

Slovenská akadémia vied
Jazykovedný ústav L. Štúra SAV

Natural Language Processing, Corpus Linguistics, Lexicography

Eighth International Conference
Bratislava, Slovakia, 21–22 October 2015
Proceedings

Editors
Katarína Gajdošová
Adriána Žáková

RAM-Verlag
2015

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

The articles can be used under the
Creative Commons Attribution-ShareAlike 3.0 Unported Licence.



Slovak National Corpus
E. Štúr Institute of Linguistics
Slovak Academy of Sciences
Bratislava, Slovakia 2015
<http://korpus.sk/slovko.html>

© by respective authors, 2015
Editors © Katarína Gajdošová and Adriána Žáková, 2015
Typography © Vladimír Radik and Radoslav Brída, 2015
Cover © Vladimír Benko, 2015

This edition © RAM-Verlag, 2015

Table of Contents

Foreword <i>Mária Šimková</i>	7
Úvod <i>Mária Šimková</i>	8
Fran: The Next Generation Slovenian Dictionary Portal <i>Kozma Ahačič, Nina Ledinek, and Andrej Perdih</i>	9
Recent Status of Development of the ALEXIS Dictionary Writing System <i>Kamil Barbierik, Martin Bodlák, Zuzana Děngeová, Vladimír Jarý, Tomáš Liška, Michaela Lišková, Josef Nový, and Miroslav Virius</i>	17
Corpus of Dialects of the Slovak National Corpus <i>Katarína Gajdošová, Radovan Garabík, and Mária Šimková</i>	26
A New Dialect Corpus: DIALEKT <i>Hana Goláňová</i>	36
The Automatic Identification of Type of Adverbials in Syntactically Annotated Texts <i>Milena Hnátková</i>	45
The Use of Acronyms in Different Communication Modes (a Corpus-Based Study) <i>Zuzana Komrsková</i>	63
Evaluating Automatic Idiom Annotation in Spoken Corpora: the Case of Somatic Idioms <i>Marie Kopřivová</i>	72
SK ANTIPLAG: Five Years After <i>Július Kravjar</i>	77
New Tools for Working with the ORAL Series Corpora of Spoken Czech: AchSynku and MluvKonk <i>David Lukeš</i>	90
Traitements automatiques en lexicographie de langues « non dotées » <i>Patrice Pognan and Miloud Taiïfi</i>	102
Identifying Corpus-specific Collocations: The Case of Spoken Slovene <i>Senja Pollak and Špela Arhar Holdt</i>	117
Myslím, já myslím, myslím, že... Forms and Functions of the Verb <i>myslet</i> in Spoken Czech <i>Petra Poukarová</i>	126

Czech Deadverbial Prepositions Used in Postposition to a Noun (Exemplified in Corpus SYN 2010)	
<i>Aksana Schillova</i>	133
On Building the Slovak Automatic Semantic Role Labeling System	
<i>Ján Staš, Daniel Hládek, Stanislav Ondáš, and Jozef Juhár</i>	141
Slavic Languages in Universal Dependencies	
<i>Daniel Zeman</i>	151

Foreword

Slovko 2015 – this year's edition entitled *NLP, Corpus Linguistics, Lexicography* – represents a follow up of previous autumn meetings in Bratislava. Organisers, both from the Slovak National Corpus of the L. Štúr Institute of Linguistics, Slovak Academy of Sciences, and from the Slovak Centre of Scientific and Technical Information, are honoured to welcome participants from five countries: Austria, Czech Republic, France, Slovakia and Slovenia.

Two conference days offer 18 presentations, including two plenary talks. Not all papers registered for presentation were also published – current programme comprises also two presentations that cannot be found in the proceedings. Members of the programme committee carefully reviewed every paper sent with the registration (two reviewers for each text) and thus contributed to the overall quality of the scientific event and of this publication, for which we would like to express our sincere gratitude.

The 8th edition of the biannual conference *Slovko 2015* experiences the increase of papers dealing with corpus linguistics including lexicography. On the other hand, computationally oriented papers are in a minority. There is a significant shift from presenting new written corpora and their analyses to the issues concerning the building and research of spoken, even dialect corpora. We believe that this focus of papers will also become a source of inspiration both for conference participants and readers of the proceedings in their further work in the area of NLP, corpus linguistics and related research in Slovakia and neighbouring countries.

We wish all the participants of *Slovko 2015* an enjoyable stay in the Slovak Centre of Scientific and Technical Information and in Bratislava in particular to those who came from abroad. We would also like to invite you to *Slovko 2017* that will be focusing, besides NLP and corpus linguistics, on computational terminology and terminography.

Mária Šimková
Translated by Jana Levická

Úvod

Slovko 2015 – tentoraz s podtitulom *počítačové spracovanie prirodzeného jazyka, korpusová lingvistika, lexikografia* – nadväzuje na predchádzajúce jesenné stretnutia v Bratislave. Organizátori zo Slovenského národného korpusu Jazykovedného ústavu Ľudovíta Štúra Slovenskej akadémie vied a Centra vedecko-technických informácií Slovenskej republiky vítajú na tomto podujatí účastníkov z piatich krajín: Česká republika, Francúzsko, Rakúsko, Slovensko a Slovinsko.

Počas dvoch dní rokovania odznie celkovo 18 príspevkov, z toho dve plenárne prednášky. Nie všetky príspevky prihlásené na prezentáciu sú aj publikované – v aktuálnom programe odznejú aj tri príspevky, ktoré sa v tomto zborníku nenachádzajú. Členovia vedeckého výboru starostlivo posúdili každý prihlásený článok (jeden text dvaja posudzovatelia) a prispeli tak ku kvalite celého podujatia a tejto publikácie, za čo im patrí naše poďakovanie.

Na 8. ročníku bienálnej konferencie *Slovko 2015* sa zvýšil počet príspevkov z oblasti korpusovej lingvistiky vrátane lexikografie. Technicky orientovaných príspevkov je tentoraz menej. Výrazný posun nastal od prezentácií budovania a analýz písaných korpusov smerom k riešeniu otázok tvorby a výskumov hovorených, dokonca aj nárečových korpusov. Veríme, že aj takto zamerané príspevky budú pre poslucháčov a diskutujúcich na konferencii, ako aj pre čitateľov publikovaných štúdií prínosom a inšpiráciou do ďalších prác v oblasti počítačového spracovania prirodzeného jazyka, korpusovej lingvistiky a súvisiacich výskumov na Slovensku i v okolitých krajinách.

Všetkým účastníkom konferencie *Slovko 2015* želáme príjemný pobyt v Centre vedecko-technických informácií, mimobratislavským účastníkom aj v Bratislave vôbec. Zároveň vás už teraz pozývame na *Slovko 2017*, ktoré bude okrem počítačového spracovania prirodzeného jazyka a korpusovej lingvistiky primárne venované počítačovej terminológii a terminografii.

Mária Šimková

Fran: The Next Generation Slovenian Dictionary Portal

Kozma Ahačič, Nina Ledinek, and Andrej Perdih

Scientific Research Centre, Slovenian Academy of Sciences and Arts,
Ljubljana, Slovenia

Abstract. The article presents Fran, the new Slovenian language portal at the Fran Ramovš Institute of the Slovenian Language of the Research Centre of the Slovenian Academy of Sciences and Arts, containing all key Slovenian monolingual dictionaries. As a whole it contains 22 dictionaries and linguistic atlases describing the Slovenian language in its contemporary, historical and dialectal forms, Slovenian terminology in various fields, and includes language counselling and hyperlinks to corpora and other language resources. Great emphasis was placed on simple navigation and clear visualization. To the general users, the Fran portal offers contents they most frequently search for, and helps them to understand the contents. To the advanced users, it brings advanced search options, the possibility to change settings, and offers additional content.

1 Introduction

At the Fran Ramovš Institute of the Slovenian Language, 33 dictionaries have been published in the last 15 years, therefore there has been a great need to offer a larger amount of language resources integrated on the web portal and at the same time to make the concurrent use of several resources as informative and as clear as possible. For this reason the main goal was creating a simple navigation system with clear and intuitive graphical representation of the data.¹ The minimalistic graphic approach reducing the amount of data presented made it possible to publish older dictionaries (primarily published in print) as well as newer dictionaries (primarily built for use in electronic form). Since the former dictionary portal of the Institute was used by very different user groups,² it is reasonable to expect that Fran will be used by various users with various demands. Therefore, different help functions are offered to general users, while additional functionalities and contents are offered to professional users.

The Fran portal (full title: *Fran: The Dictionaries of the Fran Ramovš Institute of Slovenian Language of the Research Centre of the Slovenian Academy of Sciences and Arts*)³ was published on October 14, 2014.⁴ At the time of writing it contains 22 dictionaries and linguistic atlases and several other language resources. Among them, several dictionaries are published in electronic version for the first time. The portal offers information on

¹ The Fran portal was designed by the editor Kozma Ahačič with help of the editorial board (Helena Dobrovoljc, Nina Ledinek, Andrej Perdih, Marko Snoj, Peter Weiss, Mojca Žagar Karer) and Alenka Porenta who prepared the information about dictionaries. The portal was built by Amebis, d. o. o., and graphically designed by Hruška, d. o. o.

² Website address: bos.zrc-sazu.si. Until recently, it was the only site where the general monolingual dictionary *Slovar slovenskega knjižnega jezika* and the currently valid orthographic dictionary (*Slovenski pravopis 2001*) were freely available (later published at www.termania.net as well [25]), therefore it was used by general users as well as professional users.

³ Original title: *Fran: Slovarji Inštituta za slovenski jezik Frana Ramovša ZRC SAZU*. The portal got its name from Slovenian linguist Fran Ramovš.

⁴ In January 2015, the portal was updated to version 2.0.

lexemes of the contemporary Slovenian language as well as information on its historical development and dialectal usage at one place.⁵ All but one dictionary⁶ are freely available.

2 Content

On the Fran portal, information on the Slovenian language can be obtained from the view of contemporary standard language as well as from the historical view from the 16th century until the present or across all Slovenian dialects. At the time of writing, 22 dictionaries and linguistic atlases are available, as well as language counselling and other language resources, divided into 6 sections:

- **general** dictionaries,
- **historical** dictionaries,
- **terminological** dictionaries,
- **dialect** dictionaries,
- language **counselling**,
- **other resources**.

The group of **general** dictionaries includes dictionaries most frequently consulted by users, i.e. general monolingual dictionaries [28], [29], dictionaries of neologisms [17], [26], and orthographic dictionaries [27], [31]. Among **historical** dictionaries bilingual dictionaries published until 1895 are included [24], [34] as well as recently published monolingual dictionaries and a glossary describing historical stages in language development [1], [21], [33]. The largest group is made up of **terminological** dictionaries [2], [3], [6], [8], [9], [10], [11], [23], [32], [35]. The group of **dialect** dictionaries currently contains only the newest linguistic atlas [30]. Dialect dictionaries, primarily published in print edition, are currently being prepared for publishing in electronic form. **Counselling** consists of two parts: general language counselling [4] and terminology counselling [36]. The **other resources** consist of hyperlinks to resources created at the Fran Ramovš Institute of the Slovenian Language.⁷ They are not functionally integrated into the portal.

The dictionary data is written in XML format. Although the newest dictionaries are originally written in XML format (using iLex software [5]) or exported from a relational database (SlovarRed [16]), some dictionaries were written in EVA [14] or even Microsoft Word and then transformed into XML. Other dictionaries, including *The Dictionary*

⁵ Internet portals for other Slavic languages also try to provide various linguistic resources at one place by integrating various dictionaries, other linguistic information, and resources, e.g. *Hrvatski jezični portal* (<http://hjp.novi-liber.hr/>) for Croatian, *Internetová jazyková příručka* (<http://prirucka.ujc.cas.cz/>) for Czech, *Slovenské slovníky* <http://slovník.juls.savba.sk/>.

⁶ On the same date as Fran came into existence, the second, updated and partly renewed edition of the general monolingual dictionary *Slovar slovenskega knjižnega jezika* was published [22]. This is the only dictionary on the portal that requires registration code.

⁷ These are the reference corpus *Nova beseda* and manually morphosyntactically annotated corpus *O'Beseda*, the *Besede slovenskega jezika* word list compiled from various resources, *Besedišče* (a glossary of 178,457 words not included in *The Dictionary of Standard Slovenian Language*) and a hyperlink to the old web portal of the Institute.

of *Standard Slovenian Language* [13] were retro-digitised from printed edition. At the moment, the dictionaries do not follow a common XML Schema encoding, therefore the same information can be encoded differently in different dictionaries. Terminological dictionaries, however, all use the same schema. Furthermore, some dictionaries contain delimiters in XML while others do not. Although this is not an ideal situation, offering these dictionaries online was set as a priority. The linguistic atlas is available in PDF format.

3 Functionality and Design



Fig. 1. Navigation panel

All dictionaries, the linguistic atlas and counselling websites are integrated in such a way that although only search results from selected dictionaries are presented to the user (by default the search is performed in all resources), the navigation panel gives information on the number of results in all the dictionaries, the atlas and counselling websites as well as provides links to navigate into any one of them. Additionally, direct links to search results in the reference corpora *Gigafida* and *Nova beseda*, the balanced corpus *Kres*, the corpus of spoken Slovenian *Gos*, the manually morphosyntactically annotated corpus *O'beseda* and in the word list *Besede slovenskega jezika* and the glossary *Besedišče* are available.

There are 515,183 dictionary entries altogether on the Fran portal (see the blue number of the first line in Fig. 1 “Vse na Franu”). Upon clicking or tapping the short dictionary title or counselling (under *Slovarji* “Dictionaries” or *Svetovanje* “Counselling”) all entries containing search result anywhere in the entry of the selected dictionary are shown. Clickable numbers in the blue label represent the number of entries with search results in headwords and subheadwords⁸ in the selected dictionary, the green labels represent the number of entries that contains search results outside of headwords and subheadwords.

By default, search results from the most frequently consulted dictionaries are shown first, i.e. the results from the second (or first) edition of monolingual dictionary (*Slovar slovenskega knjižnega jezika*) are shown first, followed by results from the dictionary of neologisms (*Slovar novejšega besedja slovenskega jezika*) and orthographic dictionary (*Slovenski pravopis* 2001).

The portal offers two search types: simple and advanced. By default, simple search is used. It returns the results in headwords and subheadwords (as explained above, dictionary entries with results in other dictionary elements are available from the navigation panel). Typing of the search condition is facilitated by the incremental search functionality, offering the user up to five (sub)headwords.

Using advanced search, it is possible to select a dictionary or a group of dictionaries in which the search should be performed. Additionally, it is also possible to define structure elements where the search string is to be found. If a structure element is not available in all selected dictionaries, the input field is greyed out. Depending on the dictionary structure it is possible to further specify searching for headword, definition, label, terminology, phraseology etc.

⁸ Searching for subheadwords is often found problematic [12]. By searching both in headwords and subheadwords the problem of finding subentries is made easier.

Fig. 2. Advanced search in several elements (right) of the selected dictionaries (left)

Fig. 3. Advanced search: highlighted results

In both simple and advanced search, wildcards can be used (* matches any string, ? matches any character), as well as double quotes (“ ” for phrase searching). For additional search possibilities both in simple and advanced search, Apache Lucene⁹ functionalities are provided including the use of regular expressions. The default search setting uses normalization of accentuated diacritical characters (searching for *e* finds *éêë...*), but when searching for a specific diacritical character normalization is not used (*ê* does not search for *eêë...*). Since the portal is aimed at very different types of users, i.e. general and professional ones, including linguists, two virtual keyboards for inserting special characters have been created. When using simple search, the keyboard containing 42 most common special characters is shown, while the keyboard for advanced search contains 244 special characters.

When no result is found using simple search, the search engine offers up to five similar results. First researches have shown that the technology used brings fairly good results [15].

⁹ <https://lucene.apache.org/>

For advanced users, especially for linguists, dictionary front and back matters are offered (prefaces, introductions, lists of abbreviations etc.).

Sorting of the search results can be ordered in three ways. As explained above, results found in (sub)headwords of the most frequently consulted dictionaries are ranked highest by default. Afterwards, partial results are shown, e.g. when the result is found as a part of multi-word unit. Search results can also be sorted according to the dictionary. This means that all results from one dictionary are presented together. If the user selects searching in one dictionary only, alphabetic entry sorting is also available, showing a clickable headword list (in alphabetic order) in a panel on the left hand side and presenting entries next to it. This option is useful especially when searching anywhere in the dictionary entry.

Majority of the dictionaries on the Fran portal was primarily published in print and therefore follow the traditionally developed lexicographic conventions. Understanding dictionary structures, abbreviations and symbols that require consulting front matter in printed edition, are brought to the user of the portal in the form of tooltips when hovering the text or symbols with mouse or tapping the screen. For example, *elektrotehnika* “electrotechnics” is shown for *elektr.*, and *samostalnik srednjega spola* “neuter noun” for *s*, and *Terminološko gnezdo* “Terminology section” is shown for the ♦ symbol.¹⁰



Fig. 4. Help on symbols



Fig. 5. Expanding abbreviations

General language counselling and terminology counselling sites are integrated into Fran. They had existed for several years already, but were integrated as the questions and answers they contain often complement the information in the existing dictionaries. *Jezikovno svetovanje* (general language counselling) offers counselling on general language problems, most frequently connected with orthography and grammar, not infrequently also with language development. *Terminološko svetovanje* (terminology counselling) is aimed at experts in all fields. They can get advice on solving concrete problems of naming new concepts or get help making a choice when several terms exist for the same concept and there is a need of standardization of terminology or just selecting an appropriate term.

The Fran portal offers interactivity between the portal and its users. A word list containing loanwords that have been borrowed recently was built from the existing dictionaries of neologisms. Users are offered to suggest words that could replace the loanwords if people accept the suggestion and start using it. The selected suggestions are

¹⁰ This mechanism depends on structural elements and their content. In the process of retro-digitization, some dictionaries have not yet been adequately structured, therefore some mistakes are possible. These issues will be solved in the future.

available for user voting or sharing through Twitter. In the first five months since opening the Fran portal, the favourite word seems to be *selfi/selfie* having 16 suggestions for an alternative expression, some of them actually being used in various texts.

Regarding the design it is worth mentioning that it is adjusted for five different screen sizes from the smallest (for smartphones) to larger sizes used in laptop and desktop screens.

4 Future Improvements

In 2015, an upgrade is being prepared regarding content, functionality and design of the portal. Commenting on the dictionary entries is to be enabled, a word list containing recent loanwords is to be build and new dictionaries are to be added. In this and the following upgrades some of the existing dictionaries of etymology, phraseology and several dialectal dictionaries are to be included. Upon completion in the following years, the dictionary of synonyms as well as the first part of the dictionary of the 16th century standard Slovenian language and new dictionary of the Slovenian standard language will be included. Although adding dictionaries is welcome, there is a need to keep the portal clear and easily usable, therefore it is possible that some rarely consulted dictionaries would not be shown anymore by default, however, users would have a possibility to choose their preferred dictionaries to be shown. After all, how useful can information, lost in a huge amount of other information, be?

5 Conclusion

The Fran portal is the newest Slovenian dictionary portal, offering 22 dictionaries and linguistic atlases as well as other language resources. It contains most frequently consulted dictionaries (general monolingual dictionary, orthographic dictionary) as well as terminological and historical dictionaries, a linguistic atlas and language counselling. Among them, several dictionaries are published in electronic versions for the first time. As the portal is aimed at various user groups, general users are offered help functionality and possibility to adjust search result ordering, while advanced users can benefit from advanced search and additional information on dictionaries, i.e. front matters. Beside that, the portal also offers some interactive functionalities.

References

- [1] *Besedje slovenskega knjižnega jezika 16. stoletja*. (2011). Založba ZRC, ZRC SAZU, Ljubljana.
- [2] *Botanični terminološki slovar*. (2011). Založba ZRC, Ljubljana.
- [3] *Čebelarški terminološki slovar*. (2008). Založba ZRC, ZRC SAZU, Ljubljana.
- [4] Dobrovoljc, H. and Bizjak Končar, A. (2015). Pravopisno slovaropisje na Inštitutu za slovenski jezik Frana Ramovša ZRC SAZU. *Slavia Centralis*, VIII(1):34–50.
- [5] Erlandsen, J. (2004). iLex – New DWS. In *Third International Workshop on Dictionary Writing Systems (DWS 2004)*. Accessible at: <http://nlp.fi.muni.cz/dws2004/pres/#15>.
- [6] *Farmacevtski terminološki slovar*. (2011). Založba ZRC, Ljubljana.
- [7] Fran, slovarji Inštituta za slovenski jezik Frana Ramovša ZRC SAZU, 2014–, version 2.0, URL: www.fran.si, retrieved 30 March 2015.
- [8] *Gemološki terminološki slovar*. (2005). Založba ZRC, ZRC SAZU, Ljubljana.
- [9] *Geografski terminološki slovar*. (2005). Založba ZRC, ZRC SAZU, Ljubljana.
- [10] *Geološki terminološki slovar*. (2006). Založba ZRC, ZRC SAZU, Ljubljana.
- [11] *Gledališki terminološki slovar*. (2007). Založba ZRC, Ljubljana.

- [12] Gouws, R. H. (2007). Sublemmata or main lemmata. A critical look at the presentation of some macrostructural elements. In *Dictionary Visions, Research and Practice*, pages 55–69, John Benjamins Publishing Company, Amsterdam/Philadelphia.
- [13] Hajnšek Holz, M. (1993). Leksikografski problemi prenosa knjižne oblike Slovarja slovenskega knjižnega jezika v računalniško. In *Jezik tako in drugače*, pages 420–432, Društvo za uporabno jezikoslovje Slovenije, Ljubljana.
- [14] Jakopin, P. (1995). EVA – A Textual Data Processing Tool. In *Proceedings of the first TELRI seminar: Language Resources for Language Technology*. Tihany.
- [15] Kosem, I. (2014). Fran: pameten in intuitiven? *Slovenščina* 2.0, 2(2):161–193.
- [16] Košmrlj-Levačič, B. and Seliškar, T. (2004). Uporabniški računalniški program SlovarRed 2.0 = SlovarRed 2.0 – a terminological application. In *Terminologija v času globalizacije: zbornik prispevkov s simpozija Terminologija v času globalizacije, Ljubljana, 5.–6. junij 2003 = Terminology at the time of globalization: collected papers from the Scientific Conference Terminology at the Time of Globalization, Ljubljana, 5th-6th June 2003*, pages 179–199, Znanstvenoraziskovalni center SAZU, Založba ZRC, Ljubljana.
- [17] Krvina, D. (2014). *Sproti slovar slovenskega jezika*. ZRC SAZU, Ljubljana.
- [18] Lew, R. (2012). How can we make electronic dictionaries more effective? In *Electronic Lexicography*, pages 343–363, Oxford University Press, Oxford.
- [19] Müller-Spitzer, C., Kopenig, A., and Töpel, A. (2012). Online dictionary use: Key findings from an empirical research project. In *Electronic Lexicography*, pages 425–458, Oxford University Press, Oxford.
- [20] Lew, R. and Mitton, R. (2011). Not the Word I Wanted? How Online English Learners' Dictionaries Deal with Misspelled Words. In *Electronic Lexicography in the 21st Century: New Applications for New Users. Proceedings of eLex 2011, Bled, 10–12 November 2011*, pages 165–174, Institute for Applied Slovene Studies, Trojina, Ljubljana.
- [21] Novak, V. (2006). *Slovar stare knjižne prekmurščine*. Založba ZRC, ZRC SAZU, Ljubljana.
- [22] Perdih, A. and Snoj, M. (2015). SSKJ². *Slavia Centralis*, VIII(1):5–15.
- [23] *Planinski terminološki slovar*. (2002). Založba ZRC, ZRC SAZU, Ljubljana.
- [24] Pleteršnik, M. (1894–1895). *Slovensko-nemški slovar*. Knezoškofijstvo, Ljubljana .
- [25] Romih, M. and Krek, S. (2012). Termania – prosto dostopni spletni slovarski portal. In *Zbornik Osmo konference Jezikovne tehnologije*, pages 163–166, Institut Jožef Stefan, Ljubljana, URL: http://www.trojina.org/slovenscina2.0/arhiv/2014/2/Slo2.0_2014_2_09.pdf.
- [26] *Slovar novejšega besedja slovenskega jezika*. (2013). Založba ZRC, ZRC SAZU, Ljubljana.
- [27] *Slovar pravopisnih težav*. (2014–). ZRC SAZU, Ljubljana.
- [28] *Slovar slovenskega knjižnega jezika*. (1970–1991). SAZU, Inštitut za slovenski jezik, Državna založba Slovenije, Ljubljana.
- [29] *Slovar slovenskega knjižnega jezika. Druga, dopolnjena in deloma prenovljena izdaja*. (2014). Cankarjeva založba, Ljubljana.
- [30] *Slovenski lingvistični atlas 1*. (2011). Založba ZRC, ZRC SAZU, Ljubljana.
- [31] *Slovenski pravopis*. (2001). Založba ZRC, ZRC SAZU, Ljubljana.
- [32] *Slovenski smučarski slovar*. (2011). Založba ZRC, Ljubljana.
- [33] Snoj, M. (2006). *Slovar jezika Janeza Svetokriškega*. Založba ZRC, Ljubljana.
- [34] Stabej, J. (1997). *Slovensko-latinski slovar po: Matija Kastelec – Gregor Vorenc, Dictionarium Latino-Carniolicum (1608–1710)*. ZRC SAZU, Založba ZRC, Ljubljana.
- [35] *Terminološki slovar avtomatike*. (2014). Založba ZRC, Ljubljana.
- [36] Žagar Karer, M. (2015). Terminologišče – kraj, kjer terminolog išče. *Slavia Centralis*, VIII(1):22–33.

Recent Status of Development of the ALEXIS Dictionary Writing System

Kamil Barbierik, Martin Bodlák, Zuzana Děngeová, Vladimír Jarý, Tomáš Liška, Michaela Lišková, Josef Nový, and Miroslav Virius

Institute of the Czech Language, Academy of Sciences of the Czech Republic, Prague, Czech Republic

Abstract. The paper focuses on the new dictionary writing system called Alexis that has been currently developed at the Institute of the Czech Language of the Academy of Sciences of the Czech Republic, v. v. i. We review the used software technologies and the structure of the system that consists of the list of entries and the editing, output, editorial, and administrative modules. Then we describe the features that were developed during last year. We start with modifications of the editorial and output modules that include the introduction of user roles and the introduction of the system of messages that can be sent to editors. After that we present the updates of the list of entries module that now supports sharing of lists of lemmas. Finally, we present a brand new web interface, intended to be used by the public. We also mention some of the planned new modules, including a revision control module.

1 Introduction

The dictionary writing system (DWS) *Alexis* has been developed as software support for writing a monolingual dictionary of contemporary Czech – *Akademický slovník současné češtiny* [5] that is being prepared at the Department of Contemporary Lexicology and Lexicography of the Institute of the Czech Language of the Czech Academy of Sciences of the Czech Republic, v. v. i. *Alexis* has already been presented [4], [3] but it has been considerably developed since then. The dictionary is expected to be a medium-sized dictionary with approximately 150 000 lexical units. It is targeted at native speakers with secondary education.

First we will briefly discuss the motivation for development of a new system. Then we will focus on technical details of the system and will review the main modules that model both the microstructure and the macrostructure of the dictionary. We will focus on the features that were implemented during last year, including editorial tools, user management, and a public web interface. We will conclude with evaluation of the current status of the project.

2 Motivation

At first, we considered to adopt (with possible modification) some of the already existing dictionary writing systems, commercial or open-source. However, after some discussion between the lexicographic team and the programmers included in the project [1] we decided to develop a customer software from scratch, mainly for these reasons:

1. We will have full control over the source code and development cycle.
2. We will be able to model specific requirements of the lexicographers.
3. The lexicographers will work with system that suits their needs.

On the other hand, writing a brand new system requires a lot of effort, at least at the beginning of the development process.

3 Used Technologies

According to the requirements of the lexicographers, Alexis should enable simultaneous work of a number of users from various locations. In order to simplify deployment of the system, we decided to implement it as a multi-layered web application. Since dictionary data are highly structured, but mostly of textual character, we decided to use the *MySQL* relational database as a data layer. The database structure that models the dictionary is rather complex with approximately 100 tables built around the *lemma* table. Each lemma in the dictionary is uniquely identified by its identification number *id* in the database. Since multiple variants and explanations of the meaning can correspond to the same lemma, they are implemented as one-to-many relationship in the database.

The application layer that lies between the data layer and the presentation layer is mainly implemented in the PHP scripting language and is deployed on the Apache HTTP server. The presentation layer is based on the *HTML 5* and *CSS 3* technologies with client side scripting implemented in the *JavaScript* language with *jQuery* and *jQuery UI* frameworks. Therefore the lexicographers can use the system from any device with a web browser connected to the Internet.

As the system is developed by several programmers, it was necessary to use some kind of distributed revision control and a source code management system. We adopted the branching model of the *Git* system proposed by Driessen, V. (2010) and based our development on the *GitHub* platform. We use two main branches: a master branch that represents a stable code and a *devel* branch that corresponds to the development code. When a serious problem is detected in the stable branch, a new *hotfix* branch is created, the problem is solved in this branch, and the code of the hotfix is then merged both in the master and the *devel* branches. Recently, we have introduced a *refactor* branch which will be mainly used to implement performance improvements.

In order to simplify the communication between the lexicographers and the programmers, we have decided to make use of the Mantis issue tracking system.

4 Overview of the System

The presented dictionary writing system can be divided into the following interconnected modules:

- administration module
- list of entries module
- editing module
- output module
- editorial and message modules
- public web interface.

The list of entries module and the editing module have already been presented [1] but many new features have been recently implemented in the other modules.

4.1 Administration

Each user accessing Alexis needs to authenticate before he or she starts using the system. The user account management is currently the most important part of the administration module. To each user account one of the following roles is assigned:

- The *administrator* has full control of the system and he can use the administration module to create or alter the other user accounts. He can also delete any lemma from the dictionary.
- The *editor* can view and edit dictionary entries and use the editorial module.
- The *domain expert* can only view assigned lemmas and use editorial tools on these lemmas.
- The *visitor* can only view lemmas assigned to his account.

During the process of building a new dictionary, cooperation with specialists is a very important part of our work. First, it is appropriate to work together with linguistic specialists – in order to check grammatical information, pronunciation, regional distribution and stylistic markers. Also it is necessary to collaborate with experts from other scientific disciplines. Experts in physics, biology, chemistry, social sciences, religious studies and many others check lexemes as special terms. Preview of the data is also available to other people – grant auditors (NAKI), auditors of the Academy of Sciences, or journalists. When providing selected data to experts and other users, the database system Alexis is a very powerful tool to use. Depending on the type of the user account, users can read, comment or update the data directly in the database system, without any need of additional administration that was required previously (collecting/transferring the data in .pdf/.doc format/transcript to the database). The user “administrator“ assigns selected data, the user “domain expert“ can view selected data and can post comments to the assigned lemmas and parts of lexemes. The lexicographer (the user “editor“) analyses and processes the comments of experts. The user “visitor“ can only read selected data – it is primarily the status for auditors.

The method of direct communication with external consultants directly in the database system Alexis is a highly appreciated means of facilitation of work. It is not necessary to export data to .doc or .pdf format (however, this possibility is still available), pass them personally, write comments in a document and incorporate them back into the database. Other user roles can be easily implemented, depending on needs of the lexicographers.

The second important part of the administration module serves for viewing and editing of a fixed item list. We will return to this issue in the subsection dedicated to the editing module.

zař.	heslo	hom.	sl. druh	varianty	zpracovatel	vytvořeno	změněno	kor.
1	atlas	1	m. než.		svetla	10.10.2013	31.03.2014	✕
2	atlas	2	m. než.		svetla	10.10.2013	23.04.2014	✕
3	atlas	3	m. než.		svetla	10.10.2013	20.11.2013	✕

Fig. 1. List of entries module

4.2 List of Entries Module

The list of entries is a web page that is displayed when a user successfully authenticates. The list represents the macro-structure of the dictionary and can be regarded as kind of an

index that displays all lemmas in the dictionary in the alphabetical order. As it is expected that the final version of the dictionary will contain approximately 150,000 lemmas, the list is divided into pages of fixed length (the number of items per page may be adjusted in the user interface) and we have implemented a possibility of navigation between pages.

Each item in the list displays a lemma, its variants, and other features requested by the lexicographers, such as the word class, the date of creation of the lemma or the status of the lemma as shown in Fig. 1. The lemma is retrieved from the database by using its unique id. Other information is joined to this id from other tables.

The lexicographers can limit the number of shown entries by using the *Quick search* and the *xFilter* tools (see Fig. 2.). The Quick Search tool operates within the entire database, it is possible to search almost any field of the micro-structure. Depending on the type of the searched category, the user is either presented a text input with support for AJAX based auto-completion and wild-card asterisk convention (e.g. when searching for a specific form of a variant) or with a drop-down list of possible values (e.g. when searching for lemmas with a given word class).

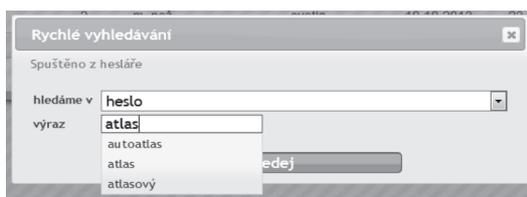


Fig. 2. Quick search tool with auto-completion

The *xFilter* is a web form that is used to design complex queries that can consist of multiple sub-queries joined by using logical conjunctions. The results of these two tools are combined, i.e. the *xFilter* will work only on the list of entries that match the condition entered in the *Quick Search* tools. The constructed queries can be saved for possible future reuse and even shared with other lexicographers. Moreover, several special quick filters are also prepared: it is possible to list all entries created by a logged user or all entries modified within a certain period of time.

As very important from the viewpoint of the lexicographer, there proven to be the use of quick filters:

1. *My entries* (it opens the list of all entries created by the logged user).
2. *Changed from – to / Created from – to* (it opens the list of all entries created or edited during a certain time period).
3. *Multi-variants entries* (it opens the list of all entries with more variants of the lemma).
This filter is currently frequently used for revisions of multi-variant entries. In the end, it can be replaced by another filter, depending on the character of revisions.

4.3 Editing Module

The editing module is launched when an editor clicks on an item in the list of entries module. It is used to edit lexical units; it represents the micro-structure of the dictionary. The module is implemented as a large HTML form that can be divided into the following four sections:

- header section
- variants section
- sense section
- section of cross references.

In the *header section*, some meta-information about a lemma can be entered, including the state of the lemma (new, completed ...), the date of its creation and the date of the last modification, the assigned lexicographer, or the lexicographer's notes. It is also possible to set the output of the lemma in the header. The lemma can be included in the printed and the electronic output, in the electronic output only, or excluded from the output. The differences between the types of the output will be discussed in the following subsection. The header also contains a list of opened corrections attached to the given lemma.

The *section of variants* serves for entering micro-structure elements related to the variants of the lemma, such as word class, morphology, origin, or pronunciation of the lemma. As variants of a lemma often have similar characteristics, we have implemented a possibility to copy the existing variant into a new one.

In a similar way, the *sense section* (see Fig. 3.) is used to edit microstructure elements describing meanings of the lemma and its exemplification. Concerning meanings, it is possible to add, remove or reorder them. However, the lexicographers prefer excluding a meaning from the output (but keeping it in the database) to deleting it. In order to save space, each meaning card is by default collapsed to display only the most important fields and can be dynamically unpacked to its full size using JavaScript (client-side) functions.

In the last section, references to the other lemmas can be defined: a link to idioms, diminutives, female equivalents of lemmas, a link to equivalents of perfective / imperfective verbs and to iterate verbs, and a link to verbal substantives.

As it was already mentioned, the editing module is a very complex form that consists of various widgets such as standard text input fields, multi-line edits, check boxes and radio buttons, and also custom widgets such as fixed item lists. These lists are used to edit elements that can be selected from predefined values, e.g. the word class. For these elements, the administrator prepared a list of predefined values; the lexicographers then use this list to choose one value (or more). In case a new value, not included in the list, is needed, the lexicographers can use a field called "*other value*". If the administrator finds out that some other value is frequently used, he or she may add it to the fixed list in the administration module.

To input a pre-formated text that is used for example in explanation of the meaning or exemplification, we use the *TinyMCE* WYSIWYG (i.e. What You See Is What You Get) editor implemented by means of the JavaScript language. We also use several standard widgets provided by the *jQuery UI* library such as calendars or popup dialogues.

In order to improve the performance of the editing module and to decrease load of the database server, it is planned to utilize asynchronous AJAX calls to load less frequently used parts of the editing form on demand. From the editing module, the user can return to the list of entries module or it is possible send the current lemma into the output module.

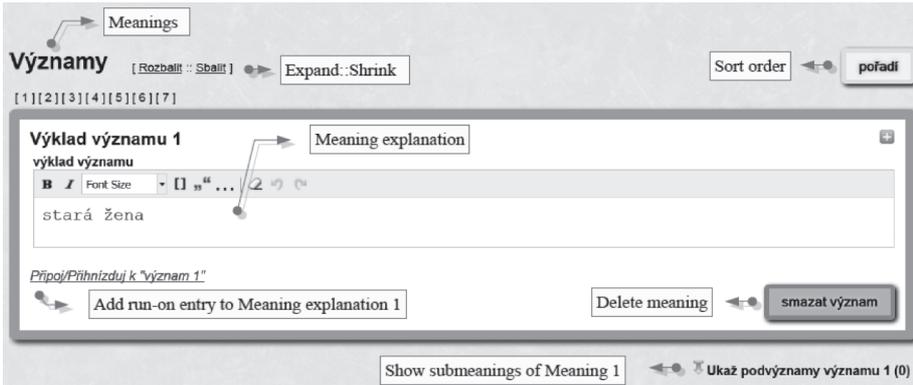


Fig. 3. Packed meanings section

5 New Features

Most of the new features have been implemented in the output and editorial modules. Besides, a brand new public web interface has been implemented (finished at the end of 2014).

5.1 Output Module

The output module can be launched either from the list of entries or from the editing form. If launched from the list of entries, all items in the current selection are printed. If launched from the editing form, only the current lemma is printed. The output in the output module is based on complex rules that describe the way in which all elements of the micro-structure of the lemma are displayed.

Two main types of the output are available: electronic and printed. The electronic output displays all information provided in the editing form. This output is implemented as an interactive web page. By clicking on any item, a new correction dialogue box is shown (see the following section for more details).

The printed output is implemented as a PDF file; we use *mPDF* PHP library to generate PDF documents. However, at present we are also investigating a possibility of replacing this library by the *wkhtmltopdf* tool. The tool uses the webkit rendering engine which provides more HTML and CSS standard compliant results. First performance tests revealed that the tool is almost twice as fast as the *mPDF* library.

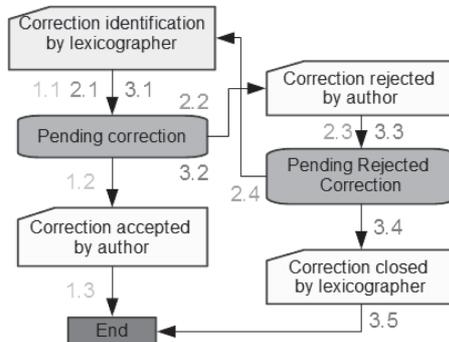


Fig. 4. Lifetime of the correction

The printed output displays items in the same format in which they will be published in the final printed dictionary. Some fields, such as the lexicographer's note, are not included in the printed output and the same information from multiple variants of the same lemma is reduced in the printed output. Additionally, editors can decide to exclude certain lexical units from the printed output. Furthermore, we use the output to MS Word format in the internal communication.

5.2 Editorial and Messages Modules

In the dictionary making process it is necessary to have a tool to work with corrections. We have developed an efficient editorial module that provides a notification system, history tracking, or targeted corrections.

A correction may be appended on any element of the micro-structure of the dictionary. The lifetime of a correction implements a state machine shown in Figure 4. The process is described in more detail in [4]. Recently, we have added a possibility to append a comment to any state of active correction. It is also possible to aim the correction at chosen editors. This allows editors to specialize in certain parts of the micro-structure (e.g. one editor will correct problems with morphology, whereas another focuses on exemplification). Additionally, targeted correcting usually shortens the time necessary for closing the correction.

We have also developed a messenger module that displays messages from the editorial tool targeted to the logged-in user. Each step of the correcting process is recorded in the database, it is possible to retrieve the entire history of the process. This system is also used to display complete history of all corrections of a given lemma.

The editorial and messages modules are used for various types of corrections and suggestions that can be targeted at any field of the database form, e.g. grammatical information, pronunciation, synonymy. Besides particular suggestions it is possible to add some other comments (explaining, expressing uncertainty, etc.). The correction can be accepted or rejected by the author of the dictionary entry, or executed by a corrector (co-worker, terminologist, editor). Here follow several examples of authentic comments that send authors of the dictionary articles among themselves. The examples were chosen from these three areas: inclusion in the lemma list, stylistic markers and exemplification.

1. Inclusion in the lemma list. Lemma **bohaprázdnost** ‚neuctivé chování před Bohem‘ – correction: *Nezařazovat, malá doloženost (SYN 2x, Newton 2x).*; correction accepted. Lemma **bankér II** ‚hráč držící bank ve hře‘ – correction: *Doplnit přechýlené slovo bankéřka II, znám z úzu.*; correction rejected: *V úzu asi existuje, doklady ale nenalezeny, nezaloženo.*
2. Stylistic markers. Lemma **dopsat** zast. ‚dodatečně napsat, vytvořit, sestavit (literární, vědecký ap.) text nebo jeho další součást, a přidat‘ – correction: *Nejde o zastaralý význam!*; correction accepted: *Souhlasím, byl to omyl.* Lemma **digitálky** kolokv. ‚digitální hodinky‘ – correction: *Místo kval. kolokv. dát kolokv. ▲*; correction rejected: *A proč?*; next correction: *Na základě analýzy materiálu v SYNu + srov. hodnocení ASCS, SSSJ (hovor).*; correction accepted.
3. Exemplification. Lemma **drncat** ‚drkotavě, třaslavě jet‘: vlak drncal; terénní vůz drncal po kamenité cestě; kočárek drncal na dlažebních kostkách. – correction: *Neopakovat ve všech třech příkladech stejný slovesný tvar.*; correction accepted.

Idiom **vypouštět pokusné balonky** ‚zkoumat, zjišťovat, zjistit postoje lidí k něčemu, reakci lidí na něco‘: zde se vypouštějí pokusné balonky studií budoucnosti a podle ohlasu publika následuje rychlá realizace. – correction: *Příklad není srozumitelný, najít lepší.*; correction pending.

5.3 Public Web Interface of the Dictionary

At the end of the year 2014, a closed test of a newly implemented public interface of the dictionary was launched. The interface is also implemented as a web application, additionally native applications for iOS and Android platforms are planned.

The public interface displays only a subset of lexical units, marked as completed by the lexicographers. The interface contains alphabetically sorted list of variants of lemmas. By clicking on a variant, a complete dictionary entry is displayed according to formatting rules similar to rules defined in the output module of the dictionary. Cross references between lemmas are implemented as hypertext links. Users can also export a selected dictionary entry into the PDF file.

The list of variants can be filtered by a quick filter tool that displays only variants containing a desired sub-string; a wild-card convention is enabled. Besides, a complex filter is also available; it allows users to combine search criteria over different elements of the micro-structure represented by different database tables using logical conjunctions AND (conjunction) and OR (disjunction). For example it is possible to retrieve lemmas with the adjective as their word class and at the same time are of Latin origin. In the background, the system prepares a corresponding query in SQL language that is then used to retrieve desired lexical units.

6 Summary and Outlook

The above described modules are already deployed on the production server and are used by the lexicographic team on daily basis. Approximately 10% of the supposed 150,000 lemmas have already been entered into the system.

The project is supported by a grant of the Ministry of the Culture of the Czech Republic within the National and Cultural Identity (NAKI) applied research and development program. This part of the dictionary project is finishing at the end of the year 2016. Complete source codes of the final version of Alexis will be published under an open-source licence.

At present a new revision control module is being prepared. This module will allow lexicographers to store the current status of a lexical unit into an archive and will also make possible to recover previous stored versions. In the near future we are going to develop an automatic lemma processing tool.

However, even the above described modules are subject to changes. For example, currently, we implement live preview to the editing module. This tool will allow to see changes to lemma right at the time when it is edited.

References

- [1] Barbierik, K. et al. (2013). A New Path to a Modern Monolingual Dictionary of Contemporary Czech: the Structure of Data in the New Dictionary Writing System. In Gajdošová, K. and Žáková, A., editors, *Proceedings of the 7th International Conference Slovko*, pages 9–26, Jazykovedný ústav Ľudovíta Štúra, Slovenská akadémia vied.
- [2] Barbierik, K. et al. (2013). Development of dictionary writing software. In *Proceedings of the 39th Conference Software Development*, pages 133–147, College of Polytechnics, Jihlava.
- [3] Barbierik, K. et al. (2014). The Dictionary-Making Process. In *Proceedings of the 16th International Conference EURALEX*, pages 125–135, Institute for Specialized Communication and Multilingualism at the European Academy of Bolzano/Bozen (EURAC).
- [4] Jary, V. et al. (2014). Dictionary writing system: Output and Editorial modules. In *Proceeding of the IT for Practice 2014*, pages 17–23, VŠB Ostrava.
- [5] Kochová, P. et al. (2014). At the Beginning of a Compilation of a New Monolingual Dictionary of Czech (A Report on a New Lexicographic Project). In *Proceedings of the 16th International Conference EURALEX*, pages 1145–1151, Institute for Specialized Communication and Multilingualism at the European Academy of Bolzano/Bozen (EURAC).
- [6] Driessen, V. (2010). A successful Git branching model. Accessible at: <http://nvie.com/posts/a-successful-git-branching-model/>, retrieved 15 January 2015.

Corpus of Dialects of the Slovak National Corpus

Katarína Gajdošová, Radovan Garabík, and Mária Šimková

L. Štúr Institute of Linguistics, Slovak Academy of Sciences, Bratislava

Abstract. The authors present the project of the Corpus of Dialects of the Slovak National Corpus and the concept of its creation. The paper gives an overview of current text sources included in the corpus, the format of the metadata records, description of information about the speakers and text transcriptions, the particulars of converting the transcriptions into a unified format, tagging and querying.

1 Introduction

In Slovakia, there is an apparent dichotomy in spoken language use – dialects are used by the autochthonous population of respective dialect areas in everyday social and often working relations, but communication in dialect is often marked with a social stigma. Especially in urban areas and white collar job positions people tend to use standard Slovak¹, an interregional language register of higher prestige [6].

Slovak dialects are passed down from one generation to the next only in their verbal form, they are almost never used in written communication (except in some fictional settings). There is a strong process of levelling going on, especially concerning the replacement of autochthonous vocabulary by standard language, but also of the younger generation speaking the standard language as their L1. This process does not occur only in Slovakia with respect to its dialects but also in other countries where dialects used to be spoken in various communication situations until recently. Nowadays, they keep disappearing from everyday communication especially in towns and cities due to socio-economic changes. At the same time, there are more than a handful of enthusiasts who monitor their native regions, including the usage of dialects and then they use their vernacular also on the internet – discussing current news, sharing jokes, writing blogs, joining social networks etc.

There is a linguistic continuum between Czech (Moravian) and Slovak dialects; in the north, Slovak morphs into transitional Polish dialects (*góral*), in the east, there is a continuum to Rusyn. Traditionally, the linguistic border between Slovak and Czech is drawn at the Moravian-Slovak borderline, Rusyn is invariably considered separate from eastern Slovak; *góral* dialects are sometimes taken for Slovak dialects, but the predominant position of current dialectology is to treat them as transitional dialects and not as a part of Slovak dialect area [1].

This brought forward the issue of naming of the corpus – in the draft of the project, the original name used to be the Corpus of Slovak Dialects. The project, however, envisages to include also the dialects from the border regions of Slovakia as well as from Slovak diasporas abroad, where large compact groups settled the past. These dialects have their own characteristic features which used to be identical with the dialect in their

¹ It is a variety of language relatively close to the prescribed form of the *spoken literary* Slovak, but differs in some phonological and lexical feature. If spoken as officially prescribed, literary Slovak is perceived as distinctly marked, even comical, and is usually not used anywhere in normal communication [3].

native region but they gradually changed due to coexistence with languages and dialects in their new setting (e. g. in Hungary, Romania, Serbia). By including examples of abovementioned group of dialects into the corpus entitled the Corpus of Slovak Dialects its content would not be consistent with the delimitation of Slovak dialects in the Slovak dialectological tradition that is why it was necessary to find a different denomination. (On further development of the name of this corpus, cf. following text.)

Slovak dialects are divided into three basic groups:

- The western Slovak dialects are spread throughout the Trenčín, Nitra, Trnava, Myjava areas and other regions.
- The central Slovak dialects are spoken in the regions of Liptov, Orava, Turiec, Tekov, Hont, Novohrad, Gemer and in the Zvolen area.
- The eastern Slovak dialects can be found in the regions of Spiš, Šariš, Zemplín and Abov.²

These groups are further divided into a variety of subdialects, especially in mountainous regions. Slovak dialects are the basic source of information on historical Slovak grammar as well as the source of information about the life in past as such. However, several sources exist only in a paper form or only in form of audio recordings and, therefore, they are practically unavailable for the general public (e. g. the sources in the Archive of the Department of Dialectology of L. Štúr Institute of Linguistics, Slovak Academy of Sciences). Other sources are scattered in different books or journals, many of them having been irretrievably lost already. In order to preserve this part of the Slovak cultural heritage and to make it available for general public as well as research community, several staff members and some finances were allocated in the framework of the project Building of the Slovak National Corpus and the Digitalization of **Linguistic Research in Slovakia** – 3rd phase, which has been co-financed by the Ministry of Culture of the Slovak Republic, Ministry of Education, Science, Research and Sport of the Slovak Republic and the Slovak Academy of Sciences.

2 Concept of the Corpus of Dialects of the Slovak National Corpus

The project of the Corpus of Dialects of the Slovak National Corpus (hereinafter referred to as CD SNC) started to be drafted at the SNC of L. Štúr Institute of Linguistics, Slovak Academy of Sciences in 2013. The aim of this subproject of the SNC is to draw on the experience gathered by building previous types of corpora and tools developed in the SNC in order to build a corpus of dialect utterances and their transcriptions that could be made available on-line for dialect phenomena query. In identifying available resources and testing their computational processing into a corpus with standard query tools, a number of differences are present with respect to previously built corpora:

- text sources for dialect corpus – published transcriptions require digitalization;
- each dialect has its own specific features (speech sounds/letters, assimilation, palatalization, elision, etc.);
- (nearly) each transcriber uses his/her own method of transcription (especially the character repertoires for pronunciation transcription differ);
- the transcription method should be unified and the conversion chart for each type of transcription should be made;

² <http://korpus.sk/dialect.html>

- current audio recordings usually lack transcription;
- student made transcriptions differ in quality, their verification is desirable, sometimes re-transcription is necessary;
- older records contain incomplete metadata, the names of localities and districts changed over the decades;
- given the great diversity of words and word forms, automatic lemmatisation and morphological annotation are not feasible, etc.

In 2014, the concept of the CD SNC and the related project called the Archive of Dialects of the Slovak National Corpus have been discussed on a number of meetings within the SNC, consulted with the members of the Department of Dialectology, as well as presented and discussed on several professional meetings. Respecting the Slovak dialectological tradition which does not consider transitional dialects used in border areas of Slovakia and foreign diasporas to be Slovak, the original name the Corpus of Slovak Dialects was altered to the Corpus of Dialects. As the Department of Dialectology also gathers dialect material, the corpus built at the SNC bears the name of the SNC Department in order to distinguish them. The name Archive of Dialects of the SNC was created by analogy.

Given the need to process electronically as much of the specific dialect material as possible, the availability or rather unavailability of dialect audio recordings and transcriptions and the lack of available staff members who would make transcriptions and corrections, several decisions had to be taken:

- a) to find and gather published dialect transcription, to process them electronically for the inclusion in the written version of the CD SNC – the corpus will comprise only texts without audio recordings, since there were either no transcriptions (there was only a handwritten transcription), or the transcriptions have not been preserved;
- b) to build the Corpus of Spoken Dialects of the SNC separately and make it gradually publicly available; it will be predominantly composed of present-day dialect recordings and their transcriptions – to complete this version with an archive, find and include recordings made by institutions working in Slovak studies [2].

Dialect recordings and transcriptions will be processed and presented in three ways:

1. text corpus consisting of published dialect transcriptions that will be gradually supplemented with other transcriptions featuring existing audio recordings and that will also become a part of the spoken corpus of dialects;
2. spoken corpus of dialects containing audio recordings linked to their transcriptions;
3. archive of dialects containing only audio recordings without transcriptions; in case there are staff members available to make the transcriptions, these recordings can be further processed and included in the corpora.

The corpora will be publicly available in the form of query interface only, similarly to other SNC resources; the archive will be available for research only at the SNC department. In the next stage, we plan to process seminar papers and diploma theses, including selected transcriptions of the audio recordings that are part of the Archive of Dialects of the SNC.³ All the issues related to the gathering of recordings, transcription, processing

³ The Archive of Dialects of the SNC preserves dialect audio recordings on different types of media. Audio recordings are digitised at the SNC, processed into a unified format and added with e. g. metadata on the origin of the recording, its quality, dialect area of speakers. The Archive of Dialects of the SNC represents a valuable central repository of dialect recordings that have been virtually unavailable, until recently being kept at various university departments.

and making them accessible for scientific purposes were discussed and consulted with the lawyers from the Office for Personal Data Protection of the Slovak Republic. Their recommendations helped to formulate relevant provisions in the license agreement or to set up precise procedures for processing the dialect material, especially personal names.

3 The First Version of the Corpus of Dialects of the SNC

The aim of the first, building phase of the CD SNC is to gather existing text transcription of dialect audio recordings or handwritten transcriptions, especially those already published, to process them in a unified way using a corpus methodology and tools and finally to make them available to the public thus enabling the research of dialect phenomena. The pilot version of the CD SNC was finished in March 2014, but it was accessible only as an internal resource. Its release in the form of a publicly available NoSketchEngine interface (cf. part 3.5) could be made only after solving the license issues concerning dialect recordings and their transcriptions as well as the extent of application of the Act on Personal data Protection (especially the issue of publication / coding of personal names of respondents). The first version *dialekt-1.0*, containing almost 73 855 tokens, underwent minor changes and was made public in September 2014. Its current – second version *dialekt-2.0* – was made public in August 2015 featuring 328 907 tokens⁴.

3.1 Source Texts of the CD SNC

The first phase of the CD SNC comprises the corpus treatment of dialect audio recordings or transcribed recordings published in monographs, journals, diploma thesis etc. The version *dialekt-2.0* comprises dialect texts originating from 11 sources⁵.

Table 1 contains bibliographical data of source texts, structure <source>, referring to the respective source, and the number of tokens in each source. Sources are sorted by size (data in the last column) in descending order.

Text resource	Structure of the CD SNC	Number of tokens
Habovštiak, Anton: Oravci o svojej minulosti. Reč a slovesnosť oravského ľudu. Martin: Osveta 1983, s. 23 – 358.	<doc source="osm">	159 892
Múcsková, Gabriela – Muziková, Katarína – Wambach, Viera: Praktická dialektológia. Vysokoškolská príručka na nárečovú interpretáciu. Wien: Facultas Verlags- & Buchhandels AG Wien, 2012. 138 s.	<doc source="prir">	33 513

The fact that they were recorded on older media resulted in progressive degradation of their quality. Unfortunately, many of them could not be saved any more – they got lost due to moving or retirement etc. (see [2])

⁴ <http://korpus.sk/dialect.html>

⁵ Four dialect texts provided by the Department of Dialectology, L. Štúr Institute of Linguistics, Slovak Academy of Sciences are annotated individually according to their respective location.

Text resource	Structure of the CD SNC	Number of tokens
Habovštiak, Anton: Oravské nárečia. Bratislava: Slovenská akadémia vied 1965, s. 355 – 396.	<doc source="oravn">	30 244
Buffa, Ferdinand: Šarišské nárečia. Bratislava: VEDA, Vydavateľstvo Slovenskej akadémie vied 1995, s. 318 – 373.	<doc source="sarnar">	25 306
Jóna, Eugen: Novohradské nárečia. Ed. P. Žigo. Bratislava: Veda 2009. 164 s.	<doc source="nov">	22 980
Kováčová, Viera: Sotácke nárečia na západoslovansko-východoslovanskom jazykovom pomedzí. Bratislava: Slovenská akadémia vied v Bratislave, Slavistický ústav Jána Stanislava 2005, s. 123 – 144.	<doc source="vkov">	11 976
Ripka, Ivor: Dolnotrenčianske nárečia. Bratislava: Veda 1975, s. 216 – 246.	<doc source="dolntrn">	11 671
ANT DO JÚĽŠ 19/12 – Dolný Hričov. Archív nárečových textov Dialektologického oddelenia Jazykovedného ústavu Ľ. Štúra SAV.	<doc source="ant">	10 195
Múcsková, Gabriela: Nárečie a spisovný jazyk v bežnej hovorenej komunikácii obyvateľov Gelnice. Dizertačná práca. Bratislava: Jazykovedný ústav Ľudovíta Štúra SAV 2006, s. 92 – 100.	<doc source="gmucs">	7 403
Buffa, Ferdinand: Nárečie Dlhej Lúky v Bardejovskom okrese. Bratislava: Vydavateľstvo Slovenskej akadémie vied 1953, s. 116 – 128.	<doc source="nadhhl">	5 612
Pauliny, Eugen: Nárečie zátopových osád na hornej Orave. Spisy Jazykovedného odboru Matice slovenskej. Série B. Zväzok 3. Turčiansky Sv. Martin: Matica slovenská 1947, s. 99 – 115.	<doc source="zatopos">	5 006
ANT DO JÚĽŠ 50/28 – Ábelová. Archív nárečových textov Dialektologického oddelenia Jazykovedného ústavu Ľ. Štúra SAV.	<doc source="ant">	2 535
ANT DO JÚĽŠ 52/42 – Klenovec. Archív nárečových textov Dialektologického oddelenia Jazykovedného ústavu Ľ. Štúra SAV.	<doc source="ant">	1 380
ANT DO JÚĽŠ 72/13 – Čemerné. Archív nárečových textov Dialektologického oddelenia Jazykovedného ústavu Ľ. Štúra SAV.	<doc source="ant">	1 198

Table 1. Source texts in the CD SNC – version *dialekt-2.0*

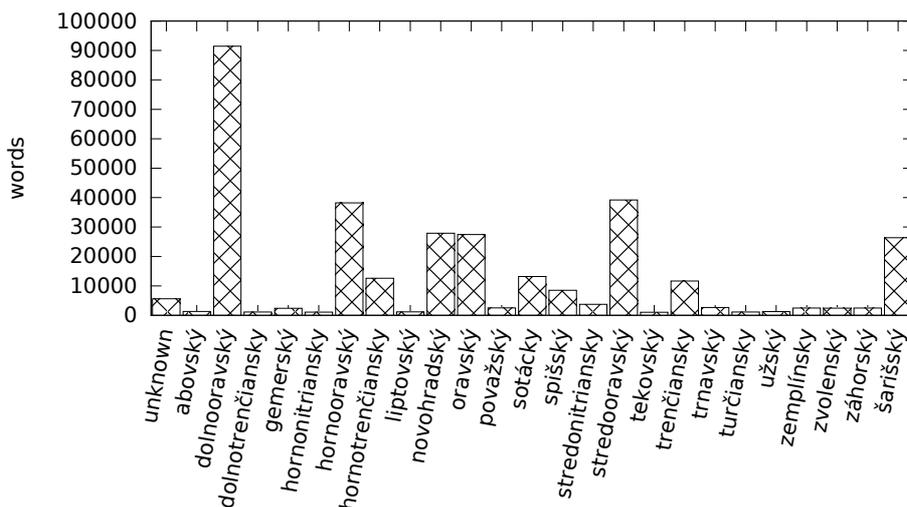


Fig. 1. The number of words from each dialect included in the corpus. The histogram shows the disparity of recorded information – the western Slovak dialects received very little attention.

3.2 Metadata on the Text

The version *dialekt-2.0* includes more detailed metadata about the text sources. Some of the data are easily derived from the name of the publication (e. g. the Orava dialect group), part of the data has been meticulously recorded by the authors, however, there are many of them that have to be searched for or completed according to the current circumstances (names and territorial division of localities have undergone several changes from the mid-20th century).

Basic data items include the information on the source (source), district (district), location (location/location2013), dialect group (dialect), dialect subgroup (subdialect), name and surname of the field researcher (explorator), date of the recording (exploredate), place of recording (exploreplace), type of the text (type), bibliographical data on the origin of the text transcription (bibl) and commentary (comment).

Moreover, the metadata also include specific information concerning the recordings originating from the Archives of the dialect texts of the Department of Dialectology: code of the recording (code), name of the recording (name), number of the recording (textnumber), text page (textpage), name of the transcriber (transcriber1), date of the transcription of the audio recording (datetrans1), name of the person who transcribed the text into digital format (transcriber2), date of the transcription into digital format (datetrans2), name of the proofreader (correction), date of proofreading (datecorr).

Not all the sources have all the records filled in, depending on whether those data could be found in the source text. Classification of a text as belonging to a specific dialect group and subgroup is based on the data recorded in each text source or on the expert advice from the Department of Dialectology. Location (obec/mesto), representing the origin of the text and the district to which the location belongs is fully compatible with

the list of locations⁶ used by the Department of Dialectology. Due to the fact that the location and district classification date back to different years of the previous century, we always record also the name of the respective location according to the territorial division of Slovakia by 31 December 2013, which can be found in the item called `location2013`⁷.

Text transcriptions come from various sources and feature variable quality and detail, therefore it is essential to preserve also the information about the type of the source text. Currently we distinguish following items: monograph (`mon`), monograph on national history and geography (`vlmon`), handbook (`hnd`), PhD thesis (`dis`), master thesis, bachelor thesis (`dpl`), study (`std`), seminar paper (`ref`), unpublished texts, manuscripts (`npu`).

3.3 Metadata on Speakers

Text sources contain the information on speakers depending on the accessibility of data and the custom of the transcriber. Our aim is to treat the information in a unified way when creating the corpus. Therefore we include into the metadata the name and surname of the speaker (`name`), initials of the name and surname used in the transcription (`acronym`), gender (`sex`), date of birth (`birth`) or age (`age`), birthplace (`birthplace`), usually identical with the place of the recording, and the information if the speaker is the field researcher or the respondent (`field researcher`: values `y/n`).

Although primarily presented as a synchronic corpus, it has also some diachronic features – the sources of documents were published in previous decades, they are themselves based on recordings and transcriptions made from the 1930's onwards and the interviews were often conducted with elderly people (see Fig. 2). Thus the corpus offers a unique insight into the past of rapidly disappearing landscape of Slovak dialects. Notably, the region of Bratislava is missing from the corpus – the language situation has been dramatically changing since the end of World War II which represented the end of pre-war trilingualism and the specific dialect of the capital city was not deemed worth investigating in serious linguistic circles, though this attitude is already changing [3], [5].

3.4 Structural Annotation

With respect to the specificities of a dialect text and different ways of the transcription the CD SNC is neither lemmatised nor morphologically annotated.

Besides basic structures `<doc>`, `<spk>`, `<s>` a `<p>` known also from other types of corpora it features a novelty – the structure `<rem>` – remark that includes several specific values.

⁶ http://korpus.sk/attachments/dialect_file/DIALEKT-tab-district-location.txt

⁷ http://korpus.sk/dialect_file.html

Structural tag	Acronym	Value	Explanation	Example
remark	<rem>		information on dialectological or non-dialectological form of a respective token/tokens that <rem> refers to	<rem dial="n" var="" val="">A čo ste jej vtedy ovedali?</rem> – utterance of the field researcher in standard Slovak
		dial="y/n"	information on dialect text with the value var	<rem dial="y" var="fčil" val="">fčiléky</rem>
			information on dialect text with the value val	<rem dial="y" var="" val="lajblíg bez rukávoϋ">bekeše</rem>
		var=""	variant of the respective token from the text	<rem dial="y" var="fčil" val="">
		val=""	explanation of the respective token from the text	<rem dial="y" var="" val="lajblíg bez rukávoϋ">bekeše</rem>

Table 2. Structural tags used in the CD SNC

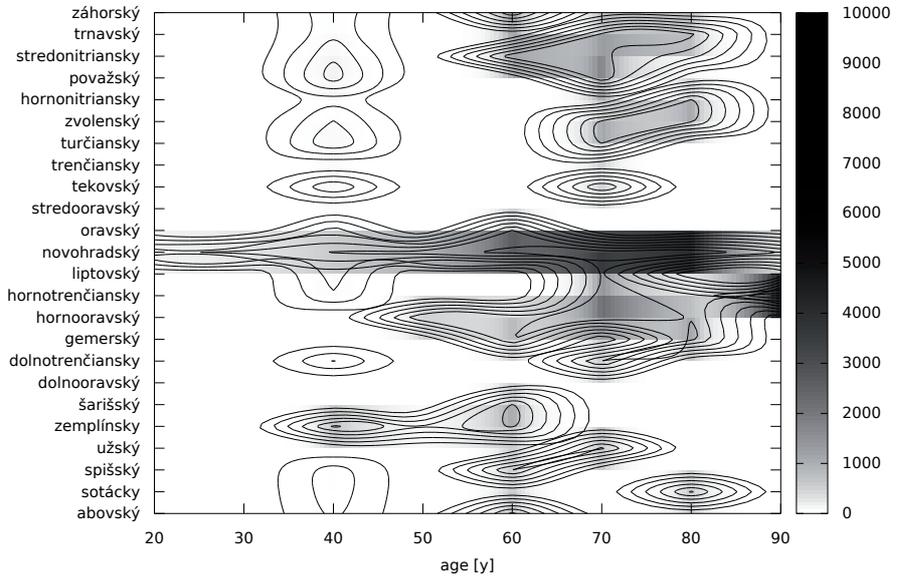


Fig. 2. Distribution of number of words (contours and shade of grey) in dialects (vertical axis, sorted approximately by geographical position from western (top) to eastern (bottom) Slovakia) by age (horizontal axis).

3.5 Transcription

Transcription systems in Slovak dialectology are generally based on Slovak orthography with several added characters and diacritical marks; the International Phonetic Alphabet is virtually unknown. Unfortunately, the sources of texts included in the version *dialekt-2.0* differ in some important details in their transcriptions. The differences lie not only in the characters used to transcribe the phonemes, but more importantly in the amount of finer details recorded and in the feature depth of phonemic versus phonetic analysis. Since one of the goals of the corpus was to keep as much information about the dialects as possible, we have chosen a transcription system that is a superset of all the transcriptions used in the source texts. The transcriptions are automatically converted into this common format (which means mostly just a simple character or string substitution), but the information is not changed. This way, the transcription in the corpus remains unified and readable within the same system, but the texts from separate sources contain different information. E.g. the word “nej” ([ɲɛj̥] inIPA) could be transcribed as both *ňej* or *ňej̥*, depending on the depth of phonemic analysis. To facilitate query in the corpus, a specialized virtual keyboard (named SNC-DIALECT) with the special characters used in the transcriptions is available in the NoSketch Engine interface, since the version *dialekt-2.0*.

'v' : 'v',	'úN' : 'úñ',
'v' : 'v',	'áN' : 'áñ',
'l' : 'l',	'z' : 'z',
'p' : 'p',	'x' : 'x',
'k' : 'k',	'u' : 'u',
'm' : 'm',	'ñ' : 'ñ',
'oN' : 'oñ',	'ã' : 'ã',
'aN' : 'añ',	'ä' : 'ä',
'eN' : 'eñ',	'û' : 'û',

Fig 3. Sample of a number of conversion charts used to unify transcriptions of dialect texts

3.6 Query

The CD SNC is available via the NoSketch Engine corpus manager [4] to all registered users of the SNC⁸. The querying is possible by using the attribute *word* and regular expressions. As for the regular expressions the user can employ the operators *within* and *containing* in order to query in different texts according to the annotation available.

4 Further Perspectives for the CD SNC in Terms of Size and Quality

Slovak dialects have been recorded as the research object of the Slovak dialectology roughly from the 1930s onwards firstly by means of a handwritten transcription, later on reel-to-reel magnetic tapes and compact cassettes and nowadays also using modern digital media. In the framework of the project Building the Slovak National Corpus and Digitalization of the Linguistic Research enabled the gathering, digitalization and processing of available written and spoken dialect sources so that they can be made

⁸ <http://korpus.sk/usage.html>

publicly available for the research of dialect phenomena. The primary aim of the corpus in its initial phase is to collect existing (often published) texts in transcribed Slovak dialects, systematically annotate and analyse the texts and index them in a text corpus. Existing Corpus of Dialects of the SNC comprises in its second version almost 330 000 tokens from 11 sources. The processing of other already published texts is ongoing (they are scanned, proofread or transcribed) and in 2016 they will be publicly available in the third version of the Corpus. Thanks to the cooperation with institutions and departments working in the field of Slovak studies providing their archive material sources, also the Archive of the Dialects of the SNC keeps growing. It offers the possibility to analyse gathered dialect recordings within the SNC. Future tasks that will require a substantial financial means and especially human resources include: transcriptions of audio recordings, correction and technical processing of the transcriptions, segmentation and linking the sound and the transcription. However, already the current version of the Corpus of Dialects of the SNC as well as the Archives of the Dialects of the SNC represent significant sources for those who are interested in Slovak dialectological research.

References

- [1] Kapičáková – Szczerbová, M. (2010). Oravské goralské nárečia vo svetle kolonizácií a teórie jazykových kontaktov. In Múcsková, G., editor, *Varia. XX. Zborník plných príspevkov z XX. kolokvia mladých jazykovedcov*, pages 273–282, Slovenská jazykovedná spoločnosť pri Jazykovednom ústave Ľudovíta Štúra SAV, Bratislava.
- [2] Karčová, A. (2015). Archív nárečí Slovenského národného korpusu. Východiská jeho tvorby, súčasný stav a perspektívy. In *Variety jazyka a jazykovedy. Východoslovenské nárečia v minulosti a dnes*. Filozofická fakulta Prešovskej univerzity v Prešove, 12 pp. (in press).
- [3] Ondrejovič, S. (2007). Výskum hovorenej podoby spisovnej slovenčiny po 40 rokoch. *Sociolinguistica Slovaca*, 6:9–17.
- [4] 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.
- [5] Satinská, L. (2015). „Keď sme mali Taschengeld, tak sme si kúpili jeden krémes.“ Prešporáčtina ako špecifický mestský sociolekt. In Barnová, K. and Chomová, A., editors, *Varia. XXI. Zborník plných príspevkov z XXI. kolokvia mladých jazykovedcov*, pages 437–450, Katedra slovenského jazyka a komunikácie Filozofickej fakulty Univerzity Mateja Bela v Banskej Bystrici – Slovenská jazykovedná spoločnosť pri Jazykovednom ústave Ľudovíta Štúra, Banská Bystrica.
- [6] Šimková, M., Garabík, R., Gajdošová, K., Laclavík, M., Ondrejovič, S., Juhár, J., Genčí, J., Furdík, K., Ivoríková, H., and Ivanecký, J. (2012). *The Slovak Language in the Digital Age*. Springer-Verlag, Berlin Heidelberg.

A New Dialect Corpus: DIALEKT*

Hana Goláňová

Faculty of Arts, Charles University in Prague, Czech Republic

Abstract. A new dialect corpus is currently under construction in the Spoken corpora section of the Institute of the Czech National Corpus, the first of its kind in this context: DIALEKT. The aim of the present paper is to introduce the DIALEKT corpus in general, the methodology of dialect data collection, the sociolinguistic parameters of the corpus and its transcription scheme. A subsequent part of the article deals with the preparation of base maps for the corpus and their incorporation into an interactive web environment designed for analyzing data from all types of spoken language corpora. This interactive web application will make it possible to jointly access linguistic information from both dialect corpora and traditional spoken corpora, a useful feature for the research community and laypeople alike.

1 Introduction

In the past, the Institute of the Czech National Corpus focused solely on building corpora of spoken language whose purpose was to collect everyday language spoken by people of different sociological backgrounds in various communication situations. The ORAL series corpora (ORAL2006 [1], ORAL2008 [2] and ORAL2013 [3]) is especially notable as a sustained effort of continuous data collection.¹ The ORAL2013 corpus already captures language data from speakers from all over the Czech Republic, but dialect data are not its primary target. This is why a new series of dialect corpora has been devised to complement the traditional spoken language corpora. DIALEKT will be the first corpus in this series; it will concentrate exclusively on collecting dialectal language material, serving not only professional dialectologists, but also the broader linguistics community and laypeople. One of the key differences from other spoken language corpora is the methodology of dialect data acquisition.

The process of compiling a dialect corpus gradually led to the idea of putting together unified base maps for all spoken corpora and integrating them into an interactive web application. This application should enable users to jointly query data from all spoken and dialect corpora and visualize them on a map. The article offers a more in-depth discussion with examples later on, but its main goal remains an introduction to the DIALEKT corpus and its opportunities of use.

2 The DIALEKT Corpus

The DIALEKT corpus will cover the traditional regional dialects across the territory of the entire Czech Republic; it encompasses two layers of dialect material. The older layer consists of recordings made from the close of the 1950s up to the 1980s; the newer one

* This paper resulted from the implementation of the Czech National Corpus project (LM2011023) funded by the Ministry of Education, Youth and Sports of the Czech Republic within the framework of Large Research, Development and Innovation Infrastructures.

¹ Future plans include creating a publicly accessible corpus merging the entire ORAL series into a single corpus. The successor of the ORAL series is the ORTOFON spoken language corpus, one of whose aims is full territorial coverage of the Czech Republic.

contains recordings starting from the 1990s up to the present day. Part of the older layer is carried over from the language material collected by the dialectological department of the Institute of the Czech Language of the AS CR, v. v. i., which was published as part of the *Addenda to the Czech Linguistic Atlas* [4]. Apart from these pre-existing older sources provided by various individuals and institutions, new dialect material is also being collected at the Institute of the Czech National Corpus. Both these layers contain linguistic material which captures archaic dialectal features of spoken language on all levels of linguistic description, spanning all the dialect areas of the Czech-speaking language territory.

One of the goals of the new material layer is to include recordings of speakers coming from geographical locations close to the ones used for the older material layer, so as to enable users to follow language as it evolves in a particular dialect region or otherwise defined area. As far as data collection methodology is concerned, the usual rules developed within Czech dialectology are observed. In the current phase of data collection, we concentrate primarily on capturing the traditional territorial dialects in their most conservative state. Dialect fieldwork for both corpus layers is therefore characterized by focusing exclusively on representatives of the oldest generation (quite homogeneous so far), in order to capture above all the original dialect features. Speakers are mostly local natives from a rural background; they belong to the settled, non-mobile stratum of the population, tied to an agricultural way of life or to a particular trade or craft. Speakers over 60 years were considered for inclusion, i.e. people born at the close of the 19th century and in the first half of the 20th century. The recordings are predominantly monological and often convey interesting topics, painting a picture of the old rural world and life in it (farming, trades and crafts, customs, legends etc.).

2.1 Sociolinguistic Aspects of the Corpus

The DIALEKT corpus recordings will also be richly sociolinguistically annotated;² this metalinguistic information will be useful e.g. for defining subcorpora. Let us mention at least some of these sociolinguistic metadata to get a rough idea. Apart from being classified according to the traditional system of dialect-based territorial division of the Czech Republic (dialect region / *nářeční oblast*, dialect subgroup / *nářeční podskupina*, dialect area / *nářeční úsek*, dialect type / *nářeční typ*),³ the recordings can also be sorted according to region (Bohemia, Moravia and Silesia) or country (part of the material originates from Polish locales). Another option is to filter dialect data based on whether they are urban or rural, possibly selecting the target size of the dwelling. A fundamental piece of information for each recording is its year of origin and its source. The type of the communication situation (monological or dialogical) and the main conversation topic are also noted.

As for speakers, it is compulsory to specify their sex, age, place of childhood residence (until 15 years of age) and place of longest residence (in our case, these two entries are generally the same, since the selection process favours local natives), and the latter's parent unit in the system of dialect-based territorial division. Apart from that, information about the speaker's education and longest occupation is also recorded if possible.

² The DIALEKT corpus keeps track of more sociolinguistic data about recordings and speakers than the Dialect corpus of the Slovak National Corpus, primarily because it strives for data compatibility with the rest of the spoken corpora of the CNC.

³ The system of dialect-based territorial division employed for the DIALEKT corpus is based (with a few emendations) on the *Czech Linguistic Atlas* [5] and the *Encyclopedic Dictionary of Czech* [6].

Místo nahrávání | **Mluví** | **Areál** | **Nářeční oblast** | **Nářeční úsek** | **Ukázka** ✕

korpus DIALEKT

Nahrávka: 001-C-SVC-1971-SA

Rozbor ukázky:

Hláskoslovi:

- úženi *é > í*
- změna *ý > v* v dvojhlasce *ej*
- protetické *v (w)*
- skupiny *bn > mn, dn > nn > n*
- výslovnost hlásky *v* jako *w*, nebo jako *u*
- krácení samohlásky *i* v slovtvorných příponách
- prodloužení kmenových samohlásek v přízvučné slabice (*barabizna*)
- pozůstatky výslovnosti *í*
- *ž*anik *j* na počátku slov (*itřnice*)
- zjednodušení skupiny hlásek *eja > á* (*řejakí > řiák*)

Tvaroslovi:

- infinitiv končící na *ř*

Lexikon:

- *wotporazít'* „porazit z více kusů dobytka jeden“

Jiné:

na ortografické úrovni se používá pauzová interpunkce (pauzy .. / předěly .)

Poznámka:

v přepisu celé sondy se nacházejí další nářeční prvky typické pro podještědský a podkrkonošský nářeční typ, z archaických rysů např. prodloužení u slabikotvorných *l* a *r* (*plno, itřnice*)

DIAL (dialektologický přepis)

v Jablonci, gde je teř modlitebna českosloveckí cirkwe, bejvala stará barabizna a tam řežničít nákej Honcira. Franta Honcira.

no, wono tenkrát to nebilo žáni umřeňi náki to dobitče wotporazít' a nadělat z něj itřnic, náki ti salámi anebo jak se to všechno menuje.

ORT (ortografický přepis)

v Jablonci .. kde .. je teř .. modlitebna .. československý cirkwo .. bojvala stará barabizna .. a tam řežničil .. nějakéj Honcira .. Franta Honcira ..

no wono tenkrát .. to nebylo žádný uměni .. nějaký to dobytče wotporazít' .. a nadělat z něj itřnic .. nějaký ty salámy anebo .. jak se to všechno menuje ..

Fig. 1. Thematic tab with a recording excerpt, dialectological (DIAL) and orthographical (ORT) transcripts and analysis (see also Sec. 3.1)

2.2 Transcription of the Linguistic Material

The ELAN transcription program⁴ [7] is used for segmenting and transcribing the linguistic data for the DIALEKT corpus. ELAN makes it easy to create a two-tier transcript of the recordings.

The dialect corpus will, therefore, have two tiers, featuring a dialectological and an orthographical transcript⁵ (see Fig. 1). The basic transcript is the dialectological one, which is assembled according to the rules for transcribing dialectological texts.⁶ This approach to transcription follows the usual conventions in the field of Czech dialectology, i.e. both consonants and vowels are transcribed using symbols employed in dialectological transcripts in order to capture the actual pronunciation (*wono*, *řežničít*, *být*, *kark* etc.). All

⁴ Accessible at: <http://tla.mpi.nl/tools/tla-tools/elan/>.

⁵ The ORTOFON spoken corpus, which is currently under construction, also has a two-tier transcript, the basic orthographical tier being complemented by a phonetic one.

⁶ We mostly follow the *Rules for the Scientific Transcription of Dialectological Records of Czech and Slovak* [8], but also take some inspiration from *Czech Dialect Texts* [9] and the *Addenda to the Czech Linguistic Atlas* [4].

types of voicing assimilations are systematically recorded, as are alterations and elisions in the pronunciation of consonant clusters and vowel length modulations. In contrast, word boundaries are kept as they appear in standard orthography and we use unrestricted syntactic punctuation. Capital letters appear only at the beginning of proper names. Suprasegmental phenomena are left out.

The second tier of the transcription of dialect data consists of an orthographic transcript based on the general rules in use for the spoken corpora of the CNC. Words are transcribed in forms identical or close to their usual spellings (the guidelines are based on standard orthography) while retaining features of the spoken language and regional specificities. Vocalic length is unified according to standard spelling (even in the case of the systematic shortening typical of Silesia); quality changes are recorded as pronounced (*čítit – cejtít – cétit*). Morphological phenomena, e.g. endings of all types of declension (*sinoj, perletěj*) and conjugation (*nosijó, mosíja*), and needless to say, all dialectal or regional lexical peculiarities (*rulík, trávnica, ostat*), are also honoured. With respect to consonants, we keep *v*- and *h*-prothesis but leave out voicing assimilations and variations in the pronunciation of consonant clusters – these cases are covered by the dialectological tier. Capital letters and word boundaries are left as per the rules of standard orthography. In contrast with the dialectological tier, pausal punctuation is used (a distinction is made between actual pauses and prosodic boundaries), which often helps detect in retrospect e.g. voicing assimilation mismatches on the dialectological tier. Thanks to the unifying orthographic tier, the body of linguistic data which is thus taking shape will be searchable in much the same way as the remaining spoken corpora.

In addition to the two previously mentioned transcription tiers, the recordings contain also metalinguistic annotation tiers which keep track of information about the non-verbal sounds made by speakers (*laughter, loud yawning* etc.) or ambient sounds pertaining to the entire communication situation (*TV or radio noise, dog barking* etc.).

Users will be able to browse both basic transcription tiers according to their segmentation, probably along with chunks of the original sound recordings. In addition to viewing concordance results with a limited context, we expect to provide access to full transcripts as well, because the conversation topics are often captivating and the material will thus be able to accommodate different research methodologies.

3 Cartographical and Geographical Processing of Data from the DIALEKT Corpus and Other Spoken Corpora

A crucial aspect of the compilation and analysis of spoken corpora is classifying the speakers according to the system of hierarchical territorial division. This system of geographical division of the Czech linguistic territory was created based on isoglosses capturing the distributional boundaries of differential dialectal phenomena; dialectologists have refined it over the past several decades. The boundaries of dialect regions and smaller areas used at the CNC have been determined mainly based on the *Czech Linguistic Atlas* and other similar manuals.⁷ In order to be able to process data from all spoken corpora, we worked to increase the accuracy of dialect region and area maps in collaboration with a cartographer.⁸

⁷ See mainly [10], [11], [12] and [13].

⁸ Our thanks are due to K. Kupka for his expert work on preparing the map data.

3.1 An Interactive Web Application

In the present day, modern cartographical methods can be used not only for statically displaying dialect region and smaller territorial units, but also for describing the individual locales and linguistic phenomena under investigation. The initial base maps were created using the ArcGis software,⁹ which makes it possible to export them for specific uses in a variety of formats. In the next step, we made a prototype application using the Google Maps service, which is more user-friendly and offers more flexible features, in order to test the various ways spoken corpus data can be accessed via a map interface. One of the options that emerged is an interactive map (currently in the stage of planning) displaying the points of the individual recordings in the DIALEKT corpus, whose role is to offer an intuitive spatial overview of the data with respect to the system of territorial division, and simultaneously to provide additional pop-up information¹⁰ about each point upon clicking on it. The information is organized into thematic tabs offering e.g. a list of characteristic features for the dialect region or area to which the given point belongs; a short excerpt from a transcript along with an analysis of it; or the possibility to listen to the recording corresponding to the excerpt (see Fig. 1). Our ideal goal is an integrated environment which will allow access to both these static descriptions and dynamically generated visualizations based on the user's queries in the DIALEKT corpus and the rest of the spoken corpora of the CNC. The user should thus be able to make a connection between speakers and recordings from the same or a nearby locale across all available corpora, easily confronting what these various information sources say about the territorial spread of a given feature. The maps should also allow visualizing the proportions of speaker contributions based on their childhood place of residence or longest place of residence (see [14] for details).

3.2 Visualizing Linguistic Phenomena on a Base Map

As far as visualizing users' corpus queries on a map is concerned, it is already possible to display the proportions of target variants as represented in spoken corpora, divided according to the traditional dialect regions, using the SyD corpus variant exploration tool [15].

Thanks to the language data contained in the DIALEKT corpus, the available spoken corpora, and possibly other dialect corpora in the future, users will be able to devise their own corpus queries and use the map to visualize e.g. the distribution of various phone-level phenomena (*v*-prothesis, narrowing, various types of assimilation etc.), morphological and word-formational phenomena, competing lexical variants etc. This visualization can be implemented either using Google Maps or another similarly user-friendly environment, or the aforementioned SyD corpus variant analysis tool.

⁹ Neither the publicly accessible web interface (<http://www.arcgis.com/explorer/>) nor the full paid version satisfied our further requirements for working with map data and custom visualizations.

¹⁰ This notion of providing a more detailed description of recordings and speakers from a given locale in the context of a map was inspired by the German database <http://www.dialektkarte.de/> and the English database <http://sounds.bl.uk/Sound-Maps/Accents-and-dialects>. The proof-of-concept implementation based on our data was made by D. Lukeš using the Google Maps API (Available at: <https://trnka.korpus.cz/~lukes/maps/>).

It will also be possible to confront the current distribution of various linguistic phenomena with the older status quo as recorded e.g. by the *Czech Linguistic Atlas*, thereby creating models with the aim of tracing development tendencies in contemporary Czech.

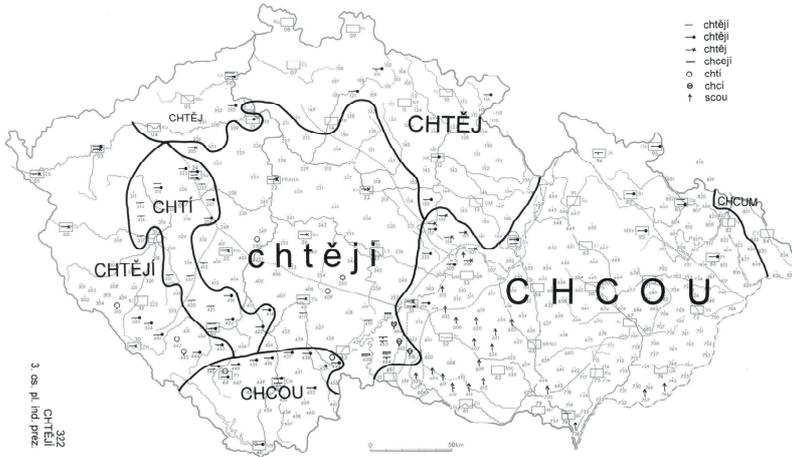


Fig. 2. Variants of the 3rd pers. pl. pres. ind. of the verb *chtít* (to want) in the recordings of the *Czech Linguistic Atlas*. Figure borrowed from [10, p. 459].

As an example, let us compare the maps depicting the distribution of variants of the 3rd pers. pl. pres. ind. of the verb *chtít* (to want) as captured by the *Czech Linguistic Atlas* [10, p. 322], and the results of a corresponding query using the SyD corpus variant exploration tool, which is based on the synchronic data of the CNC's spoken corpora¹¹ (see Figs. 2 and 3). The basic standard form of the 3rd pers. pl. pres. ind. is *chtějí*, which is also the basic form appearing in written language. However, spoken language offers a much more diverse picture of the variants in use. Thanks to a dialectological survey covering the entire territory of the Czech Republic, the *Czech Linguistic Atlas* has been able to capture additional forms of the verb: *chtějí*, *chcejí*, *chtěj*, *chtí*, *chcí*, *chci*, *chcou*, *chcú*, *scou*, *scú*, *chcum*. The corresponding query in the SyD variant exploration tool did not return occurrences of all these dialect variants of the verb form, but it revealed the current state of distribution of the individual variants that were found.

Inspecting a map visualization of the result (Fig. 3), we discern several shifts and development tendencies in the language: in Bohemia, the variant *chtěj* dominates in spoken language (it was originally located in the Northern Bohemian dialect region and along the river Ohře; it was also found among urban speakers in Central and Southern Bohemia, together with the form *chtějí*). In accordance with the original dialectological survey, the variant *chcou* is still present in Moravia and Silesia (Eastern Moravia still preserves the form *chcú*). In the Southern Bohemian dialect region, the situation is more varied: apart from the widely disseminated variant *chtěj*, the (original) standard form *chtějí* is used,

¹¹ The presented data are currently based on the ORAL2006 [1], ORAL2008 [2] and ORAL2013 [3] corpora. We welcome the fact that this pool of material will soon be augmented with data from the new spoken corpora, which will enable a more accurate description and confrontation of language phenomena.

the shortened form *chtěji* (previously attested primarily in the Central Bohemian dialect region), and to some extent the dialect form *chtí*. On the other hand, the form *chcou*, which had in the past been present in Southern Bohemia as well as in Moravia, is attested only sporadically in the recordings of the ORAL series corpora. The variants *chcejí*, *chcí*, *scou*, *scí* and the Silesian form *chcum* are disappearing from current

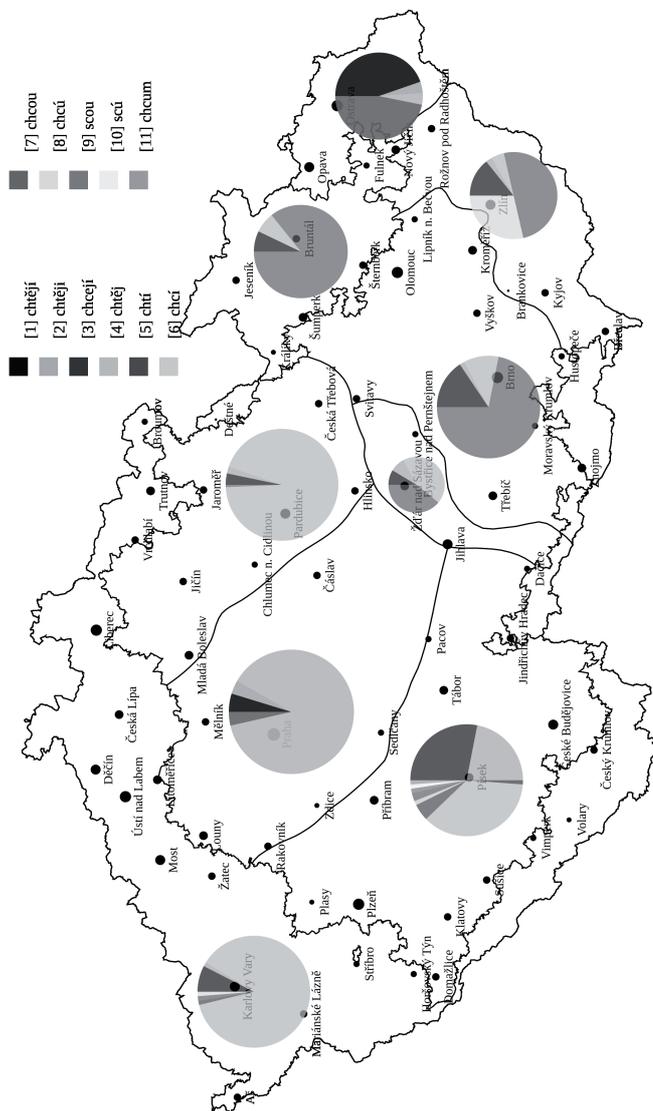


Fig. 3. Variants of the 3rd pers. pl. pres. ind. of the verb *chtít* (to want) in the ORAL series corpora, as visualized in the SyD tool. The visualization was adapted for greyscale printing; the results of the query can be inspected in more detail and in full colour at <https://syd.korpus.cz/2DOXMDB4>.

active use; apart from these, we also had to exclude the variant *chci* from the query.¹² In Silesia, a relatively high percentage was scored by the standard variant *chtějí*, which might correspond to the tendency of speakers from the Silesian region to replace dialectal variants with standard forms.

4 Conclusion

The DIALEKT corpus will be the only publicly accessible dialect corpus capturing data from all over the territory of the Czech Republic in the Czech linguistic context. In its final shape, it will probably contain around 200,000 running words.¹³ In the first place, it should serve the research community: dialectologists, who might use it as a complementary source of information for the compilation of a comprehensive *Dictionary of the Dialects of the Czech Language* (it will enable them to augment the dictionary entries with an appropriate dialectal context), but also linguists in general who are interested in tracking development tendencies of the language in the locales under investigation, or the evolution of individual language phenomena (see Sec. 3.2). It could also provide useful information to researchers from related fields (historians, ethnographers, sociologists etc.). Last but not least, it will be a flagship product for raising awareness about dialect data and the possibilities of corpus research.

The integrated web environment for working with the CNC's spoken corpus data (currently under preparation) will undoubtedly also be attractive for the lay public. One of the first already accessible results of this effort is the improved version of the SyD corpus variant exploration tool. One of the great advantages of the SyD tool and similar web applications (based on the Google Maps API or libraries with a similar functionality) is that they allow working with maps and the available language data in an intuitive and interactive fashion; compared to traditional static map-based depictions of linguistic phenomena¹⁴ (e.g. in linguistic atlases and other sources), they also afford the possibility of exporting and further processing the result. An interactive web environment will also constitute a valuable teaching resource to be used in schools; the tabbed pop-ups characterizing a given locale, speaker and recording excerpt could be particularly useful in this regard. This is an especially beneficial feature in the case of a dialect corpus, because it offers a detailed analysis of the dialect features in a selected locale (see Fig. 1 above).

¹² The form *chci* is simultaneously homonymous with the 1st pers. sg. pres. ind. of the verb *chtít*, which means that including it in the SyD map visualization would be misleading.

¹³ Currently however, we are considering the possibility of publishing a first version of the corpus at roughly half this size, i.e. about 100,000 words, in order to test the viability of our methodology. Since the corpus is under construction and continuously being updated, we refrain in this article from presenting descriptive statistics concerning the number of recordings processed, the total length of raw sound material, the speakers, the proportional contribution of the individual dialect regions, the number of collaborators involved in data collection and annotation, etc.

¹⁴ The Dialect Corpus of the Slovak National Corpus currently only provides public access to a static map displaying the Slovak dialect regions and dialect groups.

References

- [1] Kopřivová, M. and Waclawičová, M. (2006). *ORAL2006: korpus neformální mluvené češtiny*. Ústav Českého národního korpusu Filozofická fakulta Univerzity Karlovy, Praha. Accessible at: <http://www.korpus.cz>.
- [2] Waclawičová, M. et al. (2008). *ORAL2008: sociolingvisticky vyvážený korpus neformální mluvené češtiny*. Ústav Českého národního korpusu, Filozofická fakulta Univerzity Karlovy, Praha. Accessible at: <http://www.korpus.cz>.
- [3] Benešová, L., Křen, M., and Waclawičová, M. (2013). *ORAL2013: reprezentativní korpus neformální mluvené češtiny*. Ústav Českého národního korpusu, Filozofická fakulta Univerzity Karlovy, Praha. Accessible at: <http://www.korpus.cz>.
- [4] Balhar, J. et al. (2011). *Český jazykový atlas. Dodatky*. Academia, Praha.
- [5] Balhar, J., Jančák, P. et al. (1992, 1997). *Český jazykový atlas 1, 2*. Academia, Praha.
- [6] Karlík, P., Nekula, M., and Pleskalová, J., editors (2002). *Encyklopedický slovník češtiny*. Nakladatelství Lidové noviny, Praha.
- [7] Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., and Sloetjes, H. (2006). ELAN: a Professional Framework for Multimodality Research. In *Proceedings of LREC 2006, Fifth International Conference on Language Resources and Evaluation*, pages 1556–1559.
- [8] Dialektologická komise České akademie věd a umění (1951). *Pravidla pro vědecký přepis dialektických zápisů českých a slovenských*. Česká akademie věd a umění, Praha.
- [9] Lamprecht, A., Michálková, V. et al. (1976). *České nářeční texty*. Státní pedagogické nakladatelství, Praha.
- [10] Balhar, J. et al. (1999, 2002, 2005). *Český jazykový atlas 3, 4, 5*. Academia, Praha.
- [11] Bělič, J. (1972). *Nástin české dialektologie*. Státní pedagogické nakladatelství, Praha.
- [12] Voráč, J. (1976). *Česká nářečí jihozápadní I*. Nakladatelství Československé akademie věd, Praha.
- [13] Voráč, J. (1976). *Česká nářečí jihozápadní II*. Nakladatelství Československé akademie věd, Praha.
- [14] Kopřivová, M., Klimešová, P., Goláňová, H., and Lukeš, D. (2014). Mapping diatopic and diachronic variation in spoken Czech: the ORTOFON and DIALEKT corpora. In Calzolari, N. (Conference Chair), Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 376–382.
- [15] Cvrček, V. and Vondříčka, P. (2011). Výzkum variability v korpusech češtiny. In Čermák, F., editor, *Korpusová lingvistika Praha 2011 – 2. Výzkum a výstavba korpusů*, pages 184–195, Nakladatelství Lidové noviny, Praha.

The Automatic Identification of Type of Adverbials in Syntactically Annotated Texts¹

Milena Hnátková

Faculty of Arts, Charles University in Prague, Czech Republic

Abstract. The aim of this paper is to produce a list of adverbial multiword units (for different types of adverbials) using an automatic search for collocations and subsequent manual editing of the results, and also automatic identification of types of adverbials according to this list by the exploitation of the Franta program that automatically identifies collocations in corpus data. The comprehensive list of ways of expression of temporal noun adverbials (they can be expressed by nouns, noun phrases or prepositional phrases) was made on the basis of automatic search and manual processing of the results.

1 Introduction

An adverbial is a modifier dependent on a verb, adjective or adverb. It expresses the circumstances or relations in which an action described primarily by a verb occurs. Adverbials are divided into: local adverbials, temporal adverbials, adverbials of manner, cause, purpose, condition and concession.

The basic semantic classification of adverbials and nouns was made in order to support parsing of Czech (especially for distinguishing adverbials from objects). For this classification, the Czech WordNet ([1]) does not provide us with data classified in a sufficiently fine-grained way that we would need and that is contained in extensive corpora of Czech. Therefore, we made a new classification based on corpora of Czech written texts, the corpora being part of the Czech National Corpus. We divided the adverbials into three classes:

- (1) temporal adverbials
- (2) local adverbials
- (3) other adverbials (especially adverbials of manner, cause, purpose, condition).

Each of these classes is further divided into:

- (a) one-word adverbials (temporal, local, other)
- (b) multiword adverbials (temporal, local, other).

The multiword adverbials are further divided into:

- (b1) the fixed (phraseme) multiword adverbials
- (b2) general multiword adverbials.

We will demonstrate the method of processing on temporal adverbials, the processing of other types is similar. We will present the results of the automatic identification of temporal adverbials as a working version of a manually annotated corpus called Etalon containing 550,000 tokens.

Temporal adverbials can be regarded as answers to the following kinds of questions:

When? – *Jednoho dne musel otec odcestovat.* (The father had to travel **someday**.)

Since when? – *Situace se však od té doby nezlepšila.* (However, the situation has not improved **since then**.)

¹ This article was financially supported by the grant Grammar-based Treebank of Czech (Treebank češtiny na základě gramatiky – P406/13-27184S).

Until when? – *Zůstal jí věrný do konce života.* (He remained faithful to her **till the end of life.**)

How long? – *Podle expertů by likvidační práce měly trvat šest měsíců až rok.* (According to experts, the liquidation work should take **from six months to one year.**)

How often? – *Hrát se bude každý den až do konce července.* (We/One will play **every day** until the end of July.)

A temporal adverbial is expressed:

by an adverb: *Co děláte **dneska** večer?* (What are you doing **tonight**?)

by a noun: *Počkejte **chvilku**.* (Wait **a minute.**)

by a prepositional phrase:

*Bylo to **na sklonku léta**, když jedné teplé noci přišel **kolem desáté hodiny** ke svému domu.* (It was **in the late summer**, when on one warm night he came to his house **around ten o'clock.**)

*Nechal **dramatickou pauzu**, **během níž** jsem se rozhodl.* (He let a dramatic pause, **during which** I decided [what to do].)

by a subordinate clause: *Pohledávka vznikne, **když vám někdo dluží**.* (The claim arises **when someone owes you something.**)

2 Automatic Identification of Temporal Adverbials

2.1 Single-word Temporal Adverbials

For purposes of identification of single-word temporal adverbials expressed by adverbs a list of single-word adverbs was made in an automatic way from the source file of automatic morphological analysis. A manual classification of adverbs was also made in this list and temporal adverbs were marked/annotated. The resulting list comprised a total of 280 items of single-word temporal adverbs (for example: *předečtím* ‘a day before yesterday’, *kdykoli* ‘whenever’, *kdys* ‘although’, *dneska* ‘today’, *mezitím* ‘meanwhile’).

2.2 Multiword Temporal Adverbials

These adverbials are divided into fixed (phraseme) multiword temporal adverbials and general multiword temporal adverbials.

2.2.1 The Fixed (Phraseme) Multiword Temporal Adverbials

For purposes of identification of temporal adverbials a list of all fixed multiword temporal adverbials was created on the basis of the lists exploited by the program **FRANTA** (Ph(F) Raseme **AN**notation and **T**ext **A**nalysis (Hnátková 2011: 171)) for searching idioms/phrasemes. This list made it possible to automatically make a list of all non-inflectional phrasemes (for example: *na poslední chvíli* ‘at the last minute’, *dříve nebo později* ‘earlier or later’, *do soudného dne* ‘until doomsday’). Out of these idioms/phrasemes temporal adverbials were then manually selected, they are 287 at present (examples: *na sklonku roku* ‘at the end of the year’, *jednou za uherský rok* ‘once in a blue moon’, *při každé příležitosti* ‘at every opportunity’).

2.2.2 General Multi-word Temporal Adverbials

In our project, we mainly concentrated on the classification of temporal adverbials expressed by nouns, prepositional phrases with a single-word preposition, prepositional

phrases with a composite prepositional expression and prepositional phrases containing an „event/action“ noun. Semantic classification of adverbs is carried out on the basis of the Manual for Analytic Layer Tagging of the Prague Dependency Treebank (PDT) ([3]).

The processing of multiword nominal temporal adverbs was divided into the following steps:

(a) Creation of the list of temporal nouns

The list of temporal nouns was made in order to identify temporal adverbials expressed by nouns (or noun phrases). These nouns denote a period of time (for example *den* ‘day’, *únor* ‘February’, *jaro* ‘spring’) or event (for example *jednání* ‘action’, *výstava* ‘exhibition’, *koncert* ‘concert’) (see 2.2.2.1 below)

(b) Automatic identification of frequent nominal collocations containing these nouns in the corpus

Part-of-speech patterns of temporal adverbials were made (the list of prepositional expressions in the Dictionary of Czech Phraseology and Idiomatics (Čermák 1988) was used, for example: *po skončení* ‘after the end’, *před koncem* ‘before the end’, *po začátku* ‘after the start’) and a table for the program that searches multiword units containing temporal nouns was automatically created. After these multiword units had been found and duly annotated/marked in corpus data, the list of collocations thus found was automatically created and its items were subsequently manually categorized (see 2.2.2.2).

(c) Classification of the results, classification of types of expression of temporal adverbials

(d) Automatic markup of temporal adverbials in the texts of the czech national corpus

We now describe individual processing steps (a) and (b).

2.2.2.1 Making the List of Temporal Nouns

The list of temporal nouns (CAS) contains in addition to the classical temporal data – the names of units of time (*minuta, hodina*), the names of day parts (*rozbřesk, svítání, ráno, večer...*), the names of weekdays (*pondělí, úterý...*), the names of months (*leden, únor...*), the names of the seasons (*jaro, léto...*) – also the names of national and religious holidays and rites (*svátek, Velikonoce, Vánoce, mše, advent, posvícení, masopust, obřad*). It also includes words that can designate a time interval or time data:

the speech act: *konzultace, pohovor, rozhovor, proslov, hádka, promluva*; the phases of human life: *narození, dospívání, dětství, dospělost, mládí, manželství, rozvod, rozchod, důchod, stáří, smrt*; the name of periods of time: *období, století, milénium, desetiletí, tisíciletí, epocha, etapa, středověk, novověk, třetihory*; work activities and parts of a working day: *schůzka, šichta, činnost, zasedání, shromáždění, školení*; weather: *bouře, dešť, kalamita*; leisure and social activities: *festivál, radovánky, odpočinek, lov, koncert, veselí, slavnost, oslava*; sports activities: *cvičení, duel, utkání, zápas, závod, turnaj*; food consumption: *jídlo, večeře, snídaně, oběd*; medical procedures and treatments: *nemoc, epidemie, terapie, prohlídka, hospitalizace, operace, léčení*; the names of action/event: *hlasování, natáčení, malování, podnikání, mučení, čekání, čtení, sezení, jednání, vzhazování, obchodování* etc. The list contains about 350 temporal nouns.

The CAS list was further complemented by the data contained in the large synchronous written corpora of the SYN series[5], where prepositional phrases with the typically temporal preposition *během* ‘during’ were automatically found (such phrases can immediately be identified as temporal adverbials). The nouns contained in these phrases

were included in the CAS list. Furthermore, we manually verified automatically annotated prepositional phrases with prepositions *při* ‘after’, *při* ‘when’ and in case a temporal adverbial was found the relevant noun was again included in the CAS.

A table of other potentially temporal multiword units was also created and the FRANTA program for the search by the morphological tags was used for identifying all potentially temporal multiword units in the corpus data. These units were subsequently manually classified by the type of adverbial. It turned out that the composite prepositional phrases (for example: *v době* ‘at a time’, *před skončením* ‘before the end of’) are a very good guide for the identification of temporal adverbials.

2.2.2.2 The Part-of-Speech Pattern Search (POSP)

A list of part-of-speech patterns for the search of temporal adverbials was made on the basis of their occurrences in the Czech written corpora of the SYN series [5], which contains some general part-of-speech patterns (POSP) for finding further variants of temporal adverbials.

For the search and annotation of noun phrases containing temporal nouns by POSP, a special program FRANTA for automatic search of multiword units in morphologically annotated corpora was used. Individual words of a potential adverbial being found are annotated by the lemma of the pattern and the list of all adverbial connections that were found can be made from the search result along with the POSP type.

In examples of POSP listed below, parts of speech are denoted as follows: R – preposition, A – adjective, A2 – adjective in the genitive case, N – noun, PD2 – demonstrative pronoun in genitive, CI2 – cardinal numeral in genitive, N2 – noun in genitive, Ca4 – indeterminate numeral in accusative, Ca7 – indeterminate numeral in instrumental, Cn4 – cardinal numeral in accusative.

An example of automatic search according POSP:

<i>po mnoho desetiletí</i> ‘for many decades’	R_Ca4_N2
<i>před několika hodinami</i> ‘a few hours ago’	R_Ca7_N7
<i>v jednu chvíli</i> ‘at one moment’	R_CI4_N4
<i>do dvou dnů</i> ‘within two days’	R_CI_N
<i>na dva měsíce</i> ‘for two months’	R_CI_N
<i>za dva roky</i> ‘in two years’	R_CI_N
<i>v deset hodin</i> ‘at ten o’clock’	R_Cn4_N2

In addition to the cases provided in the overview in the Manual for PDT, we also included the search by some general POSP into the search of prepositional phrases containing temporal nouns. Using a program searching multiword units the following temporal adverbials were found in the sample corpus Etalon for the POSP R_A_A_N : *během celého vzrušeného veselí* ‘during the whole excited glee’, *po dlouhotrvající náročné operaci* ‘after a long difficult surgery’. In the search results for POSP R_N_R_N the following temporal adverbial was found: *Přesně na den od smrti mé dcery mě začaly trápit denní nevolnosti*. ‘A daily nausea began to torment me **the exact day after the death of my daughter.**’) Other examples of finding non-temporal prepositional phrases in the genitive case based on the list of temporal nouns are the search result of the noun phrases in the genitive case in the Etalon corpus according to the selected POSP: CI2_N2 – *jednoho dne* ‘someday’, PD2_N2 – *toho dne* ‘that day’, A2_N2 – *příštího dne* ‘(on) the following day’, CI2_A2_N2 – *jednoho krásného dne* ‘one beautiful day’.

3 The Evaluation of the Results of Automatic Search of Temporal Adverbials

The multiword units being found can be directly described as adverbials of a given type for a „certain“ POSP with the „temporal“ nouns (for example, the prepositional phrase with the preposition *během* ‘during’ and a temporal noun), in other cases the results must be manually verified and according to the lexical units complying with the POSP it must be determined whether a **temporal adverbial** was found. Only these cases can then be identified in the text as adverbials of the given type, but for the some collocations of the given pattern it cannot be automatically determined whether the given collocation is an adverbial, or whether it is an object, for example (see examples at the end of the paper). These occurrences identified in the data can, however, be manually checked and thus the type of a phrasal constituent can be identified on the basis of its particular occurrence in the sentence.

3.1 The Classification of Types Expressing a Temporal Adverbial

The individual items of the following classification are illustrated on examples of the actual use of the collocation in the Etalon corpus, along with the specification of other particular temporal adverbials of a given type.

The original classification contained in the Manual for Analytic Layer Tagging of the Prague Dependency Treebank (PDT) [3] was extended with other cases/examples, especially with structures containing secondary prepositions and with other nominal or prepositional phrases. Temporal adverbials as answers to the question: When? are expressed by the following means:

3.1.1 Temporal Adverbial Expressed by a Noun (or by a Non-Prepositional Noun Phrase)

3.1.1.1 Temporal Adverbials Expressed by a Noun in the Genitive Case (N2)

Examples: *Roku 2003 si zaregistrovala nakladatelskou živnost.* ‘She registered the publishing business **in 2003.**’

jednoho dne ‘one day’, *loňského roku* ‘last year’, *té první noci* ‘that first night’, *osudné noci* ‘on the fateful night’

3.1.1.2 Temporal Adverbial Expressed by a Noun in the Accusative Case or by an Accusative Noun Phrase (N4)

Examples: *Zastavil a chvíli nás pozoroval.* ‘He stopped and he watched us **for a while.**’

*Diskutovali už **drahnou dobu.*** ‘They were discussing **for a long time.**’

celý den ‘all day’, *delší dobu* ‘long time’, *druhý den* ‘the next day’, *nějakou dobu* ‘some time’, *hodnou chvíli* ‘a long while’, *pár dní* ‘a few days’

This group includes also special cases of a noun phrase in the accusative case with inverted word order: *každých deset minut* ‘every ten minutes’, *každých čtrnáct dní* ‘every fortnight’, *posledních dvacet minut* ‘the last twenty minutes’, *pouhých devět minut* ‘just nine minutes’, *pěkných pár roků* ‘good few years’

3.1.1.3 Temporal Adverbial Expressed by a Noun in the Instrumental Case (N7)

Examples: *Chodím **poslední dobou** soustavně pozdě.* ‘I am **these days** constantly late.’

časem, chvílemi ‘times’, *každým rokem* ‘each year’, *osmým rokem* ‘the eighth year’

3.2 Temporal Adverbial Expressed by a Prepositional Phrase

The survey of possible prepositions: *během, do, k, kolem, koncem, na, nad, mezi, o, od, okolo, počínaje, po, pod, pro, před, přes, při, uprostřed, v, z, za.*

3.2.1 Temporal Adverbial Expressed by a Prepositional Phrase with the Temporal Preposition *během* ‘during’ and by a Noun (or a Noun Phrase) in the Genitive Case (*během*+2)

Examples: *Nainstalovali přístroj během půlhodiny.* ‘They installed the device **within half an hour.**’

Setkáváme se s tím během života. ‘We encounter it **during our lives.**’
během několika minut ‘within a few minutes’, *během roku* ‘during the year’,
během podnikání ‘during the business’

3.2.2 Temporal Adverbial Expressed by a Prepositional Phrase with the Preposition *do* ‘until’ and by a Noun (or a Noun Phrase) in the Genitive Case (*do*+2)

Examples: *Do roka by měla dost peněz.* ‘**Within a year/In a year** she would have enough money.’

do rána ‘till morning’, *do dnešního dne* ‘till today’, *do smrti* ‘till death’, *do dvou dnů* ‘in two days’, *do jara* ‘until spring’, *do té doby* ‘until that time’

3.2.2a Temporal Adverbial Expressed by a Noun Phrase with the Prepositional Expression *do konce* ‘to the end’ and by a Noun (or a Noun Phrase) in the Genitive Case (*do konce*+2)

Examples: *Chcete to tak nechat do konce života?* ‘Do you want to keep it that way **to the end of life?**’

do konce života ‘to the end of life’, *do konce roku* ‘to the end of the year’

3.2.2b Temporal Adverbial Expressed by a Noun Phrase with the Prepositional Expression *do začátku* ‘to the beginning’ and by a Noun (or a Noun Phrase) in the Genitive Case (*do začátku*+2)

Examples: *Musí ho vybudovat do začátku jara.* ‘It must built **till the beginning of spring.**’

3.2.3 Temporal Adverbial Expressed by a Prepositional Phrase with the Preposition *k* ‘toward’ and by a Noun (or a Noun Phrase) in the Dative Case (*k*+3)

Examples: *Klid nastal až někdy k ránu.* ‘There was calm sometime **toward morning.**’

k půlnoci ‘towards midnight’, *k večeru* ‘toward evening’, *k poledni* ‘toward noon’

3.2.3a Temporal Adverbial Expressed by a Noun Phrase with the Prepositional Expression *ke konci* ‘at the end of’ and by a Noun (or a Noun Phrase) in the Genitive Case (*ke konci*+2)

Examples: *Ke konci války se do Ruska vypravil malý německý cirkus.* ‘A small German circus traveled to Russia **at the end of the war.**’

ke konci minulého století ‘at the end of the last century’

3.2.3b Temporal Adverbial Expressed by a Noun Phrase with the Prepositional Expression *k závěru* ‘to the conclusion of’ and by a Noun (or a Noun Phrase) in the Genitive Case (*k závěru*+2)

Examples: *Lidí na úřadech práce k závěru roku přibývá.* ‘People in the labor offices is increasing **to the conclusion of the year.**’ *Kino k závěru února ukončilo*

promítání. ‘The cinema terminated projections **to the conclusion of February**.’

k závěru roku ‘to the conclusion of the year’, *k závěru dne* ‘to the conclusion of the day’.

Most occurrences of the prepositional phrase *k závěru* ‘to the conclusion of’ in the corpus data are in collocations: *blížit se k závěru* ‘to approach the conclusion’, *spět k závěru* ‘to come to the conclusion’, *chýlit se k závěru* ‘to draw near to the conclusion’. The collocations with this prepositional expression can be found in the Etalon corpus only in the expression: *dojít/dospět k závěru, že ...* ‘to come to the conclusion that ...’. These cases cannot always be automatically distinguished, therefore the identification of these cases as adverbials must be performed manually in the data.

3.2.4 Temporal Adverbial Expressed by a Prepositional Phrase with the Preposition *kolem* ‘around’ and by a Noun (or a Noun Phrase) in the Genitive Case (*kolem*+2)

Examples: *Kolem půlnoci celá ubikace burácela smíchem*. ‘The whole dormitory roared with laughter **around midnight**.’ *Bobule dozrávají kolem 24. června*. ‘The berries ripen **around the 24th of June**.’

kolem poledne ‘around noon’, *kolem desáté hodiny* ‘around ten o’clock’

3.2.5 Temporal Adverbial Expressed by a Prepositional Phrase with the Preposition *koncem* ‘at the end of’ and by a Noun (or a Noun Phrase) in the Genitive Case (*koncem*+2)

Examples: *Koncem května kvetou jeho žluté kočičky*. Its yellow catkins bloom at the end of May.

koncem roku ‘at the end of the year’, *koncem století* ‘at the end of the century’, *koncem jara* ‘at the end of spring’, *koncem podzimu* ‘at the end of autumn’

3.2.6 Temporal Adverbial Expressed by a Prepositional Phrase with the Preposition *mezi* ‘between’ and by a Noun (or a Noun Phrase) in the Instrumental Case (*mezi*+7)

Examples: *Mezi svátky k nám chodívali vždycky jedno odpoledne*. ‘They used to come to us **between the holidays** always on one afternoon.’

mezi jídlem ‘between meals’, *mezi půlnocí a jednou hodinou* ‘between midnight and one o’clock’

3.2.7 Temporal Adverbial Expressed by a Prepositional Phrase with the Preposition *mimo* ‘outside’ and by a Noun (or a Noun Phrase) in the Accusative Case (*mimo*+4)

Examples: *Mimo pracovní dobu obcházel konkurzy na vedlejší role*. ‘He was taking part in the competitions for minor roles **outside working hours**.’

mimo ordináční hodiny ‘outside surgery hours’, *mimo jednání* ‘outside the meeting’, *mimo určený čas* ‘outside the specified time’.

3.2.8 Temporal Adverbial Expressed by a Prepositional Phrase with the Preposition *na* ‘on’ and by a Noun (or a Noun Phrase) in the Accusative Case (*na*+4)

Examples: *Na druhý den dal ženich zapřáhnout a posadil nevěstu do kočáru*. ‘The groom gave a yoke **on the second day** and took his bride into the carriage.’

na chvíli ‘on a moment’, *na celý život* ‘on a lifetime’, *na podzim* ‘on the autumn’, *na poslední chvíli* ‘on the last moment’, *na pár dní* ‘on a few days’

3.2.8a Temporal Adverbial Expressed by a Prepositional Phrase with the Prepositional Expression *na konec* ‘for the end of’ and by a Noun (or a Noun Phrase) in the Genitive Case (*na konec+2*)

Examples: *Mimořádná valná hromada je svolána na konec ledna.* ‘The extraordinary general meeting is convened **for the end of January.**’

na konec roku ‘for the end of the year’, *na konec srpna* ‘for the end of August’

3.2.8b Temporal Adverbial Expressed by a Prepositional Phrase with the Prepositional Expression *na závěr* ‘for the conclusion of’ and by a Noun (or a Noun phrase) in the Genitive Case (*na závěr+2*)

Examples: *Kladno na závěr soustředění v Turecku vyhrálo.* ‘Kladno won **for the conclusion of the training camp** in Turkey.’

na závěr roku ‘for the conclusion of the year’, *na závěr dne* ‘for the conclusion of the day’, *na závěr podzimu* ‘for the conclusion of the autumn’.

3.2.8c Temporal Adverbial Expressed by a Prepositional Phrase with the Prepositional Expression *na zbytek* ‘for the rest of’ and by a Noun (or a Noun Phrase) in the Genitive Case (*na zbytek+2*)

Examples: *Soud ho poslal na zbytek života do vězení.* ‘The court sent him to prison **for the rest of his life.**’

na zbytek dne ‘for the rest of the day’, *na zbytek odpoledne* ‘for the rest of the afternoon’.

3.2.8d Temporal Adverbial Expressed by a Prepositional Phrase with the Prepositional Expression *na zlomek* ‘for a fraction of, in a split’ and by a Noun (or a Noun Phrase) in the Genitive Case (*na zlomek+2*)

Examples: *V závěru musel v kopci na zlomek sekundy úplně zastavit.* ‘He had to stop completely at the hill **for a split second.**’

na zlomek vteřiny ‘for a split second’, *na zlomek okamžiku* ‘for a fraction of the moment’, *na zlomek času* ‘for a fraction of time’.

3.2.9 Temporal Adverbial Expressed by a Prepositional Phrase with the Preposition *na* ‘on’ and by a Noun (or a Noun Phrase) in the Locative Case (*na+6*)

Examples: *Mluvila o tom hned na začátku.* ‘She talked about it right **at the beginning.**’

na jaře ‘in spring’, *na zpáteční cestě* ‘on the way back’.

3.2.9a Temporal Adverbial Expressed by a Prepositional Phrase with the Prepositional Expression *na konci* ‘at the end of’ and by a Noun (or a Noun Phrase) in the Genitive Case (*na konci+2*)

Examples: *Na konci šichty mě bolelo celé tělo.* ‘The whole of my body ached **at the end of the shift.**’

na konci roku ‘at the end of the year’, *na konci války* ‘at the end of the war’

3.2.9b Temporal Adverbial Expressed by a Prepositional Phrase with the Prepositional Expression *na počátku* ‘at the beginning of’ and by a Noun (or a Noun Phrase) in the Genitive Case (*na počátku+2*)

Examples: *Bílé kvítky ostružin se objevují hned na počátku léta.* ‘White flowers of the blackberries can be seen **at the beginning of summer.**’

na počátku zimy ‘at the beginning of the winter’, *na počátku podnikání* ‘at the beginning of the business’.

3.2.9c Temporal Adverbial Expressed by a Prepositional Phrase with the Prepositional Expression *na prahu* ‘on the threshold of’ and by a Noun (or a Noun Phrase) in the Genitive Case (*na prahu+2*)

Examples: *Ani na prahu smrti nepřestávám být marnivá.* ‘I continue to be vain even on the threshold of death.’

na prahu jara ‘on the threshold of spring’, *na prahu zajímavého životního období* ‘on the threshold of an interesting period of life’.

3.2.9d Temporal Adverbial Expressed by a Prepositional Phrase with the Prepositional Expression *na přelomu* ‘on the turn of’ and by a Noun (or a Noun Phrase) in the Genitive Case (*na přelomu+2*)

Examples: *Na přelomu padesátých let tam působil francouzský odborník.* ‘A French expert worked there at the turn of the fifties.’

na přelomu dubna ‘at the turn of April’, *na přelomu letošního roku* ‘at the turn of the year’, *na přelomu století* ‘at the turn of the century’.

3.2.9e Temporal Adverbial Expressed by a Prepositional Phrase with the Prepositional Expression *na sklonku* ‘at the end of’ and by a Noun (or a Noun Phrase) in the Genitive Case (*na sklonku+2*)

Examples: *Bylo to na sklonku léta.* ‘It was in late summer.’

na sklonku roku ‘at the end of the year’, *na sklonku kariéry* ‘at the end of the career’

3.2.9f Temporal Adverbial Expressed by a Prepositional Phrase with the Prepositional Expression *na začátku* ‘at the beginning of’ and by a Noun (or a Noun Phrase) in the Genitive Case (*na začátku+2*)

Examples: *Kvete na začátku léta.* ‘It is in bloom at the beginning of summer.’

na začátku minulého století ‘at the beginning of the last century’, *na začátku skvělé mezinárodní kariéry* ‘at the beginning of a great international career’.

3.2.10 Temporal Adverbial Expressed by a Prepositional Phrase with the Preposition *nad* ‘over’ and by a Noun (or a Noun Phrase) in the Instrumental Case (*nad+7*)

Examples: *Teprve nad ránem jsem na chvíli neklidně usnula.* ‘I slept unquietly for a moment over the morning.’

The only temporal adverbial of this type is with the noun *ráno* ‘morning’, in addition to basic form *nad ránem* lit. ‘over the morning’ also the following variants can be found in the corpus: *nad chladným ránem* ‘over the cold morning’, *nad letním ránem* ‘over the summer morning’, *nad nedělním ránem* ‘over the Sunday morning’, *nad novoročním ránem* ‘over the New Year’s morning’, *nad včerejším ránem* ‘over yesterday’s morning’, *nad časným ránem* ‘over early morning’ (Compare the variant 3.19)

3.2.11 Temporal Adverbial Expressed by a Prepositional Phrase with the Preposition *o* ‘at’ and by a Noun (or a Noun Phrase) in the Locative Case (*o+6*)

Examples: *Hlavně když přijde o půlnoci.* ‘The main thing is that he will come at midnight.’

o Vánocích ‘at Christmas’, *o víkendu* ‘over the weekend’, *o pauze* ‘at a break’.

3.2.12 Temporal Adverbial Expressed by a Prepositional Phrase with the Preposition *od* 'from' and by a Noun (or a Noun Phrase) in the Genitive Case (od+2)

Examples: *Mnoho se od té doby nezměnilo.* 'Much has changed since that time.'

od dnešního dne 'from today', *od dětství* 'from childhood', *od července do září* 'from July until/to September', *od prvního okamžiku* 'from the first moment'.

3.2.12a Temporal Adverbial Expressed by a Prepositional Phrase with the Prepositional Expression *od konce* 'from the end of' and by a Noun (or a Noun Phrase) in the Genitive Case (od konce+2)

Examples: *Intenzivně Kanyza maluje od konce sedmdesátých let minulého století.* 'Kanyza intensively paints from the end of the seventies of the last century.'

od konce ledna 'from the end of January', *od konce války* 'since the end of the war'.

3.2.12b Temporal Adverbial Expressed by a Prepositional Phrase with the Prepositional Expression *od počátku* 'from the beginning' and by a Noun (or a Noun Phrase) in the Genitive Case (od počátku+2)

Examples: *Od počátku večera se na vás bezostyšně zaměřila.* 'She blatantly focused on you from the beginning of the evening.'

od počátku roku 'from the beginning of the year', *od počátku šestého týdne* 'from the beginning of the sixth week'

3.2.12c Temporal Adverbial Expressed by a Prepositional Phrase with the Prepositional Expression *od poloviny*, *od půlky* 'since mid-' and by a Noun (or a Noun Phrase) in the Genitive Case (od poloviny+2)

Examples: *Těžba byla od poloviny ledna postupně obnovena.* 'The mining was gradually restored since mid-January.'

od poloviny roku 'since the middle of the year', *od půlky července* 'since mid-July', *od poloviny devadesátých let* 'since the mid-nineties'.

3.2.12d Temporal Adverbial Expressed by a Prepositional Phrase with the Prepositional Expression *od začátku* 'from the beginning of' and by a Noun (or a Noun Phrase) in the Genitive Case (od začátku+2)

Examples: *Od začátku sedmdesátých let bydlí v Praze.* 'He has been living in Prague from the beginning of the seventies.'

od začátku utkání 'from the beginning of the match'.

3.2.13 Temporal Adverbial Expressed by a Prepositional Phrase with the Prepositional Expression *ode dne* 'since the day of' and by a Noun (or a Noun Phrase) in the Genitive Case (ode dne+2)

Examples: *Podnikatel je povinen vést účetnictví ode dne zahájení činnosti.* 'The entrepreneur is obliged to keep the accounts since the day of the commencement of activity.'

ode dne porodu 'since the day of the birth', *ode dne nástupu* 'since the day of the entry'.

3.2.14 Temporal Adverbial Expressed by a Prepositional Phrase with the Preposition *po* and by a Noun (or a Noun Phrase) in the Accusative Case (po+4)

Examples: *Po celou dobu bojoval s traumaty z Iráku.* 'He fought with the trauma from Iraq all the time.'

po celý rok 'the whole year', *po celý dosavadní život* 'all previous life'.

3.2.14a Temporal Adverbial Expressed by a Prepositional Phrase with the Prepositional Expression *po dobu* ‘for the period of’ and by a Noun (or a Noun Phrase) in the Genitive Case (po dobu+2)

Examples: *Kdo mě bude zastupovat po dobu nepřítomnosti?* ‘Who will represent me for the period of my absence?’

po dobu natáčení ‘for the period of filming’, *po dobu nemoci* ‘during illness’.

3.2.14b Temporal Adverbial Expressed by a Prepositional Phrase with the Prepositional Expression *po řadu* ‘for many’ and by a Noun (or a Noun Phrase) in the Genitive Case (po řadu+2)

Examples: *Nejedli po řadu týdnů nic jiného.* ‘They were eating nothing for many weeks.’

po řadu let ‘for many years’, *po řadu hodin* ‘for many hours’.

3.2.15c Temporal Adverbial Expressed by a Prepositional Phrase with the Prepositional Expression *po většinu* ‘most of’ and by a Noun (or a Noun Phrase) in the Genitive Case (po většinu+2)

Examples: *Děti jsou po většinu dne venku.* ‘Children are outside most of the day.’

po většinu času ‘most of the time’, *po většinu volebního období* ‘most of the election period’, *po většinu zápasu* ‘most of the match’.

3.2.14d Temporal Adverbial Expressed by a Prepositional Phrase with the Prepositional Expression *po zbytek* ‘for the rest of’ and by a Noun (or a Noun Phrase) in the Genitive Case (po zbytek+2)

Examples: *Po zbytek dne se nedokázala soustředit na vyučování.* ‘She could not concentrate on teaching for the rest of the day.’

po zbytek života ‘for the rest of life’.

The prepositional phrase *po zbytek* ‘for the rest of’ unambiguously identifies the temporal adverbial, nouns in such multiword expressions are temporal nouns.

3.2.15 Temporal Adverbial Expressed by a Prepositional Phrase with the Preposition *po* and by a Noun (or a Noun Phrase) in the Locative Case (po+6)

Examples: *Po celodenní jízdě dorazil poutník do města.* ‘The pilgrim arrived in the city after the day of driving.’

po ránu ‘in the morning’, *po chvíli* ‘after a moment’, *po Novém roce* ‘after the New Year’.

3.2.15a Temporal Adverbial Expressed by a Prepositional Phrase with the Prepositional Expression *po letech* ‘after the years of’ and by a Noun (or a Noun Phrase) in the Genitive Case (po letech+2)

Examples: *Po letech zkoušení to vzdala.* ‘She gave it up after the years of trying.’

po letech manželství ‘after the years of marriage’, *po letech usilovné práce* ‘after the years of the hard work’

3.2.15b Temporal Adverbial Expressed by a Prepositional Phrase with the Prepositional Expression *po řadě* ‘after the series of’ and by a Noun (or a Noun Phrase) in the Genitive Case (po řadě+2)

Examples: *Po řadě jednání jsme se dohodli.* ‘We stroke a deal after the series of meetings.’

po řadě průběžných jednání ‘after the series of continuous negotiations’, *po řadě měsíců* ‘after several months’, *po řadě šancí* ‘after many chances’.

3.2.15c Temporal Adverbial Expressed by a Prepositional Phrase with the Prepositional Expression *po skončení* ‘after the end of’ and by a Noun (or a Noun Phrase) in the Genitive Case (*po skončení+2*)

Examples: *Po skončení obřadu jsme se odebrali na oběd.* ‘We went for lunch **after the end of the ceremony.**’

po skončení války ‘after the end of the war’, *po skončení uplynulého roku* ‘after the end of last year’

3.2.15d Temporal Adverbial Expressed by a Prepositional Phrase with the Prepositional Expression *po týdnu*, *po týdnech* ‘after a week of’ and by a Noun (or a Noun Phrase) in the Genitive Case (*po týdnu+2*)

Examples: *Po týdnech hledání jsem ji našla.* ‘I found her **after weeks of searching.**’

po týdnu namáhavé práce ‘after a week of hard work’

3.2.15e Temporal Adverbial Expressed by a Prepositional Phrase with the Prepositional Expression *po ukončení* ‘after finishing’ and by a Noun (or a Noun Phrase) in the Genitive Case (*po ukončení+2*)

Examples: *Čeká ho rozhodnutí, jakou cestu zvolí po ukončení základní školy.* ‘The decision awaits him, which path he chooses **after finishing the primary school.**’

po ukončení pátého ročníku základní školy ‘after finishing the fifth year of primary school’, *po ukončení svého školního vzdělávání* ‘after completion of his school education’.

3.2.16 Temporal Adverbial Expressed by a Prepositional Phrase with the Preposition *počátkem* ‘at the beginning’ and by a Noun (or a Noun Phrase) in the Genitive Case (*počátkem+2*)

Examples: *Počátkem roku schválila rada města pokračování projektu.* ‘The City Council approved the continuation of the project **at the beginning of the year.**’

počátkem letošního roku ‘at the beginning of this year’, *počátkem září* ‘in early September’, *počátkem minulého měsíce* ‘at the beginning of the last month’

3.2.17 Temporal Adverbial Expressed by a Prepositional Phrase with the Preposition *počínaje* ‘starting from’ and by a Noun (or a Noun Phrase) in the Instrumental Case (*počínaje+7*)

Examples: *Soutěž počínaje letoškem získala na kreditu.* ‘The competition **starting from this year** gained importance.’

počínaje dneškem ‘starting from today’, *počínaje letošním rokem* ‘starting from this year’.

3.2.18 Temporal Adverbial Expressed by a Prepositional Phrase with the Preposition *před* ‘before’ and by a Noun (or a Noun Phrase) in the Instrumental Case (*před+7*)

Examples: *Před dvěma dny zažívala stejný pocit.* ‘She experienced the same feeling **two days ago.**’

před chvílí ‘a moment ago’, *před natáčením* ‘before shooting a film’, *před válkou* ‘before the war’.

3.2.18a Temporal Adverbial Expressed by a Prepositional Phrase with the Prepositional Expression *před koncem* ‘before the end of’ and by a Noun (or a Noun Phrase) in the Genitive Case (*před koncem*+2)

Examples: *Před koncem pracovní doby se u vás zastavím.* ‘I will come to you before the end of working hours.’

před koncem loňského roku ‘before the end of the last year’, *před koncem sezóny* ‘before the end of the season’.

3.2.18b Temporal Adverbial Expressed by a Prepositional Phrase with the Prepositional Expression *před skončením* ‘before the end/termination of’ and by a Noun (or a Noun Phrase) in the Genitive Case (*před skončením*+2)

Examples: *Ještě před skončením páté minuty srovnal Michal Rozsívál.* ‘Michal Rozsívál equalized before the end of the fifth minute.’

před skončením vyšetřování ‘before the end of the investigation’, *před skončením války* ‘before the end of the war’.

3.2.18c Temporal Adverbial Expressed by a Prepositional Phrase with the Prepositional Expression *před začátkem* ‘before the start of / before’ and by a Noun (or a Noun Phrase) in the Genitive Case (*před začátkem*+2)

Examples: *Před začátkem finále totiž musel na operaci slepého střeva.* ‘He had to undergo an operation of appendicitis before the start of the finals.’

před začátkem školy ‘before school’, *před začátkem soutěžního představení* ‘before the start of the competitive performance’

3.2.19 Temporal Adverbial Expressed by a Prepositional Phrase with the Preposition *přes* ‘over’ and by a Noun (or a Noun Phrase) in the Accusative Case (*přes*+4)

Examples: *Seriál měl přes dva roky obrovskou sledovanost.* ‘The series had a huge audience over two years.’

přes den ‘over the day’, *přes zimu* ‘over the winter’

3.2.20 Temporal Adverbial Expressed by a Prepositional Phrase with the Preposition *při* ‘at’ and by a Noun (or a Noun Phrase) in the Locative Case (*při*+6)

Examples: *Pozvedla jsem při obědě oči od talíře.* ‘I raised my eyes from his plate at lunch.’

při odchodu do práce ‘when leaving work’, *při práci* ‘at work’.

The multiword temporal adverbial of this type containing the names of weekdays occur, in the corpus, most frequently with the noun Monday ‘*při pondělku*’ – a total of 578 occurrences, only individual occurrences of multiword temporal adverbials contain the names of other weekdays in the corpus; the multiword adverbial with the noun Thursday: *při čtvrtku* ‘on Thursday’ are even missing in the corpus.

3.2.20a Temporal Adverbial Expressed by a Prepositional Phrase with the Prepositional Expression *při pohledu na* ‘looking at’ and by a Noun (or a Noun Phrase) in the Accusative Case (*při pohledu na*+4)

Examples: *Při pohledu na Petra zůstala stát s otevřenou pusou.* ‘She stood open-mouthed while looking at Petr.’

při pohledu na cedulku ‘looking at the notice’, *při pohledu na láhev vodky* ‘looking at the bottle of vodka’

In the prepositional phrase of this type there is no temporal noun. The preposition *při* ‘at’ associated with an action noun (for example, *hledání* ‘search’, *napřívání* ‘observation’) usually expresses a temporal adverbial; from this perspective, we can consider the word *pohled* ‘look’ as an action noun.

3.2.21 Temporal Adverbial Expressed by a Prepositional Phrase with the Preposition *s* ‘with’ and by a Noun (or a Noun Phrase) in the Instrumental Case (s+7)

Examples: *Se soumrakem přírodní živel sílil.* ‘The natural element grew stronger **with twilight.**’

The temporal adverbial of this type (with the preposition *s*, ‘with’) only occurs with certain temporal nouns: *ráno, jitro* ‘morning’, *soumrak* ‘dusk’, *rok* ‘year’, *svítání* ‘sunrise’, *úsvit* ‘dawn’ and with the names of the seasons, and most often in case the noun is modified by an active deverbal adjective: *s blížícím se ránem* ‘with the approaching morning’, *s přibývajícím červnovým soumrakem* ‘with the advancing June twilight’, *s probouzejícím se ránem* ‘with the waking morning’, and also by other adjectives: *s novým ránem* ‘with the new morning’, *s pátečním ránem* ‘with the Friday morning’, *s časným soumrakem* ‘with the early nightfall’, *s každým rokem* ‘with each year’, *s příštím jitem* ‘with the next morning’.

3.2.21a Temporal Adverbial Expressed by a Prepositional Phrase with the Prepositional Expression *s postupem* ‘during, over’ and by a Noun (or a Noun Phrase) in the Genitive Case (s postupem+2)

Examples: *Kvalita vody přírodních koupališť se s postupem léta zhoršuje.* ‘The quality of the water of natural swimming pools deteriorates **over/during summer.**’ *To se s postupem stáří stále zlepšuje.* ‘It is getting better **with ageing.**’
s postupem doby, s postupem času ‘over time’, *s postupem zimy* ‘during winter’

This variant of a temporal adverbial is sometimes used without the preposition *s*, particularly in the expression: *postupem času* ‘during time’ and *postupem doby* ‘during time’, and also with other temporal nouns: *postupem dní* ‘during days’, *postupem staletí* ‘in centuries’.

3.2.21b Temporal Adverbial Expressed by a Prepositional Phrase with the Prepositional Expression *s příchodem* ‘with the advent/coming of’ and by a Noun (or a Noun Phrase) in the Genitive Case (s příchodem+2)

Examples: *Teprve s příchodem večera kouzlo pominulo.* ‘The charm receded only **with the advent of the evening.**’
s příchodem zimy ‘with the advent of winter’, *s příchodem prázdnin* ‘with the coming of the holidays’, *s příchodem teplého počasí* ‘with the coming of warm weather’

3.2.22 Temporal Adverbial Expressed by a Prepositional Phrase with the Preposition *uprostřed* ‘in the middle of’ and by a Noun (or a Noun Phrase) in the Genitive Case (uprostřed + 2)

Examples: *Co ode mě chceš uprostřed noci?* ‘What do you want from me **in the middle of the night?**’
uprostřed zimy ‘in mid-winter’, *uprostřed léta* ‘in mid-summer’, *uprostřed zuřivé hádky* ‘in the middle of a violent quarrel’

3.2.23 Temporal Adverbial Expressed by a Prepositional Phrase with the Preposition *v* ‘at’ and by a Noun (or a Noun Phrase) in the Accusative Case (v+4)

Examples: *V jednu chvíli nemáte nic.* ‘You have nothing at one moment.’

v deset hodin ‘at ten o’clock’, *v podvečer* ‘in the evening’, *v tu chvíli* ‘at the moment’.

3.2.24 Temporal Adverbial Expressed by a Prepositional Phrase with the Preposition *v* ‘in’ and by a Noun (or a Noun Phrase) in the Locative Case (v+6)

Examples: *Lískové oříšky dozrávají v létě.* ‘The hazelnuts ripen in the summer.’

v květnu ‘in May’, *v poslední době* ‘more recently’, *v dané chvíli* ‘at a given moment’, *v dnešní době* ‘these days’, *v okamžiku* ‘at the moment’.

3.2.24a Temporal Adverbial Expressed by a Prepositional Phrase with the Prepositional Expression *v čase* ‘in time’ and by a Noun (or a Noun Phrase) in the Genitive Case (v čase+2)

Examples: *Debata o výdajích vyvrcholila v čase oběda.* ‘The debate about spending reached its peak during lunch.’

v čase krize ‘in times of the crisis’, *v čase adventu* ‘at the time of Advent’, *v čase Vánoc* ‘in the Christmas time’

3.2.24b Temporal Adverbial Expressed by a Prepositional Phrase with the Prepositional Expression *v době* ‘in/at time’ and by a Noun (or a Noun Phrase) in the Genitive Case (v době+2)

Examples: *V době války to bývala vzácnost.* ‘It used to be a rarity at the time of the war.’

v době odhalení ‘at the time of detection’, *v době velkých změn* ‘in the time of great changes’, *v době víkendových večírků* ‘during the weekend parties’

3.2.24c Temporal Adverbial Expressed by a Prepositional Phrase with the Prepositional Expression *v počátku* ‘at the beginning of’ and by a Noun (or a Noun Phrase) in the Genitive Case (v počátku+2)

Examples: *Taková předsevzetí se objevují v počátku každého roku.* ‘Such resolutions appear at the beginning of every year.’

v počátku devadesátých let ‘in the early nineties’, *v počátku utkání* ‘at the beginning of the match’.

3.2.24d Temporal Adverbial Expressed by a Prepositional Phrase with Prepositional Expressions *v polovině*, *v půlce* ‘in the middle’ and by a Noun (or a Noun Phrase) in the Genitive Case (v polovině+2, v půlce+2)

Examples: *V polovině koncertu přijde na řadu píseň.* ‘The song comes up in the middle of the concert.’

v polovině loňského roku ‘in the middle of last year’, *v polovině srpna* ‘in mid-August’

3.2.24e Temporal Adverbial Expressed by a Prepositional Phrase with the Prepositional Expression *v průběhu* ‘during’ and by a Noun (or a Noun Phrase) in the Genitive Case (v průběhu+2)

Examples: *V průběhu roku dochází k navyšování dotací.* ‘The subsidies are increased during the year.’

v průběhu večera ‘during the evening’, *v průběhu války* ‘during the war’

3.2.24f Temporal Adverbial Expressed by a Prepositional Phrase with the Prepositional Expression *v předvečer* ‘on the eve of’ and by a Noun (or a Noun Phrase) in the Genitive Case (*v předvečer+2*)

Examples: *V předvečer první adventní neděle rozsvítí vánoční strom v obci.* ‘Christmas tree lights up in the village **on the eve of the first Advent Sunday.**’

v předvečer poprav ‘on the eve of the execution’, *v předvečer odletu* ‘on the eve of the departure’, *v předvečer Dne smíření* ‘on the eve of the Day of Atonement’.

3.2.24g Temporal Adverbial Expressed by a Prepositional Phrase with the Prepositional Expression *v závěru* ‘at the end of’ and by a Noun (or a Noun Phrase) in the Genitive Case (*v závěru+2*)

Examples: *V závěru první půle viděli diváci další dvě tuctovky domácích.* ‘The spectators saw the other two surethings of domestic players **at the end of the first half.**’

v závěru poměrně klidné noci ‘at the end of a relatively quiet night’

3.2.24h Temporal Adverbial Expressed by a Prepositional Phrase with the Prepositional Expression *ve zlomku* ‘in a split’ and by a Noun (or a Noun Phrase) in the Genitive Case (*ve zlomku+2*)

Examples: *Ve zlomku sekundy skončila sezóna.* ‘The season ended **in a split second.**’

ve zlomku vteřiny ‘in a split second’, *ve zlomku času* ‘in a split time’

3.2.25 Temporal Adverbial Expressed by a Prepositional Phrase with the Preposition *za* and by a Noun (or a Noun Phrase) in the Genitive Case (*za+2*)

Examples: *Mlýn za války vyhořel.* ‘The mill burned down **during the war.**’

za trampského mládí ‘during tramping youth’, *za rozbřesku* ‘at daybreak’

3.2.25a Temporal Adverbial Expressed by a Prepositional Phrase with the Prepositional Expression *za doby* ‘during the time of’ and by a Noun (or a Noun Phrase) in the Genitive Case (*za doby+2*)

Examples: *Bylo to za doby hluboké totality.* ‘It was **during the time of deep totality.**’

za doby sucha ‘during the time of drought’, *za doby panování* ‘during the time of the reign’

3.2.26 Temporal Adverbial Expressed by a Prepositional Phrase with the Preposition *za* and by a Noun (or a Noun Phrase) in the Accusative Case (*za+4*)

Examples: *Budeme nahoře za chvíli.* ‘We’ll be at the top **in a moment.**’

za měsíc ‘in a month’, *za jedno odpoledne* ‘in one afternoon’

3.2.26a Temporal Adverbial Expressed by by a Prepositional Phrase with the Prepositional Expression *za dobu* ‘during the time of’ and by a Noun (or a Noun Phrase) in the Genitive Case (*za dobu+2*)

Examples: *Divadlo vystřídalo za dobu své existence několik názvů.* ‘The theatre changed several names **during the time of its existence.**’

za dobu svého působení ‘during the time of his tenure’, *za dobu její nemoci* ‘during the time of her illness’

3.2.27 Temporal Adverbial Expressed by a Prepositional Phrase with the Preposition *začátkem* ‘at the beginning of’ and by a Noun (or a Noun Phrase) in the Genitive Case (*začátkem+2*)

Examples: *Začátkem května je třešeň obalená bílými květy.* ‘The cherry tree is covered with white flowers **at the beginning of May.**’
začátkem července ‘at the beginning of July’, *začátkem dvacátého století* ‘at the beginning of the twentieth century’

4 Conclusion

As described above, a table of possible multiword adverbial noun and prepositional phrases for an automatic search of collocations was created for each kind of adverbial. The search results were manually processed and the list of multiword adverbials was created, which can be automatically tagged in morphologically annotated corpus data.

Via some general POSP containing temporal nouns it is possible to find interesting occurrences of temporal adverbials; this automatic search can lead to the automatic annotation of adverbials and to the automatic identification of their type, and, secondly, it leads to finding potential occurrences of adverbials, which it is necessary to manually annotate and to determine whether an occurrence is an adverbial of a given type.

We now show some examples where it is impossible to automatically determine whether a particular noun phrase in a morphologically disambiguated text is a temporal adverbial, or object; in parsed data it is possible to determine whether a given adverbial is a temporal one:

Proplakala celou noc. ‘She cried **all night through.**’ (object)

Byl celou noc vzhůru. ‘He was up **all night.**’ (temporal adverbial)

*Využili jsme **jednoho krásného letního dne** k dlouhé procházce.* ‘We used **a beautiful summer day** for a long walk.’ (object)

***Jednoho krásného letního dne** uspořádalo pár přátel společně se svými přítelkyněmi piknik.* ‘A few friends together with their friends organized a picnic **on a beautiful summer day.**’ (temporal adverbial)

The non-prepositional accusative noun phrases with temporal nouns cannot be automatically identified as temporal adverbials. It is only possible to mark their occurrence for subsequent manual checking. A manual identification of the type of adverbial must be performed primarily for those prepositional phrases, which generally may not be temporal adverbials: *Minulý rok jsem o Vánocích měla pár dnů **do porodu.*** ‘I had a few days **to birth** last year at Christmas.’

The described classification (see 3.1) is used to identify the type of adverbials in the data being already syntactically annotated, and possibly to repair automatic parsing.

Similarly as the temporal adverbials the local adverbials were also processed (for example: *venku* ‘outdoors’, *vzadu* ‘at the back’, *pod vodou* ‘under water’, *v lese* ‘in the woods’) and other, especially adverbials of manner (*krokem* ‘by step’, *zpaměti* ‘from memory’, *touto formou* ‘in this form’, *v dobré víře* ‘in good faith’ etc.), adverbials of cause (for example *pod dojmem posledních událostí* ‘under the influence of the last events’, *ze staré známosti* ‘from an old acquaintance’, adverbials of purpose (*za tím účelem* ‘for that purpose’) and conditional adverbials expressing condition (*za všech okolností* ‘in all circumstances’, *ani za zlaté prase* ‘even for golden pig’).

References

- [1] Czech WordNet 1.9 PDT. Accessible at: <https://ufal-point.mff.cuni.cz/xmlui/handle/11858/00-097C-0000-0001-4880-3>.
- [2] Čermák, F. et al. (1988). Slovník české frazeologie a idiomatiky 2. Výrazy neslovesné. Academia, Praha.
- [3] Hajič, J. et al. (1999). A Manual for Analytic Layer Tagging of the Prague Dependency Treebank. UK MFF ÚFAL Praha. URL: <https://ufal.mff.cuni.cz/techrep/tr09.pdf>.
- [4] 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 / Institute of the Czech National Corpus, Prague.
- [5] *Český národní korpus*. Ústav Českého národního korpusu, Filozofická fakulta Univerzity Karlovy, Praha. Accessible at: <http://www.korpus.cz>.

The Use of Acronyms in Different Communication Modes (a Corpus-Based Study)

Zuzana Komrsková

Faculty of Arts, Charles University in Prague, Czech Republic

Abstract. Language offers many possibilities to shorten a long word or a phrase. One of them is using an acronym, which is a shortened form of more than two words containing only their initial letters. Using them increases communication speed. Acronyms like *lol* or *wtf* originally appeared in chat conversations and from there, they spread to other computer-mediated communication genres, and eventually even into written texts and natural speech. This study compares their distribution in these three communication modes, based on data from corpora of contemporary language.

1 Introduction

Internet opens a whole new world for communication. It allows people to communicate over long distances quickly and easily. The Internet-mediated communication, in English-speaking environment known as computer-mediated communication (CMC), is very often discussed as a new communication mode between the spoken and written. While keeping some features from both traditional modes, it also introduces new elements, such as acronyms or emoticons. Generally said, an acronym is a shortened form of more than two words (a title, a phrase, or a collocation) which contains only the initial letters. Using acronyms, the communication speed grows up. Acronyms were originally created in chat and from there they widespread into other CMC genres (e.g. online discussion, e-mail). But they can also appear in contemporary written texts or natural speech. This study looks for acronyms in different types of corpora (written, spoken and corpora of CMC) and different languages (Czech, English, German, Slovak), trying to find out how acronyms are represented in these three communication modes and if they are really spreading outside the CMC.

2 Theoretical Background

2.1 Terminology: Acronyms, Abbreviations and Other Shortcuts

In linguistics, the term *acronym* is often connected with the study of word formation. This term is subordinate to the term *abbreviation*, which [1] define as “a shortened form of a word or phrase”. The acronym is considered a special type of abbreviation. [2] takes an acronym as “a word coined by taking the initial letters of the words in a title or phrase and using them as a new word”. Another subtype of abbreviation is an *initialism*, which consists of initial letters pronounced separately (e.g. BBC) [1]. The main feature of acronym, in contrast to initialism, lies in the fact that it is pronounced fluently as a word, if it is possible (e.g. NASA).

The terms *acronyms*, *abbreviations* and *shortcuts* are often mixed in the CMC discourse; their distribution depends mostly on the author’s preference. For example [3] defines abbreviations as word forms made up from initial letters which are pronounced in their full forms when read out. On the other hand [4] distinguishes acronyms, i.e.

shortcuts for common phrases (e.g. *brb* for *be right back*), from emotion acronyms (e.g. *lol* for *laugh out loud*), which should express Jakobson's phatic function.

In this paper the term *acronym* is used for all words composed of initial letters of a phrase, and the term *abbreviation* refers to shortened words.

2.2 Computer-Mediated Communication and Acronyms

The abbreviations, which are analyzed in the second part of this paper, were originally created in chat communication. Therefore, they are largely studied as a part of CMC. CMC is a relatively new field in linguistic research. It generally focuses on language in all types (genres) of electronic or Internet-mediated communication. [5] For synchronous CMC genres (like chat, instant messaging, SMS) a very important factor is the speed with which is text transferred. Shortcuts and abbreviations are results of language economy, of an effort to transmit as much information as possible in the shortest time possible. Other features of synchronous CMC especially include short turn-taking, the ignoring of written standards for punctuation, grammar or spelling, the use of the visual channel, or missing proofreading.

Acronyms and abbreviations are mainly discussed in analyses of chat or SMS, although they occur in all CMC genres. [6] Acronyms used in CMC spread from this domain to other communication modes – written and spoken, although [7] comments on their low frequency outside of CMC. It is usually a shortened phrase used very often to express one's agreement/disagreement (e.g. in Czech *jj* for *jo jo*) or to show listening, staying in contact. These types of acronyms are language-specific. [8]

There are many English acronyms used in non-English CMC. The most popular is *lol* (see [7], [9], [10]). Originally, it stands for *laughing out loud*, but it may also mean *lots of love*, or any other suitable phrase with these initial letters. Therefore, it is misleading to match this acronym only with laughter. [9] compare *lol* with other laughter variants (*haha* and *hehe*), occurring in their corpus of instant messaging, although at first it is "a signal of interlocutor involvement, just as one might say mm-hm in the course of a conversation" [9, p. 11]. A decline of *lol* according to respondent's age is explained through the "result of incremental loss of the stylized forms, *lol*, in favor of *haha*." [9], p. 13) However, [10] shows the phatic function of *lol*, too, and she compares it with *OK*, *cool*, or *yeah*.

Unfortunately, I found only one corpus-driven analysis of synchronous CMC with acronyms sorted according to frequency [9]: *lol*, *omg* (*oh my God*), *brb* (*be right back*), *tyl* (*talk to you later*), *btw* (*by the way*), *wtf* (*what the fuck*), *hwk* (*homework*), *gtg* (*got to go*), *np* (*no problem*), *lmao* (*laugh my ass off*), and *nm* (*not much*). After comparing this list with [11] and [7], I created my own list of frequently used acronyms. These will be studied in detail in the second part of this paper. My list consists of the following expressions: *btw* (*by the way*), *imho* (*in my humble opinion*), *lmao* (*laugh my ass off*), *lol* (*laugh out loud/lots of love*), *omg* (*oh my God*), *rofl* (*rolling on floor laughing*), *wtf* (*what the fuck*). When I divide these acronyms into semantic groups, I get the following: laughter (*lmao*, *lol*, *rofl*), a scale from a sigh and complaining to surprise (*omg*, *wtf*), and parenthesis (*btw*, *imho*).

3 Data and Methodology

The literature review on acronyms reveals that they have not been in the centre of researchers' attention. The major studies analyze only the data from CMC and few of them compare CMC with other forms of communication (e.g. speech). I focus on three

different modes of communication: written, spoken, and electronic (Internet-based). Data for this paper were extracted from corpora of different communication modes and different languages (Czech, English, German, and Slovak), based on the hypothesis that the chosen acronyms could be differently embedded in these modes and/or languages.

The list of corpora employed for searching acronyms (using queries such as [word=(?i) lol] within <s/>) is presented in Table 1.

To briefly investigate whether the relative frequencies differ in a mix of CMC genres and pure chat, two corpora of chat communication were used. Both chat corpora are a part of the Dortmund Chat Corpus, but they differ in size and communication area which was used for data collection. Since the Release Corpus shares a part of the data with the balanced Balack2a, I used the Release corpus only for the analysis of *imho* and the Balack2a for all the remaining expressions.

	Czech	English	German	Slovak
Written corpus	SYN	BNC	W-ÜBRIG ¹	prim-6.1-public-all
		COCA		
Spoken corpus	ORAL ²	BNC	FOLK	s.hovor 4.0
		COCA		
Web corpus	Araneum Bohemicum Maius, version 15.04	Araneum Anglicum Maius, version 14.04	Araneum Germanicum Maius, version 14.04	Araneum Slovacum Maius, version 14.04
Chat corpus	-	-	Balack2a	-
			Release	

Table 1. All corpora used in the study

Based on the suggestion that teenagers are frequent users of acronyms due to their use of synchronous CMC [10], I searched the spoken corpora of teenage language (The Bergen Corpus of LONDON Teenage Language – COLT for English and the *Corpus Oral de Lenguaje Adolescente – COLA* for Spanish), but I did not find any acronyms. This could be due to the year of data collection, which are 1993 and 2002.

4 The Use of Acronyms in Different Modes of Communication

This part summarizes the results of the acronym analysis. Table 2 shows the distribution of the seven acronyms in the corpora.

	Corpus Type	Corpus Name	BTW	IMHO	LMAO	LOL	OMG	ROFL	WTF
Czech	written	SYN	0.02	0.01	0	0	0	0	0
	spoken	ORAL	1.04	2.09	0	0.21	0	0	0.21
	web	ARA maius	4.21	1.78	0.01	1.35	0.68	0.09	0.70

¹ This corpus is a part of *Deutsches Referenzkorpus (DeReKo)*.

² This ORAL corpus comprises all the ORAL-series synchronic spoken corpora, and is not publicly available yet.

English	written	BNC ³	0.43	0.20	0	0	0	0	0
		COCA ³	0.05	0.02	0.005	0.15	0.13	0.003	0.05
	spoken	BNC ⁴	0	0	0	0.58	0	0	0
		COCA ⁴	0.03	0	0	0.15	0.07	0	0.05
web	ARA maius	5.99	1.29	0.31	12.69	2.52	0.13	1.24	
German	written	W-ÜBRIG	0.03	0.05	0	0.15	0.02	0.04	0.01
	spoken	FOLK	0	0	0	3.33	1.67	0	0
	web	ARA maius	1.24	1.13	0.006	1.98	0.74	0.14	0.39
	chat	BALACK2a	16.67	0	0	861.11	0	33.33	0
		Release	7.25	1.81	0	453.09	0	14.50	0
Slovak	written	prim-6.1- public-all	0.24	0.03	0	0.12	0.02	0.01	0.03
	spoken	s.hovor 4.0	0	0	0	0	0	0	0
	web	ARA maius	2.81	0.63	0.002	1.01	0.39	0.06	0.56

Table 2. The distribution of the seven acronyms in all corpora used in the study³⁴

4.1 CMC: Web and Chat Corpora

CMC shows the highest frequency of acronyms simply due to the fact that it is the platform where they were originally established. Table 2 also shows their progressive spread from the English language into other languages.

4.1.1 Web Corpora

In two web corpora, English and German, the semantic group of laughter significantly predominates because of the frequent use of *lol* (especially in English). Other laughter acronyms *lmao* and *rofl* are rather rare and they show the same tendencies as *lol*. The acronym *lol* occurs mainly (in 70%) within a sentence without any punctual separation from other words.⁵ It can be followed by a positive emoticon (in most cases by :) or :D). The co-occurrence of two or more *lol* expressions is rather rare (less than 0.05% of all occurrences in all languages). In English, I compared *lol* with the full phrases *laugh out loud* and *lot(s) of love*. The immediate context of *lol* and *laugh out loud* shows that this acronym allows the author to show their amusement without any greater interruption of the sentence structure, in contrast with the full phrase. Surprisingly enough, the immediate context or collocations were not helpful in distinguishing whether the acronym *lol* stands for laughter or *lot(s) of love*.

Btw is the most frequent acronym in the Czech and Slovak web corpus, whereas in English and German it holds the second place. It typically occurs at the beginning of

³ The number of frequencies consider only the written part of the corpus.

⁴ The number of frequencies consider only the spoken part of the corpus.

⁵ The comparison of *lol* with *haha* and *hehe* indicated that *lol* is a more general word for expressing amusement. The typical position of *haha* and *hehe* is at the very beginning or the end of a sentence.

a sentence, which is not surprising as the same applies for its full phrase version. The data also indicate that *btw* at this initial position is usually separated by a comma or a full stop. Words in the right context suggest that both *btw* and *by the way* open especially a question (e.g. *bych se ráda dozvěděla co nejvíc o procesu zkoušek na vysokou školu .. (btw nevíte někdo KDY zkoušky probíhají ?)*, Araneum Bohemicum) or comment with additional or new piece of information (which may not be related to the previous sentence).

Acronyms *omg* and *wtf* create another semantic group. Their meaning could be represented on a scale from a sigh and complaining to surprise. There is a typical punctuation for both acronyms: *omg* + exclamation mark, *wtf* + question mark, but it is possible to combine both exclamation and question mark to intensify the meaning of the acronym. These acronyms do not co-occur with any emoticons, except for emoticons expressing surprise. The comparison of these acronyms with their full phrases did not show any differences in their use; *omg* is followed by question words (especially *what* and *how*), *wtf* by the verb *to be* or a personal pronoun (in English). Both *omg* and *wtf* occur in all the examined web corpora together with *lol*, e.g. *What you can't do is simply realise that your argument is left in dust and resort to "omg lol! that's funny HA!"* (Araneum Anglicum).

Acronym *imho* expresses someone's meaning. As such it is mainly integrated into a sentence without any separation from the other words, but in English, there is a tendency to separate *imho* on both sides by commas just like its full phrase. *Imho* can also co-occur with *lol*, e.g. *Tomu potom zodpovedá aj prístup k zákazníkovi (vraj ich pánovi – IMHO LOL) a úroveň poskytovaných služieb či kvalita tovaru.* (Araneum Slovacum).

4.1.2 Chat Corpora

I searched the Dortmund Chat Corpus not only to compare frequencies, but also to find out how often an acronym creates its own utterance⁶. Table 2 shows that in the chat corpora only four of all studied acronyms were found: *btw*, *imho*, *lol*, and *rofl*. The most surprising is the presence of *rofl* which was the second lowest frequent acronym in the web corpus. On the other hand, *lol* is still the most frequent acronym, which corresponds with IM corpus, see [9]. The following examples show *lol*'s modification with a German prefix or adjective: *oberlol*, *heimlichlol*.

The only two acronyms that were found to create the whole utterances are *lol* (for Balack2a corpus in almost 54%) and *rofl* (for Balack2a in 66%). I agree with [10] that *lol* functions as a phatic marker, but it can also preserve its original meaning, see (1). In contrast with the web data, there are not any *lol* multiplication and co-occurrences with other laughter acronyms. The positive emoticons occur only rarely in the same utterances (in 5 cases).

(1) *Pit1: die grünen geben sich doch alle elitär (z. B flug nach bangkog in der business class)*

Walter: lol

The acronym *btw* is always situated at the beginning of a sentence and separated by comma or full stop. Almost 93% acronyms were in both chat corpora written by lower case.

⁶ An utterance refers to each chat contribution sent separately.

4.2 Written Corpora

In contrast with the web corpora, the selected acronyms were not found in all written corpora, probably due to the difference in corpus design (e.g. BNC includes e-mails) include some CMC genres (e.g. e-mails in BNC). The corpora also include texts from different time periods.⁷ Generally, the frequency of acronyms in written corpora is much lower than in the CMC corpora since the typical text type in written corpora – fiction – does not contain any acronyms. When we look at the distribution in text types, acronyms are used mainly in electronic newspapers and related discussions. Acronyms also become a new topic; they are explained and discussed as a new phenomenon and these occurrences can misrepresent the findings. Table 3 compares frequencies of all data and the so-called “real-used” data, from which were removed the acronyms occurring as a discussion topic.

		BTW	IMHO	LMAO	LOL	OMG	ROFL	WTF
BNC	All	39	18	0	0	0	0	0
	All (ipm)	0.43	0.2	0	0	0	0	0
	Real Use	39	18	0	0	0	0	0
	Real Use (ipm)	0.43	0.2	0	0	0	0	0
COCA	All	17	7	2	54	47	1	20
	All (ipm)	0.05	0.02	0	0.15	0.13	0	0.05
	Real Use	17	5	1	48	47	1	18
	Real Use (ipm)	0.05	0.01	0	0.13	0.13	0	0.05
W-ÜBRIG	All	3	5	0	14	2	4	1
	All (ipm)	0.03	0.05	0	0.15	0.02	0.04	0.01
	Real Use	0	0	0	9	0	0	1
	Real Use (ipm)	0	0	0	0.1	0	0	0.01
prim-6.1- public-all	All	162	25	2	80	11	7	22
	All (ipm)	0.24	0.03	0	0.12	0.02	0.01	0.03
	Real Use	147	16	0	43	8	1	21
	Real Use (ipm)	0.22	0.02	0	0.07	0.01	0	0.03
SYN	All	61	33	0	4	4	10	6
	All (ipm)	0.02	0.01	0	0	0	0	0
	Real Use	31	18	0	3	0	1	5
	Real Use (ipm)	0.01	0.01	0	0	0	0	0

Table 3. The frequencies of all acronyms compared with the frequencies of “real used” acronyms

⁷ The COCA corpus shows the increasing frequency according to time period. The highest increase is in 2009 and 2010.

Each of the studied acronyms preserves its function which was discussed in 4.1.1. The acronyms *btw* and *lol* are still the most frequent, while *lmao* with *rofl* are quite rare. In contrast with the CMC corpora, the multiplication of one acronym and co-occurrence of more different acronyms hardly occur. Another difference concerns the word form; acronyms in the CMC corpora show greater variability in the combination of upper and lower case, whereas in the written corpora they are either written in upper, or lower case.

Regarding the particular acronyms, there are not many differences from the CMC. *Btw* occupies the position at the beginning of a sentence, like in the CMC corpora. It is often (in 63 % of all cleaned *btw*) separated by a punctuation mark from other text. *Wtf* is mainly followed by a question mark and *omg* by an exclamation mark. The acronyms *btw* and *imho* are used in the sentences in the same way as their full phrase equivalents, e.g. *si zaplatili , alebo neprerabat nic , je imho vcelku jasne* (prim-6.1); *Naopak málo užitečné je IMHO pumpování peněz do tzv. sportovní reprezentace a sportovních klubů* . (SYN); *What has happened with Jamie Forrester , he is far more accomplished than Whelan IMHO* . (BNC); *IMHO he shud have have his own fkn company - and prolly will!* (COCA)

4.3 Spoken Corpora

The frequency of the acronyms in the spoken corpora is very low, even zero in the Slovak spoken corpus. This can be attributed to many factors, such as the conditions of recorded situation, year of recording or age of speakers. However, the presence of *lol* in three different spoken corpora might suggest the emerging use of this acronym in speech. Unfortunately, I did not find any of the acronyms in the teenage spoken corpora (see § 3).

The spoken data show both the real use of acronyms and the situations where acronyms are discussed and explained. The use of *lol* in speech is similar to the use of *OK*, e.g. Speaker 1: *Was that like sarcastic or funny? Was it rude? I don't get it.* Speaker 2: *LOL. I don't -- I don't like -- you know how I feel about all this.* (COCA); Speaker 1: *party . auch gu[t]* Speaker 2: *[°h] triff* Speaker 1: *jack johnson* Speaker 3: *lol* (FOLK).

Some corpora try to indicate the real pronunciation, e.g. *v t f co to je* (ORAL), *kann ma scho wieder nicht mehr sa[gen °h h°] °h äh oem ge* (FOLK). Despite the low frequency, the use of acronyms *btw*, *imho*, *lol*, *omg*, and *wtf* correspond with communication modes discussed so far. The multiplication of acronym was found only in the BNC: *Small boy says it 's okay . Ah ! Oh God ! Fucking move them to those ! Ah , lol lol lol lol Oh !*

5 Conclusion

This study presented a comparison of the distribution of the most frequent seven acronyms in three types of corpora (written, spoken and CMC corpora) and four different languages (Czech, English, German, Slovak). The highest acronym rate is still in their original communication mode – CMC, and the language in which they were originally created – English. Instances in the written and spoken modes are rather rare but their very occurrence indicates that they are spreading beyond CMC. No differences were found in the use of acronyms between the communication modes, which indicates the preservation of their semantics and function. A comparison among languages shows that the acronym proportion remains almost the same as in English.

Acknowledgements

The article was written under the auspices of grant project No. 15-011165 (Syntax mluvené češtiny) [Syntax of spoken Czech]), supported by the Grant Agency of the Czech Republic.

References

- [1] *Oxford Dictionaries*. Accessible at: <http://www.oxforddictionaries.com>, retrieved 17 March 2015.
- [2] Bauer, L. (2002). *The English Word-Formation*. Cambridge University Press.
- [3] Döring, N. (2002). Kurzm. wird gesendet – Abkürzungen und Akronyme in der SMS-Kommunikation. Muttersprache. Vierteljahresschrift für deutsche Sprache 2.
- [4] Varnhagen, C. K., McFall, G. P., Pugh, N., Routledge, L., Sumida-MacDonald, H., and Kwong, T. E. (2010). lol: new language and spelling in instant messaging. *Read Writ*, 23:719–733.
- [5] Herring, S. (2010). Computer-mediated conversation Part I: Introduction and overview. *Language@Internet* 7, article 2. Accessible at: <http://www.languageatinternet.org/articles/2010/2801>, retrieved 10 March 2015.
- [6] Crystal, D. (2006). *Language and the Internet*. 2nd edition, Camb. University Press, Cambridge.
- [7] Crystal, D. (2011). *Internet Linguistics: A Student Guide*. Routledge, New York.
- [8] Hård af Segerstad, Y. (2012). *Use and Adaptation of Written Language to the Conditions of Computer-Mediated Communication*. PhD. Thesis. Department of Linguistics, University of Gothenburg, Sweden.
- [9] Tagliamonte, S. A. and Denis, D. (2008). Linguistic Ruin? LOL! Instant Messaging and Teen Language. *American Speech*, 83:3–34.
- [10] Baron, N. (2009). The Myth of Impoverished Signal: Dispelling the Spoken Language Fallacy for Emoticons in Online Communication. In Vincent, J. and Fortunati, L., editors, *Emotion and ICTs*, pages 107–136, Peter Lang, London.
- [11] Baron, N. (2004). See You Online: Gender Issues in College Student Use of Instant Messaging. *Journal of Language and Social Psychology*, 23:397–423.

Corpora

- [1] Andersen, G. (2014). *COLT – The Bergen Corpus of London Teenage Language*. LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague. Accessible at: <http://hdl.handle.net/11372/LRT-368>.
- [2] Beißwenger, M. and Storrer, A. (2013). *Das Dortmunder Chat Korpus*. Dortmund: Lehrstuhl für Linguistik der deutschen Sprache und Sprachdidaktik. Accessible at: <http://www.chatkorpus.tu-dortmund.de>.
- [3] Benešová, L., Křen, M., and Waclawičová, M. (2013). *ORAL2013: reprezentativní korpus neformální mluvené češtiny*. Institute of the Czech National Corpus FF UK, Prague. Accessible at: <http://www.korpus.cz>.
- [4] Benko, V. (2014). Srovnatelné webové korpusy Aranea. Institute of the Czech National Corpus FF UK, Prague. Accessible at: <http://www.korpus.cz>.
- [5] Davies, Mark. (2008) *The Corpus of Contemporary American English: 450 million words, 1990-present*. Accessible at: <http://corpus.byu.edu/coca/>.
- [6] Institut für Deutsche Sprache (2012). *Deutsches Referenzkorpus / Archiv der Korpora geschriebener Gegenwartssprache 2012-II* (Release vom 29.08.2012). Institut für Deutsche Sprache, Mannheim. Accessible at: www.ids-mannheim.de/DeReKo.
- [7] Jørgensen, A. M. et al. (2008). *COLA: Un corpus Oral de Lenguaje Adolescente*. University of Bergen, Bergen. Accessible at: <http://www.colam.org>.

- [8] Kopřivová, M. and Waclawičová, M. (2006). *ORAL2006: korpus neformální mluvené češtiny*. Institute of the Czech National Corpus FF UK, Prague. Accessible at: <http://www.korpus.cz>.
- [9] Křen, M. et al. (2014). *Korpus SYN, verze 3 z 27. 1. 2014*. Institute of the Czech National Corpus FF UK, Prague. Accessible at: <http://www.korpus.cz>.
- [10] Schmidt, T. (2014). *Forschungs- und Lehrkorpus Gesprochenes Deutsch (FOLK)*. Institut für Deutsche Sprache, Mannheim. Accessible at: <http://dgd.ids-mannheim.de>.
- [11] Slovenský hovorený korpus – s-hovor-4.0. (2012). Jazykovedný ústav L. Štúra SAV, Bratislava. Accessible at: <http://korpus.juls.savba.sk>.
- [12] Slovenský národný korpus – prim-6.0-public-all. (2013). Jazykovedný ústav L. Štúra SAV, Bratislava. Accessible at: <http://korpus.juls.savba.sk>.
- [13] *The British National Corpus, version 2 (BNC World)*. (2001). Distributed by Oxford University Computing Services on behalf of the BNC Consortium. Institute of the Czech National Corpus FF UK, Prague. Accessible at: <http://www.korpus.cz>.
- [14] Waclawičová, M., Kopřivová, M., Křen, M., and Válková, L. (2008). *ORAL2008: sociolingvisticky vyvážený korpus neformální mluvené češtiny*. Institute of the Czech National Corpus FF UK, Prague. Accessible at: <http://www.korpus.cz>.

Evaluating Automatic Idiom Annotation in Spoken Corpora: the Case of Somatic Idioms

Marie Kopřivová

Faculty of Arts, Charles University in Prague, Czech Republic

Abstract. In this study, we aim to verify the reliability of the annotation of idioms in spoken corpora. Idioms are searched for and annotated using a special tool. Some Czech idioms come in different lengths, word order permutations and variants. These properties greatly complicate their identification. Somatic idioms are among the most common idioms in language. They can be easily retrieved by keyword (the name of the part of the human body). They are suitable for verifying the accuracy of annotation. For the evaluation, we use the well-known precision and recall measures.

1 Introduction

Corpora can be annotated on different levels. The most common one is the level of morphology, other levels include syntax, semantics etc. However, the annotation of collocations and idioms is very rare. This is the annotation level we would like to focus on in this paper.

We have developed a tool for the automatic identification and annotation of idioms called FRANTA [1]. It works with a morphologically tagged text [2], [3] and is based on a dictionary of idioms. The dictionary used by the tool is the Dictionary of Czech Phraseology and Idiomatics (DCPI) [8], [9], [10], [11] whose entries were accordingly modified for machine use.

We have experimentally annotated some parts of the ORAL series corpora of spoken Czech using this tool [4]. This annotated part will be henceforth referred to as ORALF. The ORALF corpus contains a subset of the data from the reference corpora ORAL2006 [5] and ORAL2008 [6]. It consists of 1,691,474 word forms. The data represent transcriptions of informal spontaneous utterances from the years 2002–2007.

In this paper we would like to evaluate the reliability of this automatic annotation in spoken corpora using the well-known measures precision and recall.

2 Automatic Idiom Annotation in Spoken Corpora

Spoken corpora contain the transcript of spoken utterances. Especially in the case of spontaneous utterances, the transcript contains some phenomena not seen in written language: word repetition, unfinished words, unfinished utterances and unusual word order, regional vocabulary [4]. These phenomena generally complicate machine processing of spoken language and automatic idiom annotation is not an exception.

The identification of some idioms is complicated by the fact that the same word combination can be used in its literal meaning, for example: *mávnout nad něčím rukou* – *wave a hand over something* (literal), *not to be concerned with something* (idiomatic). This problem is not basically solved by the tool FRANTA, it only uses the information (manually extracted from the corpus) whether, for a particular word combination, the idiomatic meaning is more frequent than the literal meaning or the opposite. According to this it either annotates the word combination as an idiom always or never.

3 Reference Data

For evaluating the automatic idiom annotation, we need reference data with correct annotation of idioms. It is unfortunately very difficult to manually annotate all idioms in a larger text, not only because of the length of the text: the situation is further complicated by the fact that there are different views on what is an idiom or idiom variant and what is not.

To simplify the situation, we decided to choose only one type of idioms, specifically somatic idioms. These are idioms which contain parts of the human body as components. Such idioms are frequent in most languages, they are related to the anthropocentric principle in language. Many of these idioms come from the description of gestures. We can then find the metaphorical meaning alongside the literal one in the text. Some of these idioms are repeated in the text and we can observe the influence of context on the recognition of the idiom.

We took the 40 nouns denoting parts of the human body¹ which occur in the corpus. Then we counted the number of automatically annotated idiom occurrences containing these nouns. From this set of 40 nouns, we selected 15 which were contained in at least 10 idiom occurrences. The resulting nouns are the following: oko (eye), prdel (ass), hlava (head), ruka (hand), prst (finger), huba (mouth), noha (leg), nos (nose), koleno (knee), pusa (mouth), srdce (heart), krk (neck), pata (heel), zadek (bottom), záda (back).

For these 15 selected nouns we then manually examined all their occurrences in the corpus, and for each occurrence, we decided whether it is part of an idiom. This was done without considering the dictionary. The result of this manual annotation was then used as the golden standard for the evaluation of the automatic annotation.

4 Results

In Table 1, you can see, for each of these words, the number of different idioms in the Dictionary of Czech Phraseology and Idiomatics (DCPI) [8], [9], [10], [11] (on which the FRANTA tool [12] is based), the number of different idioms found in ORALF by the FRANTA tool, the number of idiom occurrences found in ORALF by the FRANTA tool, and the number of occurrences of the lemma of the word itself in ORALF.

Part of human body	Number of different idioms in DCPI	Number of different idioms found in ORALF by FRANTA	Number of idiom occurrences found in ORALF by FRANTA	Number of corresponding lemma occurrences in ORALF
prdel	38	31	259	350
hlava	186	54	108	545
oko	176	43	94	377
ruka	169	43	85	502

¹ The range of somatic idiom nouns differs between authors and can be broader than the range we used. We considered only idioms containing parts of the human body, not body liquids or parts of animal bodies used expressively as substitutes for parts of the human body. See [7].

huba	86	20	38	89
noha	76	23	34	520
nos	65	16	19	85
koleno	13	5	17	125
pusa	15	7	16	97
srdce	84	10	15	69
krk	34	11	13	113
pata	17	8	11	36
prst	43	11	11	92
zadek	11	7	11	101
záda	28	6	10	149

Table 1. Parts of the human body in the ORALF corpus

Table 2 shows, for each of the selected parts of the human body, the number of idiom occurrences found automatically by the FRANTA tool, the number of wrong annotations made by FRANTA, the number of idiom occurrences not found by FRANTA, and the number of idiom occurrences annotated manually.

During the manual annotation of idioms, we searched for all occurrences of the corresponding lemma and for each occurrence, we decided whether it is part of an idiom or not.

Part of human body	Number of idiom occurrences found in ORALF by FRANTA	Number of wrong annotations made by FRANTA	Number of idiom occurrences not found by FRANTA	Number of idiom occurrences annotated manually	Precision	Recall
prdel	259	2	54	311	99.23	82.64
hlava	108	6	42	144	94.44	70.83
oko	94	4	32	122	95.74	73.77
ruka	85	13	44	116	84.71	62.07
huba	38	4	8	42	89.47	80.95
noha	34	3	23	50	91.18	62.00
nos	19	3	3	19	84.21	80.95
koleno	17	0	6	23	100	73.91
pusa	16	1	11	26	93.75	57.69
srdce	15	3	12	24	80.00	50.00
krk	13	1	2	14	92.3	85.71
pata	11	0	4	15	100.0	73.33
prst	11	2	3	109	81.82	75.00
zadek	11	2	3	12	81.82	75.00
záda	10	5	2	7	50.00	71.43

Table 2. Evaluation of the annotation of idioms involving parts of the human body in the ORALF corpus

Based on Table 2, it can be seen that in most cases precision is higher than recall. Precision is lowered by superfluous annotations made by the FRANTA automatic tool. In these cases, FRANTA annotated the collocation as an idiom even though it was used in its literal meaning. For example:

mít čisté ruce (to have clean hands vs. idiomatic meaning *not to be involved in wrongdoing*)
mít velké srdce (to have a big heart vs. idiomatic meaning *to be broad-minded*)

Recall is lowered by idiom occurrences which were not found by the automatic tool FRANTA. The most frequent reasons for not identifying an idiom occurrence are the following:

1. Idioms not listed in the dictionary (DCPI) at all.
2. Idioms listed in the dictionary with a different (but synonymous) word filling one of the slots.² For example: *být z ruky* (idiomatic meaning *to be located far away*) is listed in the dictionary vs. *mít to z ruky* (with the same idiomatic meaning) is not listed in the dictionary and thus not found by FRANTA.
3. Idioms listed in the dictionary with a different morphological form. For example: *spráskne rukama* (a variant of *spráskne ruce*).
4. Idioms with additional words inserted between the idiom components, usually filler words, or unfinished and repeated words.

5 Conclusion

The evaluation of automatic idiom annotation based on the family of somatic idioms showed that the results of automatic idiom annotation are quite encouraging, especially in terms of precision. In spite of lower recall values, idiom annotation is appreciated by phraseologists.

It can be further improved relatively easily by adding more idioms and their variants to the dictionary. Additional improvements may be achieved by taking regional dialectal morphology into account.

Another added benefit is that currently, the tool FRANTA is being used as part of a state-of-the-art rule-driven morphological disambiguation system for Czech, see [13, p. 167].

In our further work, we would like to explore these possibilities. Especially for spoken corpora, it is also important to cope with the words inserted between idiom components.

With verbal idioms, it is very difficult to distinguish the different idioms and their variants [14]. In order to facilitate research into idioms, it will probably be a good idea to integrate all variants, and maybe even idiom transformations, into one idiom lemma. But it is also a theoretical decision for which we do not usually have enough occurrences, as shown in the example *mít to po ruce* – *být po ruce*. Here the meaning of the idiom is the same, but the valence of the verbs and the total number of components differ.

² Idioms are usually treated as fixed multi-word expressions, however, in the case of verbal idioms, they have some flexibility. The verb used in the idiom can be an element chosen from quite a large set of synonymous verbs. The set can change in time or depending on the region and therefore all variants are not listed in the dictionary. It would be good to supplement the dictionary based on the corpus. To do this however, enough occurrences are needed in order to eliminate the possibility of random or idiolectal use.

Acknowledgements

This paper resulted from the implementation of the Czech National Corpus project (LM2011023) funded by the Ministry of Education, Youth and Sports of the Czech Republic within the framework of Large Research, Development and Innovation Infrastructures.

References

- [1] Hnátková, M. (2007). Automatické značkování frazémů a víceslovných výrazů gramatické povahy v korpusu. In *Gramatika a korpus*, pages 87–92, Ústav pro jazyk český Akademie věd České republiky, Praha.
- [2] Hajič, J. (2004). *Disambiguation of Rich Inflection (Computational Morphology of Czech)*. Vol. 1. Karolinum Charles University Press, Praha.
- [3] Petkevič, V. (2006). Reliable Morphological Disambiguation of Czech: Rule-Based Approach is Necessary. In *Insight into the Slovak and Czech Corpus Linguistics*, pages 26–44, Veda, Bratislava.
- [4] Kopřivová, M. and Hnátková, M. (2013). Identification of Idioms in Spoken Corpora. In Gajdošová, K. and Žáková, A., editors, *Proceedings of the Seventh International Conference Slovko 2013; Natural Language Processing, Corpus Linguistics, E-learning*, pages 92–99, Slovak Academy of Sciences, Bratislava.
- [5] Kopřivová, M. and Waclawičová, M. (2006). *ORAL2006: korpus neformální mluvené češtiny*. Ústav Českého národního korpusu FF UK, Praha. Accessible at: <http://www.korpus.cz>.
- [6] Waclawičová, M., Kopřivová, M., Křen, M., and Válková, L. (2008). *ORAL2008: sociolingvisticky vyvážený korpus neformální mluvené češtiny*. Ústav Českého národního korpusu FF UK, Praha. Accessible at: <http://www.korpus.cz>.
- [7] Mrhačová, E. (2000). *Názvy částí lidského těla v české frazeologii a idiomatice*. Ostravská univerzita, Ostrava.
- [8] Čermák, F. et al. (1983). *Slovník české frazeologie a idiomatiky. Přirovnání*. Academia, Praha.
- [9] Čermák, F. et al. (1988). *Slovník české frazeologie a idiomatiky 2. Výrazy neslovesné*. Academia, Praha.
- [10] Čermák, F. et al. (1994). *Slovník české frazeologie a idiomatiky 3. Výrazy slovesné*. Academia, Praha.
- [11] Čermák, F. et al. (2009). *Slovník české frazeologie a idiomatiky 4. Výrazy větné*. Leda, Praha.
- [12] Hnátková, M. and Kopřivová, M. (2014). From a Dictionary to a Corpus, *Phraseology and Culture*, pages 92–99, University of Maribor, Maribor.
- [13] Jelínek, T. and Petkevič, V. (2011). Systém jazykového značkování korpusů současné psané češtiny. In *Korpusová lingvistika Praha 2011. 3 Gramatika a značkování korpusů*, pages 154–170, Nakladatelství Lidové noviny / Ústav Českého národního korpusu, Praha.
- [14] Šichová, K. (2013). *Mit Händen und Füßen reden. Verbale Phraseme im deutsch-tschechischen Vergleich*. Julius Groos Verlag/Stauffenburg, Tübingen.

SK ANTIPLAG: Five Years After

Július Kravjar

Slovak Centre of Scientific and Technical Information, Bratislava, Slovakia

Abstract. The paper looks back at five years of operation of the national corpus of bachelor's, master's, PhD, rigorous¹ and habilitation² theses at Slovak higher education institutions – the Central Repository of Theses and Dissertations (CR) – and the follow-up Plagiarism Detection System. Both systems together are known as SK ANTIPLAG. CR along with the selected internet sources serves as a reference corpus for each thesis entering the CR. SK ANTIPLAG appeared in the everyday life of higher education institutions in late April 2010. The amendment to the Higher Education Act of October 2009 makes the use of both systems mandatory for all higher education institutions operating under Slovak law (20 public, 13 private and 3 state institutions). The deployment of SK ANTIPLAG in routine operation is an example of a unique and unprecedented implementation of such a system on a national level. The system's acquisition costs were covered by the Ministry of Education, Science, Research and Sport of the Slovak Republic which also covers its operating costs; higher education institutions do not pay to use SK ANTIPLAG. Thanks to a consistent methodology for collecting theses and metadata that applies to all Slovak higher education institutions, SK ANTIPLAG can provide various types of analytical outputs. The paper presents some of these outputs.

1 Introduction

Higher education institutions (HEIs) abroad started to use plagiarism detection systems about twenty years ago. In Slovakia in 2001, only one (private) HEI used a plagiarism detection system (PDS). Later, in 2009, two public HEIs started to use a PDS. The start of SK ANTIPLAG's routine operation on a national level in 2010 was a breakthrough, an innovation on a national as well as on an international level.

The main components of SK ANTIPLAG include:

- A central repository: the Central Repository of Theses and Dissertations (the CR);
- A plagiarism detection system (also known as PDS, originality check, the anti-plagiarism system, APS);
- A comparative corpus (consisting of theses from the CR and documents downloaded from the internet); and
- Local repositories at HEIs.

The theses are collected in a local repository and then uploaded by the CR in regular time intervals. HEI is the only channel for the delivery of the thesis to the CR. Each new thesis is registered and archived in the CR for 70 years, PDS compares it with the theses stored in the comparative corpus, an originality check protocol is generated, then the protocol is sent to a HEI and the thesis is included in the comparative corpus. The originality

¹ “Small doctorate” can be received by a person with a master's degree. It requires that a candidate passes rigorous examination and defends rigorous thesis; rigorous thesis is less valuable than PhD thesis; it is closer to master thesis than to PhD thesis.

² A prerequisite for the granting of the scientific-pedagogical degree “docent” (associate professor) is a PhD degree, a habilitation lecture, and submission and defense of the habilitation thesis.

check protocol by similarity index (or similarity percentage) indicates what percentage of the text from the submitted theses is similar to texts from other theses and documents. The protocol is not a confirmation that the thesis is a plagiarism or a confirmation that the thesis is an original – it is a basis for the decision of the Examination Committee in the matters of plagiarism.

The similarity of two texts is a statistical parameter of coexistence of similar words in these texts. A similar word is the identical word in different forms (gender, number, case etc.), a synonym in different forms etc. Index (percentage) of similarity is the ratio between the number of characters of the text identified as similar to the total number of characters of a thesis. Pictures are not evaluated. Tables are evaluated only if they are in the text form. The theses are evaluated in their entirety.

Nearly 80% of HEIs were prepared to use SK ANTIPLAG already in May 2010. The situation how HEIs started to use SK ANTIPLAG is showed in Table 1 and Fig. 1.

PDS in use	Year	Number of higher education institutions							
		Cumulative	Percentage	Public HEIs		State HEIs		Private HEIs	
				Cumulative	Percentage	Cumulative	Percentage	Cumulative	Percentage
Individual systems	2001	1	3,0%	0	0,0%	0	0,0%	1	10%
	2009	3	9,1%	2	10,0%	0	0,0%	1	10%
SK ANTI-PLAG	May 2010	26	78,8%	17	85,0%	3	100,0%	6	60%
	Sep 2010	30	90,9%	20	100,0%	3	100,0%	7	70%
	Mar 2011	33	100,0%	20	100,0%	3	100,0%	10	100%
	Jun 2014	36	100,0%	20	100,0%	3	100,0%	13	100%

Table 1. History of PDS's use and readiness of HEIs to use SK ANTIPLAG

All state HEIs were prepared in May 2010, all public HEIs in September 2010 and all private HEIs in March 2011.

Two private HEIs were founded in 2011 and one private HEI was founded in 2012. New HEIs started to use SK ANTIPLAG at the moment when their students started to submit their theses (May and June 2014). Currently, all 36 Slovak HEIs operating under Slovak legal order uses SK ANTIPLAG routinely.

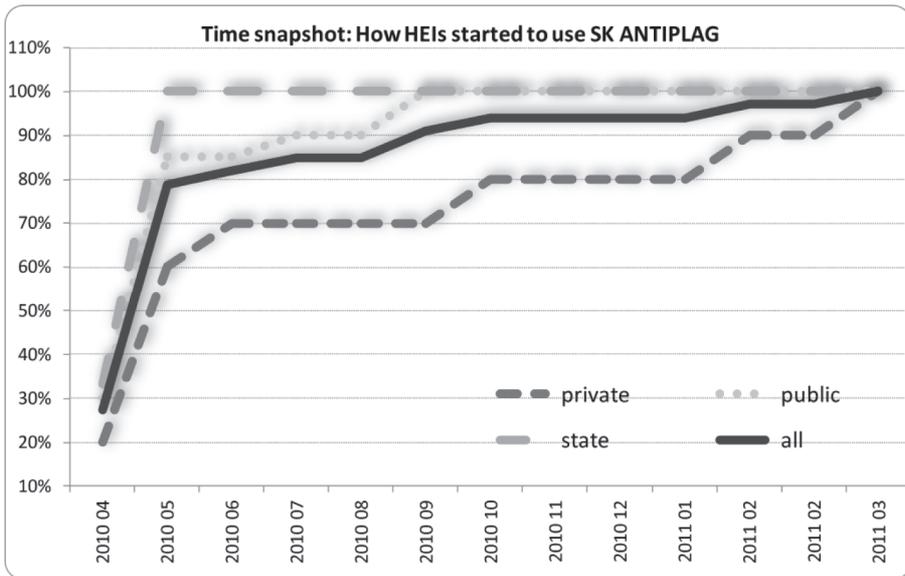


Fig. 1. Readiness of HEIs to use SK ANTIPLAG

The basic motivation that led to the establishment of SK ANTIPLAG was the need to build a barrier to uncontrolled proliferation of plagiarism. Low awareness of academic integrity, ethics, intellectual property rights, technological progress (increased computerization and internet penetration), and an enormous increase in the number of HEI students and inadequate growth in the number of HEI teachers – all these factors contributed to the spread of plagiarism.

Joint effort to suppress plagiarism and bolster academic ethics at HEIs was recorded in September 2006. At that time, the Slovak Rectors' Conference (SRC) approved two documents relating to academic ethics. The first one "Measures to Eliminate Plagiarism in the Preparation and Presentation of Bachelor's, Master's and PhD Theses" was intended for students and the second one "Code of Ethics for HEI Employees" was intended for HEI employees. They were documents of national importance, but with a subliminal impact on the academic community. Slovakia needed a real systemic measure to prevent the proliferation of plagiarism.

This systemic measure has become the implementation of SK ANTIPLAG. There was relatively little room for preparation and the task was challenging. We present only the most important milestones:

- SRC asked the Ministry of Education, Science, Research and Sport (MinEdu) in February 2008 to coordinate the activities related to the acquisition of a plagiarism detection system.
- MinEdu's decision in 2008: a plagiarism detection system on a national level will be procured; MinEdu will cover the costs. SCSTI³ was charged to define the system's

³ Slovak Centre of Scientific and Technical Information (SCSTI); in Slovak: Centrum vedecko-technických informácií SR (CVTI SR)

parameters, procure and operate it and to take care about the further development of the system.

- The Higher Education Act Amendment passed in October 2009 made the use SK ANTIPLAG mandatory for all Slovak HEIs operating under the Slovak legal order⁴; HEIs will send five types of theses (bachelor's, master's, PhD, rigorous and habilitation) to the CR in order to be checked for originality before the defense.
- In August 2009, MinEdu issued a Methodological Guidance on the formalities of theses, their bibliographic registration, originality check, archiving and disclosure. This guidance became a source for the drafting of internal directives at HEIs.
- Start of SK ANTIPLAG's routine operation: 30 April 2010.
- Public disclosure of theses registered in the CR after 31 August 2011 (metadata and full text) in accordance with license agreements on the basis of a further amendment to the Higher Education Act. The amendment provides for an obligation to keep opponent opinions in the CR.

2 Recognition Abroad

A survey was conducted in all EU countries as part of an international project "Impact of Policies for Plagiarism in Higher Education across Europe" (IPPHAE, funded by EU, 2010-2013). The project's outputs are available at <http://ippheae.eu/project-results>. A part about Slovakia Plagiarism Policies in Slovakia [6] reads as follows:

"There were some notable differences between the Slovak survey and the EU average. Almost all Slovak students (99%!) became aware of plagiarism before or during their bachelor studies. The EU average shows that 20% of students become aware of plagiarism during their masters/PhD degree or are still not sure about it." [6, p. 7]

"... Slovak students are the most aware of plagiarism among all EU countries." [6, p. 8]

"The most outstanding example of good practice is definitely the existence of national repository of theses." [6, p. 8]

"As it is run centrally and universities are obliged to upload their theses, students from all institutions have theoretically the same conditions. The other aspect is that the software tool provides just a protocol for matching with other sources. The decision about whether a given case is plagiarism or not lies with teachers and/or the examination committee and these may not always follow the same procedures." [6, p. 8]

"Compared to other countries, Slovakia should be praised for its achievements. And it already was: In June 2013, the European Commission has awarded the CVTI SR Slovakia the European Prize for Innovation in Public Administration." [6, p. 8]

"The responses from Slovak students demonstrated the highest level of understanding about plagiarism within the whole Europe. Their unwillingness (in comparison with other countries) to receive more training on plagiarism is therefore understandable. The research team of the IPPHEAE project would also like to praise Slovakia for existence of national repository of theses and built-in plagiarism detection tools." [6, p. 9]

⁴ There are 40 HEIs in Slovakia; 36 of them operate under Slovak legal order.

3 Methods of Implementing Plagiarism Detection Systems

Detecting plagiarism using information and communication technologies is one of the ways of dealing with plagiarism in higher education. The use of PDS started about 20 years ago. The most common model in the world is that HEIs provide an individual service to check the originality of student theses. This model is very likely more expensive than the model where the originality check is provided centrally for all HEIs in the country.

It is strictly defined in Slovakia what is to be sent to the CR (text of the thesis, metadata, license agreement, opponent opinions). All the theses are collected in the CR and are subjected to the originality check. SK ANTIPLAG is being developed and it is able to cover also other types of documents.

Repeated upload of theses in the CR is permitted exceptionally because repeated upload may support business like “we will provide documents for the thesis”. These words hide preparation of bachelor’s, master’s and other theses and written works on order. Such thesis does not bear the name of the actual author, but the name of the person for whom the thesis was written. Repeated uploads of theses allows a supplier in cooperation with the customer to gradually reduce the similarity index to required level.

The theses.cz system [13] operates in the Czech Republic. The system is developed and managed by the Masaryk University in Brno and is used to archive theses (bachelor’s, master’s and PhD) and to check originality. The project started in 2008 and the system is used by 50% HEIs. The repository of theses is publicly accessible. Compared to the Slovak system, HEIs are not required to use the theses.cz system. Theses may be uploaded repeatedly.

In Poland (as before 1 October 2014), 176 of 444 HEIs used the plagiat.pl system. HEIs use it in various ways, it is not necessarily used by all HEI faculties and a central repository of theses is not created. An amendment to the Higher Education Act, which entered into force on 1 October 2014 [11], changed the situation significantly. A central repository of theses was created, but the theses are not publicly accessible. Originality checks will start from the academic year 2015/2016. HEIs are obliged to send to the central repository texts of all theses (bachelor’s, master’s and PhD) with the defined metadata. By the end of 2016, HEIs have an obligation to deliver to the central repository all theses written after 30 September 2009. They will not be subjected to the originality check. According to PAP [9], February 2015, the emerging national repository contains about half a million theses. There are two main differences between the Slovak and Polish system: theses in Poland will not be publicly available [5] and they will not be checked by one plagiarism detection system.

There is prepared the repository and plagiarism detection system on a national level in Slovenia since 2009. Till today the system is not fully used by all HEIs.

In four countries – CZ, SI, PL, SK – there is a tendency to establish a national repository of HEI theses and the originality check of all theses that are uploaded in the repository. To our knowledge, this tendency cannot be seen in other parts of the world, even though it may seem so.

We have registered several sources which reported that in the country the PDS have nationwide or almost nationwide coverage. Two examples follow.

In the Rashid’s [12] paper there is written that the Turnitin plagiarism detection system has been extended to all HEIs in the public and private HEIs throughout Pakistan. It seemed as if Slovakia had a double. A more detailed analysis showed that the system does not cover all HEIs.

In 2001, the United Kingdom decided to fund the project “Plagiarism Advisory Service”, which was transformed to PlagiarismAdvice.org⁵ at the end of 2002.

“The service’s aim was to establish a national strategy to allow UK higher and further education institutions to check the authenticity of student work. Universities and colleges were given access to the Turnitin text matching software, virtually unknown in the UK at that point, at no charge for an initial three years.” [10, p. 2]

“In the UK, Turnitin is currently used by over 98%⁶ of higher education institutions and more than 44% of further education colleges and a growing number of schools. ... models of use in UK institutions vary widely, with some institutions using the software as a wholly summative tool to confirm allegations of academic malpractice on the part of a student, and with the Originality Report providing vital case processing evidence, in the majority of institutions the value of the tool as a formative aid to support teaching and learning as part of the assessment process ...” [10, p.3]

We decided to verify the information that Turnitin is used by more than 98% HEIs in the UK. There are 164 public HEIs in the UK [4] according to statistics. The publication *Privately Funded Providers of Higher Education in the UK* maps private providers of higher education:

“Our mapping research has identified a total of 674 named privately funded HE providers operating in the UK. This figure is a minimum estimate for the total number of providers, anticipating that some providers may not have been identified through the research process.” [1, p. 7]

It means the UK does not have accurate statistics on private HEIs. We approached iParadigms Europe with a question whether the 98% estimate includes both public and private HEIs. The answer was that the estimate applies only to public HEIs. Thus, the published information that “Turnitin is used by more than 98% HEIs in the UK” [10] is biased and misleading.

4 SK ANTIPLAG: A Valuable Data Source for Analyses

For the analyses, the data stored in the system for the academic years 2009/2010 to 2013/2014 were used and we did not take into account art theses, i.e. those for which the originality check protocol is not prepared. It should be noted that the years 2009/2010 and 2010/2011 are not complete due to gradual onset by HEIs (Fig. 1). The gradual onset can also be detected in the graph characterizing monthly increments of theses in the CR (Fig. 2). To the end of July 2015 in the comparative corpus there is about 10 mil. documents which costs us 5+ TB of disk space (plain texts); in the central repository there is 0.43 mil. of theses (0.85 TB of disc space, full texts).

First, SK ANTIPLAG was operated at CVTI SR, but from September 1st 2011 system is operated in the data centre, which was built in the framework of the national project “Infrastructure for Research and Development – Data Center for Research and Development”.

⁵ PlagiarismAdvice.org belongs to iParadigms Europe, which supplies the Turnitin plagiarism detection system.

⁶ The 95% value is stated in [14, p. 148], where Barrie 2008 is cited.

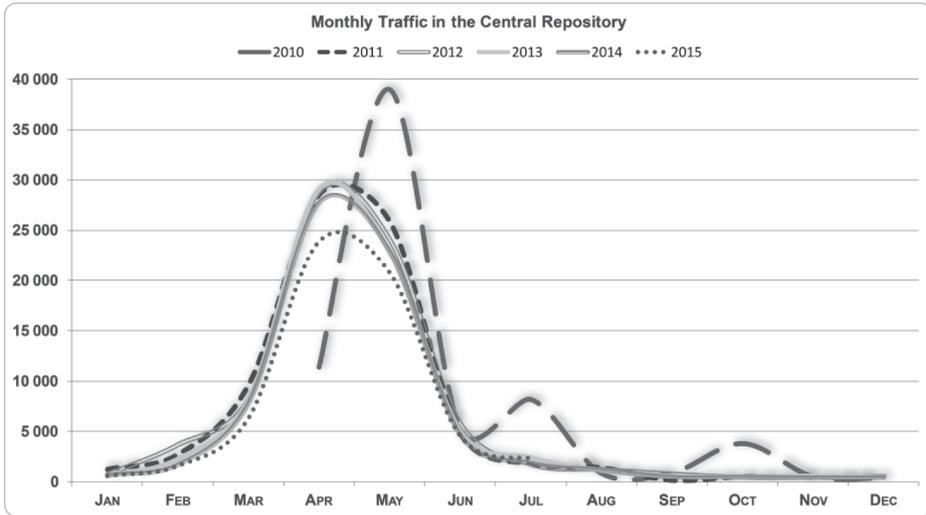


Fig. 2. Monthly traffic in the Central Repository

Given that the CR collects theses and the metadata using a consistent methodology, the system allows the provision of various analytical outputs. Some examples for illustration: the average similarity index (percentage) by academic year and by type of thesis, the share of theses with the similarity index exceeding the specified value by academic year and by type of thesis, an overview of identical thesis titles (by supervisor, by HEI), number of theses per one supervisor, the similarity index (by supervisor, field of study, department, faculty, HEI, HEI type, etc.).

HEI theses in the CR are divided as follows: 77.7% public, 19.8% private and 2.5% state. The representation by type of thesis is as follows: bachelor's 50.4%, master's 43.9%, rigorous 2.6%, PhD 2.8% and habilitation 0.3%. Bachelor's and master's theses have a dominant 94.3% share in the CR.

For the whole examined period, rigorous theses show the highest similarity index, bachelor's and master's theses are on a lower almost identical level, and PhD theses show the lowest similarity index (Table 2). The highest similarity index of rigorous theses is probably due to the fact that many rigorous theses are a continuation of master theses.

Thesis Type	Average Similarity Index
Bachelor's	6.39%
Master's	6.27%
Rigorous	10.33%
PhD	3.55%
Habilitation	4.18%

Table 2. Average similarity index by thesis type

The highest average similarity index by type of HEI and thesis type for a period of five years (Table 3) is recorded for private HEIs; state and public HEIs have significantly lower values. This may be due to several factors; it would require a more detailed analysis.

The highest similarity index is recorded for rigorous theses and the lowest for PhD and habilitation theses. The pairs of theses bachelor-master and PhD-habilitation theses are roughly on the same level.

HEI Type / Thesis Type	Bachelor's	Master's	Rigorous	PhD	Habilitation
Private	11.15%	11.50%	13.64%	6.69%	9.68%
State	5.56%	6.01%	9.71%	4.96%	4.77%
Public	4.94%	5.19%	9.28%	3.38%	3.60%

Table 3. Average similarity index by HEI type and thesis type

The highest share of theses with the similarity index exceeding 25% for a period of five years (Table 4) is recorded for private HEIs; state and public HEIs have significantly lower values.

HEI Type / Thesis Type	Bachelor's	Master's	Rigorous	PhD	Habilitation
Private	16.03%	15.77%	19.77%	6.00%	12.37%
State	5.58%	5.88%	11.21%	3.90%	6.45%
Public	4.95%	5.12%	11.94%	2.48%	3.12%

Table 4. Share of theses with the similarity index > 25% by HEI type and thesis type

The average similarity index by academic year and type of thesis shows an upward trend, which is likely related to the improvements in the comparative corpus (Fig. 3).

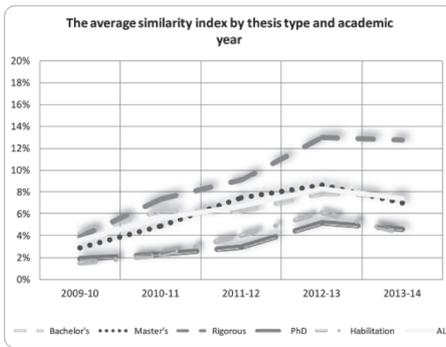


Fig. 3. Average similarity index

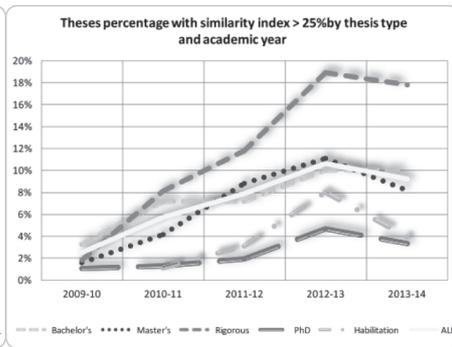


Fig. 4. Similarity index > 25%

The share of theses with the similarity index exceeding 25% by type of HEI and thesis type for a period of five years (Fig 4) shows similar trends as the average similarity index; however, it grows faster (except for PhD theses) than the average similarity index. Both indices may be to some extent an indicator of the quality of the supervisors' work with students.

For comparison of the average similarity index by fields of study we selected bachelor's theses for the academic year 2013/2014. The average similarity index for these theses was 7.61%. We filtered out the fields of study in which the number of theses was lower than 10 and the average similarity index was lower than 10% (Table 5). There are seven fields of study that have the similarity index higher than twice the average.

Field of study	Similarity index	Number of theses
Teaching special subjects	34.66%	124
Spatial planning	23.56%	28
Teaching vocational subjects	19.88%	2546
Mining	18.09%	42
Agricultural & forestry technology	17.39%	46
Landscaping	15.95%	70
Safety & health protection at work	15.46%	205
Protection of persons & property	14.75%	728
Social work	14.57%	376
Land construction	13.67%	370
Tutorage	13.10%	18
Quality of production	12.67%	205
Forestry	11.43%	83
Social services & counseling	11.30%	89
Hunting	11.06%	49
Public admin. & regional development	10.99%	811
Food hygiene	10.97%	25
Crop production	10.91%	32
Management & marketing	10.85%	11
Environmental management	10.77%	126
Postal technology	10.22%	47
Protection & land use	10.16%	173
Road transport	10.08%	42

Table 5. Fields of study and the similarity index (bachelor's theses, 2013/2014, number of theses > 9)

Table 6 shows more populated fields of study of bachelor's theses for the academic year 2013/2014 (with more than 300 theses) with the similarity index greater than 7.61% (average for bachelor's theses for the academic year 2013/2014).

Field of study	Similarity index	Number of theses
Teaching vocational subjects	19.88%	2546
Protection of persons and property	14.75%	728
Social work	14.57%	376
Land construction	13.67%	370
Public administration and regional development	10.99%	811
Jurisprudence	9.39%	1735
Public policy and public administration	8.60%	396
Industrial engineering	7.77%	337

Table 6. Fields of Study and the Average Similarity Index (bachelor's theses, 2013/2014, number of theses > 300)

Such outputs of a similar nature may serve for different levels of management as auxiliary indicators for support in the evaluation and decision-making. These outputs can be detailed to the level of HEI, faculty, department, and supervisor.

5 Similarity Index and ARRA Ranking

The Academic Ranking and Rating Agency (ARRA) assess annually the quality of HEI faculties in Slovakia. It prepares their ranking on the basis of a comparison of indicators of quantity and quality of education and research. The criteria are divided into five basic groups – education, the attractiveness of study, research, doctoral studies and grant success. The publication Evaluation of HEI faculties 2014 – Ranking of HEI faculties based on the comparison of indicators of quantity and quality of education and research⁷ [2] evaluates 112 faculties or one-faculty HEIs, comprising 104 public faculties and 8 faculties of private HEIs, in 11 groups of specializations (it does not evaluate state HEIs). Below is a list of groups of faculties and HEIs according to the publication of ARRA [2]. Groups of faculties:

Technical sciences (TECH) – civil engineering, electrical engineering, computer science, electronics, mechanical engineering and other technical fields;

Natural sciences (PRIR) – mathematics, physical, chemical and biological sciences and earth and environmental sciences;

Medical sciences (MED) – general medicine and stomatology, clinical medicine, pharmaceutical sciences, nursing and health care;

Agricultural sciences (AGRO) – agriculture, forestry, veterinary medicine and related fields;

Economic sciences (EKONOM) – economic faculties;

Other social sciences (OSTATNE SPOL) – faculties of social sciences with a focus on public administration, international relations, political and economic science, mass media communication and other related fields;

Philosophical sciences (FILOZOF) – philosophy, history, languages, literature and other related fields;

Law sciences (PRAV) – faculties of law;

Pedagogical sciences (PEDAGOG) – faculties of pedagogy;

Theological sciences (TEOLOG) – theological faculties; and

Art (UMEL) – faculties focused on music, drama, film and visual arts.

This classification into groups to some extent “homogenizes” the data for comparison, although the homogeneity of these groups is lower than the homogeneity of groups by fields of study. This classification was applied to the year under review (2013/2014) and we attempted to verify the hypothesis that the ranking of faculties by those groups will be correlated with the ranking of faculties based on the average similarity index. We assumed that a higher value of the index may mean a lower ranking by ARRA (negative correlation). The group UMEL we ignored, because for this type of theses the original check is not provided.

⁷ Hodnotenie fakúlt vysokých škôl 2014 – Ranking fakúlt vysokých škôl v SR na základe porovnania ukazovateľov kvantity a kvality vzdelávania a výskumu

In the case all languages were taken into account the hypothesis was confirmed for the following groups of faculties: PRIR and OSTATNE SPOL; the group TECH was just above the significance level. The only groups TEOLOG and FILOZOF had positive correlation. The values of correlation coefficients and their significance are shown in Table 7. Thus, it was confirmed that stronger correlation exists between exact sciences and for one group of social sciences with the ARRA ranking. In case we take into account only faculties for which the number of theses exceeds 20, the group TECH achieved the significance level 0.0144.

Group of faculties (theses in all languages)	Correlation coefficient between the similarity index and ARRA ranking	p significance of the correlation coefficient	Number of faculties
PRIR (natural sciences)	-0.8982	0.0060	7
OSTATNE SPOL (other social sciences)	-0.5712	0.0414	13
TECH (technical sciences)	-0.3774	0.0691	24
PRAV (law sciences)	-0.6407	0.1704	6
TEOLOG (theological sciences)	0.4297	0.3360	7
FILOZOF (philosophical sciences)	0.3159	0.3738	10
PEDAGOG (pedagogical sciences)	-0.2240	0.5624	9
AGRO (agricultural sciences)	-0.2903	0.6356	5
EKONOM (economical sciences)	-0.1371	0.6403	14
MED (medical sciences)	-0.1429	0.7139	9

Table 7. Correlation between similarity index and ARRA ranking (all languages)

In case we took into account the theses written only in the Slovak language, we receive three significant groups PRIR, OSTATNE SPOL and EKONOM and the group TECH was just above the significance level (Table 8). The only groups TEOLOG and FILOZOF had positive correlation. In case we take into account only faculties for which the number of theses exceeds 20, the group TECH achieved the significance level 0.0140. The significance level lower than 0.15 achieved 6 groups.

Group of faculties (theses in Slovak language)	Correlation coefficient between the similarity index and ARRA ranking	p significance of the correlation coefficient	Number of faculties
PRIR (natural sciences)	-0.8961	0.0063	7
OSTATNE SPOL (other social sciences)	-0.6399	0.0250	12
EKONOM (economical sciences)	-0.5696	0.0335	14
TECH (technical sciences)	-0.3789	0.0678	24
PEDAGOG (pedagogical sciences)	-0.5781	0.1030	9
PRAV (law sciences)	-0.6755	0.1409	6
FILOZOF (philosophical sciences)	0.2057	0.5687	10

TEOLOG (theological sciences)	0.2910	0.5758	6
AGRO (agricultural sciences)	-0.2906	0.6352	5
MED (medical sciences)	-0.1292	0.7403	9

Table 8. Correlation between similarity index and ARRA ranking (Slovak language)

6 Conclusions

SK ANTIPLAG contributed to a significant increase in students’ awareness of plagiarism, as confirmed by the international survey where the Slovak students ranked first among all European Union countries. The theses are under multiple supervision: tutor, opponents, examination committee and public. The transparency has risen significantly.

Nationwide and mandatory use of this type of software has become specific for Central Europe and it is unique on a global scale. In the Czech Republic and in Slovenia, there is the use of similar system voluntary. Poland was inspired by Slovak example and embarked on a path similar to ours after the parliamentary and governmental delegation’s visit at our institution in 2011. In 2014, they passed an act requiring HEIs to send bachelor, master and PhD theses to the central repository and subject them to the originality check. The central repository has already started operating (older theses are collected that are not subject to the originality check) and the plagiarism detection system will start its operation in the academic year 2015/2016.

A consistent methodology for the collection of theses and metadata provided Slovakia with a database that can produce outputs for the governing bodies that are unparalleled in the world. Similarity index and related indicators may serve as auxiliary indicators for ranking.

It would contribute to transparency and clarity if we had feedback from HEIs regarding the number of theses suspected of plagiarism and how the cases were resolved. Although the percentage is not large, they are definitely worth noting Table 9). The share of theses with the similarity index exceeding 40% and 60% is around 2.7% and 0.6% respectively.

Thesis type	Percentage of theses exceeding the similarity index value		Number of theses exceeding the similarity index value	
	40%	60%	40%	60%
Bachelor’s	2.87%	0.72%	5 221	1 302
Master’s	2.37%	0.44%	3 755	704
Rigorous	5.95%	1.68%	564	159
PhD	1.05%	0.27%	106	27
Habilitation	1.65%	0.27%	18	3
Number of theses			9 664	2 195

Table 9. Share and number of theses with the similarity index exceeding 40% and 60%

Plagiarism detection systems could play an active role in the elimination of inefficient spending and could save taxpayers’ money in case of appropriate implementation of such a system. US Senator T. Coburn initiated a report⁸, which quantified the opportunities

⁸ The report was prepared by US GAO (US Government Accountability Office)

to reduce duplicity in government programmes. A detailed analysis of federal programs and functional areas with unwanted duplicity, overlapping or fragmentation identified 81 areas. Senator T. Coburn estimated that USD 100 to 200 billion is spent each year on duplicate activities [7].

This paper is to some extent an analogy of the paper “SK ANTIPLAG Is Bearing Fruit” [8], which is focused mainly on analysing the indicator percentage of theses with the similarity index exceeding 25%.

References

- [1] CFE et al. (2013). Privately funded providers of higher education in the UK. Accessible at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/207128/bis-13-900-privately-funded-providers-of-higher-education-in-the-UK.pdf.
- [2] Barta, J. et al. (2014). Hodnotenie fakúlt vysokých škôl 2014 – Ranking fakúlt vysokých škôl v SR na základe porovnania ukazovateľov kvantity a kvality vzdelávania a výskumu. Accessible at: http://arra.sk/sites/arra.sk/files/file/ARRA_Sprava_2014_27_11_2014.pdf.
- [3] European Commission. (2013). European Prize for Innovation in Public Administration. Accessible at: http://europa.eu/rapid/press-release_IP-13-505_en.htm.
- [4] Higher Education Statistics Agency: (2014). Accessible at: <https://www.hesa.ac.uk/component/heicontacts/>.
- [5] Hołdyńska, E. (2014). E-biblioteka zamknięta dla żaków. Accessible at: <http://prawo.rp.pl/artykul/1132128.html>.
- [6] Foltýnek, T. (2013). Plagiarism policies in Slovakia. Accessible at: http://ippheae.eu/images/results/2013_12_pdf/D2-3-26%20SK%20IPPHEAE%20MENDELU%20Survey%20SlovakiaNarrative%20FINAL.pdf.
- [7] Kravjar, J. (2014). The path to an award-winning antiplagiarism system for higher education institutions. Accessible at: http://www.scilogs.com/scientific_and_medical_libraries/the-path-to-an-award-winning-antiplagiarism-system-for-higher-education-institutions/.
- [8] Kravjar, J. (2015). SK ANTIPLAG Is Bearing Fruit. In *PLAGIARISM ACROSS EUROPE AND BEYOND 2015*. Conference Proceedings, pages 147–163. Accessible at: http://plagiarism.pefka.mendelu.cz/files/proceedings_15.pdf.
- [9] PAP (2015). MNiSW: w ogólnopolskim repozytorium już około 0,5 mln prac dyplomowych. Accessible at: <http://naukawpolsce.pap.pl/aktualnosci/news,403712,mnsw-w-ogolnopolskim-repozytorium-juz-okolo-05-mln-prac-dyplomowych.html>.
- [10] PlagiarismAdvice.org. (2013). A National strategy for addressing student plagiarism. Accessible at: <http://www.plagiarismadvice.org/resources/institutional-approaches/item/national-strategy>.
- [11] plagiat.pl (2014). Nowe przepisy antyplagiatowe. Accessible at: <https://www.plagiat.pl/webplagiat/main.action?menu=news&id=1300>.
- [12] Rashid, A. (2012). Academic Policing via top-down implementation of Turnitin in Pakistan: students’ perspective and way forward. Accessible at: <http://www.plagiarismadvice.org/research-papers/item/academic-policing>.
- [13] theses.cz (2014). Vysokoškolské kvalifikační práce. Accessible at: <http://theses.cz/>.
- [14] Weber-Wulff, D. (2014). False Feathers. A Perspective on Academic Plagiarism. Springer-Verlag, Berlin / Heidelberg. 148 pp.

New Tools for Working with the ORAL Series Corpora of Spoken Czech: AchSynku and MluvKonk*

David Lukeš

Faculty of Arts, Charles University in Prague, Czech Republic

Abstract. This paper introduces two simple web-based tools whose aim is to make it easier to work with the ORAL series spontaneous spoken language corpora of the Czech National Corpus. Both strive to overcome and circumvent some of the limitations, either in the data themselves or in their visualization, currently faced by linguists who use them for research. AchSynku is a variant search tool which aims to compensate for the lack of lemmatization in spoken corpora by suggesting, based on a word form input by the user, a list of variant and related forms occurring in the target corpora. MluvKonk is a visualization environment which turns single-line concordances into a multi-tier layout with one speaker per tier. This makes it easier to follow the structure of a multi-party conversation, including turn-switching and overlaps. Though ultimately destined to be superseded by more systemic solutions, both applications are under active development and feedback is welcome, because these ulterior solutions will precisely take advantage of lessons learned in developing and especially using AchSynku and MluvKonk.

1 Introduction

The ORAL series of spoken language corpora compiled at the Institute of the Czech National Corpus (ORAL2006 [4], ORAL2008 [13] and ORAL2013 [2]) contain a wealth of information about spontaneous informal spoken Czech and its usage patterns. Unfortunately, certain aspects of this information are not always easy to extract using the currently available **KonText** corpus manager interface to the data.¹

Part of this is due to the fact that these corpora consist of plain transcripts only, without any additional linguistic annotation (lemmatization, morphological tagging). This makes it difficult to perform truly exhaustive searches spanning the entire paradigm corresponding to a lemma, or morphologically oriented searches (i.e. where the lemma is underspecified).

* This paper resulted from the implementation of the Czech National Corpus project (LM2011023) funded by the Ministry of Education, Youth and Sports of the Czech Republic within the framework of Large Research, Development and Innovation Infrastructures. My thanks also go to both anonymous reviewers for many useful remarks which especially helped shape the discussion around **AchSynku** and contributed to changes in its implementation.

¹ **KonText** is accessible at <https://kontext.korpus.cz/>. Full access to the source data for the ORAL corpora is not currently available nor planned, but special arrangements can be worked out on a case by case basis.

The issue is compounded by the fact that due to their transcription guidelines, the ORAL series corpora contain a much wider selection of transcription variants of a given word form than conventional written language. Indeed, this is part of the reason why they have not been lemmatized or tagged yet, the other part being due to the unruly syntax of unscripted spoken language (false starts, aposiopeses, anacolutha, *apo koinou* constructions etc.; see [1]). Both these aspects make the ORAL corpora speech transcripts difficult to process reliably using extant NLP tools (see [5] for a detailed discussion).

A second problem lies with the structure of spontaneous spoken language itself. While written language is linear², the organization of speech in informal conversations is fundamentally multi-linear or parallel in nature. Several speakers take turns at producing utterances, with overlaps and back-channelling phenomena being very frequent. Humans are well-equipped for disentangling the individual threads of conversation by ear: we are able to collate utterances and speakers based on idiosyncratic voice features, directionality etc., keeping track of the evolution of conversation in real time. By contrast, visually unravelling linearized transcripts of such interactions can be quite challenging, and we need all the help we can get to form an accurate representation of the original communication situation the transcript is supposed to represent. Disciplines such as conversation analysis crucially depend on this intuition for providing correct descriptions of recurring patterns (repair sequences, turn-taking, adjacency pairs etc.) in the structure of dialogues and multi-party interactions.

This paper presents two proof-of-concept tools which attempt at least partially to address the aforementioned issues. **AchSynku** is a variant search tool which serves to compensate for the lack of lemmatization in the ORAL series corpora; **MluvKonk** is a multi-tier visualization interface which offers a more intuitive layout for presenting spoken corpus query hits than the familiar concordance line format used by **KonText**.

2 Variant Search: AchSynku

AchSynku³ (Fig. 1) is a simple tool for maximizing the recall of users' queries in the ORAL series corpora by expanding a single form into the entire paradigm that corresponds to its lemma(s). This is a way of partially compensating for missing lemmatization in these corpora: since the transcripts try to reflect the variability found in informal spoken Czech (e.g. non-standard forms from Common Czech and regional dialects) and do nothing to regularize its often fragmentary syntax and unclear sentence boundaries,⁴ the results of applying off-the-shelf taggers and lemmatizers (trained on and/or designed for

² It is true that this linearity can be broken by typographical means using e.g. text boxes in freeform layouts, but crucially, the *processing* of these individual items remains linear and single-threaded.

³ The name is a pun on the folk song "Ach synku, synku" whose lyrics conveniently contain words that remind one both of the CNC's flagship SYN series of written corpora, which are lemmatized and tagged, and of the ORAL series corpora ("Oral jsem, oral, ale málo"), which are not.

⁴ See [6, Chap. 2] for an argument completely rejecting sentences as a unit of analysis in spoken language, in favour of loosely coupled clause complexes.

standard written language) to spoken corpora are relatively poor, and have therefore not been included in the current official releases of the ORAL series corpora.

In the background, AchSynku uses data derived from a currently non-public experimentally lemmatized version of these corpora (see [5]), which is continuously being improved, to search for related word forms. In doing so, it prioritizes:

- **recall** over **precision**, on the assumption that the user can easily remove unwanted variants from the returned list, but would find the reverse process (tracking down obscure variants that were left out) much harder
- **ease-of-use** over **configurability**

Vyhledávač variant v korpusech řady ORAL

Pokud si nejste jisti, v jakých variantách by se vámi hledaný tvar mohl v mluvených korpusech vyskytovat, můžete si pravděpodobné kandidáty najít pomocí tohoto formuláře. Výsledek se zobrazí rovnou v podobě CQL dotazu, který lze zadat do korpusového manažeru KonText. Zkuste si např. vyhledat, v jakých zápisových variantách se v korpusech řady ORAL vyskytuje slovo *protože*.

Tvar, (spisovné) lemma

Vyhledat varianty

Fig. 1. Initial appearance of the AchSynku query form

The upside is that any word form occurring in the ORAL series corpora, but also any corresponding lemma,⁵ even though it might not directly occur in the corpora in this form, can be used to seed the variant search, irrespective of letter case. The drawback is that abiding by these tenets occasionally causes several lemmas to be conflated in the search results: if a user enters the query *moc*, there is no way to determine whether s/he means the infinitive of the verb *moci/moct* (*can, to be able to*), or one of several homonymous cases of the noun *moc* (*power*), so both paradigms are returned. It could be argued that *moci* is a literary variant of the infinitive *moct*, seldom used in informal speech, and should therefore not trigger an association with the verb in spoken language contexts. However, this decision is largely a matter of a given linguist's point of view and theoretical stance; if such restrictions were incorporated into AchSynku, they would decrease recall based on implicit assumptions about the users' intentions, which is contrary to the goals stated above. A possible solution for some of these cases where different parts of speech are involved is sketched out in Sec. 2.3.

⁵ If it was correctly identified during lemmatization, of course.

is editable. Finally, three buttons provide direct links into the **KonText** corpus manager, to concordances generated from the three ORAL series corpora based on the current form of the query in the result text area.

2.2 Implementation

Internally, **AchSynku** uses an SQLite database with a `word2lemma` table (see Tab. 1) listing forms occurring in the ORAL series corpora and their corresponding lemmas as attributed by the experimental lemmatization described in [5]. In general, SQL databases are not the best solution for storing (word, lemma) correspondences; finite state automata give better performance in a fraction of the storage. For instance, the open-source MorphoDiTa POS-tagging framework [12], which provides a lean and efficient implementation of the FSA approach, could have been used instead of SQLite. However, given the relatively small size of the table involved,⁶ it was deemed counter-productive to make the setup more complex by adding a highly specialized dependency instead of the general-purpose and fairly ubiquitous SQLite.

word	lemma
...	...
ale	ale
Ale	ale
ále	ale
...	...
Honzoj	Honza
...	...

Table 1. The `word2lemma` SQL table used by **AchSynku**. Each row corresponds to a unique word form found in the ORAL series corpora, along with its lemma.

The string i entered as input by the user is lowercased and matched against the lowercase versions of the known lemmas and word forms. All word forms x are returned that satisfy either of the following properties ($lemma()$ maps between a word form and the set of lemmas that can be attributed to it; $lc()$ returns the lowercase version of its input):

- $lc(i) \in lc(lemma(x))$ (i is one of the lemmas of x)
- $|lc(lemma(i)) \cap lc(lemma(x))| > 0$ (i and x belong to a shared lemma)

⁶ Entries are generated only for (word, lemma) pairs actually seen in the corpus.

In SQL terms, this corresponds to the following query on the `word2lemma` table (where `$query_string` is the lowercased version of the input string entered by the user):

```
SELECT DISTINCT word
FROM word2lemma
WHERE lemma IN
  (SELECT '$query_string'
   UNION SELECT lemma
   FROM word2lemma
   WHERE word = '$query_string');
```

Both the `word` and `lemma` columns are initialized with the `COLLATE NOCASE` option, so that matching against them is automatically performed in a case-insensitive manner. Since the vanilla SQLite engine only implements case folding for characters in the ASCII range, the database tables and the user queries (both UTF-8-encoded) are preprocessed using Unicode Normalization Form Canonical Decomposition (NFD), which decouples base characters from combining marks. In the case of strings representing Czech words, this means we are left with plain ASCII letters (which SQLite can casefold easily), interspersed with combining diacritics (which are the same irrespective of case and SQLite leaves them alone).⁷

2.3 Discussion and Future Plans

A sensible objection to the existence of AchSynku would be, why not add this information directly into the ORAL series corpora as a positional attribute? Since the underlying corpus query engine Manatee [9] supports multi-valued attributes, the lists of related word forms could be precomputed for each position and directly queried. The main reasons why this road was not taken were the following:

- for some word forms, these lists are fairly long and repeating them at each occurrence of the given word form would be wasteful and inelegant
- the reliability of the information provided by AchSynku is still experimental and not including it in an official release of the corpora is a warning to users to be careful when working with it
- in this regard, the fact that MluvKonk allows users to manually review the list of proposed related words and exclude some of them at their own discretion is another considerable advantage

We are currently looking into ways of optionally supporting more targeted queries which would help resolve issues such as the verb vs. noun ambiguity of the input query

⁷ I am indebted to one of the anonymous reviewers of the paper for suggesting this elegant solution; in a previous implementation, pre-computed lowercase counterparts were explicitly stored in the database.

moci mentioned earlier. Still, much circumspection is needed in this respect, because we are dealing with a double-edged sword: we have to keep in mind that the user already had to learn to operate the elaborate query interface of the **KonText** corpus manager; in contrast, **AchSynku** should be a light intuitive extension with a low entry barrier. In particular, it is pointless to try to simulate externally all the functionality available in a corpus with linguistic annotations (lemmas, tags), in view of the fact that natively tagged versions of the ORAL series corpora are planned (see below).

One way to refine the variant lists returned by **AchSynku** would be to include an optional part of speech specification field in the query form. This would result in an interface which would be only slightly more complex but potentially much more powerful. The underlying implementation being simply a matter of extending the `word2lemma` table and adapting the SQL query, the only difficulty is a concise explanation of the use cases for this feature without burdening the user with implementation details.

A large corpus encompassing all three ORAL series corpora plus possibly some additional data is planned for publication by the CNC in the foreseeable future. Hopefully, it will already contain lemmatization and tagging by default, making tools like **AchSynku** obsolete. However, since the original (unannotated) ORAL series corpora will remain accessible for reference purposes, it makes sense for the application to remain operational well past that point, and even profit from the increases in lemmatization accuracy achieved during the processing of the aggregated corpus.

3 Multi-tier Visualization: **MluvKonk**

The classical format for displaying corpus concordances is one line per hit with the KWIC highlighted. This format is eminently suited for investigating written texts, because it allows easy direct comparisons of key words and their contexts, as they are vertically lined up; it is much less convenient for speech transcripts, where different speakers take turns producing utterances and overlaps can occur. The **KonText** corpus manager allows structural marks to be displayed in the concordance (see Fig. 3) which mark the turn-switching boundaries and indicate which parts of an utterance take place within an overlap. It is thus possible to recover the original structure of the dialogue (as interpreted by the transcriber), but doing it manually is error-prone and laborious. Yet perusing the concordance in the default format offered by **KonText** prevents the linguist from easily spotting recurring structural patterns which would be obvious at first glance in a multi-tier layout where each speaker has his/her own layer of text and overlaps can thus be represented truly in parallel.

This is precisely the problem **MluvKonk** sets out to solve. Taking inspiration from well-established tools such as ELAN [11] (and its Annotation Mode), EXMARaLDA [10] (and its Partitur notation), Praat [3] (and its TextGrid tiers), or the ANNIS corpus manager and query engine [14] (and its grid visualization), it takes a concordance exported from **KonText** in the `.csv` format and displays it in a layout similar to that offered by the aforementioned programs, where the utterances by each speaker are confined to his or her own visually separated horizontal tier (Fig. 4).

```

y že jo ale . ale ... ] </sp><sp num=00 překryv=ne> + [ tak kočka už to zalomila ta tam . chrupe na . ] + [ Zuzce na bafohu (smí
at hajnýho . akorát mi není jasný ] + [ proč z tý boudy . čučí ta kočka ] + [ proč třeba ta kočičí hlava ] </sp><sp num=01 překryv=
kryv=ano> + [ protože ] </sp><sp num=03 překryv=ano> [ kočka ] </sp><sp num=01 překryv=ne> + [ dyž sme došli k vysílač
0 překryv=ne> + [ ne on je to takové naše zlatíčko . ] + [ aji kočka ona je skvělá ] + [ protože ona už ví ] + [ že ně něho nesmí
ne> + [ mmm ] </sp><sp num=00 překryv=ne> + [ ... a ta kočka porodila kořata . ] + [ samozřejmě . v te stodole tam viš co . k
e> + [ hmm jako je obyčejna viš co uplně obyčejna mourovata kočka jako .. ] + [ my máme právě eee uplně strašně podobná jakou
n) jak sem tam ] + [ byla na přimačkách ] + [ (se smíchem) kočka . ] + [ fajna fak ] </sp><sp num=01 překryv=ne> + [ já vim
avda ] </sp><sp num=01 překryv=ano> [ smečky . zatímco kočka ] </sp><sp num=01 překryv=ne> + [ v žádným případě takž
ne no ] </sp><sp num=00 překryv=ne> + [ jako . je fakt že kočka nikdy . tě takhle neuzná jak ten pes ] + [ jo že ten pes jako fak
stavit dycky ] </sp><sp num=00 překryv=ano> + [ ne jako kočka ] </sp><sp num=01 překryv=ano> [ mmm ] </sp><sp num
kryv=ano> + [ mmm ] </sp><sp num=01 překryv=ano> [ kočka ] </sp><sp num=01 překryv=ne> + [ ale to je pravda ] </sp>
/sp><sp num=01 překryv=ne> + [ pověřit nebo tak ] + [ ta kočka si to přece jenom najde sama nějakou ] </sp><sp num=00 př
ekryv=ne> + [ nevím kdo by to venčil .. ] + [ takže . viš co ta kočka to normálně pustíš a ona de sama ] </sp><sp num=00 překry
=ne> + [ hmm . ty ] + [ NJ ] + [ a jak že se menuje ta vaše kočka u té . vrátnice ? ] </sp><sp num=00 překryv=ne> + [ (se sm
řít pat ty otázky ve chvíli ] + [ kdy jim jako přes cestu přeběhla kočka ] </sp><sp num=01 překryv=ano> + [ v tom rozhovoru ] </
to třeba zdráhala říc ale ] + [ vona tam byla zrovna náká nová kočka . ] </sp><sp num=01 překryv=ano> + [ a já sem říka* . ] </
mově sem čekal deset minut . vzala mě další holka ] + [ zase kočka . a hodila mě do Tábora . a vona jela na Sezimák ] </sp><sp nu
kryv=ne> + [ no máho kolegu jednou škrábla . dyž byl . malej . kočka . ] + [ a von se vůbec se nepřiblíží ke kočkám . já sem ti dala ř
i> [ škaredě no ] </sp><sp num=05 překryv=ano> + [ to je kočka ? ] </sp><sp num=02 překryv=ano> [ no je ] </sp><sp num
e maminka ] </sp><sp num=00 překryv=ano> [ jo ? . hmm . kočka ] </sp><sp num=01 překryv=ne> + [ to si ty víš ? .. v tom kr

```

Fig. 3. A classical KWIC visualization of a concordance from the ORAL2013 corpus in the **Kon-Text** corpus manager. The `<sp> . . . </sp>` structures indicate stretches of text uttered by a given speaker; the `num` attribute uniquely identifies the speaker of the corresponding stretch, and the `překryv` attribute indicates whether the stretch was uttered in overlap with another stretch ("ano" for yes, "ne" for no).

As in the case of *AchSynku*, bug reports, comments and feature requests can be filed via **MluvKonk**'s GitHub repository: <https://github.com/dlukes/mluvkonk>.

3.1 Interface

The interface to **MluvKonk** consists of a series of tabs (see Fig. 5). The main tab is the *Concordance* tab, which allows users to upload their own concordances. It also presents two sliders which allow the user to page through the output and set the number of hits to display per page.

Concordance hits are displayed along with any available meta-information, such as speaker gender, education, age or region of origin, which can also be accessed on the *Statistics* tab (see Sec. 3.2 and Fig. 6). Finally, a *Help* tab describes the use cases and usage guidelines for **MluvKonk**.

??	druhý deň bylo už zima								
00				no			no ta je hnusná		
01		(smích)							
02				no je					
04	no ne no to je jiná tedka fakt je tak škaredě no			to je na* bezsrstá ?			to je šíle		
05				to je kočka ?			ta je hrozná		
??	je babička . na fotce ?								
00			jo ? . hmm . kočka				to ti to slušelo --		
01			to je maminka no . to je maminka		to si ty vid' ? .. v tom kroji				
03							to mi bylo patná		
Y		(odmlčení)							
00	hmm	hmm	krása ..	sluší ti to tam .		taková kočka seš (smích)			
01	každej den ? .	každej den		je	no ..		prej seš		
03									
Y		(odmlčení)							

Fig. 4. A multi-tier visualization of a concordance from the ORAL2013 corpus using the **MluvKonk** tool. Utterances by individual speakers are confined to separate layers identified by the speakers' numbers. ?? indicates that the speaker number for the given utterance could not be determined based on the concordance as exported from **KonText**; this is a limitation stemming from the context window returned by the Manatee query engine [9] underlying **KonText**, not from **MluvKonk** itself.

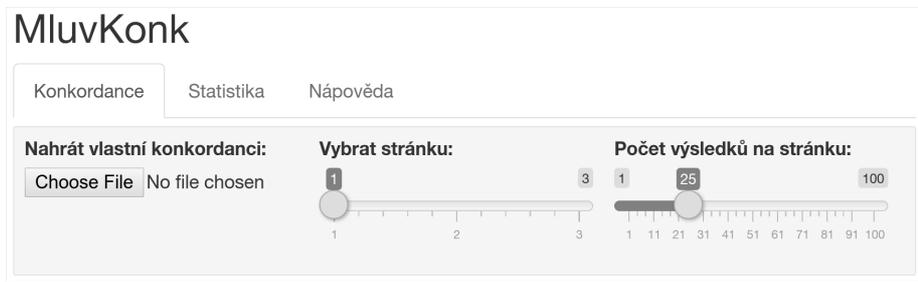


Fig. 5. **MluvKonk**'s interface, showing the three tabs *Concordance*, *Statistics* and *Help*. See text for more details.

3.2 Implementation

The CNC exposes no publicly accessible and documented API which would make it easy to programmatically access its corpus data. We therefore opted for a workflow in which a concordance is first produced using **KonText**, exported in `.csv` format and then uploaded into **MluvKonk**.⁸ A 5 MB size limit is imposed by the server on this uploaded file, because **MluvKonk** is still in beta status and might behave unpredictably on too large inputs. This

⁸ The requirements for producing a valid `.csv` export which can be read back in by **MluvKonk** are specified in detail on the *Help* tab of the application.

should not be an appreciable limitation in practice, because its purpose is to aid manual inspection of concordances, and there are only so many hits one can go through by hand.

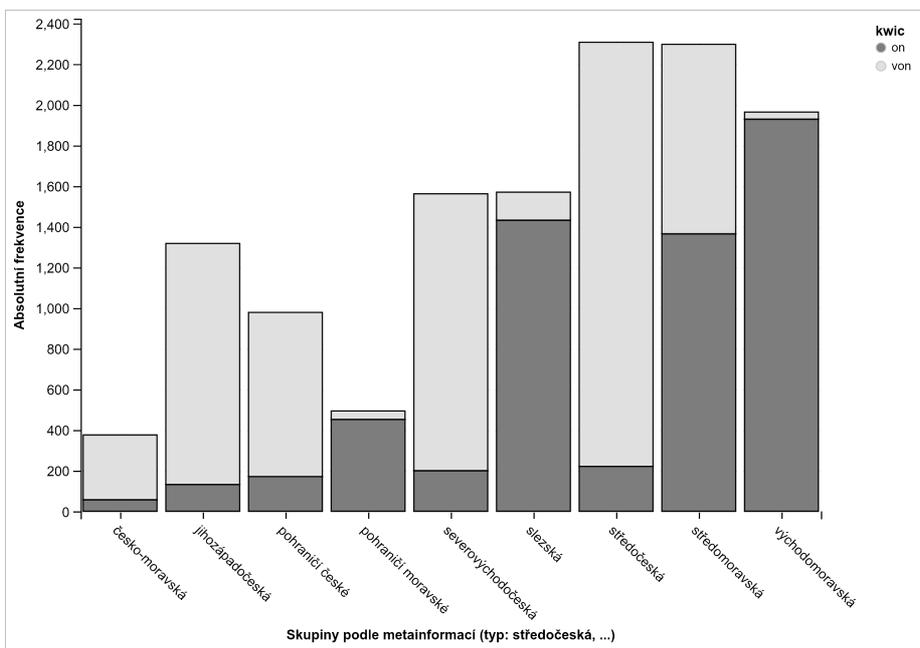


Fig. 6. A bar chart generated using **MluvKonk**'s *Statistics* tab, displaying the distribution of variants *von* and *on* (*he*) according to the speaker region of origin metadata of the concordance hits. The y-axis reads "Absolute frequency" and the categories on the x-axis are (from left to right): Bohemian-Moravian transient region, South-West Bohemia, Czech Borderlands, Moravian Borderlands, North-East Bohemia, Silesia, Central Bohemia, Central Moravia, East Moravia. As expected, variants with [v]-prothesis dominate in Bohemia, and variants without it in Moravia and Silesia (Central Moravia stands somewhere in between).

MluvKonk's backend takes advantage of the Shiny web application framework [8]. Shiny consists of a custom webserver and a set of libraries which make it easy to design, code and deploy data-driven web applications using the R language for statistical computing [7]. The Shiny framework itself takes care of making all requests for data and computation asynchronous, which results in a single-page application that feels responsive and fast to use. The backend is optimized to render only the part of the concordance (i.e. the page) which is currently being viewed by the user; the rendering itself consists simply of inserting the utterances into HTML tables with one row per speaker, with empty cells representing stretches of time where a speaker is silent.

Currently, the *Statistics* tab serves only as a modest showcase for the powerful capabilities of Shiny in terms of generating data-driven graphics. It enables the user to create

distributional bar charts on the fly based on any of the metadata available in the concordance (see Fig. 6 for an example).

3.3 Future Plans

As mentioned in Sec. 3.2, there is no official public API to the CNC's corpora. However, undocumented ways of accessing their backend (with appropriate permissions) do exist. It is therefore possible at least in theory to eliminate the slightly cumbersome necessity of going through the export step in **KonText** and reimporting the data into **MluvKonk**. Nevertheless, in practice, a simple search box would often prove insufficient, and it would be redundant to replicate all the advanced search functionality available in **KonText**. Additionally, **MluvKonk** is currently not optimized for handling fairly large concordances of several tens of thousands of entries, and the file upload size limit mentioned in Sec. 3.2 offers a convenient way to enforce an upper boundary.

A much better approach is to go the other way round and integrate **MluvKonk**'s multi-tier visualization as a display option in **KonText**. This functionality should really be part of the main corpus manager anyway, because it offers ways of interfacing to the data and analyzing it which are clearly sorely missed at present; it is only a question of time and resources until it is implemented. Once it is, **MluvKonk** will have served its purpose as a testing grounds both for the technical aspects of generating such a concordance visualization based on existing corpus data, and for the users' requirements in working with it.

4 Conclusion

This paper presented two new tools for working with the CNC's ORAL series corpora. The user interfaces to both applications are Czech-only, but this should not be a problem, since the corpora which they augment are themselves only really useful to Czech-speaking researchers. In many ways, these are only stopgap solutions which aim to provide temporary relief with respect to the specificities of spoken language corpora. In the case of **AchSynku**, this consists in providing an easier way of querying entire paradigms and variant sets at once⁹ in corpora which currently do not feature any lemmatization. Once spoken language corpora are natively lemmatized and tagged, this helper tool will not be necessary any more. Similarly, once a multi-tier visualization giving an intuitive overview of the structure of the conversation is implemented as a concordance display option in **KonText**, there will be no need for **MluvKonk** either. In the meantime, however, they will serve as useful testing grounds in their respective areas, and hopefully even facilitate actual research. This is why, in spite of their temporary nature, feedback is more than welcome in both cases (especially for **MluvKonk**, as lessons learned with it will lead to better results once the functionality is integrated into **KonText**).

⁹ This includes variants that the user might not even be aware of, because they are specific to the transcription guidelines, e.g. the following 13 variants of the word *protože* (because): *prtože, prže, pže, přče, protoe, prože, proe, poče, ptože, potože, prtže, protže, přže, prtoe*.

Additionally, both tools are open-sourced under the terms of the GNU General Public License v3,¹⁰ which means that they can be freely modified or integrated into projects adhering to the same licensing terms. In practice though, since these are enhancements bolted *post hoc* on existing corpus infrastructure, it would make more sense for new projects which require the type of functionality provided by AchSynku and MluvKonk to plan their infrastructure accordingly from the outset so that it includes these features natively.

References

- [1] Auer, P. (2009). On-line syntax: Thoughts on the temporality of spoken language. *Language Sciences*, 31:1–13.
- [2] Benešová, L., Křen, M., and Waclawičová, M. (2013). *ORAL2013: reprezentativní korpus neformální mluvené češtiny [ORAL2013: A Representative Corpus of Informal Spoken Czech]*. Praha.
- [3] Boersma, P. and Weenink, D. (2015). Praat: doing phonetics by computer. v5.4.08.
- [4] Kopřivová, M. and Waclawičová, M. (2006). *ORAL2006: korpus neformální mluvené češtiny*. Ústav Českého národního korpusu FF UK, Praha.
- [5] Lukeš, D., Klimešová, P., Komrsková, Z., and Kopřivová, M. (2015). Experimental tagging of the ORAL series corpora: Insights on using a stochastic tagger. In Král, P. and Matoušek, V., editors, *TSD 2015, LNAI 9302*, pages 388–396.
- [6] Miller, J. and Weinert, R. (1998). *Spontaneous Spoken Language: Syntax and Discourse*. Clarendon Press, Oxford.
- [7] R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Wien.
- [8] RStudio, Inc (2014). *Easy web applications in R*. URL: <http://www.rstudio.com/shiny/>.
- [9] 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.
- [10] Schmidt, T. (2004). Transcribing and annotating spoken language with EXMARaLDA. In *Proceedings of the LREC-Workshop on XML based richly annotated corpora, Lisbon 2004*. ELRA, Paris.
- [11] Sloetjes, H. and Wittenburg, P. (2008). Annotation by category—ELAN and ISO DCR. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, pages 816–820.
- [12] Straková, J., Straka, M., and Hajič, J. (2014). Open-source tools for morphology, lemmatization, POS tagging and named entity recognition. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 13–18, Stroudsburg, PA, USA. Johns Hopkins University, Baltimore, MD, USA, Association for Computational Linguistics.
- [13] Waclawičová, M. and Křen, M. (2008). ORAL2008: New balanced corpus of spoken Czech. In *Proceedings of the International Conference “Corpus Linguistics–2008”*, pages 105–112. Saint Petersburg University Press, Saint Petersburg.
- [14] Zeldes, A., Ritz, J., Lüdeling, A., and Chiarcos, C. (2009). ANNIS: A search tool for multi-layer annotated corpora. In *Proceedings of Corpus Linguistics 2009*, pages 20–23, Liverpool.

¹⁰ See <http://www.gnu.org/licenses/gpl-3.0.en.html>.

Traitements automatiques en lexicographie de langues « non dotées »

Patrice Pognan¹ and Miloud Taïfi²

¹Institut National des Langues et Civilisations Orientales, Université de Paris, Sorbonne, France

²Université Sid Mohamed ben Abdellah, Fès, Maroc

Résumé. Une coopération franco-marocaine, largement soutenue par les institutions, a permis d'élaborer non seulement le dictionnaire le plus exhaustif du berbère du Maroc, mais aussi de mettre en place un environnement propice au traitement automatique du berbère, notamment en matière de lexicographie. Cette expérience sert de départ au projet de dictionnaire de la santé cilubà – français, langue encore moins « dotée » que le berbère.

1 Introduction

Le berbère, langue importante du Maghreb, est une langue d'autant moins dotée informatiquement qu'elle comporte de l'ordre de quatorze parlers jouissant de positions très variées face à l'écrit, à la formation de la langue et, à plus forte raison, face à une éventuelle informatisation de la société.

De manière générale, le kabyle pour l'Algérie et le chleuh pour le Maroc sont les mieux placés vis-à-vis d'une production écrite et de l'intérêt pour la langue et sa transmission, orale, écrite ou informatisée.

Nous travaillons sur une troisième variété de berbère, également importante par le nombre de locuteurs et l'étendue du territoire sur lequel elle est parlée: le berbère du Maroc central. A la différence des deux précédentes, les travaux qui la concernent sont souvent anciens et nullement concernés par une quelconque automatisation.

Nous allons présenter ici les travaux réalisés pour la constitution du Dictionnaire raisonné berbère – français. Parlers du Maroc [17], les applications qui en découlent et la mise en place d'un environnement de travail informatisé qui constitue simultanément un premier fonds documentaire sous forme d'un corpus dictionnaire et le cadre adéquat à de nouveaux travaux en évitant nos erreurs passées.

Cette expérience sert à l'élaboration d'un projet similaire sur le cilubà (écrit auparavant *tshiluba*), l'une des langues de la République Démocratique du Congo, prévoyant la réalisation d'un « dictionnaire de la santé cilubà – français ». Ce projet sera présenté en conclusion du présent article.

La constitution d'un dictionnaire de référence pour le berbère du Maroc central (tamazight) est un travail de très grande envergure. Miloud Taïfi avait déjà publié en 1991 le Dictionnaire Tamazight–Français (parlers du Maroc central) [15]. Le présent dictionnaire raisonné berbère – français. Parlers du Maroc [17] n'est pas uniquement une version corrigée, augmentée et remaniée de l'ouvrage précédent, mais c'est aussi et surtout une vision très largement novatrice du système de la langue berbère. L'énorme travail d'élaboration du système linguistique et de constitution d'une écriture systémique a pris des années, soutenu par une coopération constante depuis 2004¹.

¹ Action intégrée Volubilis « Dictionnaire bilingue français-berbère. Bases de données lexicales et vocabulaires d'apprentissage », 2004–2007.

2 Transformation du dictionnaire en un corpus structuré

Les difficultés ont été immédiates dans la mesure où nous souhaitions conserver l'acquis du dictionnaire de 1991 devenu la référence pour le berbère du Maroc central. Ce dictionnaire, par des concours particuliers de circonstances, n'était pas imprimé, mais très méticuleusement calligraphié. Il a été impossible de réaliser une saisie optique avec les logiciels de l'époque, la machine repérant de nombreuses distorsions de l'écriture manuelle que l'œil d'un humain non spécialiste de la typographie ne pouvait voir.

La conséquence est que le contenu de ce premier dictionnaire a été saisi par de nombreuses petites mains dans 29 fichiers Word (un par lettre de l'alphabet). Il en a résulté un nombre assez élevé d'erreurs aléatoires éliminées progressivement par plusieurs relecteurs et surtout par l'application des programmes de traitement.

C'est donc à partir de cet ensemble de fichiers que nous avons élaboré les traitements successifs.

2.1 Du dictionnaire au corpus

Nous avons transformé les 29 fichiers Word du dictionnaire en fichiers textes bruts codés en Unicode UTF-8 que nous avons fusionnés en un corpus relativement important de 2 700 000 caractères.

A partir de ce corpus, une série de programmes produit une forme structurée du corpus comprenant les structures inscrites implicitement dans le dictionnaire grâce à l'usage de symboles spécifiques, augmenté de valeurs grammaticales calculées automatiquement.

2.2 Du corpus brut au corpus structuré

Le traitement est divisé en deux modules. Le premier a pour but de reconnaître les structures existantes du dictionnaire et le second d'enrichir ces structures de connaissances de nature morphologique à partir des indications minimales qui sont données dans le dictionnaire.

Le **premier module est divisé en 5 programmes** (en Python) qui s'enchaînent. Au départ, nous avons un texte brut en UTF-8 tel que l'exemple ci-dessous :

-
- FSP France – Maghreb « Pratiques langagières au Maghreb: corpus et applications », 2006–2009.
 - contrat AUF-LTT « Bases de données pour l'étude grammaticale et lexicale des langues dans une visée multilingue », 2009–2010.
 - PHC Volubilis « Lexicographie différentielle des langues: usage au Maroc et bases de données », 2010–2013.

BRYD

◆ aberyuḍ (u)

iberyaḍ ▶ boue, bourbe, fange. (v. aussi : aḥeryuḍ, ḥrḍ ; aluḍ, lḍ). • ituttey g uberyuḍ, il est tombé dans la boue.

BRZ

◆ brez

brez, tebraz, ur-briz ▶ razzier, piller. • berzen yiqeṭṭaen akabar, les brigands ont pillé la caravane.

◆ Tu — tubrez

tubrez, ttubraz, ur-tubriz ▶ être razié, pillé.

◆ M — mebraz

mebraz, tmebraz, ur-mebraz ▶ se soumettre réciproq. à des razzias, piller l'un l'autre. • ur da tmebraznt teqbilin igan taḍa, les tribus liées par les engagements de « taḍa » ne se soumettent pas mutuel. à des razzias.

◆ abraz (u)

ibrazn ▶ razzia, pillage ; butin.

Les structures du dictionnaire sont reconnues à l'envers en partant de la structure la plus profonde: les exemples.

Le second programme traite le niveau des significations. Il extrait et ordonne les éléments de synonymie.

Le troisième programme isole les différentes racines en les séparant les unes des autres par une ligne vide et met en évidence les différents mots (entrées) relevant de la racine en question.

Le quatrième programme traite le niveau du mot où il regroupe toutes les informations morphologiques et met entre crochets les remarques à la fin de la ligne, p. ex. [même racine que la précédente ?].

Le cinquième programme ne traite que les indications d'occurrence d'une racine dans d'autres langues, par exemple en kabyle et/ou en arabe.

La sortie du premier module se présente ainsi:

BRYD

◆ aberyuḍ (u), iberyaḍ

▶ boue, bourbe, fange

■ aḥeryuḍ, ḥrḍ

■ aluḍ, lḍ

• ituttey g uberyuḍ, il est tombé dans la boue

BRZ

◆ brez, brez, tebraz, ur-briz

▶ razzier, piller

• berzen yiqeṭṭaen akabar, les brigands ont pillé la caravane

◆ Tu — tubrez, tubrez, ttubraz, ur-tubriz

▶ être razié, pillé

◆ M — mebraz, mebraz, tmebraz, ur-mebraz

▶ se soumettre réciproq. à des razzias, piller l'un l'autre

• ur da tmebraznt teqbilin igan taḍa, les tribus liées par les engagements de « taḍa » ne se soumettent pas mutuel. à des razzias

◆ abraz (u), ibrazn

▶ razzia, pillage ; butin

Le **second module** est déterminant pour pouvoir créer les applications issues du dictionnaire (en particulier, lexiques associés au dictionnaire, base de données

correspondant au dictionnaire, exemplier, ...). Il est composé de trois scripts qui calculent des informations supplémentaires.

Le premier programme numérote les racines pour distinguer les nombreuses formes homographes.

Le deuxième programme est la clé de toute l'entreprise. Sans une analyse exacte des catégories lexicales, les projets ultérieurs auraient été irréalisables. Son but est double. Il doit d'une part pouvoir assurer un classement par racines et un classement par mots et d'autre part, enrichir toutes les informations morphologiques.

Le programme détermine les catégories lexicales – verbes simples (1.1.), verbes dérivés (1.2.) et nominaux (substantifs et adjectifs) issus d'un verbe (1.3.) ou non (2.). Il calcule toutes les formes d'annexion sur la base des formes libres.

Le troisième programme traite les exemples au sein du corpus. Il a été remplacé par le programme spécifique de construction de l'exemplier Taïfi.

Voici une illustration du corpus avant traitement des exemples:

ŞBH	ŞBH2								
ur-şbiḥ	◆ 1	şbiḥ	ŞBH2	1.1.	VN	V	şbiḥ	şbiḥ	teşbiḥ
		▶ être beau, joli ; être agréable, charmant							
		■ fulki, flk							
		■ ğuda, ğd							
		■ izill, zl							
		● ur teşbiḥ illi-s, sa fille n'est pas belle							
wuṣbiḥn	◆ 2	uṣbiḥ	ŞBH2	1.3.	VN	ADJ	uṣbiḥ	wuṣbiḥ	uṣbiḥn
tuṣbiḥt		tuṣbiḥt	tuṣbiḥin	tuṣbiḥin					
		▶ beau, joli ; agréable, charmant							

3 Lexiques annexes pour le dictionnaire

Le dictionnaire étant classé par racine suivant la tradition des études sémitiques, il a semblé opportun de proposer au lecteur la possibilité de circuler plus aisément au sein du dictionnaire par l'ajout d'un index « mot berbère – racine berbère », la détermination de la racine à partir du mot n'étant pas toujours aisée en raison d'alternances qui varient suivant les parlers. Nous avons aussi souhaité, en attendant la réalisation du Taïfi français – berbère [17], permettre l'utilisation du dictionnaire par l'intermédiaire du français grâce à un index français – racine berbère.

Les calculs effectués peuvent être aisément montrés par l'extrait ci-dessous :

BDR

◆ budr

budr, tbudur, ur-budr ▶ se maquiller, se farder ; se faire beau ; soigner son apparence. (v. aussi : helleq, hlq ; herges, hrgs). ● la ttbudur hma ad teddu ġer tmeġra, elle se maquille, elle se fait belle pour aller assister à la fête du mariage.

◆ S — sbudr

La liaison entre la racine (en rouge) et le mot (en bleu) va permettre de construire le premier lexique par l'intermédiaire de 15 tâches différentes. La liaison entre la racine en

rouge et les équivalents français en vert va permettre de créer le lexique français vers les racines berbères, ce qui nécessite 11 programmes et deux étapes manuelles.

3.1 Lexique mots berbères – racines berbères

Voici un extrait de ce lexique:

la: L ³	laližu: LŽ ¹	lbal: BL ¹¹	lbers: BRS ²
la: L ⁴	lalkul: LKL ²	lbaladiya: BLDY	lbesbas: BS ⁴
la ...la: L ⁴	lall: L ¹³	lbalaj: BLĠ ¹	lbet: LBT
labas: BS ³	lalla: L ¹³	lbaliza: BLZ ²	lbetta: BṬ ³

3.2 Lexique français – racines berbères

En voici un court extrait:

abreuvoir: SRŽ ² , ŠRŽ ³	abuser: ĠDR, XDĖ
abri: DRG, DRY ³ , NFY ² , NTL ¹ , SFL ² , SNFY ~ à bétail: NWL ² ~ pour se protéger: ŽY ¹ ~ sous roche: FR ¹⁰ se mettre à l'~: DRY ³ tenir à l'~ des regards: HŽB	~ de la confiance de: Nfq ² ~ de sa force: ŽWR ² acacia: D ²¹ ~ jaune: MRD ³ accablé: LF ³ , NBR ¹ ~ de malheurs: HDŠ ~ de soucis: ĠF ~ par les fortes chaleurs de l'été: GLF ³
abricot: MŠ ³	
abricotier: MŠ ⁸	

Ce type de lexique est important parce qu'il représente déjà un embryon de dictionnaire français – berbère ou, au moins, un substitut de ce dictionnaire. Il est intéressant de constater que la demande d'un moyen de consulter le dictionnaire par l'intermédiaire du français est une demande faite par les Marocains.

4 Extraction de l'exemplier

Les locutions et les exemples sont extraits à partir du dictionnaire en un fichier de plus de 13600 enregistrements. A partir de chaque enregistrement, nous isolons l'exemple en berbère, une éventuelle traduction littérale explicative de la forme berbère et la traduction française. Nous nous trouvons en présence d'un séparateur ambigu, la virgule, qui peut apparaître un nombre de fois quelconque dans n'importe lequel des trois segments. C'est pourquoi nous avons créé une procédure particulière qui intervient au sein du programme lorsque sont épuisées les possibilités de calcul à partir de ponctuations fiables. Elle découpe le texte en segments limités par les virgules, elle marque le premier segment en tant que segment berbère et le dernier en tant que segment français. Elle analyse le contenu de chacun des autres segments à l'aide de caractères spécifiques à chaque langue (« à'èèèîôœù'çocjpv » pour le français et « čšžđghřtze » pour le berbère) et de quelques mots-outils très courts tels que prépositions, conjonctions,...

Le fichier produit par le programme d'analyse des locutions et exemples est chargé dans une base de données en Access. L'interface utilisateur donne accès aux différents champs créés (expression berbère, traduction littérale, correspondant français, l'exemple d'origine en entier, mais aussi le type d'expression berbère: locution, expression, chant ... ou exemple).

628

EXEMPLIER TAIFI

locution

berbère
tasa n ufullus

traduction littérale
foie de poulet

français
craintif, peureux, couard

— **tasa n ufullus (foie de poulet), craintif, peureux, couard.**

Une interface de recherche permet de trouver un mot (simple, préfixé, suffixé, préfixe et suffixé) dans l'expression berbère, dans la traduction littérale ou dans l'expression française correspondante. Les résultats des recherches sont exportés et conservés au format .html.

EXEMPLIER					
rang	berbère	traduction littérale	français	phrase	type
1226	iggudy bna dem assa g ssuq		il y a beaucoup de monde aujourd'hui au souk	• iggudy bna dem assa g ssuq, il y a beaucoup de monde aujourd'hui au souk	exemple
1445	nzan ihuliy n assa g ssuq ba!l		les moutons étaient bon marché au souk aujourd'hui	• nzan ihuliy n assa g ssuq ba!l, les moutons étaient bon marché au souk aujourd'hui	exemple
1089	la tmedfaen midden assa g ssuq		les gens se bouscuaient aujourd'hui au marché, il y avait trop de gens au marché aujourd'hui	• la tmedfaen midden assa g ssuq, les gens se bouscuaient aujourd'hui au marché, il y avait trop de gens au marché aujourd'hui	exemple
5496	qqah may d ixleq rebba izmeç d assa g ssuq	tout ce que Dieu a créé était aujourd'hui au souk	il y avait aujourd'hui beaucoup de monde au souk	• qqah may d ixleq rebba izmeç d assa g ssuq (tout ce que Dieu a créé était aujourd'hui au souk), il y avait aujourd'hui beaucoup de monde au souk	expression
6864	igla imendi assa g ssuq, iga Zahennama	les céréales sont chères aujourd'hui au souk, elles sont un enfer)	les prix des céréales ont excessivement augmenté aujourd'hui au marché	• igla imendi assa g ssuq, iga Zahennama (litt. les céréales sont chères aujourd'hui au souk, elles sont un enfer), les prix des céréales ont excessivement augmenté aujourd'hui au marché	exemple
8282	ikker nnfad xef yirden assa g ssuq		le blé était très demandé aujourd'hui au marché	• ikker nnfad xef yirden assa g ssuq, le blé était très demandé aujourd'hui au marché	exemple
9378	tekker lqeyyama assa g ssuq		il y a beaucoup de monde aujourd'hui au souk	• tekker lqeyyama assa g ssuq, il y a beaucoup de monde aujourd'hui au souk	exemple

Un programme en Python permet de produire un fichier .txt (en Unicode UTF-8) à partir de fichiers .html.

Nous avons ainsi l'environnement qui nous permet d'interroger la base de données « exemplier » et d'en exploiter les résultats. Nous orientons les recherches vers le soutien à la réalisation d'une méthode d'apprentissage.

5 Constitution d'une méthode d'apprentissage

Les éléments explorés jusqu'à présent et extraits par recherche dans l'exemplier sont très simples et concernent les connaissances élémentaires qui constituent les premiers pas (expression du temps, expression du lieu), voire le « kit de survie » (oui, non; qui, quoi, où, comment,...).

Nous prendrons comme exemple l'expression simple du temps (hier, aujourd'hui, demain). Les titres de leçons pourraient avoir une forme telle que :

idelli, assa neǧd asekk

hier soir, aujourd'hui ou demain

avec différentes parties de présentation de petits textes :

ass

jour

S : ass, wass ; ussan, wussan (c'est-à-dire la racine: le mot au singulier état libre, état annexé; au pluriel état libre, état annexé)

ass-a aujourd'hui, de nos jours

ass-a zikk tôt le / ce matin

s wass de jour, en plein jour, pendant la journée, le jour

ass s wass un jour sur deux

g wass en une journée, en un jour

i wass la journée, pour une journée, par jour

ila wass toute la journée

ku ass chaque jour ; jadis, autrefois, anciennement

all ass-a jusqu'à aujourd'hui ; ce n'est qu'aujourd'hui

seg wass-a dès aujourd'hui, dorénavant

ass-nna le jour où ; quand

seg wass-nna depuis ce jour-là, depuis lors ; depuis le jour

yiwn n wass un jour ; un beau jour

ammas n wass en plein jour, publiquement

ǧart n wass après-midi, soirée (m. à m.: derrière du jour)

Le dépouillement de l'exemplier n'est qu'un des apports à la réalisation d'une méthode d'apprentissage qui nécessitera de nombreuses recherches complémentaires réalisées par une petite équipe de collaborateurs. Mais cet exemplier provenant du dictionnaire raisonné berbère – français, parlars du Maroc [17], la méthode bénéficiera d'un long travail de dépouillement des dictionnaires existants et de la littérature orale que Miloud Taïfi a réalisé des années durant.

Les paragraphes précédents se sont fait l'écho de quelques réalisations pour le dictionnaire.

Nous souhaitons maintenant présenter pour ses caractéristiques surprenantes la programmation de la typographie du dictionnaire en VBA Word.

6 VBA Word pour l'édition du dictionnaire

Si l'on faisait une enquête publique sur ce qu'est Word, il est très vraisemblable que l'on ait essentiellement des réponses du type : c'est un traitement de texte, c'est une machine à écrire améliorée, plus souple, ... Il est, par contre, peu vraisemblable que l'on trouve : c'est un environnement de programmation pour la fabrication, la modification et la mise en forme de fichiers de texte riche. Les pourcentages relatifs correspondraient vraisemblablement à l'usage généralement fait de Word, c'est-à-dire, en gros, à un centième, peut-être encore moins, de ses possibilités.

Car c'est véritablement un environnement de programmation: tout ce qui est fait manuellement dans Word peut être réalisé par programmation, mais celle-ci offre une plus grande précision (il y a par exemple 4 manières de faire une double justification pour une seule manuellement) et permet de réaliser des opérations impossibles manuellement : c'est d'ailleurs très exactement ce qui nous a amené à étudier la programmation interne à Word, en VBA (Visual Basic pour Applications). Word ne permet pas, par défaut, de faire des en-têtes de type « dictionnaire » avec la première entrée de la page à gauche dans l'en-tête et la dernière entrée de la page à droite.

La programmation en VBA permet la création, l'ouverture, la modification, la fermeture ou la destruction de fichiers externes. Elle permet également la programmation de pages web avec les options que l'on trouve habituellement avec un CMS.

Un usage standard de Word autorise la réalisation d'une première page différente des suivantes avec une numérotation dont l'origine (supérieure à zéro) peut être définie par l'utilisateur. Mais cet utilisateur n'a ensuite aucun moyen à sa disposition pour avoir des en-têtes variables à chaque nouvelle page, notamment pour obtenir une disposition de type dictionnaire où l'en-tête aura à gauche le premier élément que l'on souhaite mettre en exergue comme moyen de se repérer au sein du dictionnaire avec de l'autre côté, à droite, le dernier élément sur la page.

Suivant la tradition concernée, cela peut être un mot ou une racine. Dans le cas de notre dictionnaire, respectant la tradition des études (chamito-) sémitiques, nous souhaitons présenter dans l'en-tête la première et la dernière racine de la page avec la numérotation au centre.

Tout en-tête fabriqué de la page 2 à la page n est immédiatement dupliqué sur les n-1 dernières pages du fichier, c'est-à-dire sur toutes, sauf sur la première page pour laquelle une définition particulière est permise.

Pourquoi ?

Pour pouvoir répondre correctement à cette question, il est nécessaire d'expliquer le fonctionnement de Word. Word est un système à programmation orientée objet, c'est-à-dire que tout élément appartenant à Word est nécessairement un objet. Il y a en premier lieu, l'objet « application », c'est-à-dire Word mis en fonctionnement. De nombreux objets sont organisés en collection. C'est le cas des « documents ». Vous pouvez ouvrir simultanément dans Word de 1 à n documents parmi lesquels un seul est actif.

Il existe de nombreux moyens d'accéder au contenu d'un document : directement en prélevant l'intégralité du contenu principal (« content »), en allant de section en section (ce qui va nous intéresser ici), de paragraphe en paragraphe, de mot en mot, de caractère à caractère sachant que tous ces niveaux s'emboîtent les uns dans les autres : à l'intérieur d'une section, on peut rechercher / lister les paragraphes au sein desquels on peut rechercher des mots, etc.

Par défaut un fichier Word est défini comme une seule section où tous les en-têtes sont identiques ou comme deux sections où seule la première page est différente des pages suivantes, ce qui est notre cas.

6.1 Traitement typographique du dictionnaire en VBA Word

6.1.1 Création de sections

La première étape à réaliser est de constituer autant de sections qu'il existe de pages, ce qui permet d'affecter à chaque section les attributs souhaités extraits automatiquement de la page.

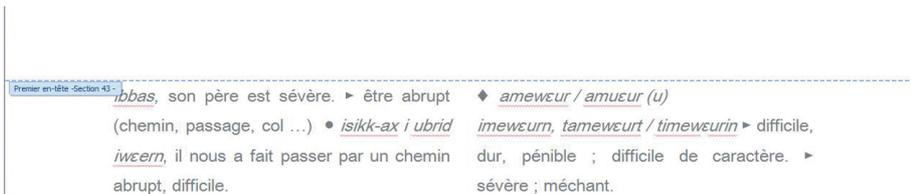
Le nombre de pages du document est fourni en donnée dans une fenêtre de dialogue. Le programme crée (ou maintient) la première page du document avec une forme particulière par rapport aux suivantes et crée une section nouvelle à chaque page :

```
.InsertBreak Type:=wdSectionBreakNextPage
.PageSetup.DifferentFirstPageHeaderFooter = True
```

Il est nécessaire ensuite de dissocier la section courante de la section précédente, afin de ne pas reprendre l'en-tête et le pied de page. Si le fichier contenait déjà un en-tête, celui-ci est supprimé:

```
With ActiveDocument.Sections(i).Headers(2)
.LinkToPrevious = False
.Range.Delete
End With
```

On peut facilement constater, en ouvrant les en-têtes et pieds de page que les sections ont bien été créées:



6.1.2 Numérotation des racines

La seconde étape va consister à affecter un numéro d'ordre aux racines homographes, et uniquement à celles-ci: les racines qui n'apparaissent qu'une seule fois dans le fichier ne sont pas numérotées.

Le traitement se fait au sein du document en entier en se servant de la suite de paragraphes. Le résultat est le suivant :

à plomb.		♦ <i>awaed (wa), iwaeden</i> ► fait de promettre,
WZR1	kb.	d'engager sa parole, promesse.
♦ <i>wawzer</i> , masc., <i>id wawzer</i> ► <u>depression</u>		♦ <i>lweeda</i> , fém., <i>lweedat</i> ► promesse (faite à
entourée de hauteurs. ► <u>partie non labourée</u>		un Saint) ; offrande pieuse (à un Saint)
d'un champ ; bande de terrain non labouré.		WED2
WZR2	ar., kb.	♦ <i>waed, waed, twaead, ur-waed</i> ► rejoindre ;
♦ <i>lewzir</i> / <i>lwazir</i> / <i>luzir</i> , <i>lwuzara</i> ► <u>visir</u> ;		aller trouver ; se diriger vers ; regagner ,
ministre.		rentrer chez soi. • <i>iwaed-t ger taddart</i> , il l'a
WZE	ar.	rejoint à la maison, chez lui. • <i>gas la tnağ d</i>
♦ <i>wezzee, wezzec, twezac, ur-wezzic</i> ► se		<i>urgaz-nes twaed may-s</i> , à chaque fois

6.1.3 Mise en exposant de la numérotation

Le traitement se fait ici aussi sur l'ensemble du document et repère les racines qui sont des mots en majuscules et en caractères gras qui sont sélectionnées grâce à ces critères sans recherche plus approfondie :

```
If .Paragraphs(i).Range.Words(1).Case = wdUpperCase _
And .Paragraphs(i).Range.Words(1).Bold Then
```

Leur partie numérique est ensuite mise en exposant :

```
If Asc(.Characters(j)) > 47 And Asc(.Characters(j)) < 58 Then
.Character(j).Font.Superscript = True
```

Ce qui nous donne :

WZR ¹	kb.	d'engager sa parole, promesse.
♦ <i>wawzer</i> , masc., <i>id wawzer</i> ► <u>depression</u>		♦ <i>lweeda</i> , fém., <i>lweedat</i> ► promesse (faite à
entourée de hauteurs. ► <u>partie non labourée</u>		un Saint) ; offrande pieuse (à un Saint)
d'un champ ; bande de terrain non labouré.		WED ²
WZR ²	ar., kb.	♦ <i>waed, waed, twaead, ur-waed</i> ► rejoindre ;

6.1.4 Mise en exergue des racines

Pour faciliter la lecture du dictionnaire, les entrées, c'est-à-dire les racines sont colorées en bleu :

```
If .Paragraphs(i).Range.Words(1).Bold _
And .Paragraphs(i).Range.Words(1).Case = wdUpperCase Then
.Paragraphs(i).Range.Words(1).Font.ColorIndex = wdBlue
```

ce qui donne le résultat suivant :

attaqué pendant la trêve.		WTR ¹
WT ⁶		♦ <i>lewter</i> , masc., ► <u>prière surégatoire</u> qui se
♦ <i>tawtat, tiwtatin</i> ► pompon (pour vêtement		fait après celle du soir, elle comporte une
d'enfant et de femme) ► morceau d'étoffe		seule génuflexion.
pour habiller une poupée.		WTR ²
WTL ¹		♦ <i>wawter</i> , masc., <i>id wawter</i> ► <u>humérus</u> .
♦ <i>watla</i> , dér. de <i>wef</i> "frapper" ?, <i>watla</i> ,	WT ¹	ar., kb.
<i>twatla, ur-watla</i> ► frapper, rosser, cogner.		♦ <i>wata, wata, twata, ur-wata</i> ► aplanir,

6.1.5 Fabrication des en-têtes

Maintenant que le corps du document est traité, nous allons pouvoir l'utiliser pour fabriquer l'en-tête de chaque page (c'est-à-dire de chaque « section » du « document »).

La première racine (sous le nom de variable *deb*), celle qui va être inscrite sur la gauche de l'en-tête, est extraite ainsi :

```
With ActiveDocument.Sections(i)
For j = 1 To .Range.Words.Count
If .Range.Words(j).Bold And _
.Range.Words(j).Case = wdUpperCase Then
deb = .Range.Words(j)
deb = Trim$(deb)
Exit For
End If
Next j
End With
```

La dernière racine (*fin*), celle qui va être inscrite sur la droite de l'en-tête, est obtenue de manière identique, mais à partir de la fin de la page en remontant:

```
With ActiveDocument.Sections(i)
For K = .Range.Words.Count To 1 Step -1
...

```

Les racines adéquates ayant été repérées, nous construisons (ligne 10) le texte de l'en-tête (.Range.Text) qui sera constitué de la première racine suivie d'une tabulation (vbTab) dont le positionnement a été calculé en (5), qui permet de positionner la numérotation au centre, et de la dernière racine qui sera située à droite positionnée par la tabulation définie en (6).

```
(1)With ActiveDocument.Sections(i).Headers(2)
(2) .Range.Delete
(3) .Range.ParagraphFormat.Alignment = wdAlignParagraphCenter
(4) .Range.ParagraphFormat.TabStops.ClearAll
(5) .Range.ParagraphFormat.TabStops.Add _
Position:=CentimetersToPoints(6.5), Alignment:=wdAlignTabCenter
(6) .Range.ParagraphFormat.TabStops.Add _
Position:=CentimetersToPoints(13), Alignment:=wdAlignTabRight
(7) .Range.Font.Name = «Arial Unicode MS»
(8) .Range.Font.Size = 12
(9) .Range.Bold = True
(10) .Range.Text = deb & vbTab & numpage & vbTab & fin
(11)End With
```

Toute inscription qui se serait trouvée dans l'en-tête est supprimée (2) ainsi que d'éventuelles tabulations (4). L'en-tête sera en police Arial Unicode (7) en corps 12 (8) et en gras (9).

L'en-tête étant constitué, nous pouvons, comme dans le corps du dictionnaire, mettre le rang des racines en exposant, ce qui donne :

WZR¹

1080

W&R

à plomb.

WZR¹

◆ *awacd* (*wa*), *iwaacden* ► fait de promettre,
kb. d'engager sa parole, promesse.

6.2 Traitement en VBA Word des deux index d'accompagnement du dictionnaire

Nous avons présenté (en 2) les deux index qui accompagnent le dictionnaire. Ils ont un défaut que nous allons corriger ici. Ils sont issus des calculs faits sur le dictionnaire où toutes les racines portent un numéro d'ordre, même lorsqu'elles sont uniques. Ceci est dû au besoin de non-ambiguïté pour la construction des relations entre tables de la base de données, mais ceci ne convient plus pour la publication de l'ouvrage et les travaux pour la typographie du dictionnaire vont nous permettre de supprimer la numérotation des racines uniques.

Nous calculerons pour chacun des index un type spécifique d'en-tête.

6.2.1 Construction d'un fichier externe de référence des racines homographes

Pour la publication du dictionnaire, les racines uniques ne sont pas numérotées. A partir de chacun des 29 fichiers constituant le dictionnaire, le programme ci-dessous extrait la première racine de chaque suite homographe et l'enregistre (Selection.Paste) dans un fichier externe unique RACINES1 de manière séquentielle (Selection.EndKey unit :=wdStory).

```

Sub chercheRacineRangUn()
Dim actuel As Document
Dim racines As Document
Set actuel = Documents.Open(ActiveDocument.Path & "\25 - W.doc")
Set racines = Documents.Open(ActiveDocument.Path & "RACINE1.doc")
With actuel
MsgBox (Range.Paragraphs.Count)
For i = 1 To .Paragraphs.Count
If .Paragraphs(i).Range.Words(1).Bold And .Paragraphs(i).Range.Words(1).Case = wdUpperCase Then
If Right$(Trim$(.Paragraphs(i).Range.Words(1)), 1) = "1" Then
.Paragraphs(i).Range.Words(1) = Trim$(.Paragraphs(i).Range.Words(1))
If Not (Asc(Left$(Right$(.Paragraphs(i).Range.Words(1), 2), 1)) > 47 And Asc(Left$(Right$(.Paragraphs(i).Range.Words(1), 2), 1)) < 58) Then
.Paragraphs(i).Range.Words(1).Copy
racines.Activate
Selection.EndKey unit:=wdStory
Selection.TypeParagraph
Selection.Paste
End If
End If
End If
Next i
End With
With racines
.Save
.Close
End With
End Sub

```

6.2.2 Traitement des deux index

Lors du traitement des deux index, nous serons en présence de deux types de racines numérotées en 1, celles qui sont isolées et celles qui sont le début d'une série. Si nous ne retrouvons pas la racine concernée dans le fichier RACINE1.doc, c'est qu'elle est unique et nous supprimerons donc sa numérotation. Voici la fonction chargée de chercher la présence d'une racine dans le fichier externe de racines :

```

Function trouverRacine(rac)
Set myRange = ActiveDocument.Content
myRange.Find.Execute findtext:=rac, Forward:=True
If myRange.Find.Found = True Then
trouverRacine = True
Else
trouverRacine = False
End If
End Function

```

C'est le programme « traiterIndex() » qui, à l'aide de la fonction « trouverRacine(rac) » présentée ci-dessus, traite la numérotation des racines dans chacun des deux index.

Voici la typographie de l'index 1 :

grurs: GRS ⁴	gund: GND ¹	g ^w er: GR ⁴	gzey: GZY ²
gres: GSR	gungey: GNGY	g ^w ez: GZ ¹	gzil: GZL
gses: GS ¹	gunzer: GNZR	g ^w eždež: GŽD ¹	gzul: GZL
gudy: GDY ¹	gunzer: NZR ¹	g ^w žem: GŽM ¹	gæed: G&D
gufsu: GFS	gurzu: GRZ ²	g ^w žem: KŠM ¹	

et celle de l'index 2 :

<u>quand</u> : DY ² , G ¹ , L ¹ , LD ¹ , LG ¹ , M ³ , MLM, MNTR, R ² , SM ¹	<u>quatorze</u> : RBE ²
~ <u>bien même</u> : MQR, MGR ¹	<u>quatre</u> : KZ ² , RBE ²
~ <u>même</u> : HL ⁴	à ~: MRBE ²
	en ~: MRBE ²

Etant donnée la spécificité de chacun des index, c'est un programme différent qui fabrique l'en-tête, « traiterEnTeteIndex1() » pour le premier, celui de la correspondance mots berbères – racines berbères :

G - gru	60	gæed - G
gru: GR ⁴	gugg: G ⁵	gusmu: GSM ¹
		gzem: GDŽM

et « traiterEnTeteIndex2() » pour le second index, celui qui associe des racines berbères à des entrées françaises :

<u>querelleur</u>	250	quotidien
se ~: NG ³	<u>quinine</u> : KN ¹¹ , QWS	
<u>querelleur</u> : ŠRN	<u>quignon</u> : QŠR ²	

7 Conclusion et perspectives

A l'issue de ce parcours, nous disposons actuellement:

- en premier lieu, d'un dictionnaire berbère – français de 1223 pages accompagné de deux lexiques, respectivement de 166 et 319 pages. Ce dernier lexique est important, car il permet d'associer des mots français aux racines berbères correspondantes dans des applications ultérieures. L'ensemble doit sortir sur les Presses de l'IRCAM² à Rabat d'ici la fin de l'année 2015.
- d'un corpus brut structuré de plus de 2 700 000 caractères qui peut être interrogé tel quel ou transformé en n'importe quel type de base de données ou dispositif équivalent.
- d'une base de données de 13 600 locutions et exemples, interrogeable, dénommée exemplier Taïfi à partir de laquelle nous avons débuté des travaux pour l'élaboration d'une méthode d'apprentissage.
- d'un ensemble de plus de 5000 verbes qui serviront de base à un conjugeur automatique.

Le corpus dictionnaire est un acquis essentiel qui va nous permettre, en premier lieu, de créer par des processus déjà testés dans la réalisation de dictionnaires et lexiques

² Institut Royal de Civilisation AMazighe

(il s'agissait de faciliter la création d'un dictionnaire français – slovaque sur la base d'un dictionnaire slovaque – français) de proposer la masse lexicale déjà traitée dans le dictionnaire berbère – français avec toutes les indications (données et calculées) jointes aux entrées suggérées. Cela ne dispense ni de la définition d'une conception adéquate du nouveau dictionnaire, ni de devoir aménager les entrées et notamment de rajouter des entrées absentes du dictionnaire berbère – français, ni d'un travail considérable de réalisation concrète du dictionnaire, mais cela l'accélère considérablement!

Le corpus dictionnaire permet également la génération automatique de plusieurs types de dictionnaires et de lexiques: dictionnaire classé par mots avec indication de la racine, dictionnaire inverse (a tergo), lexique terminologique avec classement thématique (par exemples plantes, animaux, artisanat)...

L'expérience acquise nous permet d'aborder un projet analogue dans une langue encore moins dotée. Il s'agit d'un projet de « dictionnaire de la santé cilubà – français ». Le cilubà (écrit antérieurement tshiluba) est une langue bantoue, classée L31, et parlée en République Démocratique du Congo. Ce projet sera réalisé en coopération entre l'Institut National des Langues et Civilisations Orientales à Paris (Prof. Odile Racine avec notre concours) et l'Université de Mbùjimâyi (Prof. Emmanuel Kambaja Musampa, Doyen de la faculté des Lettres et Sciences Humaines) avec le recours à d'autres institutions partenaires.

Forts de l'expérience berbère, nous avons défini un environnement informatique pour mener les travaux à bien. Nous ouvrirons en premier un portail dédié (ou associé au portail actuel Linguothèque) réalisé en WordPress 4.3 choisi pour sa relative facilité d'usage. A ce portail seront associés dès le départ des travaux sur l'écriture nouvelle (cela pourrait se faire grâce à une collaboration avec l'Université de Gand qui a déjà développé plusieurs projets concernant le cilubà dont un dictionnaire de référence en ligne, œuvre du professeur Kabuta). La transcription automatique de l'ancienne à la nouvelle écriture ne peut se faire (l'ancienne écriture ne possédant pas les tons) que par consultation du dictionnaire de Kabuta, celui de Kayoka Mudingay étant encore en écriture ancienne.

La constitution d'un corpus, préalable nécessaire aux travaux, devra s'appuyer sur cette écriture et le traitement automatique lexical des textes avec des procédures de tri adéquates. A ce corpus en formation, il est également nécessaire d'associer une structuration linguistique des données cilubà en réfléchissant au statut des racines et peut-être en proposant un traitement des substantifs ne correspondant pas à la tradition actuelle. Il devrait en résulter la possibilité de construire n'importe quel type de bases de données (ou équivalent).

Les applications privilégiées seront les dictionnaires et les lexiques pour lesquels nous rechercherons un travail coopératif sur des partitions de disques liées au portail.

Pour des raisons d'efficacité, nous retiendrons des expériences passées le caractère multifonctionnel de nos outils et veillerons à leur généralité et réusabilité.

Bibliographie

- [1] Azdoud, D. (2011). Dictionnaire berbère-français. Editions de la MSH, Paris.
- [2] Benamara, H. (2013). Dictionnaire amazighe – français. Parler de Figuig et ses régions. Presses de l'IRCAM, Rabat.
- [3] Bounfour, A. and Boumalk, A. (2001). Vocabulaire usuel du tachelhit. Tachelhit – français. Centre Tarik Ibn Ziyad, Rabat.
- [4] Kabuta, N. S. (2008). *Nkòngamyakù cilubà-mfwàlànsa*, RECALL-CIYÈM, 395 pp.

- [5] Kambaja Musampa E. (2007). Etude socioterminologique du vocabulaire médical cilubà, *Lexikos*, 17:88–113.
- [6] Mudingay Kayoka, M. (2013). *Dictionnaire Français-Tshiluba / Tshiluba- Français*, 3^{ème} édition, Editions Lulu.Com, Raleigh, 522 pp.
- [7] Oussikoum, B. (2013). Dictionnaire amazigh – français: le parler de Ayt Wirra, Moyen Atlas-Maroc, Presses de l'IRCAM, Rabat.
- [8] Pognan, P. (2006). Apprendre le tamazight dans une approche linguistique plurielle. Colloque Linguistique descriptive et didactique de l'amazigh, Agadir.
- [9] Pognan, P. (2010). Extraction automatique du syntagme nominal prépositionnel berbère. 6. Bayreuth-Frankfurt-Leidener Kolloquium zur Berberologie, Bayreuth.
- [10] Pognan, P. (2013). L'exemplier Taïfi, en hommage au maître et à l'ami. In Etudes et recherches en linguistique et littérature berbères, Fez.
- [11] Pognan, P. (2013). Un nouvel horizon pour le dictionnaire raisonné berbère-français – parlars du Maroc: l'élaboration d'une méthode d'apprentissage. Auréoles berbères, Mélanges offerts à Michel Peyron, Tetouan.
- [12] Pognan, P. and Taïfi, M. (2010). Traitements informatiques du dictionnaire raisonné berbère – français, élaboration de « l'exemplier Taïfi ». In Journées d'étude JEC « Linguistique(s) de corpus: de la constitution à l'exploitation des corpus », Paris.
- [13] Pognan, P. and Taïfi, M. (2011). Un dictionnaire en tant que corpus: traitements informatiques du dictionnaire raisonné berbère – français de Miloud Taïfi. 4^{ème} atelier international sur l'amazighe et les TIC, IRCAM, Rabat.
- [14] Rahho, R. (2005). Dictionnaire berbère-français. Parler des Beni-Iznassen. Université de Fès, Thèse de doctorat. FLSH, Dhar El Mehraz.
- [15] Serhoual, M. (2002). Lexicologie, lexicographie et sémantique berbères: 1) Lexicologie amazighe 2) Dictionnaire tarifit – français. Thèse de doctorat d'état. Faculté des Lettres et des Sciences Humaines, Tétouan.
- [16] Taïfi, M. (1991). Dictionnaire tamazight – français. Parlars du Maroc central, L'Harmattan-Awal, Paris.
- [17] Taïfi, M. (2004). Si les berbérophones ne s'entendent pas oralement, qu'ils s'écrivent! Pour une écriture grammaticale du berbère. Standardisation de l'amazighe. Publications de l'IRCAM, Rabat.
- [18] Taïfi, M. (2015). Dictionnaire raisonné berbère – français. Parlars du Maroc (1223p) et Pognan: index „Formes de mots berbères – racines“ (166p) et index „Significations françaises – racines berbères“ (319p). IRCAM, Rabat (A paraître).
- [19] Taïfi, M. and Pognan, P. (2005). Langues berbères: à la recherche du système perdu. Colloque international « Linguistique amazighe: les nouveaux horizons », Tetouan.
- [20] Taïfi, M. and Pognan, P. (2009). Traitements automatiques du « Dictionnaire raisonné berbère – français, Parlars du Maroc ». In 8^{èmes} Journées scientifiques du réseau LTT de l'AUF, Lisboa.
- [21] Willems, E. (2006). *Dictionnaire français-cilubà*, 4^{ème} édition revue, corrigée et actualisée par Pamba Kamba Kamba et Lutumbu Kapata Matthieu, Editions de l'Archidiocèse, Kananga, 413 pp.

Identifying Corpus-specific Collocations: The Case of Spoken Slovene

Senja Pollak¹ and Špela Arhar Holdt²

¹Jozef Stefan Institute, Ljubljana, Slovenia

²Institute for Applied Slovene Studies, Trojina, Slovenia
Faculty of Arts, Ljubljana, Slovenia

Abstract. In this paper we present a methodology for extracting and comparing collocations in two different corpora of the same language. In our case, we aim to identify collocations that are characteristic of spoken Slovene in relation to the reference corpus of written Slovene. In the first step, we use collocation extraction functionality available in the open-source NoSketchEngine platform. Next, based on obtained logDice scores, we compare the extracted collocations by computing the *CorpDiff* score that we introduce in this paper. In our experiment we focus on collocations with the absolute *CorpDiff* score above 2 and propose a corpus-based linguistic analysis of selected examples, as well as a quantitative evaluation of relevance of extracted candidates. We discuss advantages and deficiencies of the approach, as well as possible applications and improvements from the perspective of a more extensive experiment that is planned in future work.

1 Introduction

In this paper, we present the methodology for comparing collocations of two corpora with the aim of identifying the collocations that are corpus-specific. When it comes to defining collocation, different authors take different perspectives. In general, approaches to collocation research can be divided into frequency-based approaches and phraseological approaches (cf. [4]). *Frequency-based approaches* (e.g. [1], [2], [3], [5]) define collocations as statistically significant co-occurrences of words within a certain distance.¹ On the other hand, the *phraseological approaches* (e.g. [7], [8], [9]) accentuate grammatical and semantic properties of collocations, understanding collocations as lying between idioms and free word combinations, which are two opposites regarding the meaning compositionality. In the extraction phase of our study, we choose the frequency-based approach, while in the interpretation phase we consider other (syntactic and semantic) aspects as well.

As a case study, we explore the specificities of collocations in a corpus of spoken Slovene compared to the reference corpus of written Slovene. The assumption behind the idea is that such a comparison will facilitate insight into genre-specific language use, thus providing valuable data to the field of applied linguistics (e.g. lexicography, language teaching, translation) and language technology development.

The study serves also as a pilot study for a more extensive experiment. As part of a larger project aiming at linguistic analysis of Slovene user-generated content (in twitter messages, forums etc.), one of the tasks is to extract corpus-specific collocations and compare them with standard written Slovene. As a motivating example, let us consider the collocation *go*

¹ This understanding of collocation is closely related to n-grams, which are sequences of words (or lemmas) in a string of text. If on one side collocations can be understood simply as n-grams ranked by a lexical association measure, their units are, in contrast to n-grams, not necessarily adjacent [6].

viral, which is specific to the web and is “used in reference to Internet content which can be passed through electronic mail and social networking sites (Facebook, etc.): an image, video, or link that spreads rapidly through a population by being frequently shared with a number of individuals has *gone viral*”.² The identification of this type of expression is important from various perspectives (understanding of language development, language teaching, translation, sociolinguistics, natural language processing, etc.).

2 Related Work

From the methodological perspective, this paper presents a SketchEngine-based corpus study. One of the tools available in the SketchEngine [10] is *SketchDiff*, which is very similar to the *CorpDiff* score, we introduce below. However, there are several differences between the two procedures. In our approach we rely only on the open source version, i.e. NoSketchEngine, in which the *SketchDiff* functionality is not available. Both corpora are available through NoSketchEngine@nl.ijs.si.³ We operate on a list of collocations extracted by the function *Collocations* (available in NoSketch version) and not on word sketches. This decision follows from our interest in testing a very general approach, providing the possibility of applying it to the non-standard Slovene of user-generated web content (e.g. twitter), for which preprocessing tools are underperforming [26]. However, this decision results in losing some ease in the analysis of results.⁴ Next, we compute the difference of two collocation (logDice) scores resulting in the *CorpDiff* score and propose a threshold parameter in order to define collocations that are specific enough to be relevant for detailed analysis (we could focus only on the specific corpus or analyze collocations characteristic to either specific or reference corpus; the latter is the case of the present experiment). Last but not least, the SketchDiff in SketchEngine is available only for sub-corpora, while we use it for comparing two completely independent corpora.

Word Sketches for Slovene were presented in [11] and [12]. From terminological perspective, the tools for extracting collocations in Slovene and their analysis (focusing on noun phrases) can be found in [13] and [14]. The analysis of collocations in Slovene written language was conducted e.g. by [15] and [16]. Corpus analysis of Slovene spoken language was covered by [17] but without a specific focus on collocations. A more thematic perspective on specific expressions in spoken Slovene is proposed in [18]. The majority of the aforementioned work from the field of Slovene applied linguistics was motivated by a lexicographer perspective and focused on a single corpus resource. For other languages, the differences between the written and the spoken language have been addressed in a contrastive manner by several authors, among them [19], [20] and [21].

3 Task Definition and Methodology

Given two corpora S and R in the same language, and a set of selected lemmas L , the goal is to find for each l in L , a set of collocations C_l $\langle c(\text{locator}), l(\text{lemma}) \rangle$, which are substantially different in the two corpora. For the set of lemmas L , this results in collocation sets C_{LS} and C_{LR} of pairs of the form $\langle c, l \rangle$ characteristic for S and R , respectively.

² <http://www.urbandictionary.com/define.php?term=go%20viral>

³ <http://nl.ijs.si/noske/index-en.html>

⁴ In future work, we plan to evaluate the balance between advantages and disadvantages of the two approaches when applying them to non-standard Slovene language.

In our case, *S* refers to the corpus of spoken Slovene GOS, while *R* refers to the reference corpus of written Slovene KRES. We are more interested in the specificities of spoken Slovene, but we treat the two corpora equally, since in our perspective all lexical choices are considered to be significant, either by their frequent use or their salient absences. Similar to our use, the methodology can be applied to compare specificities of domain, user generated content, web or learners' corpora.

More specifically, for association measure we used logDice score [22]:

$$\logDice = 14 + \log_2 D = 14 + \log_2 \frac{2 f_{xy}}{f_x + f_y}$$

This score is available as one of the association scores in the lexicography-oriented program SketchEngine [10]. *D* in the formula refers to the original Dice coefficient [23] while f_x , f_y and $f_{x,y}$ refer to the frequencies of word *x*, word *y* and to the co-occurrence of both, respectively. It has several features [22] such as that it is easy to interpret (ranging between 0 and 14) and not being dependent on the corpus size since it uses relative frequencies. For initial *logDice* calculations we use function *Collocation*, available in the freely available NoSketch Engine, where we use the following settings: the window was set to -3/+3, while the minimum frequency for exporting collocations was 10 for the lemma and 5 for the collocation. Based on these initial files, we selected lemmas with logDice score above 5 and relative frequency (based on corpus size) above 20 per million. For those we computed differences (*CorpDiff*) in logDice scores:

$$CorpDiff(S, R) = \logDice(S) - \logDice(R)$$

that we inspect together with other information, such as frequencies, relative frequencies (based on total words count), logDice of each collocator, etc. High positive *CorpDiff* scores denote collocations specific to the specific corpora GOS (*S*) and negative scores the collocations characteristic of the written reference corpus (*R*), while scores around 0 mark collocations that are not specific to either of the corpora. We also accept a list of manually defined stop list words⁵ that are not relevant for collocation analysis, but appear very frequently as extracted collocators from the spoken corpus. For the analysis, we inspected the collocations in C_{LS} and C_{LR} with the score $|CorpDiff(S, R)| > 2$.

4 Corpora Presentation

4.1 GOS

Corpus of spoken Slovene GOS⁶ [24] comprises the transcripts of approximately 120 hours of speech in various situations: radio and TV shows, school lessons and lectures, private conversations between friends or within the family, work meetings, consultations, conversations in buying and selling situations, etc. All speech is transcribed in two versions – with pronunciation-based spelling and with standardized spelling – and it comprises over one million words.

⁵ In our case we used phatic discourse markers, such as *mhm*, *eee*, *eem*.

⁶ <http://eng.slovenscina.eu/korpusi/gos>

4.2 KRES

Corpus KRES⁷ [25] is sampled from a large corpus Gigafida⁸ (ibid.) and we use it as a reference corpus of (written) Slovene. It contains Slovene text of various genres, from daily newspapers, magazines, all kinds of books (fiction, non-fiction, textbooks), web pages, and similar, with a balanced genre structure. It contains almost 100 million words.

Both corpora are available for download as well as for use through NoSketch Engine³.

5 Case Study

5.1 Selection of Lemmas

First, we selected 10 most frequent nouns in each of the two corpora (KRES and GOS). The union of both lists consisted of 16 nouns for which collocations were analyzed: *bistvo* [*fact*⁹], *čas* [*time*], *človek* [*man*], *dan* [*day*], *delo* [*work*], *gospod* [*mister*], *leto* [*year*], *mesto* [*city*], *otrok* [*child*], *primer* [*example*], *red* [*order*], *slovenija* [*slovenia*], *stvar* [*thing*], *svet* [*world*], *ura* [*time*], *življenje* [*life*].

5.2 Extraction of Collocation Candidates

For each of the nouns, a list of corpus-specific collocators was generated as described in Section 3. Table 1 shows, as an example, the results for the noun *dan* [*day*]. For this lemma, eight collocators with the absolute *CorpDiff* score greater than 2 were found (the positive *CorpDiff* values indicate GOS-specific collocators and the negative values Kres-specific ones).

As mentioned in the Introduction, the frequency-based approach to collocations was adopted during the data extraction phase. The total number of extracted candidates (for all the 16 nouns) was 141 (see Table 2).

DAN	CorpDiff	LogDice (GOS)	LogDice (KRES)	Relative Freq (GOS)	Relative Freq (KRES)
želeti [(<i>to</i>) <i>wish</i>]	3	8.8	5.8	25.2	4.3
danes [<i>today</i>]	2.9	9.3	6.5	57.1	6.4
dober [<i>good</i>]	2.5	10.3	7.8	127.8	21.3
štirinajst [<i>fourteen</i>]	2.4	9.7	7.3	38.7	8.9
cel [<i>entire</i>]	2.3	9.8	7.5	56.1	10.9
svoj [<i>one's own</i>]	-7.3	0	7.3	0	24.9
začeti [(<i>to</i>) <i>begin</i>]	-8.2	0	8.2	0	24.7
nek [<i>one</i>]	-9.1	0	9.1	0	37

Table 1. Collocators for lemma *dan* [*day*] with $|\text{CorpDiff}|$ score above 2

⁷ www.korpus-kres.net

⁸ www.gigafida.net

⁹ The literal meaning of noun *bistvo* is *essence*, but in the examples in this article it is related to the expression *v bistvu* [*in fact*], therefore we translate it as *fact*. In the majority of translations we try to keep the source of the collocation extraction transparent. In some cases, we provide literal meaning (lit.).

In the analysis analysis we move from strictly frequency-based approach and adopt the phraseological point of view (how syntactically or semantically bound the collocator is to the headword, see Section 5.3). We manually examined all the pairs <lemma, collocator> in the concordances of both corpora. In the first step, we discarded from the list the collocation candidates with collocators of two groups:

- **frequent function words**, such as: *ko* [*when*], *ker* [*because*], *in* [*and*], *na* [*on*]. An important exception in this group were combinations of directly adjacent (syntactically bound) prepositions forming a fixed expression, e.g. *v bistvu* [*in fact*] and *v redu* [*ok*¹⁰] were categorized as relevant candidates.
- **frequent spoken language specific words**, such as: *ne* [*isn't it*], *zdaj* [*now*], *potem* [*then*], *nek* [*one*]. While the occurrence and distribution of such lexica (e.g. discourse markers) in spoken language is relevant per se, in our evaluation this type of data was discarded, as their appearance in the potential collocations was a sole result of their general high frequency in the spoken corpora).

As shown in Table 2 (middle column) 88 out of from 141 pairs were discarded in this preliminary selection based on the above-mentioned criteria and 53 (37.6%) were kept for further analysis. In future work, we hope to successfully decrease the number of irrelevant candidates by different parameter settings (e.g. limiting the window size when part-of-speech is a function word) and by elaborating stop lists for frequent words in spoken and written Slovene.

5.3 Analysis of Collocations Candidates

The collocations, remaining after the selection step presented in Section 5.2, are of two types:

A) Combinations in which the lemma and its collocator form a **syntactically and semantically bound unit**. These corpus-specific word phrases are the core result of our study and indicate the highest potential for future use in various fields of applied linguistics, including lexicography (see further discussion on this topic below). The last column of Table 2 indicates the number of extracted collocators of this type.

B) **Syntactically looser co-occurrences** that are not collocations in the narrow sense of term, but might be relevant for linguistic research, such as discourse analysis, since they disclose different genre specificities of corpora in question.

Lemma	Initial collocation candidates	Collocation candidates after preliminary selection	Relevant candidates (final selection, Type A)
BISTVO [<i>fact</i> ⁹]	32	2	2
ČAS [<i>time</i>]	2	2	1
DAN [<i>day</i>]	8	8	5
DELO [<i>work</i>]	3	1	1
GOSPOD [<i>mister</i>]	6	2	1
LETO [<i>year</i>]	34	11	8
PRIMER [<i>example</i>]	3	1	0
STVAR [<i>thing</i>]	6	3	3

¹⁰ *v redu* (lit. *in order*) means *ok* is extracted by preposition *in* and lemma *order*.

URA[<i>time</i>]	8	7	7
SVET[<i>world/council</i>]	14	3	3
RED [<i>order</i>]	22	10	10
SLOVENIJA [<i>slovenia</i>]	3	3	2
All	141	53	43
Perc.	100.0%	37.6%	30.5%

Table 2. Evaluation of collocation candidates for lemmas (only lemmas with candidates with |CorpDiff| score above 2)

The decision on the relevancy of the results is somewhat subjective. Firstly, the evaluation was hindered by the lack of definite purpose of the data extraction. Possible applications could be for: the description of spoken language (spoken dictionary); complementing existing lexicographic resources with data on specificities of usage in spoken/written language; language teaching; discourse/genre analysis (where also syntactically looser candidates would be relevant) etc. In the evaluation, we consider different application options, however, not all results are equally relevant for all of the purposes. Secondly, the results were evaluated by a single researcher, which for future work could be improved by measuring inter-annotator agreement. Finally, data from the spoken corpus was in some instances difficult to evaluate, as the general notion of a collocation is based primarily on written language, in which syntax is more regular and predictable in comparison to the spoken language. Other specificities of the spoken discourse have to be taken in the consideration as well. For example, among the GOS-specific results occurs the expression *v redu v redu* [ok, ok]. While the initial evaluation might discard this unit as a simple repetition of a phrase *v redu* [ok], the concordances show there could be a difference in usage: while *v redu* typically implies understanding and agreement with the conversational partner, *v redu v redu* to a greater extent signals that further negotiation on the topic is not necessary and that the speaker is prepared to change the subject. In the evaluation, we perceived results of this type as relevant, nevertheless, an exhaustive qualitative analyses would be needed for further discussion on some of the candidates.

For demonstration, we enumerate 10 collocations that we consider as relevant collocations (Type A) extracted from the spoken language corpus and written language corpus, respectively:

- **GOS (spoken):** *v bistvu* [in fact], *dober dan* [hello]; *cel dan* [entire day], *dame in gospodje* [ladies and gentlemen], *X let nazaj* [X years ago], *par let* [a couple of years], *prva stvar* [the first thing], *druga stvar* [the second thing], *cel svet* [the entire world], *ful v redu* [totally ok];
- **KRES (written):** *do konca svojih dni* [ever after], *nekega dne* [one day], *X let pozneje* [X years after], *ob začetku/koncu leta* [at the beginning/end of the year], *v preteklem letu* [in the past year], *X leta Y stoletja* [in the X's of the Y century], *X leto starosti* [at the age of X], *vrstni red* [order], *dnevni red* [day's agenda], *Republika Slovenija* [The Republic of Slovenia].

5.4 Corpus Analysis of a Selected Example

For better understanding of the analyzed material and the evaluation process, we conclude the presentation of the results with a discussion on the collocation candidates for the

lemma *day*. For this lemma, we extracted eight potential collocators, based on the absolute *CorpDiff* score greater than 2 (see Table 1). In the preliminary selection phase (Section 5.2) no candidates were discarded. Among the results, five collocations were categorized as Type A and three as Type B (Section 5.3).

An expected example of a GOS-specific collocation is *dober dan* [hello] (lit. *good day*), a commonly used greeting formula {1}. The collocations *cel dan* [the entire day] and *štirinajst dni* [fourteen days] are somewhat less obvious. The phrase *cel dan/cele dneve* {2} occurs in similar context in the spoken as well as the written corpus, nevertheless it appears relatively more frequent in the spoken corpus. The phrase *štirinajst dni* {3} occurs in both corpora as well. Further analysis showed that in corpus Kres the synonymous phrase *dva tedna* [two weeks] is used significantly more frequently than *štirinajst dni* [fourteen days], while in the spoken corpus the variants are more balanced, hence the phrase was identified as GOS-specific.

- {1} učenci dobro jutro **dober dan** [students good morning hello]
- {2} delal sem **cele dneve** in eee vikende in tako naprej [I was working for entire days and eee weekends and so on]
- {3} grem grem eee jz grem čez **štirnajst dni** grem na Kubo [I'm going eee I'm going to Cuba in fourteen days]

Kres-specific collocations were identified as well. The phrase *nekega dne* [one day] occurs 3,243 times in Kres {4}, while in the corpus of spoken Slovene it only occurs twice, both times during a school lesson on Slovene language, where the speaker is referring to literary language {5}. Secondly, different types of phrases with the collocator *svoj* can be identified, among them *do konca svojih dni* [until the end of their days]. Similarly to the previous example, this phrase seems to be more typical of written language {6}, as in the spoken corpus it doesn't appear at all.

- {4} **Nekega dne** so pred vrata domače hiše prišli trije volkovi. [One day, three wolves arrived to the doorsteps of the house.]
- {5} in **nekega dne** se kaj zgodi ? // ja se eden vrne [and one day happens what? // yes one of them returns]
- {6} Odšla bom nazaj v Beograd in te nosila v srcu do konca **svojih dni** ... [I will return to Beograd and keep you in my heart until the end of my days ...]

The remaining three collocators were interpreted as Type B: not strictly phrasal units, but still indicative of corpus-specific language use. The co-occurrence *želeti-dan* [wish-day], for example, reveals expressions used in performative speech acts [27], e.g. *I/we wish you a (pleasant/nice) day*. The majority of these examples originate from the part of the corpus GOS that comprises radio program recordings, highlighting radio shows as a genre in which such greetings are especially frequently used {7}. Typical of radio shows is also the co-occurrence of words *danes* [today] and *dan* [day] in expressions introducing various “trivia of the day” {8}. On the other hand, a KRES-specific co-occurrence *začeti-dan* [(to) begin, day] reveals expressions typical of administrative language as it appears in various legal documents {9}.

- {7} nagrada je tvoja lep **dan** ti **želimo** ogromno sončka [the prize is yours we wish you a nice day a lot of sun]

- {8} **danes** je svetovni **dan** knjige [today is World Book day]
 {9} Ta zakon **začne** veljati petnajsti **dan** po objavi v Uradnem listu Republike Slovenije [this Act enters into force on the 15th day following its publication in the Official Gazette of the Republic of Slovenia] (/lit.: this Act begins on the 15th day /.../)

6 Conclusions and Future Work

In this paper, we presented a study of collocations in the corpus of spoken Slovene compared to the reference corpus of written Slovene. We used logDice-based collocation extraction available in NoSketch Engine and proposed the *CorpDiff* measure for identifying collocations characteristic only of one of the two corpora.

The extracted collocation candidates indicate differences in spoken and written Slovene on lexical and stylistic (genre-based) level. The results are potentially useful in theoretical and applied linguistics, e.g. in the fields of lexicography, translation, language teaching or discourse analysis, as well as for further development and use of natural language processing techniques for Slovene.

There are, however, some remarks to be taken into consideration. The two corpora are of very different sizes. Since our measure bases on the logDice score, appropriate for comparison of corpora of different sizes, this is not a problem per se. Anyhow, the limited size of the spoken corpus is problematic, as the number of collocations esp. for less frequent lemmas can be quickly too small to draw meaningful conclusions. This fact was one of the reasons for limiting the selection of lemmas for analysis only to the most frequent ones. Secondly, the linguistic analysis of the results has proven to be not only partially subjective (as discussed in 5.3), but also time-costly. To optimize the procedure, the number of false positive candidates should be reduced and the visualization of data improved. With this aim, we plan to sort the results also based on the collocator's part-of-speech tag, consider limiting the window for collocation extraction to -1/+1 (esp. for function words) or to apply *WordSketch* functionality using sketch grammars and evaluate the gain in precision and structuring of results compared to the price of robustness and independence of preprocessing tools. Furthermore, the presented methodology will be used for comparative extraction of collocations from other corpora, such as the corpus of Slovene user-generated content [26]. The comparison of the results will highlight the potential influence of the size and structure of the used language resources on the quality of the results.

Acknowledgements

The work described in this paper was funded by the Slovenian Research Agency within the national research project "Resources, Tools and Methods for the Research of Nonstandard Internet Slovene" (J6-6842) and infrastructure program "Centre for Applied Linguistics" (I0-0051).

References

- [1] Sinclair, J. (1966). Beginning the Study of Lexis. In Bazell, C., Catford, J., Halliday, M. A. K., and Robins, R., editors, *In Memory of J. R. Firth*, pages 410–430, Longman.

- [2] Sinclair, J. (1991). *Corpus Concordance Collocation*, Oxford University Press.
- [3] Halliday, M. A. K. (1966). *Lexis as a Linguistic Level*. In Bazell, C., Catford, J., Halliday, M. A. K., and Robins, R., editors, *In Memory of J. R. Firth*, pages 148–162, Longman.
- [4] Nesselhauf, N. (2005). *Collocations in a Learner Corpus*. John Benjamins Publishing.
- [5] Firth, J. R. (1957). Modes of Meaning. In *Papers in Linguistics 1934–1951*, pages 190–215, Oxford University Press, London.
- [6] Smadja, F. and McKeown, K. (1990). Automatically extracting and representing collocations for language generation. In *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, pages 252–259, Pittsburg, Pennsylvania.
- [7] Cowie, A. P. (1994). Phraseology. In *The Encyclopedia of Language and Linguistics* (6), pages 3168–3171, Oxford / New York.
- [8] Kjellmer, G. (1987). Aspects of English Collocations. In Meijs, W., editor, *Corpus linguistics and beyond*, pages 133–140, Rodopi, Amsterdam.
- [9] Benson, B., Benson, E., and Ilson, R. F. (1997). *The BBI Dictionary of English Word Combinations*: Revised edition. Benjamins.
- [10] Kilgarriff, A., Rychlý, P., Smrž, P., and Tugwell, D. (2004). The Sketch Engine. *Proceedings of Euralex*, pages 105–116, Loriet, France.
- [11] Krek, S. and Kilgarriff, A. (2006). “Slovene Word Sketch”. In Erjavec, T. and Žganec G. J., editors, *Proceedings of the 9th International Multiconference Information Society*, pages 62–67, Institut Jožef Stefan, Ljubljana.
- [12] Kosem, I., Gantar, P., and Krek, S. (2013). Avtomatizacija leksikografskih postopkov. *Jezikovne tehnologije, Slovenščina 2.0*, 1(2):139–164.
- [13] Vintar, Š. (2010). Bilingual term recognition revisited. *Terminology*, 16(2):141–158.
- [14] Pollak, S. (2014). Luščenje definicijskih kandidatov iz specializiranih korpusov. *Slovenščina 2.0*, 2(1):1–40.
- [15] Gantar, P. (2007). *Stalne besedne zveze v slovenščini: korpusni pristop*. Založba ZRC, Ljubljana.
- [16] Logar, N., Gantar, P., and Kosem, I. (2014) Collocations and examples of use: a lexical-semantic approach to terminology. *Slovenščina 2.0*, 2(1):41–61.
- [17] Kosem, I. and Verdonik, D. (2012). Key word analysis of discourses in Slovene speech: differences and similarities. *Linguistica*, 52(1):309–322.
- [18] Verdonik, D. (2014). Frazeološkost krščanskega izraza v vsakdanjem govoru. *Jezikoslovni zapiski*, 20(1):59–78.
- [19] Biber, D. (1988). *Variation across speech and writing*. Cambridge University Press, Cambridge.
- [20] Leech, G., Rayson, P., and Wilson, A. (2001). *Word Frequencies in Written and Spoken English*. Longman, London.
- [21] Henrichsen, P. and Allwood, J. (2005). Swedish and Danish, spoken and written language. A statistical comparison. *International Journal of Corpus Linguistics* 10(3):367–399.
- [22] Rychlý, P. (2008). A Lexicographer-Friendly Association Score. In *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2008*, pages 6–9.
- [23] Dice, L. (1945). Measures of the Amount of Ecologic Association Between Species. *Ecology* 26(3):297–302.
- [24] Verdonik, D., Kosem, I., Zwitter Vitez, A., Krek, S., and Stabej, M. (2013). Compilation, transcription and usage of a reference speech corpus: the case of the Slovene corpus GOS. *Language resources and evaluation* 47(4):1031–1048.
- [25] 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*. Zavod za uporabno slovenistiko, Ljubljana / Fakulteta za družbene vede, Trojina.
- [26] Fišer, D., Erjavec, T., Zwitter Vitez, A., and Ljubešič, N. (2014). JANES se predstavi: metode, orodja in viri za nestandardno pisno spletno slovenščino. In *Proceedings of the 9th Language Technologies Conference*, pages 56–61, Institut Jožef Stefan, Ljubljana.
- [27] Austin, J. L. (1962). *How to do things with words*. Harvard University Press, Cambridge.

Myslím, já myslím, myslím, že... Forms and Functions of the Verb *myslet* in Spoken Czech

Petra Poukarová

Faculty of Arts, Charles University in Prague, Czech Republic

Abstract. From the perspective of meaning, verba cogitandi do not seem to be an especially complicated group. They are frequently used and usually easily understood. However, if we want to describe this topic systematically, the situation is not so clear and straightforward. This paper attempts to analyse one verb from the category of verbs of thinking – *myslím* (*I think*), namely its forms and functions which can be used in common spontaneous conversation. I compare examples of this verb form with examples found in the corpus of written Czech. The study strives to employ a corpus-driven approach; the source data have been extracted from the ORAL corpus of spontaneous spoken Czech and from the corpus of written Czech SYN2010.

1 Theoretical Background

In the ORAL corpus, the infinitive form *myslet* is the most frequent verb of thinking and the form of the first person singular *myslím* is the most frequent word form of the lemma *myslet* – it accounts for 58% of the occurrences of the lemma *myslet*. Table 1 shows the distribution of lemmas and word forms of the analysed words in the ORAL corpus and in the SYN2010 corpus of written texts.

	lemma <i>myslet</i>	word form <i>myslím</i>	word form <i>myslím</i> (i.p.m.)	word form / lemma <i>myslím/myslet</i> (%)
ORAL	11,862	6,888	1158.17	58.00%
SYN2010	83,835	27,560	226.52	33.00%

Table 1. The number of occurrences of the analysed words in ORAL and SYN2010 (lemmas, word forms, i.p.m. of analysed words and ratio of the word forms to lemmas).

The high frequency goes hand in hand with the specificity of this word, which can be understood from different points of view:

Definability: Wierzbicka speaks about “indefinability” of the verb *myslet* in general, which means that there is no need to define such an evident word, because we could obscure its sense with the attempted definition [15, p. 61]. She considers the verb *myslet* a semantic primitive: “The elements which can be used to define the meaning of words (or any other meanings) cannot be defined themselves; rather, they must be accepted as ‘indefinabilia’, that is, as semantic primes, in terms of which all complex meanings can be coherently represented” [15, p. 22]. Čmejrková describes the verb *myslet*, or rather the form *myslím*, as an “unspecified expression” [3, p. 325]. We might find an explanation of this “indefinability” or “unspecifiability” in the process of thinking itself. As this internal activity takes place in our minds, we have no sensory access to it; the result of this activity is not evident.

Syntactic-semantic relations: The verb *myslet* belongs among elementary stative predicators and can fulfil the role of predicate in sentences. In general, only verbs (in their finite forms) can be used as predicates in Czech, because they are able to express certain dynamic aspects of an object (situations, states, processes and their changes) and they provide information about the communication situation (tense, mood) [13]. We can say that the form *myslím* questions this definition, because the use of the present tense does not always indicate that the speakers hold an opinion in the particular moment of their speech. Grepl and Karlík classify the verb *myslet* as a state predicate with the meaning of “mental possession” (in this case: information) [5, p. 100], whereas Nebeská describes it as a mental predicate which expresses the attitude of speakers to some information, the degree of certainty of its validity (*myslím* = I do not know it for sure, it is hypothetical vs. *vím* = absolute certainty). Based on these mental predicates, we can form some opinion about the world surrounding us [11, p. 96]. Wierzbicka also uses the term “mental predicate” [15, p. 136].

The typical valence frame of the verb *myslet* is *subject* + “*myslet* (*si*)” + *something* (information in the accusative or expressed with a subordinate clause: *myslím, že*). In real data (both written and spoken), we can observe that the most frequent variant is expressed by the structure *myslím, že*, whereas examples with the presence of the object complement occur frequently in written texts.

But both absence and occurrence of an object in a certain form enable us to “read” or recognise the meaning of the used verb form.

One form, more functions: The dictionary [7] describes the main (but marginal with respect to the frequency of occurrence, see the section with results) meaning of the verb *myslet* as follows: *vědomím postihovat skutečnost a její vzájemné vztahy; přemýšlet¹ (myslí v protikladech, myslí mu to dobře)*; this meaning captures the activity in our brain, not the state as mentioned above. Oxford Dictionaries (<http://www.oxforddictionaries.com>) describe the following meaning: “Direct one’s mind towards someone or something; use one’s mind actively to form connected ideas.” These definitions, based on an abstract concept, are relatively vague. This uncertainty in description could be the reason for the semantic bleaching of the forms *myslím, že* and *myslím* [6, p. 140] and of the shift from full semantics to:

1. an epistemic quantifier, introducing subordinate clauses [3, p. 327],
2. parenthesis,
3. an epistemic particle (which means positive modality on the one hand – something is likely to happen – and negative modality on the other – it is not sure whether it happens)
4. or lexical fillers.

These forms do not usually participate in syntactic structure; however, [4] introduces the so-called “third syntactical plan”, according to which the form *myslím* expresses the attitude towards the communicated situation from the point of view of modality [5, p. 660].

Cooperative principle: While the grammatical roles of the form *myslím* are weakened, its pragmatic aspect is evident. The speakers distance themselves from the verity of the information (they are more than 50% sure that the situation they are speaking about is true). This corresponds to the maxim of quality [8, p. 193]:

¹ Due to its considerable vagueness and a high level of abstraction, only a rough translation of this definition can be provided: *to grasp reality and its relations by using cognition, to ponder sth.*

1. Do not say what you believe to be false.
2. Do not say that for which you lack adequate evidence.

At the same time, it denies the request of a clear response.

The aim of this study is to examine the word form *myslím* in the corpus of spoken Czech and to try to describe its functions in spontaneous communication.

2 Methodology

The spoken data were extracted from the ORAL corpus of spontaneous spoken language, which includes the ORAL 2006, 2008 and 2013 corpora. Spontaneous spoken language is the means of communication typically used in the family circle, among friends and close people in general, in other words in such cases in which the speakers are minimally self-conscious about the formal attributes of their speech. It is often associated with a lack of preparation, improvisation, multi-party interaction, informality. The recordings for the ORAL corpora come from all parts of the Czech Republic (Bohemia, Moravia, and Silesia). In addition to the word form *myslím*, its pronunciation variants were also searched for, such as *myslim*, *mysím*, *mysim* etc.; a total of 6,888 occurrences were found in the corpus. For the purposes of this preliminary study, a sample of 700 occurrences was manually analysed.

There are 27,560 occurrences of the word form *myslím* in the SYN2010 corpus². Because of the prevalence (second most often used – the most common was the occurrence of a punctuation mark) of right-sided collocation candidates in the form of the epistemic quantifier “*myslím, že*” (at positions 1–3 from the KWIC there were 21,272 cases), we decided to focus on the occurrence of *myslím* on its own. After removing the cases of *myslím, že* by using a negative filter (at positions 1–3 from the KWIC, the construction , *že* was removed), we were left with a remainder of 6,526 cases of the word form *myslím*, from which we chose a random sample of 700 cases of *myslím* (this corresponds to the size of the sample from the ORAL corpus).

3 Results

3.1 *Myslím* in Spoken Texts

Myslím, že: The sample of 700 occurrences of the word *myslím* contains 485 occurrences of the phrase *myslím, že*. More than a half of all occurrences indicate subordinate clauses, where the phrase *myslím, že* is the least important part of the message, in other words it has the lowest information value (Trávníček 1951, cited in [3, p. 327]). Interestingly enough, the conjunction *že* does not always stand right next to the verb (e.g. *taky vždycky myslím, že, že po ní skočím tady; ale já si myslím, že to nedostane...*), but there can be other words between them (the linking is not so strong). The explanation might be found in the nature of the recordings in the ORAL corpus – in spontaneous speech, the syntactical structures are looser than in written texts. Several examples follow: *to si myslím skoro že tím nebude; já myslím i Lenka s Karlem že si to užili; myslím Hamburk, že tam byl.*

We can also find examples where the uncertainty about the truth value of the statement is emphasized with another expression, such as *myslím, mám tušení, že, akorát si myslím,*

² For more information about these corpora see: <http://wiki.korpus.cz/doku.php/cnk:uvod>.

že *by to chtělo* (the use of a synonym and a conditional); *no a ty tam ale no tak to si myslím jako že nemusíš asi; se snaží jezdit domu . no a . asi myslím*³. On the other hand, we can find examples which are seemingly illogical: *já si myslím že i já vim; já myslím že určitě jo; já já myslím že určitě* (the examined verb in combination with expressions of absolute certainty, such as *vim, určitě*). It could be caused by the aforementioned weakening of semantics [3, p. 327].

Myslím without an object: *Myslím* without an object adds up to 198 occurrences in the sample. It appears in the function of a particle expressing the speaker's attitude, it plays a pragmatic role and can be considered as a parenthesis. According to Karin Aijmer, "only non-factive predicates in the first person, for example, can be used parenthetically" [1, p. 7]. *Myslím* can replace a particle with the same meaning: *asi, pravděpodobně, nejspíš*; and vice versa, we can explain the meaning of *myslím* using these particles: *ale já v tom měla nějak drezink, myslím jogurtovej; je tam podepsaná myslím; myslím ten ostrov se menuje, to už sem zapomněl; železnice tam myslím není*.

Due to its particle function, the form *myslím* can occupy different positions in the sentence: between adjective and noun, between the individual members of a compound predicate etc.: *jo tak to sem se myslím dívala; někde byla, i taky myslím v těch Srbech taky byla; tak bych myslím ... vypila kdeco*. Regarding the definition of the main meaning of the word *myslet* (see above 1), no evidence was found in the sample to support this definition.

Myslím with an object: This structure has been found only in 17 cases in the sample. The complement is expressed by the accusative (*až z Horní Dobrouče myslím tydlety; no myslím jako to že*; on the level of collocation *a to myslím vážně; jo viš co myslím*), and in several cases also in the nominative (*no já myslím ta blondýna ze Sedmpan; myslím ten svatej no*). In these cases, the structure performs a similar function as a particle. The words *taky / tak / ale (myslím)* are the most frequent left-sided collocation candidates.

3.2 Myslím in Written Texts

The annotations of texts in the SYN2010 written corpus, or rather metalinguistic data about individual texts, supply us with information about the distribution of words across (among other things) text type groups. The form *myslím* can be found not only in fiction (16,680 cases from a total of 27,560 cases, 61%), journalism (7,729 cases, 28%), but also in non-fiction (3,151 cases, 11%). We originally assumed it would occur only in fiction and journalism, because of the specific context of the word form in the first person, for example: reported speech in fiction and interviews, first-person narrative, inner monologues in fiction, etc. Occurrence in scientific literature is explained by the fact that in the SYN2010 corpus, this text type group included professional and hobby magazines, which commonly feature interviews. Nevertheless, it remains the category where *myslím* is by far the least common and shows the smallest average reduced frequency, 496.49 (compared with 4985.71 in fiction and 2187.85 in journalism).

³ Punctuation marks such as . or .. indicate a short or longer pause in the speech. When a comma is used (only in the older corpora ORAL 2006 and 2008), it is motivated by syntactical structure, whereas ORAL 2013 does not contain commas.

Myslím, že: The phrase *myslím, že* was not included in the analysed sample, but its occurrence is not without interest. With the examples from the ORAL corpus, we were talking about the fact that in spoken texts, the syntactic frame is not as rigid, therefore additional words can appear between *myslím* and the conjunction *že*. However, because of the nature of spoken language, the distance between these two components is not as large as it may be in written texts. A couple of examples follow: *Furt si myslím – leda že se to nesečkává s tvým velebným souhlasem, že by to tak bylo nejúhlednější.*; “... i když myslím,” *opět zvážněl a vrátil se ke svým starostem, “že Geoffreymu mnohem víc záleží na dobytí Normandie”*. Because of the way written text is processed (we can jump back and forth in the text and the perception does not happen in “real time”), the greater distance of syntactically related components does not pose an issue.

Myslím: As far as the word form *myslím* alone is concerned, we can explain its meaning using a couple of synonyms: the particle *asi* (“maybe”; 278 occurrences from a total of 665 cases taken into account⁴: *Takhle jsem to myslím říkal.*; *pocit bezmocnosti, který je myslím jedním z nejhorších pocitů vůbec*), the phrase *mám na mysli, míním* (“what I mean is”; 107 occurrences, quite often specifying the meaning in association with one of the pronouns *to, tím, čímž*, and the whole expression appears bracketed or between hyphens: *Doufám, že Ti už nějaká spadla do klína. (To myslím symbolicky.)*; *Napadlo mě, že mezi mými známými, čímž myslím manžele svých přítelkyň, bylo jen poskrovnu mužů, kteří se mi líbili*). In 5 cases the verb *myslím* is used to express the thinking process, the workings of human mind, ex.: *mám mozek, myslím, žiju, dejchám; Jestliže myslím jen tváří v tvář smrti proč myslím?*

The meaning of the verb is expressed by adding a preposition phrase as well. There are 68 cases of combinations with the preposition *na* (in the meaning thinking of someone, reminiscing). These cases do not appear in the spoken sample at all (e.g.: *Aspoň bude vědět, že na ni myslím.*; *A také myslím na to, že je naší povinností*).

There are collocations in the sample of written texts which do not appear in the spoken sample. We will outline an overview of those we found during our text analysis, and later on we will focus in more detail on the 3 types that appeared in both the spoken and written samples, albeit with varying frequency. The following phrases appeared in the written texts: *taky myslím, to si myslím (I think so), aspoň myslím (I guess), to myslím vážně (I am serious), myslím to s ním dobře (I mean him/her well)*.

Common collocations for both kinds of samples were: *aspoň myslím, víš/víte/..., jak to myslím, to myslím + adv.*

Aspoň myslím was used to signal relativization of a previous communication only twice in spoken texts, while being used for the same purpose 26 times in written texts. (SYN2010: *a taky jsem si absolutně jistej tím, co jsem viděl. Teda aspoň myslím.*, ORAL: *to na počítačích dole nemáme tenhle program. teda aspoň myslím*).

We found an even greater difference between the following two cases: the collocation *víš/víte... jak to myslím (you know what I mean)* appeared 4 times in spoken texts, always

⁴ From the sample of 700 occurrences of the form *myslím*, we removed a few remaining instances of the construction *myslím, že* where the conjunction *že* was on the 4th or farther position from the KWIC, or in examples where the verb form *myslím* and conjunction *že* were not separated by a comma (an example from poetry: *nevycházím ze dveří Honí se ponebí chciplý pes myslím že byla to Lajka*).

preceded by a form of the verb *vědět* in the 2nd person singular. By contrast, it appeared 15 times in written texts, introduced by the verbs *vědět*, *rozumět* a *chápat*.

The phrase *to myslím + adv.* appeared only once in the spoken sample (*a to myslím vážně*), whereas it appeared 36 times in written texts. With the sole exception of *myslím to upřímně*, only *vážně* occurs in the adverbial position, even though it can be further modified in various ways: *fakt/opravdu (really) / naprosto (totally) / smrtelně (mortally) vážně*.

Last of all we will take a look at a seemingly contradictory phrase, whose occurrence we did not foresee in written texts, yet which occurred nevertheless. In our analysis we took into account all occurrences of the verb form *myslím* in the written corpus and we found 195 cases. The first place in terms of frequency was taken by the phrase *myslím, že vím* and its variant *myslím, že to vím*, which occurred 110 times. (E.g.: *myslím, že vím, jak byl spáchán*). Furthermore, the co-occurrence of the verb *myslím* and a certainty adverb is not without interest: *myslím, že určitě jich pár spadlo; myslím, že asi určitě potom přestoupí* (from the magazine *Sport*, an interview); *myslím, že se určitě mylíte* (from fiction, an expression of the speaker's politeness), *myslím, že téměř jistě vyrazím na Žilinu* (the use of the phrase *téměř jistě* indicates vagueness); *ale rozhodně si myslím, že to neznamená konec kapely* etc.

4 Conclusion

As we have seen, the objective of this really preliminary study is to investigate the particular instances of the verb form *myslím* in one written and one spoken corpus (SYN2010 and ORAL). For the purpose of comparison, we tried to compare some of the aspects of the spoken texts with the written sample. The analysis of these examples is not at all exhaustive, it is more of a roadmap of plans for future more detailed studies and comparisons.

In the sample from SYN2010, there appeared examples of *myslím* in the role of epistemic quantifier (these were not included in the random sample) and epistemic particle expressing uncertainty of the information from the speaker's point of view; these belong to the field of pragmatics. This was in accordance with spoken texts. The occurrence of phrases involving prepositions (*myslím na*) was much more common in the sample of written data than in the spoken one, as was the occurrence of set collocations, which appeared only rarely in spoken texts.

What we did not expect in the written corpus (in contrast to spoken data) was the occurrence of seemingly illogical phrases like *myslím, že vím* (this type was the most common in our sample of 195 cases). The association with certainty adverbs *jistě*, *určitě*, *bezesporu* was also common.

In conclusion, it is our opinion that it would be beneficial to carry out a more detailed comparison using statistical methods (between spoken and written texts, within the text types or text groups, and a comparison between fiction and spoken texts on the one hand and spoken texts and non-fiction on the other). It could be interesting to focus on the position of the verb form *myslím* within sentences and sound units, and take notice of the phrase *já myslím*, which is the most common left-sided collocation candidate in spoken texts, but only the 9th most common in written texts.

Acknowledgements

The article was written under the auspices of grant project No. 15-011165 (Syntax mluvené češtiny) [Syntax of spoken Czech]), supported by the Grant Agency of the Czech Republic.

References

- [1] Aijmer, K. (1997). I think – an English modal particle. In *Modality in Germanic languages: Historical and comparative perspectives*, 99(1):1–47.
- [2] Benešová, L., Křen, M., and Waclawičová, M. (2013). *ORAL2013: reprezentativní korpus neformální mluvené češtiny*. Ústav Českého národního korpusu, Filozofická fakulta Univerzity Karlovy, Praha. Accessible at: <http://www.korpus.cz>.
- [3] Čmejrková, S. (2008). Výraz myslím v psaném a mluveném textu. In *Grammar and Corpora 2007*, pages 325–344, Academia, Praha.
- [4] Daneš, F., Grepl, M., and Hlavsa, Z. (1987). *Mluvnice češtiny 3*. Academia, Praha.
- [5] Grepl, M. and Karlík, P. (1998). *Skladba češtiny*. Votobia, Olomouc.
- [6] Grepl, M. (2011). Funkce částic (partikulí) ve výstavbě českých výpovědí. In *Jak dál v syntaxi*, pages 139–144, Host, Brno.
- [7] Havránek, B. et al. *Slovník spisovného jazyka českého*. Accessible at: <http://ssjc.ujc.cas.cz/>.
- [8] Hirschová, M. (2013). *Pragmatika v češtině*. Karolinum, Praha.
- [9] Kopřivová, M. and Waclawičová, M. (2006). *ORAL2006: korpus neformální mluvené češtiny*. Ústav Českého národního korpusu, Filozofická fakulta Univerzity Karlovy, Praha. Accessible at: <http://www.korpus.cz>.
- [10] Křen, M., Bartoň, T., Cvrček, V., Hnátková, M., Jelínek, T., Koček, J., Novotná, R., Petkevič, V., Procházka, P., Schmiedtová, V., and Skoumalová, H. (2010). *SYN2010: žánrově vyvážený korpus psané češtiny*. Ústav Českého národního korpusu FF UK, Praha. Accessible at: <http://www.korpus.cz>.
- [11] Nebeská, I. (2001). Některé mentální predikáty. In *Obraz světa v jazyce*, pages 96–105, Filozofická fakulta Univerzity Karlovy, Praha.
- [12] Oxford Dictionaries. Accessible at: <http://www.oxforddictionaries.com>.
- [13] Rusínová, Z., Nekula, M., Grepl, M., and Karlík, P. (1995). *Příruční mluvnice češtiny*. Nakladatelství Lidové noviny, Praha.
- [14] Waclawičová, M., Kopřivová, M., Křen, M., and Válková, L. (2008). *ORAL2008: sociolingvisticky vyvážený korpus neformální mluvené češtiny*. Ústav Českého národního korpusu, Filozofická fakulta Univerzity Karlovy, Praha. Accessible at: <http://www.korpus.cz>.
- [15] Wierzbicka, A. (2014). *Sémantika: elementární a univerzální sémantické jednotky*. Karolinum, Praha.

Czech Deadverbial Prepositions Used in Postposition to a Noun (Exemplified in Corpus SYN 2010)¹

Aksana Schillova

Faculty of Arts, Charles University in Prague, Czech Republic

Abstract. The article deals with usage of postpositions in the Czech language. It is considered that postpositions are units fulfilling in a syntagm a function equivalent to a function of prepositions but used in postposition to a controlled noun. The research is based on the Czech national corpus (Corpus SYN 2010). The article concerns search of postpositions in this corpus and the following analysis of their frequency lists, collocation candidates lists, and concordances. Special attention is paid to deadverbial units *navzdory*, *napospas*, *vstříc*, *naproti* which according to the corpus are regularly used not only in preposition but also in postposition to a noun (or its substitutes). The subject of the analysis is frequency, combinatorics and syntagmatic function of these units used in postposition. Problematic issues and main directions of the following research of the Czech postpositions are outlined.

1 Introduction

As the term itself suggests, a position of a preposition before a noun (or its substitutes) is axiomatic in the Czech language. In the contemporary Czech grammar books this fact is not even discussed [1], [2], [3], [4]. Postpositions, i.e. prepositions used in a postposition to a noun (or its substitutes), are mentioned, for instance, in the works by F. Čermák but only as a phenomenon typical of the other languages such as Finnish, Basque etc. [5], [6].

Regarding the Czech language, the existence of postpositions in it is still an unexplored area. So far just one reference to this subject has been found in the specialized literature. This is a passage about facultative moving of the Czech prepositions to the end position in a nominal group [7]. This refers to prepositions originated from the verbs, for instance *počínaje*, *konče*, *nevýjímaje* etc. In addition to everything else, the author of the passage states that “there is no semantic, stylistic or frequency of use difference between a position before a controlled noun and a position behind it” (translation by A. S.). There are several examples to compare: *Zopakoval celou látku, počínaje antikou a konče současností* and *Stavební slohy, románským počínaje a barokem konče, se tady na malém prostoru prolínají* [7, p. 803].

Interesting “findings” also appear in defining dictionaries. For instance, the dictionary entries of a unit *vstříc*, which is defined as an adverb, include the following illustrative examples: *Paní šla svému manželovi vstříc až na první schodiště terasy*. J. John [8]; *Jít, kráčet, spěchat někomu vstříc* [9]; *Vyšel návštěvě vstříc před dům* [10]. Compare with the example of *naproti* defined as an adverb: *Vzal hůl a čepici a šel dceři naproti*. V. B. Třebízský [8]. It is evident that the above mentioned examples of *vstříc* and *naproti* are interpreted in the dictionaries as the examples of adverbs because there is no noun (or its

¹ This article was created within the project “Promotion of research projects aimed for linguistic disciplines and study of literature” and the subproject “Linguistic analysis and creation of the list of prepositions used in a postposition to a noun or its substitutes (on the basis of the Czech National Corpus)” solved at Charles University in Prague from the Specific university research in 2015.

substitutes) right after them. But it is also obvious that if in these examples we put *vstříc* / *naproti* before a noun (or its substitutes), the meaning of these constructions will not substantially change. In addition, in this context *vstříc* and *naproti* can be replaced by synonymous prepositions *k* or *směrem k*. Compare: *šla svému manželovi vstříc* and *šla vstříc svému manželovi*, i.e. *šla ke svému manželovi* or *šla směrem ke svému manželovi*; *šel dceři naproti* and *šel naproti dceři*, i.e. *šel ke své dceři* or *šel směrem ke své dceři*. Based on the above, it can be assumed that the above mentioned examples from the dictionaries are examples of prepositions used in postposition to a noun (or its substitutes) but not examples of adverbs.

Corpus data (discussed in the next section) also show that in the Czech language a group of units, which resemble postpositions by their combinatorics and function, is used. The relevant objective is to ascertain firstly, whether these units are postpositions in fact, secondly, whether they differ (semantically, stylistically, by their frequency of use etc.) from their homonymous or complementary prepositions, thirdly, to ascertain their place in the system of the Czech synsemantics. The ultimate task is to create a comprehensive list of the Czech postpositions of different types. In addition to the above mentioned postpositions originated from verbs and adverbs, some other postpositions, which have not been registered yet, are likely to be used. Likewise, there may be used not only one word but also multiple word postpositions. This needs to be researched and clarified.

Due to the vastness of this new subject, in this article we are going to focus on one type of units, which, as may be supposed, are used in a function of postpositions, namely on units homonymous to adverbs or originated from adverbs. Since these units have not been the subject of a particular research yet, the primary objective of analysing these units is to determine a set of these units, to ascertain their frequency of use in a representative corpus and to describe their combinatorics and syntagmatic functions. We will try to accomplish this task in the second part of this article using Corpus SYN 2010 [11]. The analysis will enable us to outline the basic problems connected with the usage of these postpositions in the Czech language. The following research can be aimed at solving these problems. Research perspectives will be described in the final section of the article.

2 The Frequency of Use, Combinatorics and Syntagmatic Function of the Deadverbial Postpositions

The performed corpus research showed that in a postposition to respective case forms of nouns (or their substitutes), the following units are regularly used: *navzdory*, *napospas*, *vstříc*, *naproti*. In the Czech grammar books, these units have a double interpretation. They are listed among prepositions as well as among adverbs. The main factor of disambiguation of these homonyms is the presence or absence of nouns (nominal groups) in the position behind them. Thus, if a noun (nominal group) follows the unit in the form of the respective grammatical case, it is a preposition, but if the noun (nominal group) doesn't follow the unit, it is an adverb [1], [2], [3], [4].

In the corpus SYN 2010 usages of these units are also specified either as usages of prepositions or usages of adverbs according to the same principle. Therefore, we searched for postpositions among the usages specified as usages of adverbs, i.e. through CQL queries, we searched for usages of non-prepositional case forms of nouns and pronouns at a distance of 0-5 positions before the relevant units without punctuation marks between

them. Afterwards we analysed the frequency and collocation candidates lists of obtained concordances and the usages of the units. Below we summarize the results of this analysis in relation to each of these units.

2.1 Navzdory

The unit *navzdory* has in the representative corpus SYN 2010 the following frequency: 4594 usages in the function of a preposition, 121 usages in the function of an adverb.

Among these 121 “adverbial” usages, there are 81 usages in which in the contact position in front of this unit (1L) appear nouns / pronouns in the form of the dative. The upper part of these words frequency lists is occupied by pronouns: 1. *všemu* (25), 2. *jemu* (5), 3. *všem* (4), 4. *sobě* (3), 5. *jí* (2). Followed by single usages of different nouns and other pronouns: *škůdcům, škole, šeru, zájmům, zlodějům, mu, jim, vám*, etc. E.g.: *Známe tuhle scénu z kouzelně snímaných filmů, když režisérům najednou napadne, že z fantasticky hrajících nepřátelských bratří udělají kamarády, kteří spolu půjdou cestou necestou všemu navzdory skrz tisíce dobrodružství*. For comparison: collocation *navzdory všemu* has a frequency of 109 usages, e.g.: *Svou ženu navzdory všemu miloval a respektoval, třebaže na její politickou činnost a bouřlivý milostný život sám doplácel*. Or see the following examples of usages with a noun: *obavám navzdory* – 1 usage: *Obavám navzdory se stávám vítaným hostem na nefalšované islámské svatbě – veselce plné nazdobených svatebčanů, roztomilých družiček, hudby, tance a bujarého veselí v jinak poklidné poušti* / / *navzdory obavám* – 5 usages, e.g.: *Oficiální přechod v této části hranice neexistuje, a tak jsme navzdory obavám usedli do rychločlunu černošského pašeráka, abychom jej po několika hodinách zběsilé noční jízdy vyměnili, už na území Paraguaye, za rybářskou bárku*.

It is evident that the usages of *navzdory* in a postposition to a noun / a pronoun do not semantically and stylistically differ from similar usages of the same unit in a position in front of the noun / the pronoun, i.e. in both cases, the unit performs “prepositional” function. With that, syntagmatic function of a postposition is identical to syntagmatic function of a homonymous preposition. In this case, it is an adverbial or more propositional function of the type PROP-S (about prepositional functions see [5]).

At the same time the postposition *navzdory* in all usages is in the contact position with the noun / pronoun, i.e. two-component collocation has the following structure: SUB / PRON (D) + NAVZDORY (compare with the possible distant positions of *napospas*, *vstříc* and others below). But the frequency difference between the pre- and postpositional usages of the unit *navzdory* is substantial: 4594 vs. 81.

2.2 Napospas

The unit *napospas* has a frequency in the corpus of 231 usages as a preposition and 100 usages as an adverb.

Out of 100 “adverbial” usages, there are 62 usages in which this unit appears in the postposition (contact, but also distant) to the corresponding noun / pronoun. Usages in contact postposition (SUB / PRON (D) + NAPOSPAS) count 32 cases. E.g.: *Se slzami v očích si pomyslela na všechny ženy, které jsou vydané mužům napospas*. Usages in distant postposition (SUB / PRON (D) + VERB + NAPOSPAS, SUB / PRON (D) + ADV + NAPOSPAS etc.) count 30 cases. E.g.: *Když jste se však tomu pocitu vydali napospas, už jste jej déle nevnímali jako břímě; Nemůže ho silou donutit, aby odešel, a pokud ji teď*

neposlechne, bude bezmocná a vydaná mu zcela napospas; Jestli se někdy máme osvobodit od všeho útlaku a křivd, kterým jsme vydáni napospas, musíme za své osvobození zaplatit.

The upper part of the frequency list of nominal collocates (contact – 1L and distant – 2L to 5L) of the postposition unit *napospas* is occupied by pronouns as in the previous case (see navzdory), compare: 1. *mu* (10), 2. *sobě* (4), 3. *všemu* (2), 4. *všem* (2), 5. *přírodě* (2), there are also single usages of various nouns and other pronouns (*osudu, smrti, útočníkům, větrům, vlnám, vichru, veřejnosti, jim, nám*, etc.).

Nominal collocates of the unit *napospas* are represented in its prepositional usages mainly by nouns. Compare with the top of the frequency list: 1. *osudu* (26), 2. *vlnám* (4), 3. *světu* (4), 4. *přírodě* (4), 5. *myšlenkám* (4), 6. *živlům* (3), 7. *větru* (3), 8. *smrti* (3), 9. *člověku* (2), 10. *válce* (2), etc. The most frequent collocation verbs in both cases (in the preposition and the postposition of the unit *napospas*) are the same: *vydat, nechat, ponechat, dát, vynechat*, etc. Compare: *Nechali nás útočníkům napospas ve chvíli, kdy jsme je potřebovali / / Náš doprovod se rozprchl a nechal nás napospas útočníkům, kteří nás zajali a odvedli s sebou.*

Syntagmatic function of this unit as a preposition and postposition is primarily verbal (V-S). Frequent predicative collocates of the unit *napospas* (in the preposition and postposition) are also participles derived from the above mentioned verbs: *vydaný, ponechaný*, etc.

2.3 Vstříc

The unit *vstříc* has in the corpus SYN 2010 the following frequency: 992 usages in the function of a preposition, 1270 usages in the function of an adverb.

Out of these 1270 “adverbial” usages, there are 1148 usages (i.e. 90%) in which this unit is preceded within the same clause by a noun or a pronoun in the form of dative without a preposition. The usages of these nouns / pronouns in the contact position before the unit *vstříc* (SUB / PRON (D) + VSTRĚC) count 342 cases. E.g.: *Rozběhl se k posledním dveřím, a když je prudce otevřel, stála před ním a rozběhla se mu vstříc.* The usages in the distant position (SUB / PRON in position 2L-5L from the unit *vstříc*) count 806 cases. E.g.: *Pustila závěs a krůček po krůčku mu kráčela vstříc, na rtech smutný úsměv, ale oči jak jasná ocelová zrcadla, v nichž viděl vlastní nenasytý obličej; Trautman a jeho čtyři kamarádi nevycházeli z údivu, když vešli do chýše a Mike, který ještě před několika hodinami ležel těžce nemocný a vyčerpaný na lůžku, jim vesele vyšel vstříc; Kapitán nám nyní řekl, že pro jistotu vyslal s majorem, který se nalézal u velitele bataliónu, Dusila, aby nám tento vyjel vstříc a ukázal cestu.* The most frequent structural type of these distant usages is the following type: SUB / PRON (D) + VERB + VSTRĚC – 343 usages, i.e. 42% of all cases. Compare with the top of the frequency list of this type collocations: 1. *mu vyjít (lemma) vstříc* (43), 2. *jim vyjít (lemma) vstříc* (34), 3. *nám vyjít (lemma) vstříc* (25), 4. *mi vyjít (lemma) vstříc* (18), 5. *jim vycházet (lemma) vstříc* (13), etc. Personal pronouns occupy the upper part of the frequency list of dative forms of nouns / pronouns in collocation with which appear the unit *vstříc* in the contact and distant postposition. Compare: 1. *mu* (190), 2. *jim* (139), 3. *nám* (110), 4. *mi* (92), 5. *jí* (85), etc.

The unit *vstříc* is used most frequently in the postposition in the collocations with pronouns. Compare: PRON (D) + VSTRĚC (contact and distant postposition) – 794 usages / / SUB (D) + VSTRĚC (contact and distant postposition) – 354 usages. And the most frequent noun collocates are: 1. *lidem* (13), 2. *zákazníkům* (11), 3. *smrti* (11), 4.

požadavkům (9), 5. *klientům* (7), 6. *světlu* (5), 7. *rodičům* (5), 8. *slunci* (4), 9. *občanům* (4), 10. *čtenářům* (3), etc.

The unit *vstříc* in preposition vice versa is most often used in collocations with nouns – 944 cases out of 992 (i.e. 95%). The most frequent right collocates of the unit *vstříc* are: 1. *požadavkům* (42), 2. *smrti* (32), 3. *potřebám* (26), 4. *osudu* (25), 5. *budoucnosti* (18), 6. *slunci* (17), 7. *lidem* (13), 8. *zákazníkům* (9), 9. *dobrodružstvím* (9), 10. *světu* (8), etc. Excerpts from the frequency lists show that these collocation nouns are often the same. What differs, however, is the frequency of usages of these nouns in combination with the unit *vstříc* in preposition and postposition.

Collocation verbs of the unit *vstříc* are also the same in both cases (in preposition and postposition): *vyjít, vycházet, jít, vykročit, kráčet, běžet, vyrazit, vydat se, letět, plavat, spěchat*, etc. These are the verbs of motion, this fact corresponds to the meaning of this unit: *vstříc* means ‘movement in the direction of someone or something’ [9].

Syntagmatic function of the unit *vstříc* as a preposition and a postposition is verbal (V-S).

However, a certain part of usages of the unit *vstříc*, regardless of its position in relation to the noun (pre- or post), consist of examples in which the unit *vstříc* is a component of the verbal phraseological unit *vyjít / vycházet vstříc + D*, which means ‘to make something easier for somebody or to satisfy sb’s needs / demands / desires; to try satisfy sb’s needs / demands / desires’ (see [12, p. 905]). The number of these usages can be identified only by manual analysis of the contexts because usages of the unit *vstříc* as a component of the verbal phraseological unit and usages of the unit *vstříc* as a preposition or a postposition are homonymous. E.g.: out of 190 usages of collocation *mu + vstříc* (contact and distant postposition), there are 107 usages in which the unit *vstříc* is not a postposition but a component of the verbal phraseological unit *vyjít / vycházet vstříc*. E.g.: *A jelikož Koch podpořil Merkelovou v úsilí stát se kancléřkou, ta mu chce vyjít vstříc a odměnit jeho člověka.*

A set of noun collocates of the unit *vstříc* which is specified in the corpus as a preposition (see the illustration from the frequency list above), i.e. the unit *vstříc* is used in preposition to a noun, also shows that among prepositional usages of the unit *vstříc* take place usages where this unit is a component of above mentioned verbal phraseological unit but not a preposition. Compare: (*vyjít / vycházet vstříc*) *požadavkům, potřebám, lidem, zákazníkům*, etc. (as above). E.g.: *Divadlo na Vinohradech, jež se o hru zajímalo již předtím, poté přislíbilo, že vyjde vstříc všem požadavkům autora.*

Manual “filtering” of concordances for the purpose of identifying and analysing the actual frequency of usages of the unit *vstříc* firstly in the function of a component of a verbal phraseological unit (non-spatial meaning), secondly in the function of a preposition and thirdly in the function of a postposition (direct and figurative spatial meanings) will be one of the tasks in the continuation of this research.

2.4 Naproti

The unit *naproti* has in the corpus SYN 2010 the following frequency: 4047 usages in the function of a preposition (3956 usages + dative; 91 usages + genitive), 1227 usages in the function of an adverb.

Out of these 1227 “adverbial” usages, there are 444 usages in which in the position in front of this unit (1L-5L) nouns / pronouns in the form of dative without a preposition

are used (it is important to note that we did not find usages in which the unit *naproti* is in a postposition to a noun / pronoun in the form of genitive). Usages in a contact postposition (SUB / PRON (D) + NAPROTI) count 158 cases. E.g.: *Jack vystoupil z auta a šel jí naproti*. Usages in a distant postposition (SUB / PRON (D) in a position 2L-5L from *naproti*) count 286 cases. E.g.: *S ohromující náhlostí prolomil letoun spodní hranici bouřkových mraků, takže se pod ním objevila země, která jako by se mu řtila naproti*. The most frequent structural type of these distant usages is the type SUB / PRON (D) + VERB + NAPROTI (as well as in postpositional usages of *vstříc*, as above) – 123 usages, i.e. 43% of all cases. Compare with the top of the frequency list of collocations of this type: 1. *mu jít (lemma) naproti* (9), 2. *mi přijít (lemma) naproti* (6), 3. *mu přijít (lemma) naproti* (5), 4. *jí vyjít (lemma) naproti* (5), 5. *jim jít (lemma) naproti* (5), etc.

The top of the frequency list of the nouns / pronouns in a postposition (contact and distant) to which the unit *naproti* appears, occupy personal pronouns: 1. *mu* (103), 2. *jí* (60), 3. *jim* (52), 4. *mi* (49), 5. *nám* (15), etc.

The situation of the unit *naproti* in preposition is analogic. Compare: 1. *němu* (157), 2. *ní* (118), 3. *sobě* (117), 4. *mně* (99), 5. *nám* (42), etc.

But the most frequent collocational verbs of the unit *naproti* as a preposition and collocational verbs of the unit *naproti* as a postposition differ. *Naproti* in preposition has in its collocation candidates list mostly verbs denoting location or change of location like *sedět* (234), *stát* (151), *posadit se* (100), *sednout si* (54), *bydlet* (33), etc. The number of usages of the preposition *naproti* in collocation with verbs like *jít* (33), *přijít* (16), *vycházet* (11), *jet* (11) etc. is much less. But *naproti* in postposition has only verbs of motion in its collocation candidates list. Compare: *jít* (114), *přijít* (51), *vyjít* (38), *běžet* (26), *jet* (18), *přijet* (15), *vyběhnout* (13), *vydat se* (9), etc.

The differences are observed also among collocational nouns of the unit *naproti* in preposition and in postposition. *Naproti* in preposition makes collocations with concrete inanimate nouns like *židle*, *stůl*, *křeslo*, *zed'*, *ulice*, etc. *Naproti* in postposition makes collocations with animate nouns denoting people (e.g. *synovi*, *rodičům*, *přítelkyni*, *policistům*, *dělníkům*, etc.) and also with nouns denoting abstract terms (e.g. *štěstí*, *změněnám*, *vítězství*, *životu*, etc.).

Syntagmatic function of *naproti* as a preposition is either verbal (V-S) or adnominal (S-S). Compare examples: *Sedl si naproti Bondovi a jednu nohu přehodil přes opěradlo židle* (V-S) // *Bond se rozhodl pro židli naproti guvernérovu stolu a sedl si* (S-S).

At the same time *naproti* as a preposition mainly expresses spatial meaning of location. Sometimes it is also used in the meaning of 'movement direction'. E.g.: "*Později, "prohodil Slaughter přes rameno a šel naproti šedovlasému muži s fotoaparátem a magnetofonem*. But frequency of these usages will be significantly less (see frequencies of collocational verbs of the unit *naproti* as a preposition above).

The unit *naproti* in postposition performs only verbal syntagmatic function (V-S) and always expresses the direction of movement. E.g.: *Charlie se odlepila od osušky a beze spěchu mi kráčela naproti*; *Pak vylezl na strom a pozoroval slunce, a když začalo klesat mezi kopce, vypravil se mu naproti*; *Jsem-li za volantem a vidím-li před sebou psa, okamžitě zastavím, jdu mu naproti a nabízím, že ho na druhou stranu silnice převedu, popřípadě přenesu v náručí, bude-li si to přát*.

But some of the usages of the unit *naproti* regardless of its position to a noun / pronoun in fact are usages in which this unit is a component of the verbal phraseological unit *jít / přijít naproti*, which means 'to welcome, to see smb. in' (see [12, p. 263]). Compare

examples: *Jiří šel naproti Aleně na Wilsonovo nádraží, vracela se s ostatními děvčaty z tábora / / Tomas přišel Anně naproti na nádraží, noční vlak přijel na čas.*

For this reason it is needed to filter the concordances of the unit *naproti* manually with regard to the semantics of particular clauses, in order to determine and analyse filtered sets of usages of the unit *naproti* firstly as a component of the above mentioned verbal phraseological unit, secondly as a preposition and thirdly as a postposition. This task will be solved in the continuation of the research.

3 Conclusion

This first analysis of the corpus data in the case of the units *navzdory*, *napospas*, *vstříc*, *naproti*, shows that in the Czech language postpositions have their place as well as prepositions. But due to the polysemy of these units (see e.g. *vstříc*, *naproti*) and varying degrees of their connectivity with a noun or its substitutes (compare: *navzdory* – only contact postposition; *napospas*, *vstříc*, *naproti* – contact and distant postposition, as above), there is a need for further qualitative analysis of the obtained corpus statistics, i.e. to filter manually concordances with regard to the semantics and therefore a function of the units under study.

An important direction of this analysis will also be a comparison of semantics and valency (left and right) of prepositions with semantics and valency of respective postpositions. This comparison can give a possibility firstly to determine the interrelation between homonymous prepositions and postpositions (to ascertain if they are mutually complementary or if the change of a position is just a facultative matter dependent on the actual division of the sentence etc.), secondly to find an answer to the question why there are two-valent prepositions but their homonymous postpositions are only one-valent. E.g. *naproti* is a preposition which can be connected with genitive and dative case of the noun (or its substitutes), *naproti* as a postposition goes only with dative. The same thing is observed in the units *blízko*, *blíž(e)*, which are not included in this conferential article (due to page limit of the article).

The next point to solve is the interrelation between postpositions of different origin, e.g. between postpositions formed from verbs (see the introduction) and postpositions formed from adverbs.

We plan to deepen the research in the direction of solving above mentioned questions. And we also see a perspective direction of the research in comparison of Czech language material with language material of other Slavic and, if it will be possible, non-Slavic languages, i.e. in using of parallel corpora.

References

- [1] Komárek, M., Kořenský, J., Petr, J., and Veselková, J., editors. (1986). *Mluvnice češtiny II. Tvarosloví*. Academia, Praha.
- [2] Karlík, P., Nekula, M., and Rusínová, Z., editors. (1995). *Příruční mluvnice češtiny*. Nakladatelství Lidové noviny, Praha.
- [3] Cvrček, V. et al. (2010). *Mluvnice současné češtiny I. Jak se píše a jak se mluví*. Karolinum, Praha.
- [4] Šticha, F. et al. (2013). *Akademická gramatika spisovné češtiny*. Academia, Praha.
- [5] Čermák, F. (1996). Systém, funkce, forma a sémantika českých předložek. *Slovo a slovesnost*, 57(1):30–43.

- [6] Čermák, F. (2010). *Lexikon a sémantika*. Nakladatelství Lidové noviny, Praha.
- [7] Uhlířová, L. (2011). Slovosled jmenné skupiny: Předložka jako prostředek delimitace jmenné skupiny zleva. In F. Štícha, editor, *Kapitoly z české gramatiky*, pages 785–805, Academia, Praha.
- [8] Hujer, O. et al. (1935–1957). *Příruční slovník jazyka českého I-VIII*. Státní nakladatelství, Praha.
- [9] Havránek, B. et al. (1960–1971). *Slovník spisovného jazyka českého I-IV*. Academia, Praha.
- [10] Polendová, J. et al. (2011). *Slovník současné češtiny*. Lingea, Brno.
- [11] Český národní korpus – SYN2010. (2010). Ústav Českého národního korpusu FF UK, Praha. Accessible at: <http://www.korpus.cz>.
- [12] Čermák, F., Hronek, J., and Machač, J., editors. (2009). *Slovník české frazeologie a idiomatiky*. 3. Výrazy slovesné. Leda, Praha.

On Building the Slovak Automatic Semantic Role Labeling System

Ján Staš, Daniel Hládek, Stanislav Ondáš, and Jozef Juhár

Faculty of Electrical Engineering and Informatics, Technical University of Košice, Slovakia

Abstract. This paper describes the initial research towards the development of the automatic semantic role labeling system for the Slovak language. For this purpose, a new template-based semantic corpus of Slovak sentences SEMIENKO was created and prepared for training classifiers for chunking and automatic semantic role labeling. A set of three hundred example sentences was manually labeled with a small set of semantic roles according to annotation convention adopted from the work of Emil Pálaš. A newly designed hidden Markov model classifier trained on morphosyntactic patterns in presented semantic sentences with the Viterbi decoding algorithm have been used for statistical identification of chunks and automatic semantic role labeling. This is a next step in the process of automatic transcription of spontaneous Slovak speech to its understanding by a machine.

1 Introduction

One of the most critical parts of human-machine interfaces (HCI) in robotics and intelligent systems is understanding of human language by a machine. Thanks to the rapid expansion of speech and language technologies, we are able to process and recognize human language in a sufficient way, but in-depth understanding of meaning is still a big challenge in developing human-machine interfaces for intelligent computer-based and robotic systems, including multi-modal interfaces. Increasing number of humanoid robots, artificial agents or virtual assistants force us to look at the problem of understanding human speech and language and extracting semantic information from very large text databases.

Automatic semantic role labeling (ASRL) is an important step in extracting meaning from the text by giving a small set of labels to the words or group of words in a sentence. Semantic role labels that represent meaning can be considered as an appropriate way to enable understanding human language by machines. The relationship among verbs and other words need to be understood to extract meaning from the text.

The first ASRL system was reported by Daniel Gildea and Daniel Jurafsky in 2002 [5]. They use lexical resources such as FrameNet [2] and WordNet [3] and statistical techniques for extraction of semantic information from a small initial database. Their idea was to train statistical classifiers for automatic assignment of semantic roles for a large amount of unseen text.

The problem of automatic semantic role labeling can be also understood as a kind of light or shallow semantic parsing. The goal is to identify some related phrases and assign

a joint structure in the form: “*Who did what to whom, when, where, why, how*”, to each word in a sentence [19].

Regarding to the Slovak language, significant research in automatic semantic analysis was done by Emil Páleš [12], [13]. He described the process of natural language understanding (NLU) at several layers in more detail. The result of his research was a hierarchical paraphrasing engine SAPFO (Sense Apprehending Frame Operator), which is a deterministic rule-based ASRL system based on a complex language analysis, able to paraphrase Slovak sentences. Unfortunately, Emil Páleš did not continue in his research and SAPFO engine is not available at all. Therefore, it is not possible to compare it with other systems.

The motivation for our research is the fact that the morphosyntactic layer formed by part-of-speech (POS) tags gives us just limited view of the meaning. The presented paper introduces initial research towards developing system architecture for automatic semantic role labeling in the Slovak language. The proposed architecture consists of three parts: 1. definition of an appropriate set of semantic roles; 2. annotation of a corpus or creating a new one; and 3. selection of one of the classification approaches for chunking and automatic semantic role labeling. All these tasks will be discussed further in more detail.

2 Related Works

The fundamental work in the field of semantic role labeling was given by Charles J. Fillmore in 1968 [4]. He described a hierarchical classification of semantic roles.

The FrameNet corpus is based on this concept. Sentences in it are arranged in hierarchical order and each frame refers to a concept. Frames at the top level refer to a more generic concept and frames at the lower level refer to more specific concepts. The FrameNet consists of 170,000 manually annotated sentences designed for training ASRL systems [1]. While Fillmore’s set of semantic roles consists of a few basic labels at the top level of universality, the FrameNet uses a large number of “frame elements” (8884 semantic roles in ver. 1.5).

The Proposition Bank (PropBank) is another lexical resource for ASRL designed by Martha Palmer et al. [9]. Sentences in the corpus are annotated with verbal propositions and their arguments. Moreover, PropBank is very similar to the FrameNet but differs in two major ways [1]: 1. all the verbs in the corpus are annotated and 2. each argument to a verb must be syntactic constituent. A standard set of argument labels has been defined for this purpose.

The best-known lexical database is WordNet [3]. Nouns, adjectives, verbs and adverbs are grouped into synsets (cognitive synonyms), each expressing a distinct concept. Synsets are linked by means of semantic and lexical relations.

Another lexical resource that organizes English verbs into different classes is VerbNet. Each verbal class takes different thematic roles and certain syntactic constraints describing their superficial behavior into account [14].

In terms of related languages, the most interesting lexical resource of Czech with complex interlinked morphological, syntactic and semantic annotation is the Prague Dependency Treebank (PDT). It uses 67 basic semantic roles (values of functors). Token dependencies are described in a form of tree [6].

One of the another resources that contains information about semantic roles is the Czech valency lexicon Verbalex [8]. It uses two-layer semantic labeling. There is 30 main semantic roles, which were selected from EuroWordNet Top-Ontology [18], with some modifications. At the second layer, more than thousand of roles, adopted from WordNet [3], extend semantic roles from the first layer.

Given that we have no knowledge of any freely available corpora with semantic annotation in the Slovak language, we decided to start to build a new one, what was also a challenge and motivation for our research in this area.

3 The Slovak Template-based Semantic Corpus

One of the well-described set of semantic roles designed for Slovak was created by Emil Pálež in 1994. He described a set of 66 semantic roles divided into 10 subsets [13]. We analyzed them and concluded that proposed set of semantic roles is relatively extensive and it is difficult to distinguish, in many cases, between some specific roles due to their similarity. Therefore, we divided them into two layers, where the first layer describes general semantic information and the second one extends the first. The original set of 66 semantic roles was reorganized into 44 roles on the first layer and 18 roles on the second layer. The proposed two-layer set of semantic roles for the Slovak language is described in more detail in [11].

The initial template-based semantic corpus of 300 Slovak sentences called SEMIENKO was prepared according to the proposed two-layer set of semantic roles. Sentences were manually annotated with the reduced set of 44 semantic roles. 38 of them appear in our semantically annotated corpus.

The occurrence frequency of roles in our semantic corpus is summarized in the Table 1. The corpus consists of example sentences in Slovak, acquired from the work of Emil Pálež [13] and Jolana Nižníková [10] and extended with real dialogues between people in everyday situations.

The following example demonstrates semantic roles labels for the sentence: “*Ján spoznal Máriu.*” (*John met Mary.*)

[AGSIKOG] *Ján* [VRB] *spoznal* [PACIFEN] *Máriu* .

The first semantic role before the square bracket is the role on the basic layer followed by the role on the second layer. There are three tokens: “*Ján*” (*John*), whose is an agens (AGS) or kognizant (KOG) on the second layer, then the verb “*spoznal*” (*met*) followed by token “*Máriu*” (*Mary*), which plays the role paciens (PAC) on the basic level and in more detail the role fenomenál (FEN), which means the entity that can be recognized.

1st layer	2nd layer	semantic role	cnt	1st layer	2nd layer	semantic role	cnt	1st layer	2nd layer	semantic role	cnt
ABJ		antiobjekt	0	IFR		identifikátor	2	ORI	EXL	exlokatív	-
ABT		atributív	36	IFT		identifikant	2	PAC		paciens	65
AGS		agens	69	INC		iniciál	5	PAC	DTK	deštruktant	-
AGS	AFE	afektor	-	INS		inštrument	12	PAC	FEN	fenomenál	-
AGS	DON	donor	-	KAR		kardinál	10	PAC	REZ	rezultant	-
AGS	EDI	editor	-	KAR	FRA	fraktál	-	POS		posesív	0
AGS	KOG	kognizant	-	KAR	KOL	kolektív	-	REA		realizátor	3
AGS	PCS	procesor	-	KAR	MES	mesuratív	-	REL		relátor	3
AGS	PDK	produktor	-	KAR	MUL	multiplikatív	-	RES		respektív	1
ANT		atributant	27	KAR	SPE	speciatív	-	SOC		sociatív	5
BEN		benefaktor	38	KAZ		kauzatív	1	STA		statuál	2
BEN	ADS	adresát	-	KCS		koncesív	0	SUB		substituál	0
BEN	REC	recipient	-	KND		kondicionál	3	SUK		sukcesív	1
DES		destinatív	37	KVF		kvalifikátor	0	TEM		tematív	12
DIF		diferenciál	1	KZK		konzekvent	1	TER		terminál	1
DIS		distributív	0	LOK		lokatív	64	TMP		temporál	76
DUR		duratív	5	MAT		materiál	6	TMP	PER	perspektív	-
ELE		elementív	13	MDS		modus	43	VIA		viál	5
EVO		evokátor	0	MOT		motivant	3				
FIN		finál	5	OBJ		object	110	*CNJ		conjunction	34
FRE		frekvenciál	1	ORD		ordinál	0	*VRB		verb	345
FRM		formatív	0	ORI		originatív	17	*UNK		unknown	14

* additional labels

Table 1. The occurrence frequency of semantic roles in the template-based corpus

4 Automatic Semantic Role Labeling in Slovak

The complete process of automatic semantic role labeling can be described in several steps including tokenization, named entity recognition, part-of-speech tagging, chunking and semantic role labeling, as it is depicted in Fig. 4.

The first three components were adopted from our previous research [7]. In this case, tokenization is performed by rules compiled into a single state machine using Ragel tool [16]. Rules in the proposed tokenizer identify punctuations, words, abbreviations, acronyms, list items, numbers, e-mails, and URLs. Identification of sentence boundaries is performed by disambiguation of a dot.

Named entity recognition is applied in the next step. Because no manually annotated corpora in Slovak for named entity recognition has been already done, the recognition tool uses a plain dictionary-based approach. For this purpose, a set of dictionaries with list of named entities has been created.

A special morphological classifier Dagger [7] for part-of-speech (POS) tagging has been proposed. Dagger annotation tool is based on a hidden Markov model (HMM) and the best sequence of POS tags is found using the Viterbi algorithm.

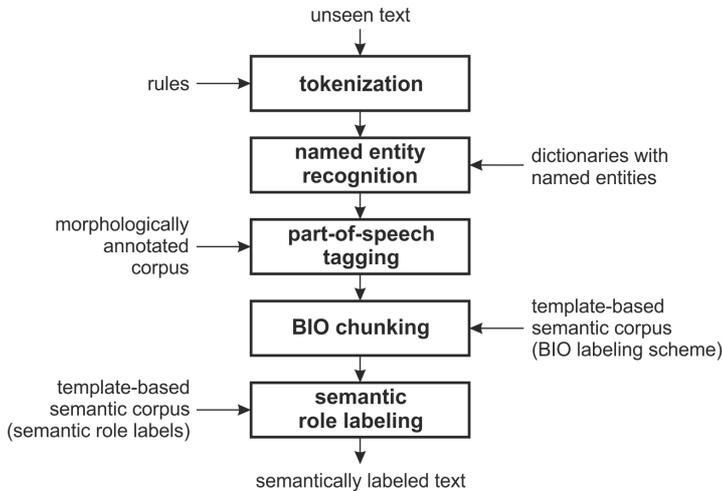


Fig. 1. The automatic semantic role labeling system architecture

4.1 Statistical Identification of Chunks

There are three approaches to automatic identification of chunks. The rule-based chunking uses hand-crafted rules that match specified context and mark it as one chunk. Rules can be defined by regular expressions or can utilize morphological tags or vocabulary. Rules have to be specified explicitly, written by an expert.

The other approach is based on statistics. The core of the system is a classification approach that can assign a set of tags (classes) to the presented context. This type of system utilizes implicit knowledge, expressed as a set of examples in a previously tagged training corpus. The classifier takes the training corpus and a model is constructed by analyzing the corpus. Hybrid approaches try to utilize both explicit (rules) and implicit (parameters of the model) knowledge.

The proposed approach uses our custom statistical tool to analyze presented semantically annotated corpus and train a classifier that is able to mark chunks in arbitrary text. The chunker is based on a modified HMM and uses the Viterbi algorithm for classification.

The classifier consists of these parts:

1. **state set** is set of all possible states that has been seen during training;
2. **state-transition model** expresses probability of a state according to occurrence of previous states. The algorithm can take two preceding states into the account;

3. **observation model** estimates probability of a state according to the presented observation. The classifier uses word and its corresponding morphological POS tag as features. The HMM model is able to utilize only one feature. Our modification uses both morphological POS tag and word to estimate state-observation probability.

The proposed tool uses standard BIO chunking with three possible states:

1. **B** marks beginning of a chunk;
2. **I** marks each other word of the same chunk;
3. **O** marks words outside any chunk - in our case verb and its dependent words.

Because of a very small size of the training set, ten-fold cross validation has been chosen as a evaluation methodology. Evaluation has been performed in ten steps and individual results have been summed together. The training set has been split into ten parts. Each round one part was used to calculate precision and confusion matrix and remaining nine parts were used to train the classifier. At the end all results were summed together and precision/recall and F1 score for each class has been calculated.

The first experiment used word and full morphological tag as a feature. Experimental results are summarized in the Table 2.

tag	precision	recall	F1 score
B	0.802759	0.868657	0.834409
I	0.633508	0.647059	0.640212
O	0.800000	0.843750	0.821293

Table 2. Statistical chunking with full morphological tag as a feature

The second experiment used word and reduced morphological tag as a feature (see Table 3). In this case only the first letter of the morphological tag has been taken, omitting other grammatical categories such as case, gender or time.

tag	precision	recall	F1 score
B	0.702069	0.805380	0.750184
I	0.756545	0.562257	0.645089
O	0.656790	0.730769	0.691808

Table 3. Statistical chunking with reduced morphological tag as a feature

The experiments show that the presented chunker does not reach state of the art precision. Taking size of the training set into the account, presented results seem to be satisfying. Surprisingly, full tag set shows lower precision.

4.2 Semantic Role Labeling

The standard machine learning techniques can be used for ASRL task. Our approach is based on HMMs that estimates observation and transition probabilities by using n-gram language models [17].

Although Venkataraman et al. applied this approach on classification of dialog acts (DAs), we suppose that it can be used for semantic role labeling task.

In this approach, the DAs are represented by the model's hidden states, whereas the utterances in the DAs correspond to the observations generated by the states. For representing of described models, two types of n-gram models are involved. The observation probabilities are estimated using n-grams trained on words for each DA class, while transition probabilities among DAs were estimated by n-gram model over DA labels and can be seen as dialogue grammar. Models are trained using standard approach from labeled data. Decoding process is performed using the Viterbi algorithm and the result is the most likely DA sequence for the string of observed words. Authors also propose "a partially supervised version" of this algorithm, where initial models are trained on a small amount of data (bootstrap data) and then unlabeled data are used in the following iterations to retrain n-gram models [17].

Authors obtained sufficient results on DAs classification task with very a small training data (only 44 utterances) which is similar to our situation. This fact was the main reason why we adopted this approach for our ASRL task.

The process for ASRL in Slovak can we described as follows:

- **observation probabilities** are estimated using bigrams for each SR label;
- **two groups of models** are trained – one group for words that belong to particular SR label and the second one is trained on morphological tags belonging to particular SR label;
- **transition probabilities** between SRs in particular utterance are estimated by a bi-gram model over sequences of SR labels. This model expresses dependences between SR labels in the utterance.

The selection of appropriate features is equally important as selection of the classification approach. As it was mentioned earlier, we have tried to use separate lexical features (Setup 1) and morphological POS tags (Setup 2). Results obtained on our initial corpus are not sufficient. We suppose that results with the presented type of parameterization can be improved by extending the corpus. The performance of the proposed setups seems to be limited, because there is need to join both type of parameters – lexical and morphological in an appropriate way, which reflects the valency of verbs.

The semantic role of particular utterance constituent depends on three main components: 1. verbs; 2. prepositions that introduce the constituent; and 3. morphological POS tags in this part of utterance.

In addition, we are also considering the use of n-gram models trained on a combination of verbs, prepositions and morphological tags of analyzed chunks that may be established in the future (in the third setup of SR classification).

In the initial experiments, bigram models for each SR label were trained. In the first SR classification setup (Setup 1), only lexical parameters (words) were applied. Utterances related to particular SR label were joined together to create a training set. In the second classification setup (Setup 2), morphological tags were used for training bigram models for each SR label. In both setups, bigram models were estimated using the SRILM Toolkit [15].

In the initial experiments, decoding was performed without using the Viterbi algorithm as follows: N-best list of the most probable SR labels was created as the result of the decoding on SR label n-grams. The evaluation was done on the testing part of the corpus. Precision was about 35% in Setup 1 and around 48% in Setup 2. Although Setup 2 gives better results than the Setup 1, it can be concluded that obtained values are weak. There are two main reasons. Small size of the corpus can be seen as the main reason. The second important reason is that using only words or morphological tags as a parameterization for the ASRL task seems to be not effective enough. We suppose that an appropriate joining of lexical and morphological features that will better reflect verb valences should bring significantly better results.

5 Conclusion

Despite experimental results, we hope that the research presented in this paper opens discussion about automatic semantic role labeling in Slovak.

In the future research, the annotated database will be significantly extended. Bigger training database should improve training of statistic classifiers. The process of statistical identification of chunks and semantic role labeling could be improved by examination of more possible features, effectively constraining the search space, or selecting different classification algorithm. Proper semantic annotation should improve natural language processing of the Slovak language.

Also, we would like to try semi-automatic learning. A suitable combination of morphological, lexical and other possible parameters would better reflect the valencies of verbs. We would like to join statistical identification of chunks and automatic semantic role labeling task together.

Acknowledgments

The research in this paper was supported in part by the Ministry of Education, Science, Research and Sport of the Slovak Republic under the research project VEGA 1/0075/15 (25%), the Slovak Research and Development Agency under the contract No. SK-HU-2013-0015 (25%) and the research project APVV-0342-11 (25%), and by the Research and Development Operational Programme funded by the ERDF under the project implementation University Science Park TECHNICOM for Innovation Applications Supported by Knowledge Technology, ITMS code: 26220220182 (25%).

References

- [1] Aher, A., Jain, K., Dan, A., and Singh, J. (2010). *A survey on semantic role labeling and dependency parsing*. Center for Indian Language Technology, Indian Institute of Technology, Bombay.
- [2] Baker, C. F., Fillmore, C. J., and Love, J. B. (1998). The Berkeley FrameNet project. In *Proc. of the 17th Intl. Conf. on Computational Linguistics, COLING'98*, pages 86–90, Montreal, Quebec, Canada.
- [3] Fellbaum, C. (1998). *WordNet: An electronic lexical database*. MIT Press, Cambridge, MA.
- [4] Fillmore, C. J. (1968). The case for case. In Bach, E. and Harms, R., editors, *Universals in Linguistic Theory*, page 25, London. Holt, Rinehart and Winston.
- [5] Gildea, D. and Jurafsky, D. (2002). Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288.
- [6] Hajič, J., Böhmová, A., Hajičová, E., and Vidová-Hladká, B. (2000). The Prague Dependency Treebank: A three-level annotation scenario. In Abeillé, A., editor, *Treebanks: Building and Using Parsed Corpora*, pages 103–127. Kluwer, Amsterdam.
- [7] Hládek, D., Staš, J., and Juhár, J. (2012). Dagger: The Slovak morphological classifier. In *Proc. of the 54th Intl. Symposium ELMAR-2012*, pages 195–198, Zadar, Croatia.
- [8] Hlaváčková, D. and Horák, A. (2005). VerbaLex - New comprehensive lexicon of verb valencies for Czech. In *Proc. of the 3rd Intl. Seminar on Computer Treatment of Slavic and East European Language, SLOVKO'05*, pages 107–115, Bratislava, Slovakia.
- [9] Kingsbury, P. and Palmer, M. (2002). From TreeBank to PropBank. In *Proc. of the 3rd Intl. Conf. on Language Resources and Evaluation, LREC'02*, pages 1989–1993, Las Palmas, Spain.
- [10] Nižníková, J. (2001). Sentence models in the Slovak language (In Slovak: Větné modely v slovenčine). Acta Facultatis Philosophicae Universitatis Prešovensis – Monographia 28.
- [11] Ondáš, S., Hládek, D., and Juhár, J. (2014). Semantic roles labeling system for Slovak sentences. In *Proc. of the 5th IEEE Conf. on Cognitive Infocommunications, CogInfoCom'14*, pages 161–166, Vietri sul Mare, Italy.
- [12] Páleš, E. (1991). Počítačový model jazyka ako celku. In Nábělková, M. and Šimková, M., editors, *Varia I: Zborník príspevkov z I. kolokvia mladých jazykovedcov*, volume 1, pages 15–19, Bratislava, Slovensko. Slovenská jazykovedná spoločnosť SAV.
- [13] Páleš, E. (1994). *SAPFO - A hierarchical paraphrasing engine (In Slovak: SAPFO - Parafrazovač slovenčiny)*. VEDA, Bratislava.
- [14] Schuler, K. K. (2005). *VerbNet: A broad-coverage, comprehensive verb lexicon*. PhD thesis, University of Pennsylvania, Philadelphia, PA, USA.

- [15] Stolcke, A. (2002). SRILM – An extensible language modeling toolkit. In *Proc. of the 7th Intl. Conf. on Spoken Language Processing, ICSLP'02*, pages 257–286, Denver, Colorado, USA.
- [16] Thurson, A. D. (2006). Parsing computer languages with an automaton compiled from a single regular expression. In Ibarra, O. H. and Yen, H. C., editors, *Implementation and Application of Automata*, volume 4094, pages 285–286, Berlin Heidelberg. Springer-Verlag.
- [17] Venkataraman, A., Stolcke, A., and Shriberg, E. E. (2002). Automatic dialog act labeling with minimal supervision. In *Proc. of the 9th Australian Intl. Conf. on Speech Science and Technology*, pages 1–6, Melbourne, Australia.
- [18] Vossen, P., Bloksma, L., Rodriquez, H., Climent, S., Roventini, A., Betagna, F., Alonge, A., and Peters, W. (1998). The EuroWordNet base concepts and top ontology. Technical Report Deliverable D017, D034, D036, WP5 EuroWordNet, LE2-4003, University of Amsterdam, Amsterdam, The Netherlands.
- [19] Yang, H. and Zong, C. (2014). A global generative model for Chinese semantic role labeling. In Zong, C., Nie, J. Y., Zhao, D., and Feng, Y., editors, *Natural Language Processing and Chinese Computing*, volume 496, page 12, Berlin Heidelberg. Springer-Verlag.

Slavic Languages in Universal Dependencies

Daniel Zeman

Faculty of Mathematics and Physics, Charles University in Prague, Czech Republic

Abstract. Universal Dependencies (UD) is a project that is developing cross-linguistically consistent treebank annotation for many languages, with the goal of facilitating multilingual parser development, cross-lingual learning and linguistic research from a language typology perspective. It is a merger and extension of several previous efforts aimed at finding unified approaches to parts of speech, morphosyntactic descriptions and syntactic dependency relations. In the present contribution we address the application of UD to Slavic languages. We devote the most space to peculiarities of pronouns, determiners, numerals and quantifiers. Other language features that are discussed include modal verbs, ellipsis, nominal predicates, and reflexive pronouns. Most of our examples are from Czech but the language features demonstrated are usually portable to other Slavic languages. We include examples from the other languages where appropriate.

1 Introduction

The general philosophy of the Universal Dependencies (UD) project¹ [14] is to provide a universal inventory of categories and guidelines to facilitate consistent annotation of similar constructions across languages, while allowing language-specific extensions when necessary. The first version of the standard was published in October 2014 and datasets for the first ten languages were released in January 2015; version 1.1 with eight additional languages was released in May 2015 and subsequent releases are currently planned after every 6 months.

Forseen applications include typological studies and cross-lingual transfer of parsing models. First experiments with the UD treebanks have already been published. For instance, [18] addressed statistical learnability of the UD dependency structures in comparison to other annotation styles; [23], [3] repeated cross-language parsing experiments that were previously done with unharmonized treebanks and the previous results were not conclusive, mostly because the diverse annotation styles were not comparable. One interesting observation is that even small datasets can be useful. While bigger is definitely better, [16] found that a “treebank” of as few as 10 sentences gave better parsing accuracy than the best-performing unsupervised method.

UD is based on an evolution of several previous efforts to find a cross-linguistically valid annotation scheme of natural language morphology and dependency syntax. These efforts have contributed to various layers of UD:

¹<http://universaldependencies.github.io/docs/>

The universal part-of-speech tags (**UPOS**) are based on the Google universal tagset [15], which has been extended and redefined from the original 12 to the current 17 tags; in addition, UD also defines a set of 17 **universal features** that can be used to describe lexical and inflectional properties of words. These features are especially useful for morphologically rich languages. The core feature set is based on Intersect [25], an interlingua for morphosyntactic tagsets. It is likely that new features or new feature values will be identified as new languages are added; therefore, the UD format allows additional language-specific features.

At the level of syntactic dependency relations, two related projects have independently tried to define a common scheme applicable to multiple languages: HamleDT (Harmonized Multi-Language Dependency Treebank) [27], [26], [17] comprises 36 languages in its version 3.0; the Universal Dependency Treebank (UDT) [13] has 11 languages in its version 2.0.

The annotation scheme used in UDT is based on Stanford Dependencies (SD) [7], [8], a popular syntactic representation that was first defined for English but later successfully adapted for various other languages. The early releases of HamleDT were based on Prague Dependencies (PD), essentially the annotation scheme of the Prague Dependency Treebank (PDT) [4]. The two projects started to converge when HamleDT 2.0 included a Stanford conversion of its trees, and became the largest collection of treebanks available in PD and SD [17]. Both teams participated in the formulation of the UD annotation guidelines and they are working on converting their data to UD; creators of treebanks for individual languages have joined the effort and either converted their existing data automatically or initiated new manual annotation. The 18 languages included in the UD 1.1 dataset [1] are Basque, Bulgarian, Croatian, Czech, Danish, English, Finnish, French, German, Greek, Hebrew, Hungarian, Indonesian, Irish, Italian, Persian, Spanish, and Swedish. With increasing coverage and popularity, UD could become a new de-facto standard in the not-so-far future.

The dependency relation inventory and guidelines of UD are based on SD and can be viewed as the next step in the evolution of SD towards a linguistically universal scheme. In the present contribution, we take a closer look at peculiarities of Slavic languages and how they can be handled in UD. We proceed from the first experiences with UD for Czech, and most examples we present come from Czech; we supplement them with examples from the other Slavic languages where appropriate.² One very relevant piece of previous work is [12], whose authors proposed several adjustments of SD for Slavic languages. As they based their work on the older (and now obsolete) version of the Stanford format, we will show that some of the issues they address have been solved in UD.

Language	Code	Treebank	Sent	Tok
Bulgarian	[bg]	BulTreeBank	13,221	196K
Church Slavonic	[cu]	PROIEL	7,818	72K
Croatian	[hr]	SETimes.HR	3,736	84K
Czech	[cs]	PDT	87,913	1504K
Polish	[pl]	IPI PAN	8,227	84K
Russian	[ru]	SynTagRus	63,000	900K
Slovak	[sk]	SNK	63,238	994K
Slovene	[sl]	SSJ500K	27,829	500K

Table 1. Dependency treebanks of Slavic languages. We use the ISO 639-1 language codes in brackets when referring to particular languages and treebanks throughout the paper.

2 Existing Treebanks

There are dependency treebanks of various sizes available for at least 8 Slavic languages. The oldest and largest of them is the Prague Dependency Treebank (PDT) of Czech [4]. It inspired annotation efforts for other languages, and about ten other languages have treebanks whose annotation style is very close to PDT, among them three Slavic languages: Slovak [19], Slovene [9], [11] and Croatian [5], [2]. Research teams in several other countries have created treebanks in different annotation scenarios, namely for Bulgarian [20], Russian [6] and recently also Polish [24]. In addition, the PROIEL project³ provided syntactically annotated texts in Old Russian and Church Slavonic [10]; a new corpus called TOROT for Old Russian and Church Slavonic has recently been launched in Tromsø.⁴ The Russian and Slovak treebanks have no standard distribution channels so far; the other treebanks mentioned above are freely downloadable and available for non-commercial research purposes. Table 1 summarizes the Slavic treebanks and their sizes.

3 Pronouns and Determiners

The UPOS tagset includes a tag for determiners, which is a category routinely distinguished in English and in Romance languages, but it is not used in the grammatical tradition of Slavic languages (among others). Determiners encompass definite and indefinite articles (which do not exist in Slavic languages, at least not as independent words), as well as other functional words; in Slavic grammars, these words are covered by the term *pronoun*.

The current definition of the borderline between pronouns and determiners in UD is drawn along syntactic properties, that is, it focuses on the function of the word rather than

² At the time this manuscript was submitted for review, Czech was the only Slavic language whose treebank had been converted to UD; later on, Bulgarian and Croatian were added.

³ <http://proiel.github.io/>

⁴ <http://site.uit.no/slavhistcorp/files/2015/04/Eckhoff.pdf>

its form.⁵ This principle essentially follows the recommendation of EAGLES (see Sections 8.3.1 of [21] and 6.2.2 of [22]). Pronouns are heads of noun phrases, while determiners are those function words that cannot stand alone but need a head (nominal, pronominal) to form an NP.

Many/DET party-goers prefer wine to beer.

Many/PRON disagreed to the leader's speech.

While this general guideline may look easy to apply at first glance, the matter is complicated by ellipsis. Consider the sentence

Moje auto je větší než tvoje. “My car is bigger than yours.”

In contrast to English, Slavic languages do not use different word forms for self-standing possessive pronouns (*yours*) and for possessive determiners (*your*). It is natural to view the sentence as an elliptical structure with deleted second instance of *auto*: *Moje auto je větší než tvoje auto.* “My car is bigger than your car.” Therefore we propose for Slavic languages to classify all possessive pro-forms as personal possessive determiners. That is, their tag will be DET and their features will include *Poss=Yes | PronType=Prs*.

Interrogative, relative, indefinite, negative and demonstrative pro-forms can be divided to those that never behave like determiners ([cs] *kdo, co, někdo, něco, nikdo, nic*) and those that could be determiners or pronouns (*jaký, který, či, nějaký, některý, něčí, každý, žádný*). The words from the latter group inflect similarly to adjectives; we may thus be tempted to classify them as determiners without looking at their context (if they appear without a noun, we would explain it by ellipsis). Unfortunately this analysis would be wrong at least for some occurrences of relative forms, which cannot be elliptic:

Muž, kterého jsem vám ukázal “The man whom (which) I showed you” cannot be expanded to *Muž, kterého *muže jsem vám ukázal* and the pronoun *kterého* cannot be attached to *muž* because *muž* is outside the relative clause in which the pronoun acts as the direct object.

For other pro-forms it is not clear whether they should be analyzed as elliptic. The Czech pronoun *každý* “every” occurs 1023 times in PDT and 76% of the occurrences are attributive (dependency labeled *Atr*),⁶ which suggests they should be tagged DET. However, 24% occurrences independent of nouns seem quite a lot to get along with postulating an invisible deleted noun. A related word *všechen* “all” is even less pronounced: 64% attributive and 36% non-attributive.

Based on this evidence, we propose that the ellipsis explanation, used for possessive determiners, is not extended to the other categories of pro-forms. Instead, the syntactic context should be consulted. If the word modifies a nominal and if there is morphological agreement, then it is a determiner; otherwise it is a pronoun.

⁵ There is an ongoing discussion in the UD community whether the definition can be modified and based more on lexical than on functional criteria.

⁶ This is just an approximation. In addition to the *Atr* label, we should also require that the determiner agrees with the modified noun in gender, number and case, and possibly also that it occurs before the noun. That way we would exclude genitive modifications such as *nabídka všech* “the offer by all”.

4 Numerals and Quantifiers

The morphological and syntactic behavior of Czech numerals is a complex matter. Small cardinal numerals *jeden* “one”, *dva* “two”, *tři* “three” and *čtyři* “four” agree with the counted noun in case (*jeden* also agrees in gender and number; *dva* also agrees in gender). They behave as if they modify the counted noun; they are similar to adjectives in this respect. Examples:

- *Jeden muž spal, dva muži hráli karty.* “One man slept, two men played cards.”
- *Jedna žena spala, dvě ženy hrály karty.* “One woman slept, two women played cards.”
- *Jedno koťe spalo, dvě koťata si hrála.* “One kitten slept, two kittens played.”

In PDT, these numerals are attached to their counted nouns as *Atr* (attribute). UD will use the same structure, only the dependency will be labeled *nummod* (Figure 1).

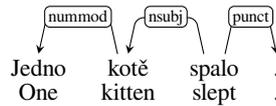


Fig. 1. One kitten slept.

Larger cardinals behave differently. They require that the counted noun be in the genitive case; this indicates that they actually govern the noun. Such constructions are parallel to nouns modified by other noun phrases in genitive. The whole phrase (numeral + counted noun) behaves as a noun phrase in neuter gender and singular number (which is important for subject-verb agreement).

- *Pět mužů hrálo karty.* “Five men played cards.”
- *Skupina mužů hrála karty.* “A group of men played cards.”

In PDT, these numerals are analyzed as heads of the counted nouns, which are attached to the numeral as *Atr*, see Figure 2.

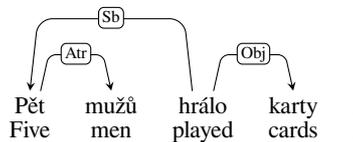


Fig. 2. Prague analysis of the numeral *pět* in nominative.

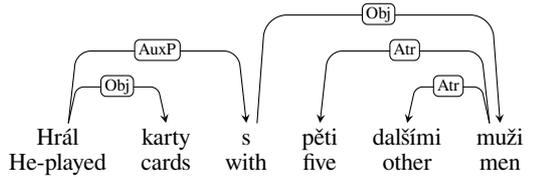


Fig. 3. Prague analysis of the numeral *pět* in instrumental.

There are both advantages and drawbacks to this solution. On the one hand, it reflects well the agreement in case, gender and number. On the other hand, it is confusing that there are two different analyses of counted noun constructions, depending on the numeric value. Moreover, the numeral does not govern the noun in all morphological cases, as shown in Table 2.

Phrase Case	Example	Numeral Case	Noun Case
Nom	pět mužů	Nom	Gen
Gen	pěti mužů	Gen	Gen
Dat	pěti mužům	Dat	Dat
Acc	pět mužů	Acc	Gen
Voc	pět mužů	Voc	Gen
Loc	pěti mužích	Loc	Loc
Ins	pěti muži	Ins	Ins

Table 2. The morphological case of a counted phrase with a high-value numeral (first column) and the consequences for the case of the parts (note that these numerals have only two distinct morphological forms, resulting in homonymy). The example phrase is *pět mužů* “five men”.

We can say that the noun has the case of the whole phrase if it is dative, locative or instrumental. The numeral then agrees with the noun in case. The numeral forces the noun to the genitive case if the whole phrase is nominative, accusative or vocative (but the vocative usage is rather hypothetical). In genitive, the noun and the numeral agree with each other; but note that the numeral uses its inflected form, as in the other cases where it agrees with the noun.

In PDT, the genitive, dative, locative and instrumental cases are analyzed in parallel to the low-value numerals, i.e. the noun governs the numeral, see Figure 3.

Pronominal quantifiers behave as high-value numerals and govern the quantified nouns:

- *Kolik mužů hrálo karty?* “How many men played cards?”
- *Několik (mnoho, málo) mužů hrálo karty.* “Several (many, few) men played cards.”
- *Tolik mužů hrát karty jsem ještě neviděl.* “I have never seen so many men playing cards.”

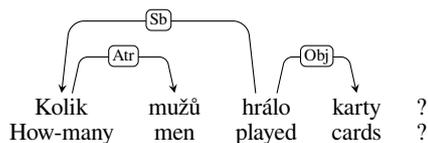


Fig. 4. Prague analysis of the quantifier *kolik* in nominative.

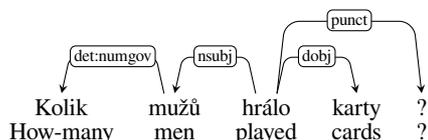


Fig. 5. UD analysis of a governing interrogative quantifier.

For Universal Dependencies we suggest to use the same tree shape for all the examples mentioned above. The counted noun will always be the head, and the numeral or quantifier will depend on it. Thus the structure will be parallel among similar phrases within one language, and also with the universal dependencies in non-Slavic languages. However, we use the UD mechanism of language-specific extended labels to preserve information about who governs the morphological case. There are four labels used and they are based on two UD labels: *nummod* and *det* (Table 3).

	Numeric	Pronominal
Noun governs	<i>nummod</i>	<i>det:nummod</i>
Numeral governs	<i>nummod:gov</i>	<i>det:numgov</i>

Table 3. Proposed language-specific dependency relation labels that distinguish quantifiers from other determiners, as well as the situations where the quantifier governs the case of the noun, from the situations where the quantifier agrees with the noun.

5 A Verb or not a Verb?

Verbal nouns ([cs] *čtení*, [ru] *чтение* “reading”) are tagged as nouns, not as verbs. But even then they may have the feature *VerbForm=Ger* to distinguish them from other nouns. (Note that the *VerbForm* feature in UD is actually not constrained to verbs.)

The active (past) participles should always be verbs (these are the forms ending with *-l*, *-la*, *-lo* etc.) Note however that occasionally there are derived adjectives with the long adjectival ending, cf. [cs] *zkrachovalý* “bankrupt”. These are tagged as adjectives, not as verbs. Passive participles and participial adjectives are told apart in a similar fashion.

If the word is used as a modifier of a noun, it should be adjective. If it is used to form the periphrastic passive, it should be verb. This boundary differs across Slavic languages, cf.

[cs] *Město bylo založeno*/VERB *Karlem IV.* “The city was founded by Charles IV.”

[cs] *Město založené*/ADJ *Karlem IV. vyhořelo.* “The city founded by Charles IV has burned down.”

[sk] *Mesto bolo založené*/VERB|ADJ? *Karolom IV.* “The city was founded by Charles IV.”

[sk] *Mesto založené*/ADJ|VERB? *Karolom IV. vyhorelo.* “The city founded by Charles IV has burned down.”

In any case, all these word forms should also have the feature `VerbForm=Part`, regardless whether their main tag is VERB or ADJ.

Transgressives (adverbial participles) such as [cs] *pomáhajíc* “helping” or [ru] *будучи* “being” should have the feature `VerbForm=Trans` and the main tag VERB. They may also have the Tense feature to distinguish present and past transgressives.

6 Auxiliary Verbs and Modal Verbs

Local equivalents of the verb *to be* are the most frequent Slavic auxiliaries, used to create periphrastic past, passive or conditional. The same verb can also be used as non-auxiliary (copula or main verb).

Some languages (e.g. Croatian) have a second auxiliary, *htjeti*, used to form the future tense. In northern Slavic languages the future is also formed using the verb *to be*.

In contrast to the Universal Dependencies applied to English and other Germanic languages, we do not recommend treating modal verbs as auxiliaries. Modal verbs are a subset of verbs that take an infinitive of another verb as complement: [cs] *můžu přijít*, [ru] *ты можешь взять / ты можешь vzjat'*, [bg] *може да бъде избран / може да бъде избран*. The morphological paradigms of Slavic modal verbs are slightly restricted but not as much as in English. They do not form passive participles⁷ and most of them also do not have imperative forms. The set is not identical to English. For instance, the Czech verb *chtít* “to want”, if used with infinitive and not with a direct object, counts as a modal verb, while its English equivalent does not. There is not much to be gained from treating the modal verbs in the same way as the auxiliary *to be*. It seems more natural to keep the modal dependency structures parallel to those of phase verbs and verbs of control, which also take an infinitival argument. That is, the infinitive will be attached to the modal verb as `xcomp`: see Figure 6.

There is usually just one modal verb to one content verb. However, two modal verbs may co-occur even if it is very rare: [cs] *bude muset chtít pracovat* “he will have to want to work”. Treating modals as content verbs has the advantage of capturing scope and hierarchy between the two modals in this example. Furthermore we also want to be able

⁷ But note that some of them have homonyms that are not used modally and that can form the passive.

to capture the scope of negation and other modifiers: [cs] *nemohl jsem přijít* “I was not able to come”, *mohl jsem nepřijít* “I was able not to come” and *nemohl jsem nepřijít* “I was not able/allowed not to come” are three semantically different expressions.

In addition to modal verbs, modality can also be expressed by modal adverbs, adjectives or nouns. In some cases they are derived from the same roots as modal verbs. These non-verbal modal expressions are particularly pervasive in Russian but other languages have them as well. Again, analyzing modal verbs as content words results in annotation that is parallel to the annotation of non-verbal modal expressions (Figure 7).

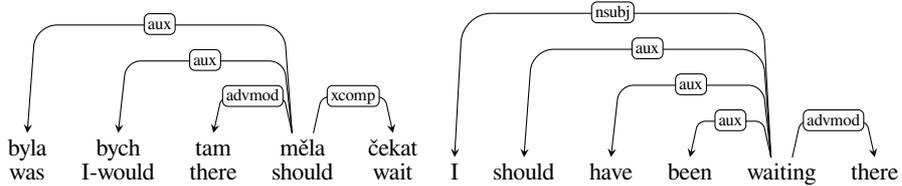


Fig. 6. Combination of modal and auxiliary verbs in Czech and English. English modals are treated as auxiliaries, Czech modals are treated as main verbs.

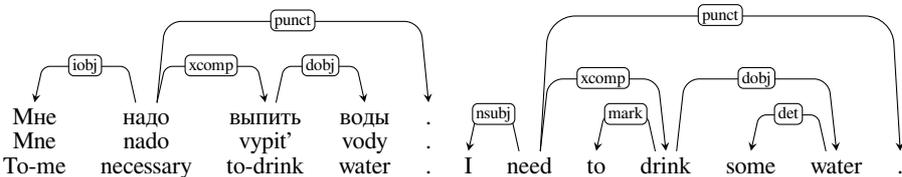


Fig. 7. Modal predicative adverb in Russian and its English translation: *I need to drink some water.*

7 Reflexive Pronouns and Verbs

Most of the time the reflexive pronoun is attached to a verb. In the case of transitive verbs, the reflexive pronoun is just another form of object (labeled *dobj* or *iobj*). The test is here whether it can be substituted with a normal personal pronoun. If it cannot be substituted, then we are dealing with an inherently reflexive verb ([cs] *smát se* “laugh”). We cannot use an object relation for these reflexives; we suggest to use a language-specific extension of the UD label *expl* (expletive) between the verb and the pronoun: *expl:reflex*.⁸

⁸ In the Czech and Croatian UD 1.1 data, we used variants of the compound relation to express that the two tokens actually form one lexeme. This was revised at a UD workshop in August 2015.

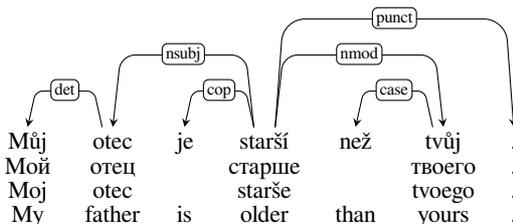


Fig. 8. Comparison with a nominal [cs, ru, ru, en]. Note that Russian omits both the comparative conjunction and the copula.

Finally, the reflexive pronoun may also be used to form the so-called reflexive passive ([cs] *to se snadno řekne* “it is said easily (= easier than done)”). The language-specific label `auxpass:reflex` should be used in this case.

Note that in Russian by convention the verb is written together with the reflexive element as one word (*смеяться / смеjat'sja* “laugh”). The general UD approach is to cut off clitics (split the token into two **syntactic words**). An often cited example from Spanish is *vámonos* “let’s go” that should be split to *vamos nos*, lit. *go us*, and each part analyzed as a separate word. This approach could be ported to Russian in cases where the clitic *-ся / -sja* can be substituted by an irreflexive pronominal object (*изменять+ся / izmenjat'+sja* “change oneself” would be parallel to *изменять его / izmenjat' ego* “change him”), and for reflexive passives. However, it does not seem a good idea to extend this approach to inherently reflexive verbs such as *смеяться / смеjat'sja*, where the reflexive morpheme does not have its own syntactic function.

8 Comparative Constructions

The UD guideline for comparisons is that the comparative complement is attached to the adjective or adverb that denotes the feature being compared. If the complement is a clause, the relation is labeled `advcl`. If it is a bare nominal, it is labeled `nmod`. Some Slavic languages use a comparative conjunction parallel to English *than*: [cs] *Můj otec je starší než tvůj*. “My father is older than yours.” [cs] *Ten hotel je větší, než jsme čekali*. “The hotel is bigger than we expected.” In other languages, the conjunction is not used and the complement is in genitive: [ru] *Мой отец старше твоего*. / *Moj otec starše tvoego*. “My father is older than yours.” Some Slavic languages use periphrastic comparative of adjectives while others largely prefer the morphological comparative. See Figures 8 and 9 for illustration.

In order to make the data more similar to other languages (including Bulgarian), we accepted the `expl(itive)`-based solution.

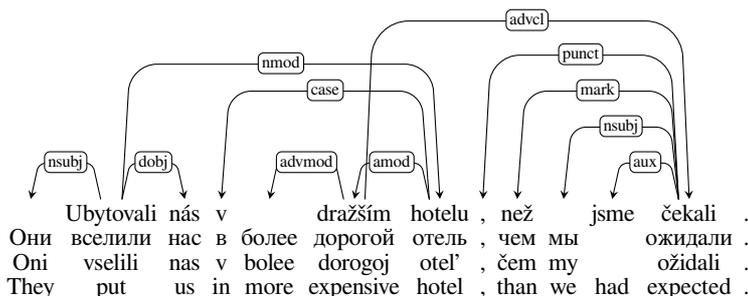


Fig. 9. Comparison with a clause [cs, ru, ru, en]. Note that Czech uses morphology to form a comparative adjective, while Russian and English form it periphrastically. Also note that Czech is a pro-drop language and omits the subjects.

9 Conclusion

We briefly introduced the concept of Universal Dependencies and listed a number of morphological and syntactic phenomena that occur in Slavic languages and their treatment in UD may not be apparent or straightforward. For each of the issues we discussed its context and proposed how it should be treated in UD. Even though in theory the UD mechanism of language-specific extensions enables treating them differently for different Slavic languages, it would go against the general spirit of UD. We argue that most of these features apply (even if with some variation) in most Slavic languages and thus they should be treated in all these languages in a unified way.

Acknowledgements

The author is grateful for the valuable comments provided by the anonymous reviewers. The work on this project was supported by the Czech Science Foundation grant no. GA15-10472S. 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] Agić, Ž., Aranzabe, M. J., Atutxa, A., Bosco, C., Choi, J., de Marneffe, M.-C., Dozat, T., Farkas, R., Foster, J., Ginter, F., Goenaga, I., Gojenola, K., Goldberg, Y., Hajič, J., Johannsen, A. T., Kanerva, J., Kuokkala, J., Laippala, V., Lenci, A., Lindén, K., Ljubešić, N., Lynn, T., Manning, C., Martínez, H. A., McDonald, R., Missilä, A., Montemagni, S., Nivre, J., Nurmi, H., Osenova, P., Petrov, S., Piitulainen, J., Plank, B., Prokopidis, P., Pyysalo, S., Seeker, W., Seraji, M., Silveira, N., Simi, M., Simov, K., Smith, A., Tsarfaty, R., Vincze, V., and Zeman, D. (2015). Universal dependencies 1.1, <http://hdl.handle.net/11234/LRT-1478>.

- [2] Agić, Ž. and Ljubešić, N. (2014). The SETimes.HR linguistically annotated corpus of Croatian. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1724–1727, European Language Resources Association (ELRA), Reykjavík, Iceland.
- [3] Agić, Ž. and Ljubešić, N. (2015). Universal Dependencies for Croatian (that work for Serbian, too). In *Proceedings of BSNLP/RANLP 2015 (in press)*.
- [4] Bejček, E., Hajičová, E., Hajič, J., Jínová, P., Kettnerová, V., Kolářová, V., Mikulová, M., Mírovský, J., Nedoluzhko, A., Panevová, J., Poláková, L., Ševčíková, M., Štěpánek, J., and Zikánová, v. (2013). Prague dependency treebank 3.0, <http://hdl.handle.net/11858/00-097C-0000-0023-1AAF-3>.
- [5] Berović, D., Agić, Ž., and Tadić, M. (2012). Croatian dependency treebank: Recent development and initial experiments. In Calzolari, N., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 1902–1906, European Language Resources Association (ELRA), İstanbul, Turkey.
- [6] Boguslavsky, I., Iomdin, L., Petrochenkov, V., Sizov, V., and Tsinman, L. (2013). A case of hybrid parsing: Rules refined by empirical and corpus statistics. In Gerdes, K., Hajičová, E., and Wanner, L., editors, *Computational Dependency Theory*, volume 258, pages 226–240, IOS Press, Amsterdam, Netherlands.
- [7] de Marneffe, M.-C. and Manning, C. D. (2008). Stanford typed dependencies manual.
- [8] de Marneffe, M.-C., Silveira, N., Dozat, T., Haverinen, K., Ginter, F., Nivre, J., and Manning, C. D. (2014). Universal Stanford dependencies: A cross-linguistic typology. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4585–4592, European Language Resources Association (ELRA), Reykjavík, Iceland.
- [9] Džeroski, S., Erjavec, T., Ledinek, N., Pajas, P., Žabokrtský, Z., and Žele, A. (2006). Towards a Slovene dependency treebank. In *Proceedings of the Fifth International Language Resources and Evaluation Conference, LREC 2006*, pages 1388–1391, European Language Resources Association (ELRA), Genova, Italy.
- [10] Haug, D. T. T. and Jøhndal, M. (2008). Creating a parallel treebank of the old indo-european bible translations. In Choukri, K., editor, *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008)*, pages 27–34.
- [11] Krek, S., Erjavec, T., Dobrovoljc, K., Holz, N., Ledinek, N., and Može, S. (2015). Training corpus – dependency treebank annotation. URL: <http://eng.slovenscina.eu/tehnologije/ucni-korpus>.
- [12] Marszałek-Kowalewska, K., Zaretskaya, A., and Souček, M. (2014). Stanford typed dependencies: Slavic languages application. In *PoITAL 2014, LNAI 8686*, pages 151–163, Springer International Publishing Switzerland.
- [13] McDonald, R., Nivre, J., Quirmbach-Brundage, Y., Goldberg, Y., Das, D., Ganchev, K., Hall, K., Petrov, S., Zhang, H., Täckström, O., Bedini, C., Bertomeu Castelló, N., and Lee, J. (2013). Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 92–97, Sofia, Bulgaria.
- [14] Nivre, J. (2015). Towards a universal grammar for natural language processing. In *CICLING 2015: 16th International Conference on Computational Linguistics and Intelligent Text Processing*, pages 3–16, Springer International Publishing Switzerland, Cairo, Egypt.
- [15] Petrov, S., Das, D., and McDonald, R. (2012). A universal part-of-speech tagset. In Calzolari, N., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Odijk, J.,

- and Piperidis, S., editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012*, pages 2089–2096, European Language Resources Association (ELRA), İstanbul, Turkey.
- [16] Ramasamy, L. (2014). *Parsing under-resourced languages: Cross-lingual transfer strategies for Indian languages*. PhD thesis, Charles University in Prague.
- [17] Rosa, R., Mašek, J., Mareček, D., Popel, M., Zeman, D., and Žabokrtský, Z. (2014). HamleDT 2.0: Thirty dependency treebanks stanfordized. In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., and Mariani, J., editors, *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, pages 2334–2341, European Language Resources Association, Reykjavík, Iceland.
- [18] Silveira, N. and Manning, C. (2015). Does Universal Dependencies need a parsing representation? An investigation of English. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 310–319, Uppsala, Sweden.
- [19] Šimková, M. and Garabík, R. (2006). Sintaksičeskaja razmolotka v slovaccom nacional'nom korpusе. In *Trudy meždunarodnoj konferencii Korpusnaja lingvistika – 2006*, pages 389–394, St. Petersburg University Press, Sankt-Peterburg, Russia.
- [20] Simov, K. and Osenova, P. (2005). Extending the annotation of BulTreeBank: Phase 2. In *The Fourth Workshop on Treebanks and Linguistic Theories (TLT 2005)*, pages 173–184, Barcelona.
- [21] Teufel, S. (1996a). ELM-DE: EAGLES specifications for German morphosyntax. Lexical specification and classification guidelines, technical report EAG-CLWG-ELMDE/F.
- [22] Teufel, S. (1996b). ELM-EN: EAGLES specifications for English morphosyntax. Lexical specification and classification guidelines, technical report EAG-CLWG-ELMEN/F.
- [23] Tiedemann, J. (2015). Cross-lingual dependency parsing with Universal Dependencies and predicted pos labels. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 340–349, Uppsala, Sweden.
- [24] Wróblewska, A. and Woliński, M. (2012). Preliminary experiments in Polish dependency parsing. In Bouvry, P., Kłopotek, M., Leprévost, F., Marciniak, M., Mykowiecka, A., and Rybiński, H., editors, *Security and Intelligent Information Systems*, volume 7053 of *Lecture Notes in Computer Science*, pages 279–292. Springer Berlin Heidelberg.
- [25] Zeman, D. (2008). Reusable tagset conversion using tagset drivers. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., and Tapias, D., editors, *Proceedings of the Sixth International Language Resources and Evaluation Conference, LREC 2008*, pages 28–30, European Language Resources Association (ELRA), Marrakech, Morocco.
- [26] Zeman, D., Dušek, O., Mareček, D., Popel, M., Ramasamy, L., Štěpánek, J., Žabokrtský, Z., and Hajič, J. (2014). HamleDT: Harmonized multi-language dependency treebank. *Language Resources and Evaluation*, 48(4):601–637.
- [27] Zeman, D., Mareček, D., Popel, M., Ramasamy, L., Štěpánek, J., Žabokrtský, Z., and Hajič, J. (2012). HamleDT: To parse or not to parse? In Calzolari, N., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 2735–2741, European Language Resources Association (ELRA), İstanbul, Turkey.

Appendix



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, MERCHANTABILITY, 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.

Creative Commons is not a party to this License, and makes no warranty whatsoever in connection with the Work. Creative Commons will not be liable to You or any party on any legal theory for any damages whatsoever, including without limitation any general, special, incidental or consequential damages arising in connection to this license. Notwithstanding the foregoing two (2) sentences, if Creative Commons has expressly identified itself as the Licensor hereunder, it shall have all rights and obligations of Licensor.

Except for the limited purpose of indicating to the public that the Work is licensed under the CCPL, Creative Commons does not authorize the use by either party of the trademark "Creative Commons" or any related trademark or logo of Creative Commons without the prior written consent of Creative Commons. Any permitted use will be in compliance with Creative Commons' then-current trademark usage guidelines, as may be published on its website or otherwise made available upon request from time to time. For the avoidance of doubt, this trademark restriction does not form part of the License.

Creative Commons may be contacted at <http://creativecommons.org/>.

Natural Language Processing, Corpus Linguistics, Lexicography

Editors Katarína Gajdošová and Adriána Žáková

Cover Design by Vladimír Benko
Typeset by Vladimír Radik and Radoslav Brída

Printed and published by RAM-Verlag
Stüttinghauser Ringstrasse 44, 58515 Lüdenscheid, Germany

170 pages

First edition at RAM-Verlag
Lüdenscheid 2015

ISBN 978-3-942303-32-3

<http://www.ram-verlag.eu>

