjaa, S., Hládek, D., Cerňak, M., Papco, M., Sabo, R., Pleva, M., Ritomský, M., and Lojka, M. (2012). Slovak automatic transcription and dictation system for the judicial domain. In *Proc. of the 5th Language and Technology Conference Human Language Technologies as a Challenge for Computer Science and Linguistics*, pages 365–369, Poznań, Poland.

[10] Staš, J., Hládek, D., and Juhár, J. (2013). Building organized text corpora for speech technologies in the Slovak language. *Linguistics studies (Jazykovedné štúdie)*, XXXI. to be published.

[11] Staš, J., Hládek, D., Trnka, M., and Juhár, J. (2011). Automatic extraction of multiword expressions using linguistic constraints for Slovak LVCSR. In *Proceedings of the 6th International Conference on Natural Language Processing, Multilinguality*, pages 138–145, Modra, Slovakia.

Verb Valency and Argument Non-correspondence in a Bilingual Treebank

Jana Šindlerová, Zdeňka Urešová, and Eva Fučíková

Faculty of Mathematics and Physics, Charles University in Prague, Czech Republic

Abstract. In this paper we present a contrastive study of one interesting non-correspondence between deep syntactic valency structures of two different languages. On the material of the Prague Czech-English Dependency Treebank we observe sentences in which an Addressee argument in one language is linked translationally to a Patient argument in the other one, particularly we aim our attention at the class of judgement verbs (in a broad sense). Considering this class of verbs, we analyze the relevant examples and discuss the nature of "the third argument" in the valency structure. As a result, we reconsider the conventions of argument labelling with the aim of achieving better consistency of annotation and we suggest possible ways of adjusting the valency theory itself to the needs of multilingual data.

1 Introduction

Modern approaches to applied linguistics take the advantage of a great number of annotated corpora, covering different depth and width of linguistic description, a wide range of content domains, and above all an impressive scale of world languages. Many of these corpora are accompanied by additional resources, such as valency lexicons. Parallel valency lexicons, accompanying multilingual corpora, satisfy the call for capturing complex lexical information, i.e. the information on both verbal translational equivalents and their valency slot realizations. Having resources of this kind at our disposal gives us a perfect opportunity to study similarities and differences between languages on the syntax-semantics interface.

In this paper we will focus on Czech and English deep syntactic valency structures in a contrastive perspective. The assumption we follow is that the deeper we look into the linguistic structure, the more similar should the structures appear, an idea shared e.g. by [5]. This idea stands behind numerous attempts to create an interlingua for machine translation systems from various types of deep syntactic structures, or semantic representations. Our goal within the research is to look at the points of non-correspondence between deep syntactic structures of Czech and English parallel sentences, to analyze them and categorize according to the syntactic and semantic properties of the utterances they represent. Questions we would like to ask in this paper are the following:

- Are there any semantic (or syntactico-semantic) criteria, rather than mere morphological hints, to let us distinguish clearly between possible variants of argument labelling?
- Does the cross-linguistic perspective offer a better insight into the nature of differences between the individual frames?
- Does the cross-linguistic perspective help us in deciding about possible theoretical amendments and in making the annotation practice more uniform?

2 Methodology and Data

We took the advantage of the existence of Czech-English parallel data, namely the Prague Czech-English Dependency Treebank (PCEDT) [4]. It is a collection of about 50,000 sentences, taken from Penn Treebank-Wall Street Journal section, translated manually to Czech, transformed into dependency trees and annotated at the level of deep syntactic relations (called the tectogrammatical layer). In short, the tectogrammatical layer contains mostly content words (with several defined exceptions) connected with oriented edges and labelled with syntactico-semantic functors according to the Functional Generative Description approach (FGD), see [6]. Ellipsis and anaphora resolution is also included, as well as an automatic alignment of corresponding nodes [12].

The PCEDT 2.0 [3] is annotated according to the FGD valency theory and two valency lexicons (one for each language) are part of the release of the treebank. The PDT-Vallex [15], [16] has been developed as a resource for annotating argument relations in the Prague Dependency Treebank [1]. Valency frames in the PDT-Vallex roughly correspond to individual verb meanings. Valency frames consist of participant slots represented by tectogrammatical functors. Each slot is marked as obligatory or optional and its typical morphological realization forms are listed. Frame entries are supplemented with illustrative sentence examples.

The Engvallex was created as an adaptation of an already existing resource of English verb argument structure characteristics, the Propbank. The original Propbank argument structure frames have been adapted to the FGD scheme, so that it currently bears the structure of the PDT-Vallex, though some minor deflections from the original scheme have been allowed in order to save some important theoretical features of the original Propbank annotation.

Currently, there is a project aimed at interlinking PDT-Vallex and Engvallex in the sense of gaining a database of frame-to-frame, and subsequently, slot-to-slot pairs for the purposes of machine translation experiments [17], extending a similar project held in the past [14].

In the project we also deal with semantic categories and verb classes. Since this topic is not covered within the FGD theory, we have consulted other available resources of native speaker's intuition regarding valency characteristic of English verbs: the Propbank [9], the Framenet [13] and Levin's classification (as stated in [10]).

3 Argument Labelling in PDT 2.0

In the FGD, five *actants*, i.e. main universal and typical arguments of a valency frame, are recognized: ACT (Actor), PAT (Patient), ADDR (Addressee), EFF (Effect) and ORIG (Origin). In the theoretical framework, it is declared that ACT and PAT stand for more general concepts of "the first" and "the second" argument in the valency structure, in other words, these positions are described more syntactically than semantically. On this theoretical background, the concept of "shifting of cognitive roles" has been adopted. According to this rule, if the verb has only two arguments, semantic Effect, semantic Addressee and/or semantic Origin are being shifted to the Patient position. With respect to such definition, we may encounter several difficulties in our research. Typically, if we look for the differences in argument labelling, we may easily be confused by different labelling of (semantically the same) Addressee argument due to a different number of valency positions

in the corresponding frames, as in (1).

(1a) John.ACT blamed Mary.ADDR of stealing.PAT his car.
(1b) John.ACT shouted at Mary.PAT.

Another problem is tied with cases of three (and more) recognizable participants in the valency structure, where ADDR in fact appears as the (syntactically) second argument (often expressed with accusative), whereas the PAT label is left for another argument role. Though much attention is paid to the criteria for the theoretical distinction of actant and free modifier roles, and for the distinction of obligatory and optional positions in FGD, surprisingly little is said about the nature of individual actant roles per se. It is somehow taken for granted that native speaker intuition in this respect recognizes semantic aspects of the actant roles well. For example, PDT guidelines [12] describe PAT as an "affected object" in a broad sense and offer an illustrative (non-exhaustive) list of its possible semantic modifications, but only for PAT as a second argument, leaving out the (for our research) interesting cases where PAT takes a third position in the valency structure[1], syntactically realized as a prepositional phrase following a direct object ADDR, like in (1a).

The authors of the PDT 2.0 annotation guidelines also confess there is a certain degree of uncertainty about the character of arguments with certain verbs and explicitly mention several borderline cases [12]. As a typical case, they offer an example of the Czech verb *bránit (protect)*, with the following possible interpretations of the three available arguments (Protector, Protected, Harm/Enemy):

(2a) Petr.ACT bránil majetek.PAT před zloději.EFF.
Petr.ACT protected his property.PAT from thieves.EFF.
(2b) Petr.ACT bránil děti.ADDR před nebezpečím.PAT.
Petr.ACT protected the children.ADDR from danger.PAT

The resolution of this problem in the annotation guidelines is based on a morphosemantic feature of animacy. If the defended entity is represented in a majority of corpus occurrences by an animate noun (or, more precisely, by an animate entity), the position in the valency lexicon should be labelled Addressee, otherwise it is assigned a Patient label. Still, it is a common phenomenon that a degree of interannotator disagreement is noticed when dealing with similar cases.[2]

The investigation of a contrastive language material shows that such cases are frequent and tightly connected to the semantic class of the verb. Moreover, it appears that native speaker intuition differs in the contrastive point of view.

4 Places of Non-correspondence

When searching the PCEDT, we have encountered five major verb classes that show inconsistencies in the annotation of valency structures, mainly concerning the Addressee

[1] Our numbering of argument positions in the paper is given by purely syntactic properties, i.e. subject first, (direct) object(s) second (or second and third), prepositional phrases third (fourth) etc. In English, this also corresponds to a standard word order; however, we also use it for Czech.

[2] It is also interesting that other researches working within the FGD framework do not operate with animacy in this respect, cf. [11].

role: verbs of judgement, verbs of attempting suasion (and causation), a joint group of several classes semantically expressing permission or accessibility granting, verbs of assistance and verbs of commercial exchange.[3] For each of the verb classes, we have consulted several resources of valency structure description. Apart from the PDT-Vallex and the Engvallex, we have searched the Framenet and the Propbank. For the reference to semantic classes of verbs, we have also consulted Levin's classification. Throughout the five mentioned semantic groups, we have encountered several patterns of frame concurrence, the most frequent being the following:

- ACT ADDR PAT x ACT PAT CAUS
- ACT ADDR PAT x ACT PAT EFF or ACT PAT AIM
- ACT ADDR PAT x ACT PAT REG
- ACT ADDR PAT x ACT PAT MEANS

This concurrence of frames appeared both in the cross-linguistic comparison (a source language sentence is annotated differently than the target language sentence) and within different verbs of the same semantic group in one language (two verbs of one language, which are semantically close, or even synonymous, are annotated differently).

All the above mentioned alternative frame variants consist of an all-actant interpretation (ACT ADDR PAT) and an interpretation involving an adjunct on the third syntactic position (ACT PAT CAUS). For the purposes of simplification we do not operate here with the notion of obligatoriness or optionality, leaving this complex issue to a separate, more elaborate study. Nevertheless, it must be said that this issue is of supreme interest to us, since it is tightly connected to the elementary question of argumenthood. Although FGD theoretically allows free modification functors to appear as a part of valency frame in case they are obligatory, in reality, this accounts only for a few members of the list of possible free modifications, usually for directionals, temporals and manner adjuncts. On the other hand, it is not a common practice e. g. to label an obligatory argument CAUS, even if its semantic incorporates causal interpretation, in this case, usually an actant label is given priority.

4.1 Judgement Verbs in Czengvallex

There are two ways in which argument non-correspondence in a bilingual corpus can be considered. Either there is a different argument labelling between a particular sentence and its translation, or there is a difference in argument labelling between verbs of the same verb class within a particular language. Both types of argument non-correspondence manifest in PCEDT among the verbs of judgement and communicating judgment. In Framenet, these two categories are considered separate, for our purposes it seems convenient to treat them jointly, e.g. as they appear in Levin's classification [10]. In the analysis, we will refer to them as *judgement verbs*.

In our sample, we have looked at the following verbs: *accuse, blame, charge, chastise, convict, criticize, fault, reprove, sue*. According to the three resources of English verbs argument structures, these verbs share three argument roles, which can be characterized as follows: *the judge*, *the judged entity* and *the reason for judgment*, or *the fault*. In the PCEDT (and its valency lexicons), the annotation practice is divided as shown in Table 1.

[3] The naming of the classes has been roughly adopted from the Framenet nomenclature.

The individual rows of the table represent different translation verb pairs, columns show the distribution of the verbs among different frames. If both verbs of the translation pair belong to the same frame, they are both inscribed in the same cell.

ACT ADDR PAT	**ACT PAT CAUS**	**ACT PAT EFF CAUS**
obvinit – accuse		
obviňovat – accuse		
vinit – accuse		
charge (with) – obvinit		
charge (with)	obžalovat	
charge (with)	žalovat	
obvinit	blame (for)	
připisovat – blame (on)		
připisovat	blame (for)	
přisuzovat – blame (on)		
přičítat – blame (on)		
obvinit – convict		
usvědčit – convict		
convict		odsoudit
obvinit	fault	
reprove	odsuzovat	
sue	žalovat	
soudit se – sue		
vytknout	chastise	
	kárat – chastise	
	potrestat – chastise	
	kritizovat – criticize	

Table 1. Frame distribution for Czech and English Judgement Verbs in the PCEDT

Most of the verbs fall into one of the following frame variants: ACT ADDR PAT, or ACT PAT CAUS (the labels of the positions marking *the judge*, *the judged entity* and *the fault* respectively). All the mentioned verbs have an addressee (*the judged entity*), though in the other variant of the frame, it is labelled PAT due to the concept of shifting cognitive roles (see [12]). The split of the annotation is apparently caused by different approaches to the third argument, i. e. *the reason for judgement*, or *the fault*. Either it is interpreted as an actant, i.e. belonging to the valency structure, or it is considered an adjunct, a free modification external to the valency structure.

Clearly, the important question is, whether *the reason for judgement* position is or is not a part of the valency structure of the verb. The resources for English verbs speak straight, all regard this position as a valency argument.[4] This question is closely connected to the question of "what exactly is an argument (theta role, participant etc.) and how many of them there really are", which has not been satisfyingly answered in the literature yet. Since this type of problem is too complex to be dealt with within this paper, we will not try to answer it directly. A very nice and summarizing debate of this issue can be found e.g. in [2]. Not only does the author question the mere possibility of finding clear matching criteria for argumenthood that would apply to all types of arguments, but he also mentions an

[4] For example in the Framenet, Reason is always part of the core frame elements of the judgement frames, Propbank also includes it in the list of frame participants of the verbs in question.

important catch of the argumenthood-defining efforts. When trying to describe argument roles unambiguously, we necessarily use criteria from many levels of linguistic description, not only syntactic and semantic, but we also have to engage morphological and pragmatical hints. This on one hand helps us to specify the roles more exactly, but on the other hand leads many times to confusion and theoretical clashes.

4.2 Data Analysis

As we can see from Table 1, the annotation is to a certain extent inconsistent. There are at least four different English verbs correctly translated by Czech verb *obvinit/obviňovat*.[5] In case of *accuse* and *charge*, the third argument is PAT, whereas in case of *blame* and *fault*, it is labelled CAUS, see (3).

(3a) Industrial companies.ACT are accusing financial institutions.ADDR of jeopardizing.PAT Japan's economy.
(3b) A Campeau shareholder.ACT charged Campeau.ADDR with violating securities law.PAT.
(3c) Many investors.ACT blamed program trading.PAT for aggravating.CAUS market swings.
(3d) The former New York City mayor.ACT faults Obama.PAT for incompetence.CAUS over the Libya consulate attacks.

Since there is no significant difference in the verb semantics, this may be the result of the influence of morphological form: The for-phrase (*blame, fault*) is a typical morphological means for expressing Cause, whereas the of-phrase (*accuse*) is one typical way of expressing Patient (affected object) semantics, and the with-phrase (*charge*) is typical for Instrument interpretation, thus being tentative to less specific labelling.

Speaking about the impact of morphosyntactic form, we must point out another fact. Whereas the direct object form of *the reason for judgement* argument builds almost immediately the actant interpretation, the prepositional phrases are ambiguous with respect to possible interpretations. According to the Prague annotation style, it appears that only primary prepositions are allowed with arguments, whereas phrases with secondary prepositions are generally regarded as adjuncts. For each actant and adjunct label, the guidelines offer a list of typical prepositional phrases (in the form of preposition plus case) used with it. The *reason for judgement* can be, with some, but apparently not all the verbs, expressed in a typically adjunct morphosyntactic form, e.g. with a subordinate adjunct clause or a secondary preposition.

(4) Vyšetřovatel obvinil kvůli incidentu u mosteckého klubu Neprakta tři muže.
The investigating officer.ACT charged three men.PAT because of the incident.CAUS in the Neprakta club in Most.[6]

Nevertheless, it seems to us that such utterances are actually less acceptable since they

[5] In case of convict the translation may be considered inappropriate.
[6] Since there was no suitable example in the PCEDT data, the sentence has been taken from Czech National Corpus. The labelling has been added by the authors of the paper.

mix the intended "objective reason for judgement" semantic interpretation with the typical form of expressing a "circumstantial motivation for judgement" (like in case of (5)):

(5) Obvinila ho, protože zrovna neměla dobrou náladu.
She blamed him because she wasn't in a good temper at the time.

In such cases, the form influences the interpretation to the circumstantial one, and the reason for judgement appears as non-overt.

Note that considering the third argument itself, there are equally relevant reasons for both interpretations (Patient and Cause). The semantics of the argument in question bears causal features (Framenet e.g. names this role *Reason*). On the other hand, it is often expressed (in lexicalized alternations[7] of the verbs in question) in a direct object position, which is typical for Patient and atypical for Cause (6).

(6) Nobody.ACT would blame the global warming.PAT on a few hundred thousand hunter-gatherers.ADDR hunting mammoths and scratching around in caves.

What is even more confusing, not even the criterion of obligatoriness can support our decision between argument and adjunct interpretation, since in the PDT-Vallex, the PAT argument of the judgement verbs is often marked optional.

With the verb pair *odsoudit – convict*, the situation is even more interesting.

(7) Despite the strong evidence against Mrs. Yeargin, popular sentiment was so strong in her favor, Mrs. Ward says , that "I'm afraid a jury wouldn't have convicted her."
I přes přesvědčivé důkazy proti Yearginové bylo této učitelce veřejné mínění tak silně nakloněno, že ředitelka Wardová říká: "Obávám se, že by ji porota neodsoudila.

The Czech verb, according to the PDT-Vallex opens valency positions for *the judge* (ACT), *the judged entity* (PAT), and *the sentence* (EFF), *the reason for judgement* being considered an adjunct CAUS. On the other hand, the original English verb *convict*, as far as Framenet and Propbank annotation states, does not include *the sentence* in its argument structure at all. Once again, we may ask the question, what are the criteria for considering *the reason for judgement* an adjunct in Czech, and not (an optional) PAT, while *the judged entity* slot could be easily re-interpreted as ADDR.

Another theoretical question considering the number of valency positions is effected by the lexicalized alternations.[8] In both Propbank and Framenet, the criticized entity and the cause of critique with verbs of communicating judgement are distinguished and treated separately, so that cases like (8a) and (8b) (where any other overt realization of a for-phrase argument is unlikely) get two different frames, thus saving the syntactic and semantic difference.

(8a) He.ACT criticized him.PAT for coming.CAUS late.
(8b) He.ACT criticized his coming.PAT late.

[7] See [7]
[8] For more information on lexicalized alternations, see [8]

On the other hand, in the PDT-Vallex, these frames are often unified into a single one with an optional Cause argument, disregarding the fact that in the second case, it is hardly imaginable that another Cause argument, with the meaning of an *objective reason for judgement* should overtly appear.

5 Proposal

What we propose is that the labelling of arguments should be as uniform as possible[9] within semantically related verbs. In case of judgement verbs we find two possible variants of labelling for unification of the annotation practice.

The all-actant variant consists of ACT (*the judge*), ADDR (*the judged entity*), PAT (*the reason for judgement*), and eventually EFF (*the sentence*). Its advantages and disadvantages are listed below:

+ In the available resources for valency characteristics of the English verbs, *the reason for judgement* is considered a part of the inner argument structure of a judgement verb, disregarding its actual morphosyntactic form, the all-actant solution keeps it a part of the frame even if it is not obligatory. Since there may be different intuitions considering obligatoriness across languages, this is an advantage with respect to the task of collecting argument alignment between languages.
+ Our proposal enables us to treat uniformly all judgement verbs having both *the judged entity* and *the reason for judgement* in their argument structure. As a result, the tectogrammatical structures of parallel trees of different languages would appear more similar.
+ Such labelling enables us to treat uniformly lexicalized alternations of the type shown in (6) for individual verbs.
+ It also enables us to distinguish between the *reason for judgement* (PAT), which according to our opinion belongs into the valency frame of the judgement verbs, and the *circumstantial cause* (CAUS) which is an adjunct describing some less relevant circumstances of the situation.
– Since PAT is a semantically underspecified label, the semantics of *reason for judgement* is lost in the description.

The adjunct variant, including ACT (*the judge*), PAT (*the judged entity*), CAUS (*the reason for judgement*), and eventually EFF (*the sentence*), on the other hand, has the following implications:

+ The semantics of *the reason for judgement* stays explicit in the annotation.
– *The reason for judgement* will often be left out of the frame, or the theory must be revised in order to allow obligatory (and maybe even optional) adjuncts of the CAUS type into the frame.
– It will not be possible to maintain uniform approach to verbs of the same verb class, since *the reason for judgement* gets into the position of an affected object with some of the verbs. In a cross-linguistic comparison, this will result in mismatches and unnecessary confusion.

[9] We are aware of the fact that there are certain frame alternations that could not be unified with respect to argument labelling in the current framework, still we see a relatively large number of inconsistencies that can be repaired by clarifying the vague points in the theory.

– It will not be possible to distinguish clearly between objective and circumstantial cause of judgement.

It seems that the all-actant variant is in many respects more advantageous than the adjunct variant. Still, to make an ultimate decision, it would be necessary to make the description of argumenthood, and maybe even obligatoriness, more clear and deciding. Also, the analysis of other argument mismatches in other verb classes may help to get a more complex picture of this issue.

We would like to pin-point that our proposal does not aim at being universal or exhaustive. There may appear exceptional cases which do not fit into our description. We base our description on the assumption that judgement verbs from a class with to a great extent uniform morphosyntactic behaviour, which of course may not be the case of other verb classes.

6 Conclusion and Future Work

In this paper we have presented a cross-linguistic analysis of valency frame mismatches within one semantic class of verbs in a parallel corpus and its valency lexicons. On the example of judgement verbs, we have pointed out several weak points of the annotation rules and suggested that clarification and further specification of the theory should help in keeping the data more consistent. As an example, we have proposed a concrete way of unifying the annotation practice for the class of judgement verbs.

In the future, we would like to continue with the analysis of typical argument and frame mismatches for other argument pairs and verb classes, in order to gain a better insight into the conceptual character of argumenthood and obligatoriness.

Acknowledgements

The project has been partially supported by the grant No. GPP406/13/03351P of the Grant Agency of the Czech Republic and by the SVV project number 267 314.

This work has been using language resources developed and/or stored and/or distributed by the LINDAT-Clarin project of the Ministry of Education of the Czech Republic (project LM2010013).

References

[1] Böhmová, A., Hajič, J., Hajičová, E., and Hladká, B. (2003). The Prague Dependency Treebank. In *Treebanks*, pages 103–127. Springer.

[2] Dowty, D. (1991). Thematic Proto-Roles and Argument Selection. *Language*, 67(3):547–619.

[3] 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., and 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.

[4] 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.

[5] Hajičová, E. (2008). What We Are Talking about and What We Are Saying about It. In Gelbukh, A. F., editor, *Computational Linguistics and Intelligent Text Processing*. Springer Berlin /Heidelberg, Berlin, Heidelberg.

[6] Hajičová, E. and Sgall, P. (2003). Dependency Syntax in Functional Generative Description. *Dependenz und Valenz–Dependency and Valency*, 1:570–592.

[7] Kettnerová, V. (2012). *Lexikálně-sémantické konverze ve valenčním slovníku*. PhD thesis, Charles University, Prague, Czech Republic.

[8] Kettnerová, V., Lopatková, M., and Bejček, E. (2012). The Syntax-Semantics Interface of Czech Verbs in the Valency Lexicon. In Fjeld, R. and Torjusen, J., editors, *Proceedings of the 15th EURALEX International Congress*, pages 434–443, Oslo, Norway. Department of Linguistics and Scandinavian Studies, University of Oslo.

[9] Kingsbury, P. and Palmer, M. (2002). From Treebank to Propbank. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC-2002)*, pages 1989–1993. Citeseer.

[10] Levin, B. (1993). *English Verb Classes and Alternations*. The University of Chicago Press, Chicago and London.

[11] Lopatková, M., Kettnerová, V., Bejček, E., Skwarska, K., and Žabokrtský, Z. (2012). VALLEX 2.6.

[12] Mikulová, M., Bémová, A., Hajič, J., Hajičová, E., Havelka, J., Kolářová, V., Kučová, L., Lopatková, M., Pajas, P., Panevová, J., Ševčíková, M., Sgall, P., Štěpánek, J., Urešová, Z., Veselá, K., and Žabokrtský, Z. (2007). Annotation on the Tectogrammatical Level in the Prague Dependency Treebank. Technical Report 3.1, ÚFAL, Charles University.

[13] Ruppenhofer, J., Ellsworth, M., Petruck, M. R., Johnson, C. R., and Scheffczyk, J. (2006). Framenet II: Extended theory and practice.

[14] Šindlerová, J. and Bojar, O. (2009). Towards English-Czech Parallel Valency Lexicon via Treebank Examples. In *Eighth International Workshop on Treebanks and Linguistic Theories*, pages 185–195.

[15] Urešová, Z. (2011a). *Valence sloves v Pražském závislostním korpusu*. Studies in Computational and Theoretical Linguistics. Ústav formální a aplikované lingvistiky, Praha, Czech Republic.

[16] Urešová, Z. (2011b). *Valenční slovník Pražského závislostního korpusu (PDT-Vallex)*. Studies in Computational and Theoretical Linguistics. Ústav formální a aplikované lingvistiky, Praha, Czech Republic.

[17] Urešová, Z., Fučíková, E., Hajič, J., and Šindlerová, J. (2013). An Analysis of Annotation of Verb-Noun Idiomatic Combinations in a Parallel Dependency Corpus. To appear in the Proceedings from The 9th Workshop on Multiword Expressions, Workshop at NAACL 2013.

Determination of Czech BCT Prototypes on the Basis of Corpus Data

Tatiana Timoshchenko

Faculty of Arts, Charles University in Prague, Czech Republic

Abstract. This study sketches a corpus-driven semantic analysis of Czech BCT prototypes. In the article of Věra and Barbara Schmiedtová [5] it is argued that the following adjectives belong to Czech language center and are therefore Czech BCT: černý, bílý, červený, modrý, zelený, žlutý, šedý/šedivý, hnědý. Our purpose is to determine the prototypical meaning of this BCT categories in the sense proposed by Anna Wierzbicka [9],[10]. A practical solution in establishing the core of the colour category is the search for salient natural entities as cognitive reference points. Corpus data, e.g. citations of comparative constructions – BCT lemma[1] + jak/jako ('as') + noun or adverb + colour, e.g. červený jako řepa, slámově žlutý – indicate the possessor of the colour feature, while BCT lemma + noun in the form of Inst. (as in bílý sněhem) refers to the 'colouring property' of an entity which can be a good exemplar of the colour.

1 Corpus Characteristics

Each of the representative corpora SYN2005 and SYN2010 in the Czech National Corpus (CNK) project contains 100 million word forms (200 million altogether) taken from different types of texts: newspaper texts (33%), poetry and literature (40%) and scientific texts (27%). The texts come from the periods of 2000–2004 (SYN2005) and 2005–2009 (SYN2010). In this research we used the combination of SYN2005 and SYN2010 (named SYN2005-10) as the main source of real-language data. SYN2010 is the free online corpus, it is available at `http://korpus.cz/verejny.php`.

2 Black

For the purpose of determining of the prototypes on the basis of the corpus material three types of linguistic construction were considered: černý jako/jak ('black as') + noun (e.g. černý jako havran), adverb + černý (e.g. antracitově černý) and černý + noun (Inst) (e.g. černý prachem). The objective was to find the referents in conventionalized linguistic structures. The simile constructions are considered to be the most reliable source of the information about BCT prototypes (see studies of K. Waszakowa [8] and E. Gieroń-Czepczor [2]), the other two constructions are rather supplements to the general scheme.

Hereinafter the relevant lemma is understood as a BCT adjective cited in the form of nominative case, 1st person, Sg. All relevant citations were extracted from the SYN2005-10 corpus and manually analyzed.

[1] Lemma here is the canonical form of a set of words, for example, *černý, černou, černými* etc. are forms of the same lexeme, with *černý* as the lemma.

Table 1 presents the results for the three most frequent entities and phenomena (their lemmas) in percentage for each construction. The lemmas with the frequency less than 2 were not included in the table.

Construction	*černý* + *jako/jak* + ***noun***	***adverb*** + *černý*	*černý* + ***noun*** *(Inst)*
Number of relevant citations	428	197	16
Referents	black organic substances (*uhel/uhlí* ('coal') 34%, *smůla/smůla* ('tar') 6%) **42%**	black organic substances (*uhlově* ('coal'), *smolně* ('tar')) **65.5**%	*špína* ('dirt') **25%**
	noc, půlnoc ('night', 'midnight') **10%**	*inkoustově* ('ink') **19%**	–
	bota ('boot') **6%**	*ebenově* ('ebony') **9%**	–

Table 1. Frequencies of the referents for *černý* in SYN2005-10 corpus

Most of the referents from the SYN2005-10 corpus refer to the qualitative[2] aspect of the colour term černý. Table 1 reveals the predominance of black organic substances (used in the process of burning or being its by-product) as exemplars of blackness, with 42% and 65,5% frequency of the first two constructions.

The collocability of some referent words is partially limited. For instance in 25% of cases it is hair that is black as uhel/uhlí ('coal') and in 65% it is a black person who is black as bota ('boot').

In the similes statistics (first column), black substances are followed by noc ('night') and půlnoc ('midnight') – the sole quantitative referents in Table 1, which were subsumed to one category. Less frequent (and thus not included in the table)are the examples of quantitative blackness: mrak ('darkness'), tma ('murk'), díra ('hole'), each of them occurred once in the similes found in the Czech corpus.

Black birds such as havran (11) ('raven') and vrána (4) ('crow') are also common objects with which the BCT černý is compared. The lemma havran has a limited collocability similarly to the lemmas uhel/uhlí (see above): in 9 citations of 11 (82%) it is used to refer to human hair.

[2] Semantic analyses of achromatic colour (such as black, white and grey colours) terms differentiate between two dimensions: that of hue (or strength of colour) and that of lightness or brightness. Accordingly černý commonly functions as a term designating a type of color (qualitative aspect), as well as an extreme lack of light (quantitative aspect).

When it comes to someone who is as black as cikán(ka)/cigán (5) ('Gipsy') in Czech corpus, that may mean a person possessing black hair, dark eyes and swarthy skin or having at least one of these characteristics according to the context (see citation 1).

(1) *Já byla vždycky **černá jako cigán** a modré oči jsem zbožňovala.*
'I've always been black as a Gipsy and adored blue eyes.'

Some adverbs from adverbial citations provide an insight into negative semantics of BCT černý, for example: uhrančivě ('bewitchingly') (frequency 5), ďábelsky ('diabolically') (3), smutečně ('mournfully') (3), zlověstně ('ominously') (3)[3].

3 White

Construction	*bílý* + *jako/jak* + ***noun***	***adverb*** + *bílý*	*bílý* + ***noun*** *(Inst)*
Number of relevant citations	516	506	29
Referents	*stěna* ('wall') **17%**	dairy products (*mléčně* ('milky'), *smetanově* ('sour cream'), *tvarohově* ('curds')) **48%**	negative emotions (*hrůza* ('horror'), *strach* ('fear')) **41%**
	sníh ('snow') **15.5%**	*křídově* ('chalky') **14%**	*sníh* ('snow') **14%**
	křída ('chalk') **14.5%**	*sněhově / sněžně* ('snowy') **12%**	*bolest* ('pain') **7%**

Table 2. Frequencies of the referents for *bílý* in SYN2005-10 corpus

As the results indicate, bílý frequently denotes symptoms of intense negative emotions and the impact of negative experience reflected in facial colouration (bílý jako stěna ('white as a wall'), křídově bílý ('chalky white'), bílý hrůzou ('white with a horror')). This case may be interpreted as metonymy one's white face → white one.

The referent, whose collocability is not limited to any object class and which is at the same time relatively frequent, is sníh ('snow'): it is present in all three columns of Table 2.

Other referents with low frequency, which seem to be, nevertheless, quite interesting, are: led (1) ('ice'), jednorožec (1) ('unicorn'), polštář (1) ('pillow'), strop (1) ('ceiling'), neutralita (1) ('neutrality'): see citation 2.

[3] The frequencies of these collocates were not included in the total number of relevant citations with adverbial constructions as they don't designate any referent of yellow colour.

(2) ...žena v bílém kostýmu, ***bílém jako neutralita***, bílém jak mezistraní, jako prapor podání rukou...

'...the woman in a white dress, white as the neutrality, as the spacing, as a handshake flag... '

The quantitative dimension of whiteness appears in only two similes (with the referent světlo ('light')) and in numerous citations with adverbial construction (zářivě ('radiant') (174), oslnivě ('dazzling') (119), svítivě ('luminous') (19), třpytivě ('shimmering') (9)).

The qualitative aspect of the colour terms černý and bílý appears to be dominating over the quantitative one. In other words, the black hue is a better prototype of the black colour than the darkness for most people. The reason of this may rest in the level of categorization in our mind [3], where the black colour is more associated with the colour spectrum (as its component) rather than with the darkness. Another possible reason is that the darkness is not always impenetrable (the moon and stars give some light at night) and the light is not always purely white, so these phenomena can't be the best prototypes of the black and white colours. Bud it should be noticed that the similes in which the qualitative and quantitative aspects are intermingled sound more poetically (viz 3):

(3) ...dívka se strohým pohledem , na nohou gladiátorky a na sobě otrhané šaty, ***černé jako noc***.

'...the girl with a stern look, she had sandals and a tattered dress on, black as the night.'

4 Red

Construction	*červený* + *jako/jak* + ***noun***	***adverb*** + *červený*	*červený* + ***noun*** *(Inst)*
Number of relevant citations	92	506	17
Referents	red vegetables (*řepa* ('beet'), *rajče* ('tomato')) **18.5%**	*purpurově* ('purple') **12.5%**	emotions (*stud* ('shame'), *zlost* ('fury')) **29%**
	red flowers (*ruže* ('rose'), *pivoňka* ('peony')) **14%**	*cihlově* ('brick') **12%**	*pláč* ('weeping') / *nevyspání* ('not getting enough sleep') **12%**
	red berries (*malina* ('raspberry'), *jahoda* ('strawberry')) **10%**	*ohnivě* ('fire') **11.8%**	–

Table 3. Frequencies of the referents for *červený* in SYN2005-10 corpus

In all three constructions *červený* collocates with the names of different things. Rather frequent in the corpus are the citations of metonymic character ONE'S RED FACE → RED ONE (see also the case with *bílý* in chapter 3), for example *červený jako pivoňka* ('red as a peony') (4), *červený studem* ('red with shame') (2) etc. Thus, what makes one red is blood and emotions that are indicated by a sudden blush or reddening of the face.

In the third column of Table 3 emotions are followed by the referents *pláč* ('weeping') and *nevyspání* ('not getting enough sleep'), which have the same frequency in SYN2005-10 corpus. As the analysis of the relevant citations reveals, both referents are directly related to red eyes. That means that the collocability of the phrases *červený pláčem/nevyspáním* is limited by this word. The similar case is the simile *oči červené jako králík* ('eyes as red as a rabbit['s]') (4) – only the eyes can be "red as a rabbit's" in Czech. Of course it is not rabbit itself but its eyes that are red, the predicate (*jako králík* ***má*** ('as a rabbit **has'**)) is just skipped in this expression.

The relatively small quantity of similes with the BCT *červený* in the Czech corpus (the expected number would be a little less than the quantity of the similes with the BCT's *bílý* and *černý* and therewith higher than the quantity of ones with the BCT *zelený*) may be caused by the existence in Czech of another BCT *rudý* indicating a more saturated tinge of the red colour.

The lemma *purpurově* in the first row of the second column of Table 3 is an arguable point, because it may be taken just for a tinge of the colour *červený*, not as a prototype. Such an analysis of colour categories is open to alternative interpretations as we are dealing with conceptual phenomena, inherently fuzzy and flexible.

5 Green

Construction	*zelený* + *jako/jak* + ***noun***	***adverb*** + *zelený*	*zelený* + ***noun*** *(Inst)*
Number of relevant citations	97	381	20
Referents	greenery (*tráva* ('grass'), *listí* ('leaves'), *louka* ('meadow')) **14%**	*olivově* ('olive') **34%**	*závist* ('envy') **35%**
	moře ('sea') **9%**	*smaragdově* ('emerald') **25.5%**	negative emotions (*hrůza* ('horror'), *strach* ('fear')) **20%**
	sedma ('number seven')[4] **5%**	*brčálově* ('periwinkle') **15.5%**	greenery (*tráva* ('grass'), *osení* ('young growth') **15%**

Table 4. Frequencies of the referents for *zelený* in SYN2005-10 corpus

According to Table 4, the greenery and other natural objects prevail in the first two columns and thus can be considered as the most frequent possessors of the green colour.

Indeed the word *zelený* is etymologically derived from the word denoting vegetation [4]. Crucially, however, its semantic prototype refers to the lush greenness experienced in spring and summer, rather than the yellowish and brownish hues of autumn and winter grass and leaves.

The best colouring property (third column) belongs to negative emotional and physiological states. As this colouring property refers exclusively to human face, we deal with a similar case of (see chapters 3 and 4) metonymy ONE'S GREEN FACE → GREEN ONE (*zelený závistí / strachem, zelený jako sedma*). Since negative emotions are visible in the face and usually resemble symptoms of illness, this use of *zelený* has experiential grounding.

Several referents in the similes with the BCT *zelený* in the Czech corpus are used exclusively to describe the green colour of eyes: *Nil* ('Nile') (4), *voda* ('water') (3), *kočka* ('cat') (3) (see 4).

(4) *Ideální je, když máte navíc vlasy rudé jako oheň, pleť bílou jako sníh a oči* ***zelené jako kočka***.

'Ideal is when you also have hair red as the fire, skin white as the snow and eyes green like a cat['s]'.

The expression "*oči zelené jako kočka"* is another case of the predicate ellipsis (see chapter 4). The complete version would sound as "*oči zelené jako* ***má*** *kočka"* (literally 'eyes green as a cat **has'**).

Other examples of similes referents that sound peculiar are as follows: *víla* ('nymph') (1), *nenávist* ('hatred') (1) – a reference to the negative semantics of the BCT *zelený*, – *sen* ('dream') (1):

(5) *A zelený a hluboký záhadný plyne pod námi proud,* ***zelený jako sen*** *a hluboký jako smrt.*

'A green deep and mysterious stream flows under us, green as a dream and deep as death '.

[4] *Být zelený jako sedma* (or *být jako zelená sedma*) means to have an unhealthy pale color of the face; the etymology of the idiom comes from the card game Mariáš, the green seven is one of its playing cards.

6 Yellow

Construction	*žlutý + jako/jak +* ***noun***	***adverb*** *+ žlutý*	*žlutý +* ***noun*** *(Inst)*
Number of relevant citations	82	405	8
Referents	yellow fruit and vegetables (*citron* ('lemon'), *tykev* ('pumpkin')) **12%**	*citronově* ('lemon') **15%**	yellow plants (*lišejník* ('lichen'), *strniště* ('stubble')) **37.5%**
	yellow flowers (*šafrán* ('saffron'), *petrklíč* ('primrose')) **11%**	*kanárkově* ('canary') **13%**	–
	gold and golden things (*zlato* ('gold'), *dukátek* ('ducat')) **7%**	*slámově* ('straw') **12.5%**	–

Table 5. Frequencies of the referents for *žlutý* in SYN2005-10 corpus

What emerges from the corpus data presented in Table 5 is the diversity of referent exemplars which are based on the associations with yellow fruit, plants and gold. Thus corpus citations for *žlutý* do not indicate a clear prototypical entity, we can just state that the semantics of the most frequent referents is neutral or positive. A quite unusual simile occurred once – *žlutý jako žárlivost* ('jealousy'), which is one of the few cases of negative reference.

The positive semantics of the yellow colour evidently stem from associations with the sun; a group of collocates which corresponds to this assertion was found among the citations with adverbial constructions: *zářivě* ('brightly') (62), *svítivě* ('shiny') (17) and *oslnivě* ('dazzlingly') (9).

7 Blue

Construction	*modrý* + *jako/jak* + ***noun***	***adverb*** + *modrý*	*modrý* + ***noun*** *(Inst)*
Number of relevant citations	122	781	7
Referents	*nebe / obloha* (‘sky’) **20%**	*blankytně* / *azurově* (‘azure’) **32%**	*zima* / *chlad* (‘cold’) **57%**
	moře (‘sea’) **7%**	*nebesky* (‘sky’) **10%**	negative emotions (*zlost* (‘anger’), *vztek* (‘rage’)) **29%**
	chrpa / květ chrpy (‘cornflower’) **6%**	*ocelově* (‘steely’) / *ledově* (‘icy’) **7%**	–

Table 6. Frequencies of the referents for *modrý* in SYN2005-10 corpus

The first two columns of Table 6 show that semantics of the BCT *modrý* is primarily related to the concept of sky.

The group of similes with the water entities such as *moře* (‘sea’) (8), *voda* (‘water’) (2), *řeka* (‘river’) (1), *jezero/jezírko* (‘lake/little lake’) (2) taken together is also quite abundant.

The reason why the concept of sky is dominant over that of sea in our case may be related to the unstability of the surface of water in comparison to that of the sky (e.g. [6]), or also to the greater universality of the sky – while the sky and its colour(s) are a common feature in everyday experience of all mankind, the sea is less so for Czech people.

Referents from the third column refer to the human body or in particular to the human face via metonymy ONE’S BLUE FACE → BLUE ONE, for example *modrý zlostí* (‘blue with anger’) (see also the similar cases with other BCT’s in chapters 3, 4, 5).

8 Grey

The Czech language has two terms for grey colour: *šedý* and *šedivý,* which are etymologically related [4]. These two words are synonyms though not complete, because in some contexts they are not interchangeable (e.g. *šedá eminence* ('grey eminence'), *šedivá teorie* ('grey / insipid theory')).

Construction	*šedý + jako/jak +* ***noun***	***adverb*** *+ šedý*	*šedý +* ***noun*** *(Inst)*
Number of relevant citations	55	377	14
Referents	*popel* ('ashes') **11%**	*ocelově* ('steely') **35.5%**	*prach* ('dust') **14%**
	myš ('mouse') **9%**	*popelavě* ('ashes') **16%**	–
	prach ('dust') / *kámen* ('stone') **5.5%**	*stříbřitě* / *stříbrně* ('silvery') **14%**	–

Table 7. Frequencies of the referents for *šedý* in SYN2005-10 corpus

Construction	*šedivý + jako/jak +* ***noun***	***adverb*** *+ šedivý*	*šedivý +* ***noun*** *(Inst)*
Number of relevant citations	32	28	13
Referents	*popel* ('ashes') **12.5%**	*popelavě* ('ashy') / *ocelově* ('steely') / *stříbřitě* / *stříbrně* ('silvery') **21%**	*prach* ('dust') **23%**
	myš ('mouse') / *mrak* ('cloud') **6%**	*kovově* ('metallic') **18%**	*hrůza* ('horror') **15%**
	–	*perlově* ('pearly') **14%**	–

Table 8. Frequencies of the referents for *šedivý* in SYN2005-10 corpus

The frequency of similes and instrumental constructions with the BCT *šedý* is a little higher than such with BCT *šedivý*, but the difference in the number of adverbial constructions with these two BCT's is more essential. This serves as an evidence of the preference of BCT *šedý* in contemporary usage (according to SYN2005-10 corpus).

If we compare the occurrences of the lemmas *šedý* and *šedivý* in SYN2005-10 corpus, it has to be acknowledged that there is no major divergence between their referents. The lemmas *popel* ('ashes') and *myš* ('mouse') seem to be the main prototypes of these BCT's. The adverbs *popelavě* ('ashes'), *ocelově* ('steely') and *stříbřitě* / *stříbrně* ('silvery'), which are allocated to the three different rows of Table 7 in concordance with the frequency, in Table 8 both occupied the first row because of the equal number of occurrences. Table 8 contains the lemma *hrůza* ('horror') in the second row of the third column. Such a location indicates a certain colouring property of horror in Czech, so we deal with a case of metonymy related to negative emotions again (see also chapters 3, 4, 5 and 7): ONE'S GREY FACE → GREY ONE (in this case *šedivý* means *pale*).

Among the adverbial citations there were several constructions with semantically negative key words, e.g., *nudně* ('boring') (3), *zoufale* ('desperately') (2), *ponuře* ('gloomy') (2) etc. for *šedivý* and *mrtvolně* ('cadaverous') (6), *jednotvárně* ('unvariedly') (5) for *šedý*. These citations were not taken into account as not relevant, but they give us important information about semantic structure of the BCT's *šedý* and *šedivý*.

The data presented above indicate the similarity of the perception and semantics of *šedý* and *šedivý* that is caused by the etymological connection between them.

9 Brown

Construction	*hnědý* + *jako/jak* + ***noun***	***adverb*** + *hnědý*	*hnědý* + ***noun*** *(Inst)*
Number of relevant citations	35	473	4
Referents	*skořápka* ('shell') / *kmen* ('trunk') **6%**	*kaštanově* ('chestnut') **26%**	*bahno* ('mud') **50%**
	–	*oříškově* / *ořechově* ('(hazel)nut') **17.5%**	–
	–	*čokoládově* ('chokolate') **15%**	–

Table 9. Frequencies of the referents for *hnědý* in SYN2005-10 corpus

SYN2005-10 corpus does not render evidence of high frequency of similes or instrumental constructions for BCT *hnědý*. The highest frequency of such construction does not overcome 2 occurrences, which is certainly not enough for objective analysis. The exemplars of adverbial construction are on the contrary quite abundant and we should thus rely on the second column of Table 9.

The relatively low frequency of the citations with *hnědý* may be explained by the fact, that this BCT is likely to be the „youngest" in Czech [7] (among the colour terms that have been reviewed here), and therefore it is not quite conceived as a basic colour in the colour system of this language and does not have any specific prototype. The

referents, which occured just once in the corpus, are diverse: *čokoláda* ('chocolate'), *vopice* ('monkey'), *skála* ('rock'), *metla* ('broom'), *hlína* ('clay') and many others.

10 Adverbial Construction

Corpus data presented in Table 9 revealed a contrastively higher frequency of adverbial construction as opposed to other types of analyzed syntagmas. This construction also clearly prevails in corpus statistics of almost all Czech BCT's discussed above (except the colour categories *bílý, černý* and *šedivý*). Further analysis shows that this contrast appears to become even more remarkable as we go up the hierarchy proposed by Berlin and Kay (black and white, red, green and yellow, blue, brown, grey & rose & orange & purple[5]) [1] with some exceptions – see Figure 1.

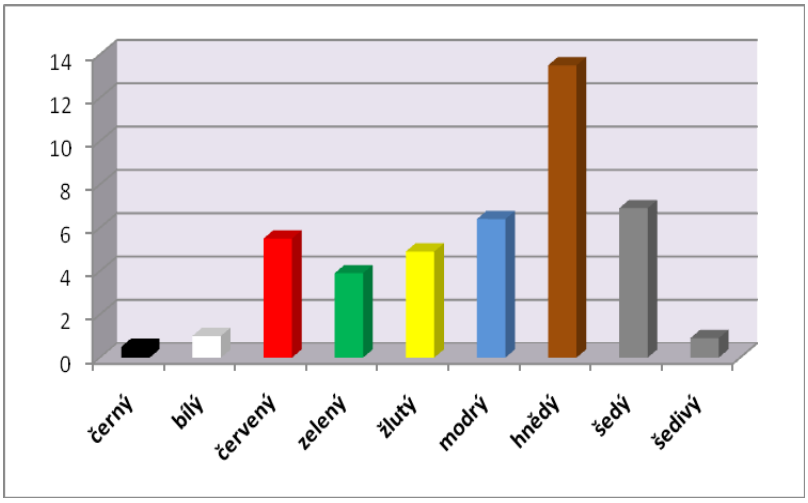


Fig. 1. The extent of prevailing of adverbial construction against the comparative one in the referent statistics for each colour term[6]

The extent of such a contrast is likely to give us the important information about the "age" of a certain colour term in Czech: the higher is the index, the younger is the relevant colour term. The BCT's *červený*, *hnědý* and *šedivý* don't fit this pattern. The BCT *hnědý*, for instance, should have been placed at the right end of the scale to make the rise even (and in this way it would deviate from the Berlin and Kay hierarchy). According to the thesis of I.Vaňková, this colour term appeared in Slavic languages later than other BCT's [7] and therefore should be placed after the term *šedý* in BCT hierarchy of these languages. The BCT *červený* seems to be another exception to the

[5] This is the chronological ordering, suggesting that the colour terms are added to the lexicon in a constrained order; the younger the certain colour term is, the closer to the right end of this hierarchy it is placed.

[6] To calculate this index the number of adverbial constructions with the certain BCT extracted from SYN2005-10 corpus was divided by the relevant number of similes.

Berlin and Kay's order, because its index is higher than ones of *zelený* and *žlutý*. But this is just the evidence of the fact, that this colour term is younger than the latter two, as opposed to *rudý*, its partial synonym, which is sometimes regarded as another Czech BCT. *Šedivý* is thus, along with *černý* and *bílý*, one of the oldest Czech colour terms.

That means, people prefer using the adverbial constructions with the younger colour terms and the similes with the older ones. The question why it is so remains open to discussion. This fact may be caused by the intention of a person to rather specify the tinge of the colour by means of its potential prototype, if there are any established ones associated with the respective colour term available[7]. The adverbial construction seems to be more suitable for such a purpose, and the simile helps to emphasize the intensity of the colour through the common prototype.

11 Conclusion

Linguistic corpora satisfy the demands of contextual investigation by providing large samples of citations and thus allowing objective research. For the purpose of determining BCT's prototypes based on corpus material the three types of linguistic constructions were considered in this study: the similes (for ex. *černý jako uhel* ('black as a coal')), adverbial (*křídově bílý* ('chalky white')) and instrumental (*šedý prachem* ('grey by dust')) constructions. An interesting distribution revealed itself in the course of investigation: in similes the BCT prototypes per se are exposed evenly, adverbial constructions analysis is efficient for description of the semantic field of BCT's and finally instrumental constructions turned out to be the source of conceptual metonymy. The BCT's are often used in metonymic descriptions of human emotional state, for instance the phrase *zelený závistí / strachem* ('green by envy / fear') is based on metonymy ONE'S GREEN FACE → GREEN ONE.

The extent of prevailing of adverbial construction as opposed to the comparative one in the referent statistics can provide the important information about the "age" of a certain colour term: the higher is the index, the younger is the relevant colour term. For instance, the colour term *červený* appeared to be younger than the terms *zelený* and *žlutý* and the term *šedivý* is apparently standing side by side with the oldest ones *černý* and *bílý* as per the time of their first occurrence in Czech.

Acknowledgements

This thesis was written in the frames of the Programme for the Developmentof Fields of Study at Charles University, No. *P11 Czech national corpus, sub-programme Czech national corpus.*

[7] The adverbial constructions, which indicate pure tinge of a colour (e.g. *matně zelený* 'dull green'), were not taken into account during the statistic analysis in this study.

References

[1] Berlin, B. and Kay, P. (1969). *Basic Color Terms. Their Universality and Evolution.* University of California Press, Berkeley, CA.

[2] Gieroń-Czepczor, E. (2011). *A corpus-based cognitive-semantic analysis of the primary basic colour terms in English and Polish*. Państwowa Wysza Szkola Zawodowa, Racibórz.

[3] Lakoff, G. (1987). *Women, Fire, and Dangerous Things. What Categories Reveal about the Mind.* The University of Chicago Press, Chicago, London.

[4] Machek, V. (1957). *Etymologický slovník jazyka českého a slovenského.* Nakladatelství Československé akademie věd, Praha.

[5] Schmiedtová, V. and Schmiedtová, B. (2006). Určení jazykové základovosti barev v Českém národním korpusu. In Čermák, F. and Blatná, R., editors, *Korpusová lingvistika: Stav a modelové přístupy*, pages 285–313, Nakladatelství Lidové noviny, Praha.

[6] Tokarski, R. (2004). *Semantyka barw we współczesnej polszczyźnie*. Wydawnictwo Uniwersytetu Marii Curie-Skłodowskiej, Lublin.

[7] Vaňková, I. (2005). Kapitoly o barvách. In *Vaňková I., Nebeská I., Saicová Římalová L., and Šlédrová J.: Co na srdci, to na jazyku*, pages 195–246, Nakladatelství Karolinum, Praha.

[8] Waszakowa, K. (2000). Struktura znaczeniowa podstawowych nazw barw. Założenia opisu porównawczego. In Grzegorczykowa, R. and Waszakowa, K., editors, *Studia z semantyki porównawczej. Nazwy barw, nazwy wymiarów, predykaty mentalne*. Część I, pages 59–72, Wydawnictwa Uniwersytetu Warszawskiego, Warszawa.

[9] Wierzbicka, A. (1990). The meaning of color terms: Semantics, culture, and cognition. *Cognitive Linguistics*, 1:99–150.

[10] Wierzbiscka, A. (1996). *Semantics: Primes and Universals*. Oxford University Press, New York, Oxford.

Veni, Vidi, Vici: The Language Technology Infrastructure Landscape after CESAR

Tamás Váradi

Research Institute for Linguistics, Hungarian Academy of Sciences, Budapest, Hungary

1 The Rationale and Origins

The present paper intends to give an overview of the CESAR Project, focusing on its impact, in particular. However, before we can come to evaluate what the language technology landscape was after the CESAR project finished, it is necessary to review what it was like before CESAR and what were the main objectives of the project. The ultimate mission of the CESAR project is to bolster the language technology capabilities of the participating countries and make them available through a Europe-wide distribution network.

The development of language technologies in the region has been going on in a fragmented, isolated manner, mostly out of self-supported initiatives by a few research centres. The current initiative is an excellent opportunity to boost language technologies in the region by systematic and comprehensive enhancement of existing language resources and language technology tools. In that sense it is the long-awaited continuation of earlier efforts at consolidating language technology infrastructure such as TELRI and CLARIN.

Language technology vitally requires vast amount of linguistically analysed datasets and tools and it is this acute need for usable, useful and easily available resources and tools for language technologies that motivated the call for application to build language technology infrastructures in a pan-European coordinated manner. That was how the CESAR Project was born along with two other regional and typological consortium to form, along with the initiators, the T4ME consortium, ending in what is now known as the META-NET network of excellence.

2 The CESAR Consortium

CESAR stands for Central and South Slavic Resources. The Geolinguistic spread of the consortium is shown in Figure 1. The consortium consists of nine partners covering six languages, five of them belong to the Slavic language family, while Hungarian belongs to the Finno-Ugric family. The number of speakers living inside the home country and across the Globe is given in Table 1. The CESAR languages represent million speakers around the world but individually with the exception of Poland they represent relatively small languages.

Fig. 1. The Geolinguistic spread of CESAR

From a language technology point of view they all belong to the so-called less-resourced languages with weak or no support for language technology. Indeed, in comparison with most of the other languages within META-NET, language technology in the respective CESAR countries is less advanced than in other parts of Europe. One of the specific objectives of the CESAR project was to help bridge the technological gap that exists between this part of Europe and the rest.

Language	Population inside homeland	Population globally
Polish	38 M	40–48 M
Slovak	5.4 M	7 M
Hungarian	10 M	16 M
Croatian	4.4 M	5.5 M
Serbian	7.3 M	9 M
Bulgarian	7.5 M	9 M

Table 1. Speaker population per language in CESAR

The partners in the Consortium are listed in Table 2. They number one, or at most two, per languages as was required by the Call. However, they are expected to represent not only themselves alone but act for the whole of the language technology community in their respective languages. They are the leading players in language technology of their language, typically Academy institutions or university departments with a track record of community building outreach within their languages.

The META-NET initiative was aimed to help the newly joined counties and in this respect the CESAR consortium can even claim to be ahead of its time in that it already

included Croatia (which joined the EU this summer) and Serbia, which aspires to be a member soon.

Partners in the consortium are not unknown to each other, in fact, they are tested and tried partners in a string of earlier EU funded projects (notably, CLARIN, MULTEXT-EAST, CONCEDE) going back to the TELRI project, which had similar objectives.

Participant no.	Participant organisation name	Participant short name	Country
1 (CO)	Nyelvtudományi Intézet, Magyar Tudományos Akadémia	HASRIL	Hungary
2	Budapesti Műszaki és Gazdaságtudományi Egyetem	BME-TMIT	Hungary
3	Sveučilište u Zagrebu, Filozofski Fakultet – University of Zagreb, Faculty of Humanities and Social Sciences	FFZG	Croatia
4	Instytut Podstaw Informatyki Polskej Akademii Nauk	IPIPAN	Poland
5	Uniwersytet Lodzki	Ulodz	Poland
6	Faculty of Mathematics, University of Belgrade	UBG	Serbia
7	Institut Mihajlo Pupin	IPUP	Serbia
8	The Institute for Bulgarian Language Prof. Lyubomir Andreychin	IBL	Bulgaria
9	Jazykovedny Ústav Ludovíta Stúra Slovenskej Akadémie Vied	LSIL	Slovakia

Table 2. Partners of the CESAR Consortium

3 The Objectives of the Project

3.1 Building an Open Linguistic Infrastructure

The central objective of the project was to produce and make available a comprehensive set of language resources and tools covering Bulgarian, Croatian, Hungarian, Polish, Serbian and Slovak. The key points of CESAR activity with respect to resources were seen not so much in the creation of new resources but rather the enhancement of resources and tools (in size, coverage, precision, recall, accuracy), the adaptation of resources and tools to become compliant with the agreed standards for interoperability as well as the upgrade of resources and tools by combining them with other resources and tools in order to achieve the foreseen level of interoperability and in adapting user-interfaces to fulfill user requirements. A special effort was taken to achieve a common standard of involved resources and tools in order to enhance and facilitate the foreseen interoperability between them, as well as to evaluate their license schemes and IPR issues (for a more detailed account of CESAR see [8] and [9]).

The resources and tools that emanated from the CESAR project were made available to META-SHARE, a pan-European distribution facility which is designed to become an important component of a language technology marketplace for HLT researchers and developers, language professionals (translators, interpreters, content and software localisation experts, etc.), as well as for industrial players, especially SMEs, catering for the

full development cycle of HLT, from research through to innovative products and services.

The resources covered by the CESAR project were delivered to the LT community at six month intervals in three batches and at the end of the project they were made available through the open digital exchange META-SHARE. Table 1. presents the statistics of these resources by partner, language and resource type.

	RILHAS	TMIT	FFZG	IPIPAN	ULODZ	UGB	PUPIN	IBL	LSIL	∑
Tools/ Services	6	3	5	19	5	6	0	16	6	66
Corpora	19	21	12	17	11	10	0	9	21	120
Lexical/ Conceptual resources	6	1	9	23	1	3	2	11	9	65
Total	**31**	**25**	**26**	**59**	**17**	**19**	**2**	**36**	**36**	**251**

Table 3. Language resources and tools submitted to META-SHARE by CESAR

Activities with the resources

As Table 3 reports, corpora as language resource were in the main focus of CESAR. Actions performed on resources consisted of three main categories: upgrading, extending as well as linking and aligning across languages. The main categories are covering a wide range of activities as listed below:

- Upgrading resources to agreed standards involving the following actions:
  - upgrade for interoperability (changing annotation format, type, tagset),
  - technology-related upgrade (wrapping, refactoring, etc.),
  - application of techniques of finding inconsistencies and errors in (automatically and/or manually built) linguistic resources, including corpora and lexica,
  - metadata-related work (creation, enhancement, conversion, standardization),
  - harmonization of documentation (conversion to open formats, reformatting, linking),
  - preparation for maintenance and deployment (debugging, cleaning, building test environments, preparing code repositories),
  - other programming tasks (e.g. standardizing API calls),
  - IPR issues.
- Extending and linking resources included the following activities:
  - adding new portions of data, enhancement of resources, interlinking resources,
  - linking existing resources across different sources,

  - providing building blocks to the existing tools (e.g. extended grammars to existing shallow parsers),
  - major restructuring,
  - integration of additional resources with existing ones to improve the quality of resulting resources.
- Aligning resources across languages consisted of the following actions:
  - introducing language-neutrality,
  - introducing cross-linguality,
  - mapping between tagsets,
  - mapping between outputs and inputs of linguistic tools for particular language,
  - synchronization of resources available for consortium languages,
  - extension of language models to embrace cross-linguality and/or promote language independence.

NooJ

A speciality of the CESAR project among the META-NET projects was its concern for the finite-state general linguistic development tool, NooJ (`http://nooj4nlp.net`). Within CESAR the Bulgarian, Croatian, Hungarian and Serbian partners had already developed substantial language resources for this platform and NooJ resources existed already for Italian, German, French, English, Spanish, Portuguese as well, to mention only the rest of the META-NET languages. However, despite the fact that there was a vibrant NooJ community actively developing valuable resources, the propagation of NooJ was faced with a difficulty due to it being limited to the .Net framework and not being open source.

The CESAR project was an excellent opportunity to remove both limitations and it was squarely in line with the mission of CESAR to undertake the task of making NooJ cross-platform (by converting the NooJ code into Java) and also to publish its source code.

The technical work was carried out by the Institute Pupin in Belgrade under the supervision of Max Silberztein, the author of original NooJ. During the project two platform independent versions of NooJ were developed: MONO and Java version of NooJ. The Java version of NooJ is also made available as open source software through META-SHARE. This landmark event opened the way for experimenting with combining NooJ with similar finite-state platforms, notably the Helsinki FS toolkit (see [3]).

IPR principles and legal issues

Promoting the use of open data and following the Creative Commons and Open Data Commons principles were in the focus of the project. This was followed by cleaning IPR for all resources involved in upload batches and arrange deposition agreements for resources coming from outside the consortium.

Activities carried out in the project were followed by META-NET wide negotiations and discussions on IPR in general and on the proposed license templates. One of the

main IPR principles was to promote the use of CC or MS (META-SHARE) licenses, considerably reduce the dominance of CLARIN licenses within CESAR (especially elaboration and checking of MS-NoRedistribution license template family). This intention was always carried out together with the idea of applying the most appropriate licence, one that was as open as possible out of the set of templates.

Resources involved were made compliant with the legal principles and provisions established and/or completed/amended by the consortium and accepted by the respective right holders.

In defining their policies on IPR CESAR was collaborating with other LRT projects that were solving their IPR issues in parallel to CESAR activities (e.g. CLARIN, ACCURAT, LetsMT! etc.).

META-SHARE

META-SHARE (`www.meta-share.org`) is an open distributed facility for the sharing and exchange of resources provided by members of META-NET. It serves as a sustainable network of repositories of language data, tools and related web services documented with high-quality metadata. Servers linked to META-SHARE form a chain of nodes, offering open access to resources for their users and editing and administrative interfaces for resource owners. META-SHARE repositories contain resource descriptions in the form of metadata conformant to the metadata model prepared by the team of META-NET with the help of CESAR project (see [5], [7]).

At the end of the project META-SHARE nodes were organized into a hierarchical structure: managing nodes are synchronized, and provide all META-SHARE metadata and resources, whilst network nodes are not synchronized (but maintained by the hosting institutes, partners), but harvested by a managing node.

After the META-SHARE server software became available in full functionality (including synchronization), CESAR partners decided to promote the original Warsaw node (hosted and maintained by IPIPAN) to become the CESAR managing node, and set up a network node at each partner's premises to provide metadata for harvesting by the CESAR managing node, which shares metadata with other META-SHARE managing nodes.

CESAR META-SHARE nodes are set up for long-term provision of the selected resources. CESAR-partners are committed to hosting and making available the selected language resources and maintaining the repository of LRs for at least 24 months after the termination of the project. Within this activity all partners are committed to giving user-support, software-based and/or human services and members of the IPIPAN team will assist the META-SHARE software development team.

At the end of the project at least one node was created for each language. 8 partners out of 9 run their own META-SHARE node, and are committed to running them beyond the end of the project were set up:

HASRIL	http://metashare.nytud.hu/
BME-TMIT	http://metashare.tmit.bme.hu/
FFZG	http://meta-share.ffzg.hr/
IPIPAN	**http://nlp.ipipan.waw.pl/metashare/**
ULodz	http://metashare.ia.uni.lodz.pl/
IBL	http://metashare.ibl.bas.bg/
UBG	http://meta-share.matf.bg.ac.rs
LSIL	https://metashare.korpus.sk/

Table 4. META-SHARE nodes in CESAR

3.2 Roadmapping

In addition to the main activities of the CESAR (enhancement of selected language resources and tools), partners carried out several activities in order to strengthen the position of carefully selected LRs.

One of the main pillars of CESAR was to map the LT community landscape and situation of the countries involved in CESAR.

The Language White Paper Series

The series of Language White Papers (LWPs) give state-of-art overview of 30 languages from the perspective of language technology community support (for more on LWP see http://www.meta-net.eu/whitepapers/overview). The core of the series are findings and general facts about the respective language, including features of orthography, morphology (rich inflectional and derivational systems; aspectual verb pairs), syntax (pro-drop, relatively free word order) that may impede the development of the LT tools and resources and the current language technology support for the respective languages. LWP gives a brief account on the LT application architectures that consist of several components to mirror different aspects of language. Second, it covers the situation in the LT research and education. Third, it concludes with an overview of the past and ongoing research programs in universities and institutions.

The White Papers shed light on the LT in Bulgarian, Croatian, Hungarian, Polish, Serbian and Slovak education including subjects and curricula, mostly at university level. Further, the White Papers mention in brief the various programs and initiatives that fund the development of the LT tools and resources for languages in question.

The documents show the results of a survey of the state-of-art of LT tools and resources for Bulgarian, Croatian, Hungarian, Polish, Serbian and Slovak in comparison with other languages covered by the META-NET initiative analysing criteria such as quantity, availability, quality, coverage, maturity, sustainability, and adaptability.

The function of the language report was to document the language community landscape for a given language community by volume and the whole Europe by the whole series. In connection with the LWPs partners identified relevant researchers and projects, policy makers, industry representatives, language communities and additional stakeholders (the elaborated list of relevant stakeholders was continuously updated during the project period).

The Language White Papers were produced as a result of close collaboration within the META-NET alliance. They were intended for distribution among stakeholders (industry, government, research community) with the aim of raising awareness on language technology – especially in countries where language technology has a weak position in various decision makers' domains.

LWPs were used extensively by the project partners to disseminate information about META-NET and CESAR at the national level to different stakeholders, primarily through the set of CESAR roadshows, one-day high-level events each dedicated to one language. These events represented an ideal opportunity to spread the LWPs as widely as possible. Also, the remaining LWPs will be used at the national and regional events to disseminate information about the META-NET, CESAR and the centres and nodes that will be functioning after the projects ended.

During the work on LWPs a joint stakeholders contact database from the consortium countries was collected and made available to all consortium partners as well as to other META-NET partners. This database covers individuals (experts), institutions (research, national-funding agencies, government) and companies (producers and important users) dealing with LRT. This database is used primarily for dissemination purposes, but it remains available for other purposes as well.

Each volume of the Language White Papers discussed the following topic for the respective language:

- *Language community*: number of speakers worldwide, number of web pages in that language, other relevant quantitative elements,
- *Role of the language in the respective country/language community*: legal framework; institutional communication and local administration; media; press etc.
- *Research community*: estimated size of the research community in the areas of NLP and ST; main gaps e.g. underdeveloped human or technical resources; activities at national level, their relevance for addressing the identified gaps.
- *Language service industry*: qualitative and quantitative analysis of the local translation, localization and interpretation industries.
- *Language technology industry*: qualitative and quantitative analysis of the local industrial landscape; estimated number of vendors and developers (companies as well as individuals); a description of the main existing LT products and services, and of their actual or potential users (public at large, business/professional users).
- *Policy makers*: list of policy makers in the respective countries.
- *Demand side*: role of language-technology products and services within the Internet, digital media and telecommunications sectors, by ways of examples.
- *Legal provisions*: national intellectual-property and digital-copyright regulations related to language resources i.e. databases and software.
- *Various types of users*: analysis of the needs of different types of users – from individual users to large multinational.
- *Contacts information*: (i.e., name, phone number, affiliation, email and postal address) on the international and especially on the national level of

representatives of the following stakeholder types: research, politics, administration, funding agencies, LT user industries, LT provider industries, journalists, language communities.

3.3 Dissemination Efforts

One of the special features of the CESAR projects was the amount of resources it dedicated to its dissemination efforts. Dissemination was interpreted not just in the usual sense as serving the purposes of communicating the results of the project. In addition to this, it was dedicated to serving one of the central mission of the META-NET projects, i.e. to raise awareness of the potential utility and indispensability of language technology in achieving the goals of the Digital Agenda (see [6]).

A detailed action plan for outreach, awareness and sustainability, that was developed at the beginning of the project, detailing awareness, community mobilisation and dissemination actions to be undertaken in each country covered by the project. It was updated in M12 to accommodate the harmonisation with the META-NET activities at large and to maximise the project's impact and ensuring its sustainability beyond the EU-supported phase.

While in the first year the action plan was focused on the first of the three main target groups, i.e. on the research community in human language technology and other related domains, in the second year the action plan and dissemination activities were oriented towards the other two target groups, i.e. industry (both language industry and other business sectors) and society (government and other public decision makers, as well as general public). The planned activities were focussed on target groups and dissemination channels (visual identity, dissemination by public appearance, media appearance: printed and electronic media).

The analysis of users' needs served as valuable input that ensured that was action plan was tailored to the users' needs as much as possible.

The research community at national level was reached by several dissemination channels (traditional or not so traditional) in order to attract the relevant players to participate in sharing resources and tools through META-SHARE platform. Also collaborative participation with national CESAR partners at different national and international conferences was favoured since this helped players at the national level and outside of CESAR consortium to present their work that might otherwise remain unseen. In this respect the CESAR project partners played the role of catalyst in transferring the information about different LT players from the national level to the European and global level. For that purpose language resources and tools from outside consortium were highly regarded and also represented a proof of good mobilisation activities at national level. Also, the role of CESAR consortium members was to convey information about he META-NET from European to the national level using defined dissemination channels.

3.4 Awareness Raising in Society and Government

Beside the defined dissemination channels that were used for both society and government awareness raising as usual, several awareness-raising activities were used at governmental/funding organisations/language councils levels – in the form of a series of

road-shows, one-day dissemination events where the CESAR project was presented at national ICT-PSP days, or at national science festivals, science days etc.).

In addition, the project was presented at various high-profile events, notably at LREC and at COLING [7]. The strategy behind the road-shows was based on the finding that experts from abroad presenting the same topic to decision makers and funders at national level usually attract far more attention than domestic experts. This was the strategy that we followed in order to make stronger impact, raise awareness and visibility at governmental and funding agencies level and thus help local experts in gaining support for the LT field from national funding. These events were organized by local organizers, i.e. consortium member institutions from Bulgaria, Slovakia, Poland, Serbia, Croatia and Hungary that play a key role in LRT at their national level.

Table 5 displays the logistic details of the series of CESAR Road-shows

Bulgaria	2nd May 2012, Sofia
Slovakia	07/08th June 2012, Bratislava
Poland	27/28th September 2012, Warsaw
Serbia	29th October 2012, Belgrade
Croatia	30th November 2012, Zagreb
Hungary	18th January 2012, Budapest

Table 5. Location and Date of the CESAR Roadshows

The events presented an excellent opportunity to showcase the results of the projects in the form of presentations and demos, as well as posters. In parallel with the presentation sessions LT products were exhibited and demonstrated by industrial partners, sponsors, and research projects at both, national and international level. Invited speakers included not only researchers of the CESAR and similar projects, but also stakeholders and politicians from the national and international pole. The database of stakeholders previously compiled proved very useful in reaching prominent members of the target audience. In some cases notably the Hungarian road-show, the event attracted a number of high-level high government officials (in the case of the Hungarian road-show, no less than five junior government ministers!), who either addressed the participants or acted as patrons of these events.

A brief statistics of the road-shows is presented as follows:

Road-shows	No. of presentations	No. of posters	No. of demos	No. of news on road-shows
HU	12	13	8	42
PL	23	24	12	8
SK	21	4	2	7
SR	12	4	6	35
BG	12	6	4	52
HR	13	4	5	10
	93	55	37	154

Table 6. Some key indicators of the Road-shows

4 Gaps and Progress in LRs of CESAR Languages

In 2011 a study entitled 'Detecting Gaps in Language Resources and Tools in the Project CESAR [11] was prepared on gaps and progresses of CESAR resources based on the categories set by META-NET in the series of Language White Papers. The analysis presented evaluations of resources on a finer set of types than those summarised in the Language White Paper series.

CESAR languages resources	Bulgarian	Croatian	Hungarian	Polish	Serbian	Slovak	Overall average
1. Reference Corpora	4.71	3.29	5.71	3.71	3.43	3.856	4.12
2. Syntax-Corpora (treebanks. dependency banks)	2.14	2.00	4.86	2.86	**0.00**	2.43	2.38
3. Semantics-Corpora	3.43	0.00	4.14	1.86	**0.00**	**0.00**	1.57
4. Discourse-Corpora	1.43	**0.00**	**0.00**	1.14	**0.00**	1.86	0.74
5. Parallel Corpora. Translation Memories	2.43	2.43	5.71	3.86	2.57	2.29	3.21
6. Speech-Corpora (raw speech data. labelled/annotated speech data. speech dialogue data)	2.29	3.00	2.57	1.86	2.86	2.86	2.57
7. Multimedia and multimodal data (text data combined with audio/video)	1.00	2.57	0.57	0.71	1.57	2.14	1.43
8. Language Models	1.57	**0.00**	4.71	1.29	2.29	2.71	2.10
9. Lexicons. Terminologies	3.57	3.29	4.00	3.29	3.14	3.14	3.40
10. Grammars	2.57	**0.00**	4.29	2.86	0.71	2.00	2.07
11. Thesauri. WordNets	4.00	2.71	3.43	3.71	3.00	2.86	3.29
12. Ontological Resources for World Knowledge (e.g. upper models, linked data)	2.00	**0.00**	2.43	1.86	0.71	**0.00**	1.17

Table 7. Evaluation of resources by type and language

The results displayed in Table 7 show that for half of the categories, there is at least one language with a score of zero, on a scale of zero to six. There are two categories, semantics corpora and discourse corpora, where half of the languages have zero values. The least resourced category was discource corpora, with an average value of less than 1.0 across the languages, and there were three categories where the average score did

not reach 2.0. The fine-grained analysis duly reflected the areas where the CESAR consortium had excellent achievements, namely, Reference Corpora, Parallel Corpora, Lexicons/Terminologies, Thesauri/Wordnets.

CESAR Language Technology (Tools, Technologies, Applications)	Bulgarian	Croatian	Hungarian	Polish	Serbian	Slovak	Overall average
1. Tokenization. Morphology (tokenization, POS tagging, morphological analysis/generation)	4.00	3.57	4.00	4.57	4.29	3.00	3.90
2. Parsing (shallow or deep syntactic analysis)	3.00	1.57	3.57	3.57	2.423	**0.00**	2.36
3. Sentence Semantics (WSD, argument structure semantic roles)	2.43	1.14	1.57	2.14	**0.00**	**0.00**	**1.21**
4. Text Semantics (coreference resolution, context, pragmatics, inference)	1.43	**0.00**	1.27	1.00	**0.00**	**0.00**	**0.62**
5. Advanced Discourse Processing (text structure, coherence, rhetorical structure etc.	**0.00**	**0.00**	**0.00**	0.57	**0.00**	**0.00**	**0.10**
6. Information Retrieval (text indexing, multimedia IR, crosslingual IR)	2.00	2.29	0.86	3.29	2.43	2.29	2.19
7. Information Extraction (named entity recognition, event/relation, extraction. opinion/sentiment recognition)	2.29	2.43	5.57	2.57	2.14	1.71	2.79
8. Language Generation (sentence generation, report generation, text generation)	1.43	1.29	**0.00**	1.14	**0.00**	**0.00**	**0.64**
9. Summarization, Question Answering, advanced information access technologies	1.86	0.29	**0.00**	1.29	0.71	1.71	**0.98**
10. Machine Translation	2.29	0.71	4.88	3.29	0.71	1.86	2.29
11. Speech Recognition	2.00	2.57	2.71	2.71	1.14	2.29	2.24
12. Speech Synthesis	2.00	3.57	3.71	4.14	3.29	3.00	3.29
13. Dialogue Management (dialogue capabilities and user modelling)	**0.00**	1.29	**0.00**	1.00	**0.00**	**0.00**	**0.38**

Table 8. Evaluation of tools and services

The analysis for tools and services yielded similar results. Here the picture is slightly worse even, in that, as shown in Table 8, in 7 of 13 categories at least one language has score 0.00 and in 5 categories the average score is less than 1.00. There is one category, advanced discourse processing, where in five out of the six language there does not exist any tool or service, and in dialogue management four languages have no tools whereas

the categories text semantics and language generation lack tools/services in half of the languages.

The two tables described above represented the initial stage of resources in the CESAR project. When the project started the initial list of tools and resources potential available for enhancement numbered 130 in total. When the project finished, CESAR partners delivered no fewer than 251 resources and tools (120 corpora, 65 lexical/conceptual resources and 66 tools/services), which, even in terms of share numbers, represent an impressive progress. While lack of time alone prevented a comprehensive and systematic repair of the detected gaps in resources, it can be confidently stated that the overall level provision of key resources for CESAR languages became higher. The most valuable progress was achieved in the field of multilingual corpora and national corpora (partners made extensive progress with enlargement and development of their national corpus).

A special effort was made to make resources and tools language independent and to support cross-linguality both in the field of resources and tools. The next section on cross-lingual alignment reports on work that was generated by the synergies created by the project objectives within the consortium.

In the field of tools, morphological analysers, lemmatisers and speech synthesis were the areas where remarkable progress was made in all covered languages. Beyond doubt, the tool that benefited the most and is likely to make the biggest impact for it, was NooJ, which has been given a new lease of life not only in terms of increased potential user base because of its platform independence and open-source status but also for the enhanced resources developed for the NooJ platform.

It is also worth mentioning that progress was not only in quantitative (extension, upgrade, new resources and tools), but also qualitative. In this respect the intellectual property right clearance and carefully prepared and detailed resource metadata should be recalled. All delivered resources were equipped with standard metadata (listed in the META-SHARE repository) and were supplied with proper license.

5 Cross-lingual Alignment for CESAR Languages

Although the CESAR activities were mainly targeted to the development for language resources and tools for individual languages, one of the technical challenges of the project was the alignment and cross-linking of resources and tools available for the consortium languages. As a result, the following resources were created:

- ***Automatic Collocation Dictionaries*** produced for all project languages on the basis of new Sketch Grammars developed for CESAR languages. The Automatic Collocations Dictionaries were produced in collaboration with Lexical Computing Ltd., based on the web corpora and word sketches of the respective languages. The entry for each collocation includes pointers to its corpus examples on the Sketch Engine website.
- ***NERosetta***: a web application for retrieval of aligned texts synchronizes NE tagged aligned texts. NERosetta aims to facilitate retrieval and comparison of named entities in a single or parallel texts. The main named entity categorization

is realized according to the Quaero annotation system and provides users with approximately 50 different search options. The initial version supports four annotation schemas (Stanford NER 3 and Stanford NER 7 for English, Krstev&Vitas for Serbian and Maurel for French) and three annotated parallel versions of Jules Verne's Around The World in Eighty Days (English-Serbian, French-Serbian and French-English).

- ***CESAR Aligned Wikipedia Headword list*** (incl. English): The 762,662 entries of the lexicon are built from the Wikipedia dumps of the six CESAR languages by using article titles and interlingual links to English and the CESAR languages. In the first phase one lexicon for each CESAR language is built after which those lexicons are merged by grouping together all entries that are connected by interlingual links. If more than one article of a language is connected to a group of articles in other languages (which are actually errors in the structure of the Wikipedias), all article titles are retained, divided by a semicolon. In the final phase category information from the English Wikipedia is added with categories divided by semicolons, and for each non-English entry the number of links to that page in the Wikipedia of the respective language is given.
- ***Multilingual Glossary of Synsets*** (incl. English): Multilingual glossary of synsets has been created by mapping several existing WordNets to the Princeton WordNet v. 3.0. It contains synsets (nouns and adjectives) in Bulgarian, Croatian, Hungarian, Serbian and Slovak, together with the links to the English WordNet.

6 The Slovak Contribution to CESAR

The Slovak contribution to the CESAR project and to META-SHARE in particular was a very welcome addition to the stock of language resources and tools in the CESAR portfolio. The language resources developed by the Ľudovít Štúr Institute of Linguistics were prime examples of unique and invaluable resources compiled in relative isolation. As is typical of most CESAR languages, the flagship resource is the Slovak National Corpus and the majority of the Slovak contribution consists of various corpora. The resource types submitted were as follows:

Corpus:	21
Lexical Conceptual Resource:	9
Tool Service:	6

Table 9. Distribution of Slovak resources by type

The most valuable Slovak corpus submitted into the META-SHARE repository is undoubtedly the Slovak National Corpus, which is the largest representative corpus of contemporary Slovak language. It contains written texts published since 1955. The texts are annotated with their genre, style and other bibliographic information. There are specialised subcorpora containing fiction, informational texts, professional texts, original Slovak fiction, texts written from 1955 to 1989, and a balanced subcorpus. The entire SNK for the public is a pseudocorpus, i.e. available through a query interface. The SNK as indeed all the other Slovak resources are impeccably annotated and cleared for IPR.

The collection of Slovak corpora include huge parallel corpora (Slovak-Czech, Slovak-English) a large part of them (containing 6 and 5.7 million tokens respectively) is available for download. The full list of Slovak corpora available through META-SHARE is displayed in

Balanced Slovak Corpus prim-5.0-vyv
Balanced Slovak Corpus prim-6.0-vyv
Corpus of Fiction prim-6.0-img
Corpus of Historical Slovak Corpus
Corpus of Informational Texts prim-6.0-inf
Corpus of Legal Texts
Corpus of Original Slovak Fiction skimg-6.0
Corpus of Original Slovak Texts prim-6.0-sk
Corpus of Professional Texts prim-6.0-prf
Corpus of Slovak Texts from the Years 1955 to 1989
Corpus of Spoken Slovak
Manually Annotated Slovak Corpus
n-grams from Slovak National Corpus
Parallelum Slovaco-Latinum Corpus
Slovak-Czech Parallel Corpus (all)
Slovak-Czech Parallel Corpus (free)
Slovak-English Parallel Corpus (all)
Slovak-English Parallel Corpus (free)
Slovak National Corpus prim-5.0
Slovak National Corpus prim-6.0
Slovak Web Corpus

Table 10. Slovak corpora contain a wider range of dictionaries (based on the web corpus of Slovak language) representing collocation dictionaries, wordnets as well as special databases (e.g. the Slovak Terminology Database with 6,000 terms from 23 fields).

They include not only monolingual resources of the Slovak language, since the Multilingual glossary of synsets contains synsets (nouns and adjectives) in Bulgarian, Croatian, Hungarian, Serbian and Slovak, together with the links to the English WordNet (Princeton WordNet v. 3.0).

Automatic Collocation Dictionary of Slovak
Dictionary of Slovak Collocations. Adjectives
Dictionary of Slovak Collocations. Nouns
Lithuanian WordNet
Multilingual Glossary of Synsets
Slovak Morphology Database
Slovak Terminology Database
Slovak Treebank
Slovak WordNet

Table 11. List of Lexical Conceptual Resources for the Slovak Language

The tools are built of several types and versions of language models of the Slovak language. The language models are based on the well-balanced Slovak National Corpus.

Language model prim-5.0-inf
Language model prim-5.0-sane
Language model prim-5.0-vyv
Language model prim-6.0-inf
Language model prim-6.0-sane
Language model prim-6.0-vyv

Table 12. Slovak language models derived from the Slovak National Corpus

7 Sustainability after CESAR

The META-SHARE infrastructure relies on the operation of interlinked META-SHARE nodes that are distributed and autonomously maintained by the participating institutions. Partners of CESAR guarantee the maintenance and the sustainability of META-SHARE infrastructures for 24 months after the end of the project and for countries involved in CESAR project.

To ensure sustainability of the technical resources developed by CESAR, partners propose the following organization of support:

- all partners are responsible for maintenance of resources and tools provided by their organizations and are thus appointed as META-CENTREs: institutions administering, supporting, updating and ensuring permanent availability of their resources (including backup),
- selected partners maintain META-NODEs, i.e. the META-SHARE applications functioning as CESAR-related points of entry for requests to access descriptions of META-NET resources and tools synchronized with other META-SHARE nodes; each partner establishing their META-NODE is responsible for maintenance of the server, applying bugfix releases and updates received from META-SHARE, providing backup of the application and data, monitoring service availability and performance etc.

The META-CENTREs and the META-NODE are maintained in 24/7 hour mode. The META-NODEs are provided by IPIPAN, FFZG, HASRIL, IBL, UBG, LSIL and ULODZ. The CESAR resource helpdesk is maintained by IPIPAN with the contribution of other partners in the consortium.

The commitments of partners to participate in the long time maintenance of the META-SHARE is confirmed in duly signed Letters of Intent in 2012.

One of the pillars of long time sustainability of the CESAR project benefits are the META-NODES established. The META-NODES are responsible for language resources and language technology tools created in their country – they feature as parts of the European open linguistic infrastructure.

The CESAR consortium has concentrated on all features of language resource that can contribute and have an impact on their sustainability (understood as future availability and usage). The consortium set up a number of requirements in order to meet the sustainability of language resources.

1) Language resources are carefully selected – a methodology and criteria that allow partners to assess the quality and importance of language resources are established and carefully followed. The aim is to ensure a balanced coverage of resources for different end users and tasks, groups of products and services.
2) Particular actions are performed to ensure quality and quantity of the selected resources – upgrading, extending and linking the resources, aligning resources across languages.
3) Language resources are made visible and accessible – META-SHARE metadata descriptions are based on established standards, best practice and users needs. Providing exhaustive metadata descriptions enables the users to find out the most suitable resource and to use it in an appropriate way.

8 Conclusions

The CESAR project specifically focused on the assembly of basic language resources for six Central and South-East European languages, all of them considered less-resourced. Building on a wide range of already existing resources and previous national or international activities, the project created, populated and operated a comprehensive language-resource platform enabling and supporting large-scale multi- and cross-lingual products and services.

The resources made available by the CESAR consortium are ready to be employed in complex LT applications built by joint initiatives of various communities in research and industry, possibly serving multiple purposes in input and intermediary modules.

During the past two years CESAR managed to achieve a number of remarkable results which raises the hope for possible continuation of the work started in this EU funded project. The most remarkable result of the work was the high number and wide range of the 251 resources and tools (presented in three batches). All partners made a great effort to upgrade, enhance or link and offer the best quality resources available for the respective six languages. The resources and tools represent a solid base which can be used in R&D of various fields of language technology industry.

The gathered language resources and tools of CESAR are offered through META-SHARE, a language resource exchange platform, which was set up by META-NET with the continuous help of CESAR. All languages concerned in the project have a META-SHARE node, which will be maintained at least for 24 months after the end of the project (1st February, 2013). The usability of the platform and the involved LRs is strengthened by cleared IPR of all involved resources. All partners expressed their will to release resources and tools only with cleared IPR.

An interesting part of the work in the project was the mobilization of the research community and stakeholders for using and propagating the products of language technology. Several dissemination channels were used to reach the research community

at national and international level in order to attract the relevant players to participate in sharing resources and tools through META-SHARE platform.

The real success of the project will be seen in its long-term impact. Its success can be measured by the usage of META-SHARE and involved language resources and tools as well as by the remaining interest of the community in maintaining and using the chosen resources.

References

[1] Garabík, R., Koeva, S., Ogrodniczuk, M., Tadić, M., Váradi, T., and Vitas, D. (2011). Detecting Gaps in Language Resources and Tools in the Project CESAR. In Vetulani, Z., editor, *Proceedings of the 5th Language & Technology Conference*, LTC2011, pages 37–41, Poznań.

[2] Krauwer, S., Maegaard, B., Choukri, K., and Jørgensen, L. D. (2004). *Report on BLARK for Arabic*. NEMLAR, Center for Sprogteknologi, URL: `http://medar.info/The_Nemlar_Project/Publications/BLARK-final_190906.pdf`.

[3] Lindén, K., Silfverberg, M., Axelson, E., Hardwick, S., and Pirinen, T. A. (2011). HFST —Framework for Compiling and Applying Morphologies in Systems and Frameworks for Computational Morphology 2011. *Communications in Computer and Information Science* (100):67–85.

[4] Maegaard, B. (2004). The NEMLAR project on Arabic language resources 9th EAMT Workshop, "Broadening horizons of machine translation and its applications", 26-27 April 2004, Malta, pages 124–128.

[5] Ogrodniczuk, M., Garabík, R., Koeva, S., Krstev, C., Pęzik, P., Pintér, T., Przepiórkowski, A., Szaszák, Gy., Tadić, M., Váradi, T., and Vitas, D. (2012). Central and South-European language resources in META-SHARE. *Infotheca*, 12(1).

[6] Rehm, G. and Uszkoreit, H., editors. (2013). META-NET Strategic Research Agenda for Multilingual Europe 2020. Springer, Heidelberg, New York etc.

[7] Tadić, M. and Váradi, T. (2012). Central and South-East European Resources in META-SHARE. In *Proceedings of COLING2012*, ACL, pages 431–438.

[8] Váradi, T. (2011). Introducing the CESAR Project. *Infotheca* XII(2):71–74.

[9] Váradi, T. (2013). Serving Multi-Lingual Europe: The CESAR Project. S. Koeva, editor, Sofia.

Czech Subjectivity Lexicon: A Lexical Resource for Czech Polarity Classification

Kateřina Veselovská

Faculty of Mathematics and Physics, Charles University in Prague, Czech Republic

Abstract. This paper introduces Czech subjectivity lexicon – the new lexical resource for sentiment analysis in Czech. The lexicon is a dictionary of 4,947 evaluative items annotated with part of speech and tagged with positive or negative polarity. We describe the method for building the basic vocabulary and the criteria for its manual refinement. Also, we suggest possible enrichment of the fundamental lexicon. We evaluate the current version of the dictionary by implementing it to the classifiers for automatic polarity detection and compare the results of both plain and supplemented system.

1 Introduction

The main goal of sentiment analysis is the detection of a positive or negative polarity, or neutrality of a sentence (or, more broadly, a text). Most often this takes place by detecting the polarity items, i.e. words or phrases inherently bearing a positive or negative value. These words (phrases) can be found by training probabilistic models on manually annotated data. However, it seems profitable for classification to employ a set of the most frequent domain-independent polarity indicators as well. The polarity items are usually collected in the so-called subjectivity lexicons, i.e. corpora of lexical items carrying an intrinsic positive or negative meaning. The implementation of polarity items from the subjectivity lexicon into the data is the first step towards sentiment analysis.

2 Related Work

The issue of building a subjectivity lexicon is described e.g. in [14] or more specifically in [2]. Here the authors use a small set of subjectivity words and a bootstrapping method of finding new candidates on the basis of a similarity measure. The authors get to the number of 4,000 top frequent entries for the final lexicon. Other method for gaining a subjectivity lexicon – translation of an existing foreign language subjectivity lexicon – is described in [2]. Mostly, the authors use subjectivity lexicons and sentiment analysis in general for machine translation purposes. They are interested in how the information about polarity should be transferred from one language to another, if the polarity can differ in the corresponding text spans and if it is possible to compile a subjectivity lexicon for the target language during the translation.

There is a number of papers dealing with the topic of building the subjectivity lexicons for particular languages (see e.g. [1], [5], [9] or [12]). But to our knowledge, in spite of the fact that there exists an ongoing research on sentiment analysis in Czech language (see [16] or [6]), there is no publicly known subjectivity lexicon available for Czech which would help to improve the task and to reach the state-of-the-art results.

3 Czech Subjectivity Lexicon

The core of the Czech subjectivity lexicon has been gained by automatic translation of a freely available English subjectivity lexicon downloaded from `http://www.cs.pitt.edu/mpqa/subj_lexicon.html`. This lexicon is a part of the OpinionFinder, the system for subjectivity detection in English. The clues in this lexicon were collected from a number of both manually and automatically identified sources (see [13]). For translating the data into Czech, we used parallel corpus CzEng 1.0 (Bojar and Žabokrtský, 2006) containing 15 million parallel sentences (233 million English and 206 million Czech tokens) from seven different types of sources automatically annotated at surface and deep layers of syntactic representation.

By this method, we gained 7,228 evaluative expressions. However, some of the items or the assigned polarities appeared rather controversial. For this reason, the lexicon has been manually refined by an experienced annotator. After excluding the clearly non-evaluative items, the lexicon has been manually checked again for other incorrect entries. Below we mention the most significant types of inappropriate entries, revealed in the checking phase by an experienced annotator.

The most common problem was including items that are evaluative only in a rare or infrequent meaning or in a specific semantic context whereas mostly they represent non-evaluative expressions (e.g. *bouda* is in most cases used as a word for a "shed", though it can also mean "dirty trick"). The main criterion for marking the given item as evaluative was its universal usability in a broader context. Thus we excluded most of the domain-dependent items. The non-evaluativeness of the item was sometimes caused by wrong translation of the original English expression. In case they had not been presented in the lexicon yet, the correct translations were added manually.

On the other hand, we found a lot of items with twofold polarity. These were mostly intensifiers like *neuvěřitelně* ('incredibly'), quantifiers like *moc* ('too'), general modifiers or words which are frequently connected both to positive and negative meaning (like *[dobré/špatné] svědomí* – '[clear/guilty] conscience'). The different polarities should be distinguished later on by recording such words in the lexicon together with their prototypical collocations. Other instances also fall under this category of dual polarity, such as ambiguous words which can be used both in positive and negative meaning – e.g. *využít někoho*, meaning 'to abuse somebody' (negative), and *využít příležitosti*, 'to take the opportunity' (positive). We put these expressions aside for further research of their semantic features and corpus analysis of their collocations, since they seem to be crucial for more fine-grained sentiment analysis (see also [4]).

A particular problem appeared to be words with an incorrect polarity value assigned. These could be divided into several categories. One of them are e.g. diminutives marked with positive polarity although they are very often used in negative (mostly ironic) sense – e.g. *svatoušek* – 'goody-goody'. Another large group consists of incorrect translations of negated words like *nečestný* – 'not honest', *nemilosrdný* – 'not forgiving' etc. In this case, the system did not take into account the negative particle preceding the given word and assigned positive polarity to all of them.

In the end we gained the final set of 4,947 evaluative expressions. The most frequent items in the final set were nouns (e.g. *hulvát* – 'a boor', 1,958) followed by verbs (e.g.

mít rád – 'to like', 1,699), adjectives (e.g. špatný – 'bad', 821) and adverbs (e.g. *dobře* – 'rightly/well/correctly', 469).

4 Data Sets

To test the credibility of the lexicon, we used several datasets on which we had previously trained the original classifiers (see [16]). Firstly, we worked with the data obtained from the Home section of the Czech news website Aktualne.cz (`http://aktualne.centrum.cz/`) manually identified as evaluative. At the beginning, there were approximately 560,000 words in 1661 articles, which have been categorized according to their subjectivity. Than we identified 175 articles (89,932 words) bearing some subjective information, 188 articles (45,395 words) with no polarity, and we labelled 90 articles (77,918 words) as "undecided". There still remain 1,208 articles which have not been classified yet. The annotators annotated 410 segments of texts (6,868 words, 1,935 unique lemmas). These segments were gained from 12 randomly chosen articles. Secondly, we used the data from Czech movie database, CSFD.cz (`http://www.csfd.cz/`). The data contained 405 evaluative segments annotated on polarity. Moreover, as both sets of the manually annotated data were pretty small, we also used auxiliary data, namely domestic appliance reviews from the Mall.cz (`http://www.mall.cz/`) retail server. We have worked with 10,177 domestic appliance reviews (158,955 words, 13,473 lemmas) from the Mall.cz retail server. These reviews were divided into positive (6,365) and negative (3,812) by their authors.

5 Testing the Lexicon

In our sentiment analysis experiments, we use the Naive Bayes classifier, a discriminative model which makes strong independence assumptions about its features, as minutely described in [16] with best results for the Mall.cz data. To test the subjectivity lexicon performance, we added two new features to the classifier, saying how many of the evaluative items of which polarity the given segment contained. So far, we have seen some slight improvement in identifying evaluative sentences on Aktualne.cz data when employing the 10-fold cross-validation (see tables 1 and 2).

Test result average:	precision	recall	f-score
POS	0.39	0.36	0.33
NEUTRAL	0.92	0.86	0.89
BOTH	0.00	0.00	0.00
NEG	0.61	0.71	0.65
average	0.84	0.81	0.82

Table 1. Aktualne.cz without subjectivity lexicon

Test result average:	precision	recall	f-score
POS	0.24	0.33	0.26
NEUTRAL	0.93	0.85	0.89
BOTH	0.00	0.00	0.00
NEG	0.64	0.74	0.68
average	0.85	0.81	0.83

Table 2. Aktualne.cz with subjectivity lexicon

On the other hand, we have not seen any significant improvement neither on the Mall.cz nor on CSFD.cz data (see tables 3, 4, 5 and 6) so far.

Test result average:	precision	recall	f-score
POS	0.93	0.92	0.92
NEG	0.85	0.88	0.87
average	0.90	0.90	0.90

Table 3. Mall.cz without subjectivity lexicon

Test result average:	precision	recall	f-score
POS	0.93	0.92	0.92
NEG	0.86	0.88	0.87
average	0.90	0.90	0.90

Table 4. Mall. cz with subjectivity lexicon

Test result average:	precision	recall	f-score
POS	0.66	0.79	0.71
NEUTRAL	0.70	0.57	0.63
NEG	0.62	0.62	0.61
average	0.67	0.65	0.65

Table 5. CSFD.cz without subjectivity lexicon

Test result average:	precision	recall	f-score
POS	0.63	0.80	0.70
NEUTRAL	0.68	0.53	0.60
NEG	0.63	0.62	0.62
average	0.66	0.64	0.64

Table 6. CSFD.cz with subjectivity lexicon

The low performance might be caused by the very small size of the data and its domain-specificity (statistically, the classifier did not reach many hits in any of the data sets). As for the future, it could be useful to test the lexicon on much bigger evaluative data.

6 Future Work

In order to improve the automatic polarity classification, it would also be advantageous to enhance the subjectivity lexicon by several methods. Firstly, we could use the dictionary-based approach as described by [8] or [10] and grow the basic set of words by searching for their synonyms in Czech WordNet [11].

Secondly, we could employ the corpus-based approach based on syntactic or co-occurence patterns as described in [7]. Also, we can extend the lexicon manually by Czech evaluative idioms and other common evaluative phrases. Moreover, it would probably be useful to add back some special domain-dependent modules for the different areas of evaluation. Hereby we plan to verify the hypothesis that increasing the size of the corpus could further improve the classification.

7 Conclusion

We have built and tested a subjectivity lexicon for sentiment analysis in Czech. Comparing to the previous results reached in the field, we observed that the very first version of the lexicon did not help to improve the polarity classification significantly, so its refinement needs to be a subject of the further research. However, we introduced the unique Czech subjectivity lexicon which can still serve as a lexical resource e.g. for semantic analysis or evaluative language research.

Acknowledgement

The work on this project has been supported by the GAUK 3537/2011 grant and by SVV project number 267 314. This work has been using language resources developed and/or stored and/or distributed by the LINDAT Clarin project of the Ministry of Education of the Czech Republic (project LM2010013).

References

[1] Bakliwal, A., Piyush, A., and Varma, V. (2012). Hindi Subjective Lexicon: A Lexical Resource for Hindi Adjective Polarity Classification. In *Proceedings of the 8th Language Resources and Evaluation Conference (LREC 2012)*.

[2] Banea, C., Mihalcea, R., Wiebe, J., and Hassan, S. (2008). Multilingual subjectivity analysis using machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, pages 127–135.

[3] Banea, C., Mihalcea, R., and Wiebe, J. (2008). A Bootstrapping Method for Building Subjectivity Lexicons for Languages with Scarce Resources. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*.

[4] Benamara, F., Cesarano, C., Picariello, A., Recupero, D. R., and Subrahmanian, V. S. (2007). Sentiment analysis: Adjectives and adverbs are better than adjectives alone. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*.

[5] De Smedt, T. and Daelemans, W. (2012). Vreselijk mooi! (terribly beautiful): A subjectivity lexicon for dutch adjectives. In *Proceedings of the 8th Language Resources and Evaluation Conference (LREC 2012)*.

[6] Habernal, I., Ptáček, T., and Steinberger, J. (2013). Sentiment Analysis in Czech Social Media Using Supervised Machine Learning. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 65–74.

[7] Hatzivassiloglou, V. and McKeown, K. R. (1997). Predicting the semantic orientation of adjectives. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, Association for Computational Linguistics.

[8] Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM.

[9] Jijkoun, V. and Hofmann, K. (2009). Generating a Non-English Subjectivity Lexicon: Relations That Matter. In *Proceeding of EACL 2009, 12th Conference of the European Chapter of the Association for Computational Linguistics*.

[10] Kim, S. and Hovy, E. (2004). Determining the sentiment of opinions. In *Proceedings of the 20th international conference on Computational Linguistics*, Association for Computational Linguistics.

[11] Pala, K. and Ševeček, P. (1999). The Czech WordNet, final report. Masarykova univerzita, Brno.

[12] Perez-Rosas, V., Banea, C., and Mihalcea, R. (2012). Learning Sentiment Lexicons in Spanish. In *Proceedings of the 8thth international conference on Language Resources and Evaluation (LREC 2012)*.

[13] Riloff, E. and Wiebe, J. (2003). *Learning extraction patterns for subjective expressions*. EMNLP.

[14] Taboada, M., Brooks, J., Tofiloski, M., Voll, K., and Stede, M. (2011). Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistics*, 37(2):267–307.

[15] Veselovská, K. (2012). Sentence-level sentiment analysis in Czech. In *Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics (WIMS 2012)*.

[16] Veselovská, K., Hajič, J., and Šindlerová, J. (2012). Creating Annotated Resources for Polarity Classification in Czech. In *Proceedings of KONVENS*, pages 296–304.

[17] Wiebe, J., Wilson, T., Bruce, R., Bell, M., and Martin, M. (2004). Learning subjective language. *Computational Linguistics*, 30(3):277–308.

A Corpus-based Analysis of the Functionality and the Meaning of Infinitive "Frustrative Construction" in Czech and Slovak

Uliana Yazhinova

Department of Slavic Studies, Humboldt University of Berlin, Germany
Department of Slavic Studies, University of Regensburg, Germany

Abstract. This article presents the quantitative and qualitative corpus-based analysis of a specific Infinitive construction in Czech and Slovak – *n(i)e a n(i)e + Inf-cxn.* The aim of this study is to contribute empirical evidence to support certain claims as reported in the academic literature and to contribute to the discussion with factual evidence. The aim of this analysis is to determine the "grammatical rules", semantic domains of meaning and the actual patterns of use for this construction and as well as to compare the results in Czech and Slovak.

1 Introduction and Phenomenon

In Czech and Slovak we can find relatively young infinitive construction *n(i)e a n(i)e + Infinitive*, which has a unique semantic and structure in comparison with other Slavonic and non-Slavonic languages. The following examples from the parallel corpus Intercorp (exp. 1, 2) and Slovak-Czech Parallel Corpus (exp. 3) illustrate some uses of this construction.

(1)
(CZ)

Slunce	*se*	*už*	*sklán-í*	*k*	*obzoru*
sun	refl	already	set - PRS.3SG	to	Horizont
a	*smrt*	***ne***	***a***	***ne***	***přijít.***
and	death - NOM.	not	and	not	come - [PFV]INF.

'The sun is setting but death has not yet come.'

(2)
(SK)

Sede-l	*a*	*rozmýšľa-l*	*a*	*rozmýšľa-l,*	*až*
sit - PST.3SG	and	think - PST.3SG	and	think - PST.3SG	until
mu	*iš-la*	*hlava*	*puknú-ť ,*	*ale*	*na*
him	go - PST.3SG	head	crack - INF.	but	to
nič	*skvelé*	***nie***	***a***	***nie***	***prísť.***
nothing	great	not	and	not	come - [PFV]INF.

'He sat and thought and thought , until his head nearly burst , but no bright idea came to him.'

(3)
(CZ)

Ne	***a***	***ne***	***najít***	*zpátečku.*
not	and	not	find - [PFV]INF.	back.

(SK)

Nie	***a***	***nie***	***nájsť***	*spiatočku.*
not	and	not	find - [PFV]INF.	reverse.

'I just could not find the way back.'

As the examples show, the *n(i)e a n(i)e + Inf-cxn* construction is used in Czech and Slovak to express a subjective, often negative response on the part of the actor or involved observer to real or perceived failure, an unexpected or undesired result of an action. This meaning is expressed by a concatenation of morphological components (syndetic reduplication[1] of negative particles and infinitive verb), which partially contravenes the usual grammatical order.

The first scholar who discussed the specific morphology, syntax and semantics of this infinitive construction in Czech was B. Hansen [7] in terms of constructional grammar. In his study "Another piece of Infinitive puzzle the Czech frustrative construction ne ne zaprazet" he suggested the term "frustrative construction", already commonly used in linguistic typology should be used for Czech also. In my work I adopted the term **"frustrative construction"** for this type of construction and want to compare results of corpus data to provide a more detailed analysis in both languages.

The following examples illustrate the use of this construction in both languages (Czech (exp. 4, 5) and Slovak (exp. 6, 7)), which is established in different communicative domains:

(4)
(CZ)

Trénuj-í		*třikrát*	*týdně,*	*jsou*	*zpocen-í*	*jako*
exercise - [IMPF]-PRS. 3.PL		three times	week	are	sweaty - [PFV]3.SG	as
myš-i,	*ale*	*panděro*	***ne***	***a***	***ne***	***splasknout.***
mouse - PL	but	belly - NOM.	not	and	not	deflate – [PFV]INF

'They are exercising three times a week, are sweaty as hell, but their pot bellies just refuse to deflate.'

(sub-corpus – SYN; Mladá fronta DNES, 12. 12.1998, publicistic)

(5)
(CZ)

Auto	***ne***	***a***	***ne***	***nastartovat.***
car - NOM	not	and	not	start - [PFV]INF.

'The car just won't start.'

(sub-corpus – SYN; Mladá fronta DNES, 6. 1. 2009, publicistic)

[1] In terminology of [17], [18]

(6)
(SK)

V	*tú*	*noc*	*minister*	***nie***	***a***	***nie***	***zaspať.***
in	that	night	minister - NOM.	not	and	not	sleep - [PFV]INF.

'In that night the minister just cannot sleep.'
(sub-corpus – prim-5.0-public-all; Poviedky z jedného i druhého vrecka, 1958)

(7)
(SK)

Ma-li	*sme*	*veľkú*	*prevahu,*	*aj*	*šance,*
have - [IMPF]-PST.1PL	REFL	totally	dominance	and	chance

no	***nie***	***a***	***nie***	***dať***	*gól.*
but	not	and	not	give - [PFV]INF.	goal.

'We totally dominated the field and had chances, too, but no matter what we did we just were not able to score a goal.'
(sub-corpus – prim-5.0-public-all, Jeho „gól" nebol gól 2009.04.27)

In the corpus data you can find two construction types *nebude a nebude/nehubnu a nehubnu* (with repetition of the finite verb) and *"ne a ne shubnout"* (with the repetition of negative particles plus infinitive), which is not sensitive. The findings of the present study focus on the nature of infinitive construction, which will be used later for the comparable research of both constructions. The examples above illustrate well the use of this type of infinitive construction in both langauges (if the speaker is correct in assuming frustrative meaning to the situation).

The next section presents theorethical background and problems defining the terms "frustrative construction" and "frustrative in lingustic typology". In sections 3 and 4, I test functional and morphological restrictions of *n(i)e a n(i)e + Inf-cxn* construction and present results of qualitative and quantitative corpus-based analysis. Finally the conclusions of the preliminary analysis of the corpus data and perspectives for the future studies will be drawn.

2 State of Research

In this section, before I present the results of the corpus-data analysis, I refer to a) the most important "notions" from traditional grammar of Czech and Slovak, b) Hansen's work on "frustrative construction", and c) linguistic typology point of view to make use of them as a background for a new theoretical framework.

In the traditional grammar of Czech and Slovak, the construction is merely noted in passing. For example, in the [6] it was noted, that the construction types with repetition (reduplication) of negative predicats like *Já ne a ne si na to vzpomenout* ('I just don't remember it') or *On pracovat nebude a nebude* ('He just not start to work') are used to express **"nemožnost"** (impossibility) or **"neschopnost"** (inability (failure) of action). In the [5] Petr Karlík presented a study of the Czech infinitive, and designated the construction *ne a ne zapršet ('It's just not rain')* (henceforth *ne a ne + Inf-cxn)* as an

"autonomous group" ("skupina samostatná") [3]. Other short descriptions by [15], [13] and [19] include at best an indication **of expressivity with a note that it is an elliptical subtype of reiterative infinitive** - "Dvojčlenné infinitivní věty oznamovací"[2]. All such observations are significant, but cannot simply be accepted as semantically and grammatically valid descriptions unless they are justified on the basis of morpho-syntactical principles and semantic modification.

More detailed analysis was proposed by [7]. He turned his attention to the morpho-syntactical specificity of this construction in Czech and the effects this construction has on the semantic interpretation of the rest of the sentence. B. Hansen [7, p. 169] argues that "the Czech *ne a ne Inf-cxn* has a specific morpho-syntax, a non-compositionally derived meaning and a specific pragmatic profile. We would propose to call it frustrative which is a meaning label for a complex function expressing both the dynamics of the action and the attitude of the speaker and, thus, combining aspectual features with speaker's emotional stance".

Moreover he summarised the paradigmatic properties and sketched it in the profile of prototypical instance[3] of "frustrative Czech construction".

2.1 Background of the Term "Frustrative"

The term "frustrative" is already an established term in language typology ([16], [14], [11], [2], [12], [10], [3], [4] etc.). Generally, "frustrative" is used to describe the semantic modification of the verb in a number of languages and it might be expressed morphologically ("frustrative morphemes") or syntactically. The morphological element gives some additional information and in all the cases indicates degree of subjectivity. However, its application is still a matter open to discussion. The following examples show some of the uses:

1) Initiated action was to no avail[4]
– "it is used in clauses to express that an action is unsuccessful or in vain"
Example in Mawayana [4, p. 144]

(7)
(Mawayana)

anumalë	*tïtëi*	*inëlë*	*koko-psik*	*tï-të- i_lep*
tomorrow	COREF - go - INF	DP.ANIM.ANA	night - DIM	COREF - go - NF_FRUST

'The next day he left, he left early in the morning (but he didn't shoot any game.)'

2) One's intention was blocked
"As far as the meaning is concerned, it expresses frustration due to unrealizable goals or a pleasant outcome of a situation due to fortunate circumstances"[5]

3) "The frustrative marker can also be marked on **nouns to express that the referent of the noun is lacking** in at least one semantic feature of the noun, or that **the Object expressed by the noun is not used for its inherent purpose"**[6] [4, p. 139].

[2] See [19]
[3] See [7]
[4] See [8]
[5] See [16]
[6] See [4]

(8)
(Mawayana)

paila	*tëkalëi*	*pïlëo*	*malë*	*i-të-top-kom*
bow	he.gave	arrow	also	3POSS - go - TMP.NOM-PL

tëhem	*we-top_lëp*
meat	shoot MP.NOM_FRUST

'The next day he left, he left early in the morning (but he didn't shoot any game).'

'He gave him a bow and arrows for their journey, a means for shooting game animals' (but this man didn't use them to shoot meat: they were an instrument for shooting but were never shot) '

4) Negative or possibly "counter-expectation" "frustrative indicates that the action was done to no avail – that is, the desired result was not achieved"[7]
5) unexpected or surprising outcome of one action[8]
6) the failure of one action or the failure to complete an action

Finally, it should be noted, that the term frustrative has the same semantic domains with Neighbour terms like "conative", "avertive", "antiresultative". In this work, however, I have decided to use term "frustrative construction" [1] as a label for this TYPE of infinitive construction. And the basic semantic ingredients of this construction are the following:

1) subjectivity (the response of an actor or an involved observer to failure)
2) conative aspect (as an attempt to complete the action)
3) anti-result (the desired effect was not achieved)
4) counter-expectation (an effect contrary to the desired effect)

Negative rating/negative evaluation of the result of an action ("bad" or "frustrating" registers that the speaker or possibly an involved observer experiences a negative emotion regarding the unproductive outcome of a given action).

3 Preliminary Results of Corpus-based Analysis

The data for the synchronic analysis used in this study was gathered from two types of resources:
1) Corpus data: from the Czech National Corpus (Český národní korpus – CNK)[9], and the Slovak National Corpus (Slovenský národný korpus – SNK[10]) and 2) Internet data.

[7] See [1]
[8] See [2], [10]
[9] http://ucnk.ff.cuni.cz/
[10] http://korpus.sk/index_en.html

I examined all sub-corpora to ascertain whether they contain the construction *ne a ne + Infinitive*[11] and detected "frustrative construction" in 12 sub-corpora. For the results, see Table 1 below. The queries gave **3,323 hits**. The infinitival phrase occurs in two variants: pre- or post-positioned. This variation, however, has no bearing on the meaning.

Preliminary analysis of the text data from the Slovak national corpus yielded many more contexts with the *nie a nie + Infinitive*[12] construction than in Czech, the queries gave **8,752** from 4 sub-corpora (see Table 2).

Variants	Written corpora								Spoken corpora			
	Syn-series	Fsc 2000	Ksk-dopisy	Link	orwell	Schola 2010	czeslplain	skript 2012	Oral 2006	Oral 2008	pmk	bmk
Ne a ne	**2,860**	**249**	8	2	0	0	1	10	1	1	1	2
Ne a ne	**172**	**15**	0	0	0	0	0	1	0	0	0	0

Table 1. CNK – number of hits in different corpora

Variants	Written corpora			Spoken corpora
	prim-6.0-public-all	r-55az89-3.0	Web corpus	s-hovor-4.0
Nie a nie	3,230	729	4,187	2
Nie a nie	298	74	231	1

Table 2. SNK – number of hits in different corpora

The first analysis for usage of this construction in both languages shows, that it is a productive pattern (*n(i)e a n(i)e + Inf-cxn*) in both languages.

Although if the frequency of this type of construction by corpora data occurs more often in written language then in spoken, it could not mean, that is a prevalent in written language only. Because of the specific semantic meaning of construction this we need to analyse some special type of conversations, but Czech and Slovak corpora do not have enough documents of this type. That's why it could be helpful to understand the communicative situation as (domains) in which construction could be used in.

3.1 Distribution of Frustrative Construction in Czech and Slovak

The next step was to manually select all relevant occurrences of the above-mentioned construction (the constructions like *ne a ne a ne/ ne, ne a ne* etc.). The aim of the occurrence frequencies analysis was to find some tendency of distribution (*n(i)e a n(i)e + Inf-cxn* construction) in written and oral texts in Czech and Slovak. This was important

[11] The basic query was: [word="ne"] [word="a"] [word="ne"] [word="Ne"] [word="a"] [word="ne"][word="ne"] [word="a"] [word="ne"] [] {0,5} [tag="Vf.*I"] within <s>, for the PFV Verb query was [word="ne"] [word="a"] [word="ne"] [] {0,5} [tag="Vf.*P"] within <s>, jinak [word="ne"] [word="a"] [word="ne"].

[12] [word="nie"] [word="a"] [word="nie"] and [word="Nie"] [word="a"] [word="nie"]

1) to reduce the number of repetitions (to work with real data) and 2) to sort texts sorts (genres). For the quantitative analysis I used the statistical SyD[13] – only available for Czech (for Slovak I analysed corpora data manually). This includes texts from three sub-corpora: Sub2010 for the written language – (401 hits for *ne a ne Infcxn*); ORAL2006 (a transcription of 221 recordings from 2002–2006 taken from spoken Czech across all regional dialect areas); and ORAL 2008 (also spoken Czech, but including more recorded material from the whole of Bohemia in 2002–2007. However, there is no dissection between the latter two corpora). For Slovak were analysed 3 types of sub-corpora: prim-6.0-public-all (2,819 hits for *nie a nie+ Infcxn*) for written language, prim-6.0-public-sk (2008 hits for *nie a nie+Infcxn*); s-hovor-4.0 (1 hit).

When one looks at a wide data base with different kinds of text, it is clear that *n (i)e a n(i)e + Inf-cxn* may occur in all discourse types. The frequency in reports, stories, journalistic texts, and web-conversation indicates the narrative character of this construction. The following tables give the data for Czech and Slovak. The table below shows the relative distribution according to type of text in Czech and Slovak:

Kind of text	**Czech**	**Slovak**
Povídky (stories)	22%	37%
Básně (poetry)	21%	10%
Drama a scenáře (drama and scenarios)	21%	15%
Román (novels)	13%	25%
Lit. faktu (non-fiction)	13%	16%
Písně (songs)	2%	2%
Populárně-naučná (non-fiction and academic or scientific literature)	5%	3%
Publicistika (journalism)	3%	16%

Table 3. Relative distribution according to type of text in Czech and Slovak

3.2 Frequency of Frustrative Construction in Czech

It was interesting to analyse the difference in the occurrence of "frustrative construction" **for publications** (newspapers, magazines, fuellitons, online material)[14], here and after PUB texts **and in literature texts** (books, novels, essays, songs, poems, dramaturgy), here and after BEZ (and TOT is the sum of BEZ and PUB). For this reason I decided to use Word per Million analysis.

$$F = \frac{N counts}{dimension\ of\ subcorpus} \times 1\,000\,000$$

The following figure shows dimensions of the sub-corpora as a function of year. Each sub-corpora has been divided into PUB (blue diamonds) concerning the publicist entries of the SYN corpus from CKN, and into BEZ (green circles) concerning the NON publicist entries. The total number of entries is also displayed (red triangles TOT). All the data are expressed in million of tokens (= words).

[13] `http://syd.korpus.cz/`

[14] coding in CNK corpus: PUB – newspapers and magazines; popular literature, BEZPUB – non-publicistic texts, NOV – romans, COL – stories; POP – non-fiction literature; VER – poetry; SCI – songs, SCR – drama and scenarios.

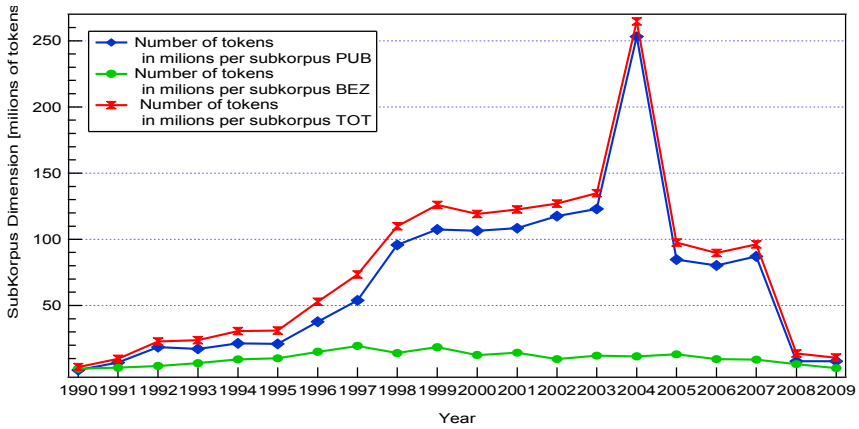


Fig. 1. Dimensions of the sub-corpora

We considered the time range from 1990 up to 2009 and divided the SYN corpus from CNK in sub-corpora covering a one year time period. The other principle was – 1 Author=one Text= in the same year. This art of analysis was possible also only for Czech, because the WPM (Word Per Million) analysis requires the number of hits per year per sub-corpus and this information is not yet available in Slovak corpora. It will be interesting to find out whether or not the Slovak examples correspond with the same tendencies as in Czech.

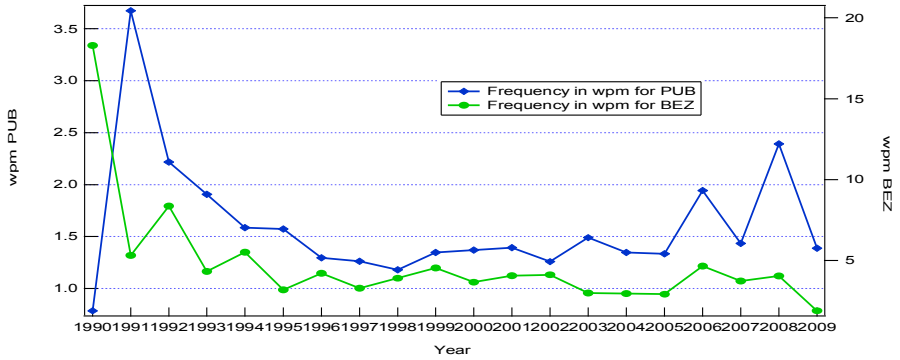


Fig. 2. Behaviour of the frequency in PUB and BEZ

Figure 2: To compare the behaviour of the frequency in PUB and BEZ the data have been plotted on top of another. Attention: the reference Y-axes for PUB (blue) is the left one, while the reference Y-axes for BEZ (green) is the right one. The two axes have different ranges: left Y — [0, 3.7] wpm ; right Y — [0, 21] wpm.

According to the diagram, the number of hits (of *n(i)e a n(i)e* + *Inf-cxn* construction) in the non-publicist texts is higher than in publicist texts. Within the picture is contained can be observed that the frequency tendency started to increase steadily in 90's (The years 1992 and 1993 show a particularly high frequency which peaks by 1994). The graphs show a plateau at the level of 1995–2005 and flatten out level in 2005. There is a clear parallel (same tendency) for publicist and non-publicist texts.

The explanation for this unexpected high frequency in the non-publicist texts is that the sub-corpora SYN includes a lot of translation texts, (not original Czech texts), which are marked as non-publicist (BEZ PUB) and this influences on the results of frequency. On the other hand, it means that this construction will be often used in the translations like a very productive pattern. That is the other reason, why this construction should be analyzed in more detail.

3.3 Morphological and Grammatical Facts about the Infinitive Verb

These infinitive constructions involve a nominative–subject and are generally restricted to the perfective aspect of infinitive verb. In this chapter I present evidence showing how the use of a grammatical feature corresponds with situational context in both languages.

The manual analysis of the data from all sub-corpora in Czech and Slovak shows that, in both languages, there is a significant preference for perfective infinitives. For example, in Czech, the contexts with *ne a ne + INF PERF* rendered in SYN sub-corpus 2,349 tokens (without duplicates and not relevant constructions). For *ne a ne + IMPF INF*[15] there were only 84 tokens. The same tendency is true for Slovak. For example, *nie a nie + PERF INF* in sub-corpus prim-6.0-public-all 1,477 tokens and for *nie a nie + IMPF INF* 58 tokens.

The table below lists the verbs, which occur most frequently PFV with FC construction in both languages.

CZECH Frequently PFV (sub-corpus SYN)	hits	SLOVAK Frequently PFV (sub-corpus prim-6.0-public-all)	hits
přijít	148	prísť	109
najít	121	nájsť	84
dát	65	prestať	67
dostat	63	zaspať	63
přestat	50	dať	50
skončit	40	dostať	50
padnout	30	pochopiť, skončiť, spomenúť si, trafiť,	33–32
trefit, dostavit se, zbavit se, pochopit, objevit, dostat se, prosadit, naskočit, zabrat, vstřelit, vyjet, zmizet, umřít etc.	less then 20	padnúť, pohnúť, pustiť, odísť, zbaviť, zísť, streliť, zmiznúť, vyjsť, vojsť, naskočiť, dočkať, zastaviť, vyhrať, spomenúť etc.	less then 20

Table 4. Frequently PFV with FC construction

See examples 1–7 for the using of PFV in *n(i)e a n(i)e* construction.

Imperfective aspect:

For the query in the Syn sub-corpus [word="ne"] [word="a"] [word="ne"][]{0,3} [tag=".....I.*"] it was totally for Czech 84 hits.
For Slovak was the query [word="nie"] [word="a"] [word="nie"] [tag="V.e.*"]. Number of hits: 58.

[15] [word="ne"] [word="a"] [word="ne"][]{0,3} [tag=".....I.*"]

CZECH Frequently IPFV (sub-corpus SYN)	hits	SLOVAK Frequently IPFV (sub-corpus prim-6.0-public-all)	hits
být	14	ísť	4
jít	11	veriť	4
padat	5	cítiť, kochať, fungovať, vychádzať, nakladať, svedčať, útočiť, udržiavať etc.	1-2

Table 5. Frequently IPFV

Some examples for IMPF:

(CZ)

" Co tak ječíš , Zdeňo ? Tohle už není k vydržení ! "
'Why are you screaming, Zdeno? Its beyong enduarence! '

Hlava	*mě*	*bolí*	k	*prasknutí*	a	vy
head	my	ache	to	crack	and	You

ne	***a***	***ne***	***být***	*zticha.*
not	and	not	be	quiet

'Why are screaming, Zdeno? This is unbearable. I have a splitting headache and you just won't be quiet'

(Háj, Felix, Školák Kája Mařík, Praha, 1990, beletrie)

(SK)

Len	*to*	*koliesko*	*na*	*myš-i*	***nie***	***a***
just	that	scroll-wheel	on	mouse	not	and

nie	***fungovať.***	*a*	*pritom*	*kolieskové*	*myš-i*	*už*
not	work - [IMPF]INF	and	moreover	scroll-wheel	mouses	yet

ne-vyráb-a	*iba*	*Microsoft...*
not produce - PRS.3PL	just	Microsoft...

'Just that scroll-whell will not work. Moreover, only Microsoft produce it...'

(PC REVUE 2001/03, genr.MIX)

Close analysis of the verbal forms used in the written corpora and in internet shows the clear division of the aspect. The narrative texts signalled by sentences with perfects. The infinitive constructions normally serve as reminders of the inescapable end, but with this constructions "end" just "not and not" coming.

4 Conclusion

The purpose of this paper was to cite and comment on a number of examples of frustrative constructions from Czech and Slovak in order to explore in greater detail the functional features of this construction.

This section summarises the main findings of this paper and indicates some further perspectives. Formally this construction is very compact – infinitive construction with reduplication of negative particles. The remaining *n(i)e a n(i)e + Inf-cxn* is then used to express the degree of emotion (emotionality) and to convey "frustration", but where does this come from? The answer is that it comes from the structure of this construction: concatenation of syndetic reduplication of negative particles plus infinitive verb, which are generally perfective. Nearly analysis of this components shows, that the repetition of two negations intensify a "conative meaning": the effort of the agent more times to performing the activity which are expressed by perfective verb. And it might be that some examples do not have a very strong negative rating (though negativity is always present). Furthermore, analysis of the infinitive verbs showed that there are some verb classes, which are more frequently by building of this constructions. And the next step is to make the classification of these verbs and to describe their properties.

On the other hand, detailed quantitative usage-based characterization of this innovative construction shows some structural and functional restrictions of this construction, which demonstrate some same tendencies in both languages. For example, this pattern *n(i)e a n(i)e + Inf-cxn* is frequent in written language like narrative reporting, and everyday mundane conversation, particularly in Slovak. The quantitative analysis for Czech demonstrate the same development tendency in different text types, too.

Further research could be directed toward a detailed corpus-based analysis of other morph-syntactic features from a diachronic perspective in order to explain the development of the lexicalized components of this construction. It would be also of interest to isolate semantically identical or closely comparable constructions in other Slavonic languages. The explanation of the Czech facts may be relevant for Slavonic as well as certain non-Slavonic languages.

References

[1] Aikhenvald, A. Y. (2008). *The Manambu language of East Sepik, Papua New Guinea*, Oxford, p. 126.

[2] Butorin, S. S. (2006). O frustrativnom komponente semantiki časticy 'kaj' v ketskom jazyke. In *Gumanitarnye nauki v Sibiri*, 4:23–27.

[3] Carlin, E. (2006). Feeling the Need. The Borrowing of Cariban Functional Categories into Mawayana (Arawak). In Aikhenvald, A. and Dixon, R., editors, *Grammars in Contact*, pages 313–332.

[4] Carlin, E. (2009). Truth and knowledge markers in Wayana (Cariban), Suriname. In The Linguistics of Endangered Languages: Contributions to Morphology and Morphosyntax, pages 134–150, Utrecht, URL: `http://lotos.library.uu.nl/publish/articles/000323/bookpart.pdf`.

[5] *Encyklopedický slovník češtiny*. Nakladatelství Lidové noviny, Praha.

[6] Grepl, M. and Karlík, P. (1998). *Skladba češtiny*. Vyd. 1. Votobia, Olomouc.

[7] Hansen, B. (2010) Another piece of the Infinitive puzzle: the Czech frustrative construction ne a ne zapršet. In Bičan, A., Klaška, J., Macurová, P., and Zmrzlíková, J., editors, *Karlík a továrna na lingvistiku. Petru Karlíkovi k šedesátým narozeninám*, pages 166–179, Masarykova univerzita, Brno, URL: `http://epub.uni-regensburg.de/23356/1/Hansen2010_Frustrative_CZ.pdf`.

[8] Kallfell, G. (2010). *Grammatik des Jopara. Gesprochenes Guarani und Spanisch in Paraguay*. Peter Lang, Frankfurt.

[9] Karlík, P. and Veselovská, L. (2009). Infinitive Puzzle. In Ziková, M. and Dočekal, M., editors, *Czech in Formal Grammar*, pages 197–213, Lincom, München.

[10] Kuznecova, J. (2010). Frustrativ kak nedostajushiji element antiresultativnoi paradigm. URL: `www.ling.whelsinki.fi/uhlcs/LENCA/LENCA`.

[11] Nordhoff, S. (2004). Nomen/Verb Distinktion in Guarani. URL: `http://publikationen.ub.unifrankfurt.de/frontdoor/index/index/docId/22366`.

[12] Overall, S. E. (2008). Frustrative: A verbal category of Amazonia. Paper read at the 18th Congress of International Linguistics, Parallel session: Tense, aspect, and modality, Seoul 21–26 July 2008.

[13] Porák, J. (1961) Dvojčlenné infinitivní věty v češtině. In *Acta Universitatis Carolinae – Philologica 3, Slavica Pragensia III*, pages 137–150.

[14] Reilly, E. M. A Survey of Texistepee Popoluca Verbal Morphology. Unpublished Undergraduate Thesis, Carelton College, Northfield, Minnesota. URL: `http://www.cog.jhu.edu/grad-stucnts/rcily`.

[15] Ružička, J. (1956). *Skladba neurčitku v slovenskom spisovnom jazyku*. Vydavateľstvo Slovenskej akadémie vied, Bratislava.

[16] Sparing-Chávez, M. (2003). *I want to but I can't: the frustrative in Amahuaca*. SIL Electronic Working Papers 2003.

[17] Stolz, Th. (2009). Total reduplication: syndetic vs asyndetic patterns in Europe. *GLS* 71:99–114.

[18] Stolz, Th., Stroh, C., and Urdze, A. (2010). *Total reduplication. The areal linguistic of a potential universal*. (Stud. Typological, 8), Akad-Verl. Berlin.

[19] Svoboda, K. (1962). *Infinitiv v současné spisovné češtině*. Nakladatelství Československé akademie věd. Praha.

Appendix

creative commons

Attribution-ShareAlike 3.0 Unported

CREATIVE COMMONS CORPORATION IS NOT A LAW FIRM AND DOES NOT PROVIDE LEGAL SERVICES. DISTRIBUTION OF THIS LICENSE DOES NOT CREATE AN ATTORNEY-CLIENT RELATIONSHIP. CREATIVE COMMONS PROVIDES THIS INFORMATION ON AN "AS-IS" BASIS. CREATIVE COMMONS MAKES NO WARRANTIES REGARDING THE INFORMATION PROVIDED, AND DISCLAIMS LIABILITY FOR DAMAGES RESULTING FROM ITS USE.

License

THE WORK (AS DEFINED BELOW) IS PROVIDED UNDER THE TERMS OF THIS CREATIVE COMMONS PUBLIC LICENSE ("CCPL" OR "LICENSE"). THE WORK IS PROTECTED BY COPYRIGHT AND/OR OTHER APPLICABLE LAW. ANY USE OF THE WORK OTHER THAN AS AUTHORIZED UNDER THIS LICENSE OR COPYRIGHT LAW IS PROHIBITED.

BY EXERCISING ANY RIGHTS TO THE WORK PROVIDED HERE, YOU ACCEPT AND AGREE TO BE BOUND BY THE TERMS OF THIS LICENSE. TO THE EXTENT THIS LICENSE MAY BE CONSIDERED TO BE A CONTRACT, THE LICENSOR GRANTS YOU THE RIGHTS CONTAINED HERE IN CONSIDERATION OF YOUR ACCEPTANCE OF SUCH TERMS AND CONDITIONS.

1. Definitions

a. **"Adaptation"** means a work based upon the Work, or upon the Work and other pre-existing works, such as a translation, adaptation, derivative work, arrangement of music or other alterations of a literary or artistic work, or phonogram or performance and includes cinematographic adaptations or any other form in which the Work may be recast, transformed, or adapted including in any form recognizably derived from the original, except that a work that constitutes a Collection will not be considered an Adaptation for the purpose of this License. For the avoidance of doubt, where the Work is a musical work, performance or phonogram, the synchronization of the Work in timed-relation with a moving image ("synching") will be considered an Adaptation for the purpose of this License.

b. **"Collection"** means a collection of literary or artistic works, such as encyclopedias and anthologies, or performances, phonograms or broadcasts, or other works or subject matter other than works listed in Section 1(f) below, which, by reason of the selection and arrangement of their contents, constitute intellectual creations, in which the Work is included in its entirety in unmodified form along with one or more other contributions, each constituting separate and independent works in themselves, which together are assembled into a collective whole. A work that constitutes a Collection will not be considered an Adaptation (as defined below) for the purposes of this License.

c. **"Creative Commons Compatible License"** means a license that is listed at http://creativecommons.org/compatiblelicenses that has been approved by Creative Commons as being essentially equivalent to this License, including, at a minimum, because that license: (i) contains terms that have the same purpose, meaning and effect as the License Elements of this License; and, (ii) explicitly permits the relicensing of adaptations of works made available under that license under this License or a Creative Commons jurisdiction license with the same License Elements as this License.

d. **"Distribute"** means to make available to the public the original and copies of the Work or Adaptation, as appropriate, through sale or other transfer of ownership.

e. **"License Elements"** means the following high-level license attributes as selected by Licensor and indicated in the title of this License: Attribution, ShareAlike.

f. **"Licensor"** means the individual, individuals, entity or entities that offer(s) the Work under the terms of this License.

g. **"Original Author"** means, in the case of a literary or artistic work, the individual, individuals, entity or entities who created the Work or if no individual or entity can be identified, the publisher; and in addition (i) in the case of a performance the actors, singers, musicians, dancers, and other persons who act, sing, deliver, declaim, play in, interpret or otherwise perform literary or artistic works or expressions of folklore; (ii) in the case of a phonogram the producer being the person or legal entity who first fixes the sounds of a performance or other sounds; and, (iii) in the case of broadcasts, the organization that transmits the broadcast.

h. **"Work"** means the literary and/or artistic work offered under the terms of this License including without limitation any production in the literary, scientific and artistic domain, whatever may be the mode or form of its expression including digital form, such as a book, pamphlet and other writing; a lecture, address, sermon or other work of the same nature; a dramatic or dramatico-musical work; a choreographic work or entertainment in dumb show; a musical composition with or without words; a cinematographic work to which are assimilated works expressed by a process analogous to cinematography; a work of drawing, painting, architecture, sculpture, engraving or lithography; a photographic work to which are assimilated works expressed by a process analogous to photography; a work of applied art; an illustration, map, plan, sketch or three-dimensional work relative to geography, topography, architecture or science; a performance; a broadcast; a phonogram; a compilation of data to the extent it is protected as a copyrightable work; or a work performed by a variety or circus performer to the extent it is not otherwise considered a literary or artistic work.

i. **"You"** means an individual or entity exercising rights under this License who has not previously violated the terms of this License with respect to the Work, or who has received express permission from the Licensor to exercise rights under this License despite a previous violation.

j. **"Publicly Perform"** means to perform public recitations of the Work and to communicate to the public those public recitations, by any means or process, including by wire or wireless means or public digital performances; to make available to the public Works in such a way that members of the public may access these Works from a place and at a place individually chosen by them; to perform the Work to the public by any means or process and the communication to the public of the performances of the Work, including by public digital performance; to broadcast and rebroadcast the Work by any means including signs, sounds or images.

k. **"Reproduce"** means to make copies of the Work by any means including without limitation by sound or visual recordings and the right of fixation and reproducing fixations of the Work, including storage of a protected performance or phonogram in digital form or other electronic medium.

2. Fair Dealing Rights. Nothing in this License is intended to reduce, limit, or restrict any uses free from copyright or rights arising from limitations or exceptions that are provided for in connection with the copyright protection under copyright law or other applicable laws.

3. License Grant. Subject to the terms and conditions of this License, Licensor hereby grants You a worldwide, royalty-free, non-exclusive, perpetual (for the duration of the applicable copyright) license to exercise the rights in the Work as stated below:

a. to Reproduce the Work, to incorporate the Work into one or more Collections, and to Reproduce the Work as incorporated in the Collections;

b. to create and Reproduce Adaptations provided that any such Adaptation, including any translation in any medium, takes reasonable steps to clearly label, demarcate or otherwise identify that changes were made to the original Work. For example, a translation could be marked "The original work was translated from English to Spanish," or a modification could indicate "The original work has been modified.";

c. to Distribute and Publicly Perform the Work including as incorporated in Collections; and,

d. to Distribute and Publicly Perform Adaptations.

e. For the avoidance of doubt:

i. **Non-waivable Compulsory License Schemes**. In those jurisdictions in which the right to collect royalties through any statutory or compulsory licensing scheme

cannot be waived, the Licensor reserves the exclusive right to collect such royalties for any exercise by You of the rights granted under this License;

ii. **Waivable Compulsory License Schemes**. In those jurisdictions in which the right to collect royalties through any statutory or compulsory licensing scheme can be waived, the Licensor waives the exclusive right to collect such royalties for any exercise by You of the rights granted under this License; and,

iii. **Voluntary License Schemes**. The Licensor waives the right to collect royalties, whether individually or, in the event that the Licensor is a member of a collecting society that administers voluntary licensing schemes, via that society, from any exercise by You of the rights granted under this License.

The above rights may be exercised in all media and formats whether now known or hereafter devised. The above rights include the right to make such modifications as are technically necessary to exercise the rights in other media and formats. Subject to Section 8(f), all rights not expressly granted by Licensor are hereby reserved.

4. Restrictions. The license granted in Section 3 above is expressly made subject to and limited by the following restrictions:

a. You may Distribute or Publicly Perform the Work only under the terms of this License. You must include a copy of, or the Uniform Resource Identifier (URI) for, this License with every copy of the Work You Distribute or Publicly Perform. You may not offer or impose any terms on the Work that restrict the terms of this License or the ability of the recipient of the Work to exercise the rights granted to that recipient under the terms of the License. You may not sublicense the Work. You must keep intact all notices that refer to this License and to the disclaimer of warranties with every copy of the Work You Distribute or Publicly Perform. When You Distribute or Publicly Perform the Work, You may not impose any effective technological measures on the Work that restrict the ability of a recipient of the Work from You to exercise the rights granted to that recipient under the terms of the License. This Section 4(a) applies to the Work as incorporated in a Collection, but this does not require the Collection apart from the Work itself to be made subject to the terms of this License. If You create a Collection, upon notice from any Licensor You must, to the extent practicable, remove from the Collection any credit as required by Section 4(c), as requested. If You create an Adaptation, upon notice from any Licensor You must, to the extent practicable, remove from the Adaptation any credit as required by Section 4(c), as requested.

b. You may Distribute or Publicly Perform an Adaptation only under the terms of: (i) this License; (ii) a later version of this License with the same License Elements as this License; (iii) a Creative Commons jurisdiction license (either this or a later license version) that contains the same License Elements as this License (e.g., Attribution-ShareAlike 3.0 US)); (iv) a Creative Commons Compatible License. If you license the Adaptation under one of the licenses mentioned in (iv), you must comply with the terms of that license. If you license the Adaptation under the terms of any of the licenses mentioned in (i), (ii) or (iii) (the "Applicable License"), you must comply with the terms of the Applicable License generally and the following provisions: (I) You must include a copy of, or the URI for, the Applicable License with every copy of each Adaptation You Distribute or Publicly Perform; (II) You may not offer or impose any terms on the Adaptation that restrict the terms of the Applicable License or the ability of the recipient of the Adaptation to exercise the rights granted to that recipient under the terms of the Applicable License; (III) You must keep intact all notices that refer to the Applicable License and to the disclaimer of warranties with every copy of the Work as included in the Adaptation You Distribute or Publicly Perform; (IV) when You Distribute or Publicly Perform the Adaptation, You may not impose any effective technological measures on the Adaptation that restrict the ability of a recipient of the Adaptation from You to exercise the rights granted to that recipient under the terms of the Applicable License. This Section 4(b) applies to the Adaptation as incorporated in a Collection, but this does not require the Collection apart from the Adaptation itself to be made subject to the terms of the Applicable License.

c. If You Distribute, or Publicly Perform the Work or any Adaptations or Collections, You must, unless a request has been made pursuant to Section 4(a), keep intact all copyright notices for the Work and provide, reasonable to the medium or means You are utilizing: (i) the name of the Original Author (or pseudonym, if applicable) if supplied, and/or if the Original Author and/or Licensor designate another party or parties (e.g., a sponsor institute, publishing entity, journal) for attribution ("Attribution Parties") in Licensor's copyright notice, terms of service or by other reasonable means, the name of such party or parties; (ii) the title of the Work if

supplied; (iii) to the extent reasonably practicable, the URI, if any, that Licensor specifies to be associated with the Work, unless such URI does not refer to the copyright notice or licensing information for the Work; and (iv) , consistent with Section 3(b), in the case of an Adaptation, a credit identifying the use of the Work in the Adaptation (e.g., "French translation of the Work by Original Author," or "Screenplay based on original Work by Original Author"). The credit required by this Section 4(c) may be implemented in any reasonable manner; provided, however, that in the case of a Adaptation or Collection, at a minimum such credit will appear, if a credit for all contributing authors of the Adaptation or Collection appears, then as part of these credits and in a manner at least as prominent as the credits for the other contributing authors. For the avoidance of doubt, You may only use the credit required by this Section for the purpose of attribution in the manner set out above and, by exercising Your rights under this License, You may not implicitly or explicitly assert or imply any connection with, sponsorship or endorsement by the Original Author, Licensor and/or Attribution Parties, as appropriate, of You or Your use of the Work, without the separate, express prior written permission of the Original Author, Licensor and/or Attribution Parties.

d. Except as otherwise agreed in writing by the Licensor or as may be otherwise permitted by applicable law, if You Reproduce, Distribute or Publicly Perform the Work either by itself or as part of any Adaptations or Collections, You must not distort, mutilate, modify or take other derogatory action in relation to the Work which would be prejudicial to the Original Author's honor or reputation. Licensor agrees that in those jurisdictions (e.g. Japan), in which any exercise of the right granted in Section 3(b) of this License (the right to make Adaptations) would be deemed to be a distortion, mutilation, modification or other derogatory action prejudicial to the Original Author's honor and reputation, the Licensor will waive or not assert, as appropriate, this Section, to the fullest extent permitted by the applicable national law, to enable You to reasonably exercise Your right under Section 3(b) of this License (right to make Adaptations) but not otherwise.

5. Representations, Warranties and Disclaimer

UNLESS OTHERWISE MUTUALLY AGREED TO BY THE PARTIES IN WRITING, LICENSOR OFFERS THE WORK AS-IS AND MAKES NO REPRESENTATIONS OR WARRANTIES OF ANY KIND CONCERNING THE WORK, EXPRESS, IMPLIED, STATUTORY OR OTHERWISE, INCLUDING, WITHOUT LIMITATION, WARRANTIES OF TITLE, MERCHANTIBILITY, FITNESS FOR A PARTICULAR PURPOSE, NONINFRINGEMENT, OR THE ABSENCE OF LATENT OR OTHER DEFECTS, ACCURACY, OR THE PRESENCE OF ABSENCE OF ERRORS, WHETHER OR NOT DISCOVERABLE. SOME JURISDICTIONS DO NOT ALLOW THE EXCLUSION OF IMPLIED WARRANTIES, SO SUCH EXCLUSION MAY NOT APPLY TO YOU.

6. Limitation on Liability. EXCEPT TO THE EXTENT REQUIRED BY APPLICABLE LAW, IN NO EVENT WILL LICENSOR BE LIABLE TO YOU ON ANY LEGAL THEORY FOR ANY SPECIAL, INCIDENTAL, CONSEQUENTIAL, PUNITIVE OR EXEMPLARY DAMAGES ARISING OUT OF THIS LICENSE OR THE USE OF THE WORK, EVEN IF LICENSOR HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

7. Termination

a. This License and the rights granted hereunder will terminate automatically upon any breach by You of the terms of this License. Individuals or entities who have received Adaptations or Collections from You under this License, however, will not have their licenses terminated provided such individuals or entities remain in full compliance with those licenses. Sections 1, 2, 5, 6, 7, and 8 will survive any termination of this License.

b. Subject to the above terms and conditions, the license granted here is perpetual (for the duration of the applicable copyright in the Work). Notwithstanding the above, Licensor reserves the right to release the Work under different license terms or to stop distributing the Work at any time; provided, however that any such election will not serve to withdraw this License (or any other license that has been, or is required to be, granted under the terms of this License), and this License will continue in full force and effect unless terminated as stated above.

8. Miscellaneous

a. Each time You Distribute or Publicly Perform the Work or a Collection, the Licensor offers to the recipient a license to the Work on the same terms and conditions as the license granted to You under this License.

b. Each time You Distribute or Publicly Perform an Adaptation, Licensor offers to the recipient a license to the original Work on the same terms and conditions as the license granted to You under this License.
c. If any provision of this License is invalid or unenforceable under applicable law, it shall not affect the validity or enforceability of the remainder of the terms of this License, and without further action by the parties to this agreement, such provision shall be reformed to the minimum extent necessary to make such provision valid and enforceable.
d. No term or provision of this License shall be deemed waived and no breach consented to unless such waiver or consent shall be in writing and signed by the party to be charged with such waiver or consent.
e. This License constitutes the entire agreement between the parties with respect to the Work licensed here. There are no understandings, agreements or representations with respect to the Work not specified here. Licensor shall not be bound by any additional provisions that may appear in any communication from You. This License may not be modified without the mutual written agreement of the Licensor and You.
f. The rights granted under, and the subject matter referenced, in this License were drafted utilizing the terminology of the Berne Convention for the Protection of Literary and Artistic Works (as amended on September 28, 1979), the Rome Convention of 1961, the WIPO Copyright Treaty of 1996, the WIPO Performances and Phonograms Treaty of 1996 and the Universal Copyright Convention (as revised on July 24, 1971). These rights and subject matter take effect in the relevant jurisdiction in which the License terms are sought to be enforced according to the corresponding provisions of the implementation of those treaty provisions in the applicable national law. If the standard suite of rights granted under applicable copyright law includes additional rights not granted under this License, such additional rights are deemed to be included in the License; this License is not intended to restrict the license of any rights under applicable law.

Natural Language Processing, Corpus Linguistics, E-learning

Editors Katarína Gajdošová and Adriána Žáková

Cover Design by Vladimír Benko
Typeset by Radoslav Brída and Ján Mášik

Printed and published by RAM-Verlag
Stüttinghauser Ringstrasse 44, 58515 Lüdenscheid, Germany

303 pages

First edition at RAM-Verlag
Lüdenscheid 2013

ISBN 978-3-942303-18-7

http://www.ram-verlag.eu