Corpus of Dialects of the Slovak National Corpus

Katarína Gajdošová, Radovan Garabík, and Mária Šimková

Ľ. 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 Slovak[1], 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 Ľ. 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 Ľ. 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 ofa 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 tokens[4].

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 sources[5].

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, Ľ. 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, Ferdinad: 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="nadlhl">	5 612
Pauliny, Eugen: Nárečie zátopových osád na hornej Orave. Spisy Jazykovedného odboru Matice slovenskej. Séria 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*

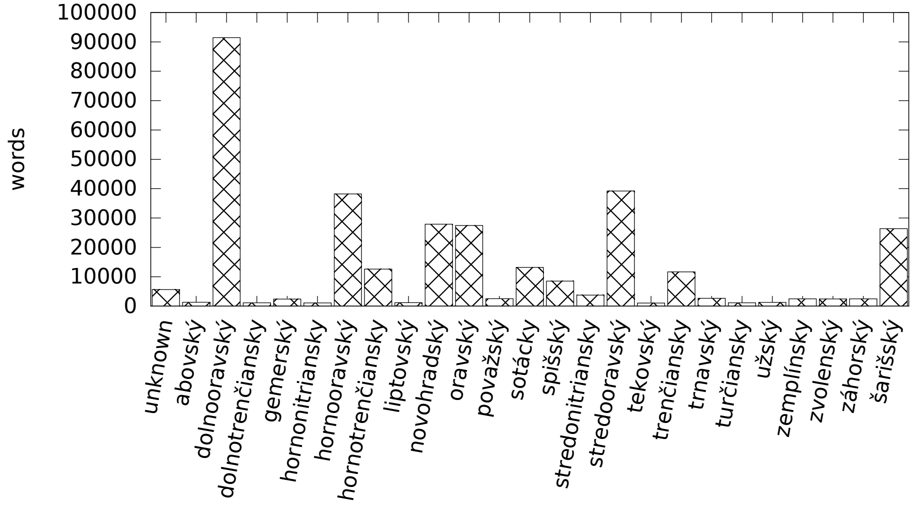


Fig. 1. The number of words from each dialect included in the corpus. The histogram shows the disparity of recorded information – the western Slovak dialects received very little attention.

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), localion (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 localion belongs is fully compatible with the list of locations[6] 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>	dial="y/n"	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
			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="lajbľíg bez rukávou̯">*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="lajbľíg bez rukávou̯">bekeše </rem>*

Table 2. Structural tags used in the CD SNC

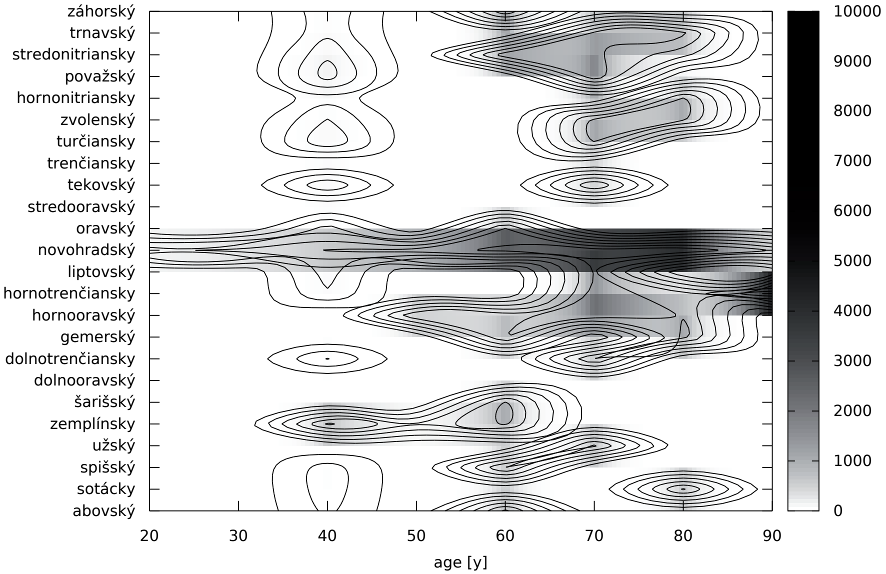
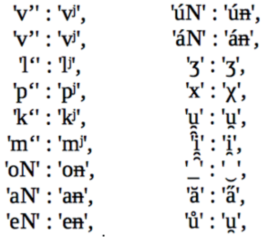


Fig. 2. Distribution of number of words (contours and shade of grey) in dialects (vertical axis, sorted approximately by geographical position from western (top) to eastern (bottom) Slovakia) by age (horizontal axis).

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" ([ɲẹi̯] inIPA) could be transcribed as both ňej or ňei̯, 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’' : 'vʲ',      'úN' : 'úꞥ',
'v’' : 'vʲ',      'áN' : 'áꞥ',
'l‘' : 'lʲ',       'ʒ' : 'ʒ',
'p‘' : 'pʲ',       'x' : 'χ',
'k‘' : 'kʲ',       'u̯' : 'u̯',
'm‘' : 'mʲ',       'î̯' : 'i̯',
'oN' : 'oꞥ',       '_̂' : '‿',
'aN' : 'aꞥ',       'ă' : 'ắ',
'eN' : 'eꞥ',       'ů' : 'u̯',
```

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 SNC[8]. 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., Genči, J., Furdík, K., Ivoríková, H., and Ivanecký, J. (2012). *The Slovak Language in the Digital Age*. Springer-Verlag, Berlin Heidelberg.