SYN2015: Representative Corpus of Contemporary Written Czech

Michal Křen, Václav Cvrček, Tomáš Čapka, Anna Čermáková, Milena Hnátková, Lucie Chlumská, Tomáš Jelínek, Dominika Kováříková, Vladimír Petkevič, Pavel Procházka, Hana Skoumalová, Michal Škrabal, Petr Truneček, Pavel Vondřička, Adrian Jan Zasina
Charles University in Prague, Faculty of Arts
Nám. Jana Palacha 2, 116 38 Praha 1, Czech Republic
{michal.kren, vaclav.cvrcek, tomas.capka, anna.cermakova, milena.hnatkova, lucie.chlumska, tomas.jelinek, dominika.kovarikova, vladimir.petkevic, pavel.prochazka, hana.skoumalova, michal.skrabal, petr.trunecek, pavel.vondricka, adrian.zasina}@ff.cuni.cz

Abstract

The paper concentrates on the design, composition and annotation of SYN2015, a new 100-million representative corpus of contemporary written Czech. SYN2015 is a sequel of the representative corpora of the SYN series that can be described as traditional (as opposed to the web-crawled corpora), featuring cleared copyright issues, well-defined composition, reliability of annotation and high-quality text processing. At the same time, SYN2015 is designed as a reflection of the variety of written Czech text production with necessary methodological and technological enhancements that include a detailed bibliographic annotation and text classification based on an updated scheme. The corpus has been produced using a completely rebuilt text processing toolchain called SynKorp. SYN2015 is lemmatized, morphologically and syntactically annotated with state-of-the-art tools. It has been published within the framework of the Czech National Corpus and it is available via the standard corpus query interface KonText at http://kontext.korpus.cz as well as a dataset in shuffled format.

Keywords: Czech, language corpus, representativeness

	size (# of running words)	type	publication year
SYN2000	100 mil.	general, representative	mostly 1990–1999
SYN2005	100 mil.	general, representative	mostly 2000–2004
SYN2006PUB	300 mil.	newspapers and magazines	1989–2004
SYN2009PUB	700 mil.	newspapers and magazines	1995–2007
SYN2010	100 mil.	general, representative	mostly 2005–2009
SYN2013PUB	935 mil.	newspapers and magazines	2005–2009
SYN2015	100 mil.	general, representative	mostly 2010–2014

Table 1: The SYN-series corpora of contemporary written Czech.

1. Background

The Czech National Corpus (CNC) aims at extensive and continuous mapping of the Czech language in the whole spectrum of its varieties and forms. This effort results in compilation, maintenance and providing access to a number of corpora (synchronic/diachronic, written/spoken, monolingual/parallel etc.), including corpora of contemporary written Czech making up the SYN series. The SYN-series corpora can be described as traditional (as opposed to the web-crawled corpora), featuring cleared copyright issues, well-defined composition, reliability of annotation and high-quality text processing (Hnátková et al., 2014).

2. Representative Corpora of the SYN Series

Currently, the SYN series consists of four 100-million representative corpora of written Czech (SYN2000, SYN2005, SYN2010, and SYN2015; the number denotes the corpus publication year) and three large newspaper corpora (SYN2006PUB, SYN2009PUB, SYN2013PUB). As all the corpora are disjoint, i.e. any document can be included only into one of them, their total size exceeds 2 billion tokens.

The representative SYN-series corpora cover four consecutive time periods in a regular five-year interval (with the exception of SYN2000 which covers the period of ten years, i.e. 1990–1999) and they contain a large

variety of written genres in proportions based on language reception studies (Králík & Šulc, 2005). Their design, comparability, strengths and weaknesses are described in more detail in Křen (2013:46–53).

The aim of this paper is to introduce SYN2015, a new representative corpus of contemporary written Czech published in December 2015. SYN2015 is a sequel of the representative corpora of the SYN series, but at the same time, it reflects necessary methodological and technological enhancements outlined below. The paper concentrates mainly on the design decisions, composition and annotation, with emphasis on the improvements that have been done in the context of the SYN series.

3. Design of SYN2015

The terms representativeness and balance describing corpora are sometimes used interchangeably. We will adhere to the following definitions:

- *representative* corpus contains a large number of texts that cover all the varieties the corpus aims to represent;
- *balanced* corpus contains these varieties in proportions that correspond to the reality of a (sub)language in question.

Although the notions of representativeness and balance are often challenged, they are essential whenever it is necessary to generalize research findings (based on a corpus as a sample) to larger population, i.e. the given (sub)language. That is why SYN2000, SYN2005 and SYN2010 were designed to be both representative and balanced. However, there are a number of considerations, related to corpus balance in particular, that have been taken into account while designing SYN2015:

- the population of texts which is to be represented by any general corpus is unknown;
- it is virtually impossible to measure the real proportions of language varieties in use (results of sociological surveys are often based on respondents' imprecise estimates);
- in practice, corpus composition is influenced largely by other factors, mainly the text classification scheme and selection of texts within the individual categories, rather than the exact balance;
- even ideal demographically balanced corpus (if ever compiled) does not guarantee true proportions of any particular language phenomena;
- corpus-based studies are increasingly aimed at more restricted varieties rather than language as a whole;
- user interfaces make it possible to easily examine the corpus composition and to make use of it effectively (tailor-made subcorpora).

Our opinion is that a general-language corpus should primarily attempt to cover the variety of existing texts and their well-designed and documented classification rather than trying to estimate their (very unstable and variable) proportions in a language. Our approach to SYN2015 corresponds to Biber's notion of representativeness in terms of "texts as products" (Biber, 1993:245). As a result, SYN2015 is designed as representative, but not claimed to be balanced.[1]

In particular, SYN2015 is designed as a representation of contemporary printed language of the last five-year period, i.e. 2010–2014. As the borders of synchronicity vary across the registers, the following criteria for inclusion of the individual texts into SYN2015 have been adopted (based on the three top-level categories, cf. below):

- fiction: publication date within the last 25 years and first publication date within the last 75 years;
- non-fiction: first publication date within the last 25 years;
- newspapers and magazines: publication date within the given five-year period.

The specific language of the internet (discussion forums, blogs etc.) is kept separately and will be covered by a newly-established NET corpus series.

The original text classification scheme of the SYN series has been updated and revised; both original and revised classifications are based on text-external criteria that reflect predominant function of a text. The revision has been made with respect to comparability with the original scheme, with the most significant change made to the sub-classification of non-fiction adopted from the Czech National Library and more detailed classification of newspaper texts (cf. Table 2).

In line with its predecessors, SYN2015 contains a large variety of texts from various publishers within the given classification category. A category is defined by a combination of two variables: *text type* and *genre*. Proportions of the particular categories in SYN2015 are set arbitrarily, yet close to the original figures. The proportions will be fixed and observed also in future representative corpora of the series. For instance, the three top-level categories of fiction / non-fiction / newspapers and magazines will share one third of the corpus each.

Next to the text type and genre, metadata related to the text classification and available for every document also include *medium* (book, journal, textbook etc.), *periodicity* (daily, weekly, monthly, less than monthly, non-periodical) and *audience* (general, children/youth). Standard division of the newspapers into the individual articles is also supplemented by their separate classification into 13 sections (politics, economics, sports, culture, leisure, commentaries etc.) and information about the author that is available for all prominent newspaper titles.

SYN2015 is hierarchically structured into *documents* (<doc>), composed from at least one text (<text>; newspaper articles, book chapters etc.). Texts are further divided into *paragraphs* (<p>) and *sentences* (<s>) so that every token is included into one of them. Each of these

1 Detailed rationale of these design decisions can be found in Cvrček et al. (in print).

text type	genre	category	proportion
fiction (FIC)			**33.33 %**
NOV		novels	26 %
COL		short stories	5 %
VER		poetry	1 %
SCR		drama, screenplays	1 %
X		other	0.33 %
non-fiction (NFC)			**33.33 %**
SCI - scientific POP - popular PRO - professional	HUM	humanities [sub-classified into: ANT - anthropology, THE - theatre, PHI - philosophy and religion, HIS - history, MUS - music, LAN - philology, INF - library and information science, ART - arts and architecture]	7 %
	SSC	social sciences [sub-classified into: ECO - economics, POL - politics, LAW - law, PSY - psychology, SOC - sociology, REC - recreation, EDU - education]	7 %
	NAT	natural sciences [sub-classified into: BIO - biology, PHY - physics, GEO - geography and geology, CHE - chemistry, MED - medicine, AGR - agriculture]	7 %
	FTS	technical sciences [sub-classified into: MAT - mathematics, TEC - technology, ICT - information and communications technology]	7 %
	ITD	interdisciplinary	1 %
MEM		memoirs, autobiographies	4 %
ADM		administrative texts	0.33 %
newspapers and magazines (NMG)			**33.33 %**
NEW	NTW	nationwide newspapers – selected titles [equal shares of HN, LN, MFD, Právo]	10 %
	NTW	nationwide newspapers – other	5 %
	REG	regional newspapers	5 %
LEI		leisure magazines [sub-classified into: HOU - hobby, LIF - life style, SCT - society, SPO - sports, INT - curiosities]	13.33 %

Table 2: Composition of SYN2015 in terms of the major classification categories.

structures is characterized by a set of attributes that always includes a unique identifier. Apart from these basic hierarchical structures, there are also two additional ones: *highlight* (<hi>; font style, emphasis etc., if available in the source text) and *line break* (<lb>; verse boundary). An overview of the information contained in the metadata and available for every individual corpus token is shown in Figure 1.

SYN2015 can be searched via KonText[2] interface which enables users also to examine corpus composition and to make use of the wide variety of included texts intuitively and effectively. In particular, it is possible to create subcorpora according to any selected combination of the

2 Corpus query interface developed at the CNC as a fork of the NoSketch Engine and based on Manatee as the backend (Rychlý, 2007; Machálek & Křen, 2013); KonText is available at http://kontext.korpus.cz/.

doc.title:	Q. E. D.	**doc.subtitle:**	Krása matematického důkazu
doc.author:	Polster, Burkard	**doc.issue:**	
doc.publisher:	Dokořán	**doc.pubplace:**	Praha
doc.pubyear:	2014	**doc.first_published:**	2014
doc.translator:	Pick, Luboš	**doc.srclang:**	en: angličtina
doc.authsex:	F: žena	**doc.transsex:**	M: muž
doc.txtype_group:	NFC: oborová literatura	**doc.txtype:**	POP: populárně naučná literatura
doc.genre_group:	FTS: formální a technické vědy	**doc.genre:**	MAT: matematika
doc.medium:	B: kniha	**doc.periodicity:**	NP: neperiodická publikace
doc.audience:	GEN: obecné publikum	**doc.isbnissn:**	978-80-7363-532-9
doc.biblio:	Polster, Burkard (2014): Q. E. D. Překlad: Pick, Luboš. Praha: Dokořán.	**doc.id:**	polst_qedkrasama
text.section:		**text.section_orig:**	
text.author:		**text.id:**	polst_qedkrasama:1
p.type:	normal	**p.id:**	polst_qedkrasama:1:192
s.id:	polst_qedkrasama:1:192:1	**hi.rend:**	bold italic

Figure 1: An example of the structural attributes with their values as shown by KonText.

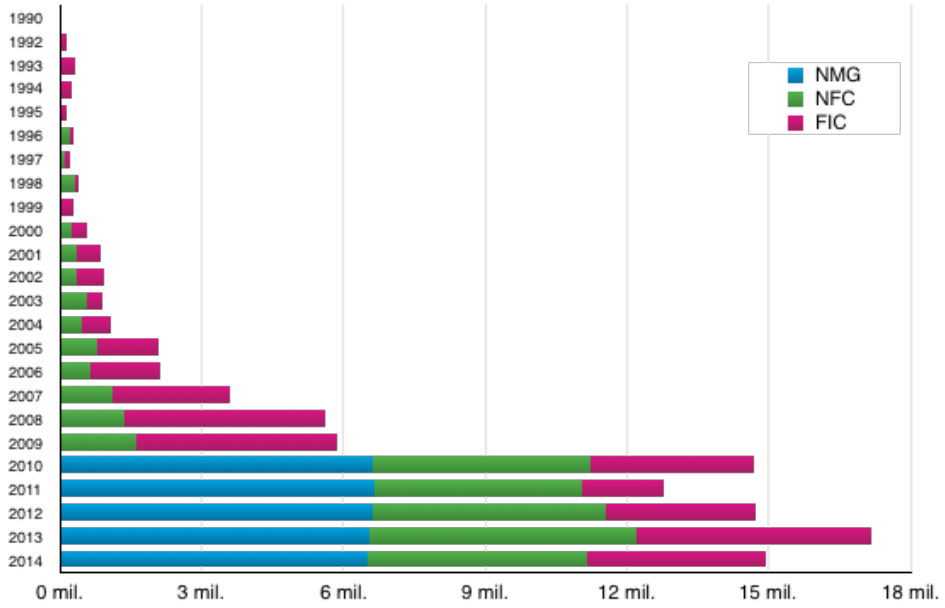

Figure 2: Proportion of fiction, non-fiction, newspapers and magazines in each year.

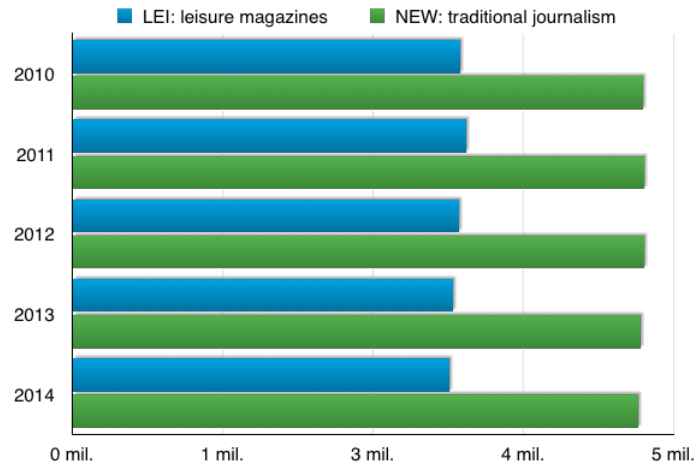

Figure 3: Proportion of traditional and leisure journalism within the newspapers and magazines in each year.

metadata in a step-by-step manner while observing constantly changing sizes of further available subparts. Moreover, a functionality for automatic selection of texts given by custom proportions of the individual categories is currently under development. It would give users the possibility to "balance" their own subcorpus in KonText. This feature reflects our effort to provide well-defined and reliably annotated corpus data so that users can easily make their choices about the (sub)language to study.

4. Text Processing Enhancements

Tools used for processing the SYN-series corpora combine fully automatic steps (foreign languages detection, de-duplication etc.) with human-supervised and even manual ones (text classification interface). This is necessary to keep high quality standards that are not compromised despite the growing amount of the data.

However, most of the tools have been in use for more than ten years and are thus already outdated. This is why a brand new system called SynKorp has been implemented and used for the production of SYN2015. SynKorp is an integration of the internal text administration and annotation database with the text processing toolchain into a single environment. Its main features include:

- full transition to XML and UTF-8; this allows e.g. retaining the original appearance (font style, footnotes etc.) of the texts wherever possible;
- complete redesign of the text processing toolchain using standard and up-to-date tools;
- modularity, i.e. easy substitution of the individual tools;
- revised architecture of the text administration database;
- new user interface;
- processing of the individual texts made configurable and reproducible, with easy inspection of the intermediate results.

SynKorp can be viewed as a control panel that allows its operator to group the individual texts into batches, to select the tools to process them with and to inspect the results, as well as to carry out all the necessary text classification and bibliographic annotation. The text processing is done basically in the following three steps:

- text acquisition and its registration in the internal database;
- conversion of the selected texts from the original format (doc, pdf, epub etc.) into the common intermediate format; this includes optional de-duplication on a document level (Onion; Pomikálek, 2011), paragraph-level foreign languages detection (Cavnar and Trenkle, 1994), detection and cleanup of paragraphs with prevailing contextless content (in-house tool);
- bibliographic annotation and text classification in a web-based user interface.

The output is an XML file with complete metadata stored in the database that is ready for further processing, e.g. lemmatization and morphological tagging. As a result, the text processing is much easier and faster with SynKorp while retaining the present quality.

5. Lemmatization and Morphological Tagging

The update includes also a major enhancement of the tokenization, sentence segmentation, lemmatization and morphological tagging including their adaptation to the new intermediate format. The lemmatization and morphological tagging consist of three main components (please refer to Hnátková et al., 2014 for details): a comprehensive morphological dictionary, rule-based disambiguation component and a stochastic tagger. The morphological dictionary has been continuously updated and enhanced in order to increase its coverage, with special attention to the performance on SYN2015 texts. The rule-based component includes rules that make use of syntactic properties of Czech (and operate on word chunks, phrases, or whole sentences), rules that identify and disambiguate phrasemes, and heuristic rules. The heuristic rules make use of specific contexts or they employ the frequencies of possible lemmas in the text. Tokens not completely disambiguated by the rule-based component are finally processed by the stochastic tagger.

D: delimiter (punctuation etc.)	16,5 %
M: morphologically unambiguous	21,3 %
R: rule-based disambiguation component	40,1 %
T: stochastic tagger	22,1 %

Table 3: Shares of the individual values of the *proc* attribute.

SYN2015 also includes a specific attribute *proc* that denotes the step of the tagging process responsible for full disambiguation of each token. Table 3 shows the contribution of every component to the final disambiguation of all tokens in the corpus (with the delimiters marked separately). The rule-based component not only fully disambiguates almost a half of the non-delimiters in the corpus, but table 4 shows that it also significantly decreases the ambiguity of tokens that have not been disambiguated in full.

	before	after
lemmas per word token	1.30	1.06
tags per word token	12.02	2.91
lemmas per ambiguous word token	1.40	1.17
tags per ambiguous word token	15.98	5.67

Table 4: Average number of lemmas and tags per word token (delimiters excluded) before and after application of the rule-based component.

6. Syntactic Annotation

SYN2015 is annotated with a new dependency syntactic markup based on formalism of the analytical layer of the Prague Dependency Treebank (PDT; Bejček et al., 2013). The PDT has been chosen over the newer Universal Dependencies[3] standard because it is closer to the traditional Czech syntax and therefore more familiar to users.

In dependency syntax, every token in a sentence is represented by a node in a tree graph, and depends either on another node in the sentence, or on an artificial external node representing the whole sentence (in this case, it is the root of the tree; usually the root is the finite verb in the main clause). Figure 4 shows an example of such a dependency representation of the sentence:

Téhle podobnosti se v rodině směje me.
'This similarity$_{DAT}$ refl. in family laugh$_{PL}$.'
We laugh at this similarity in our family.

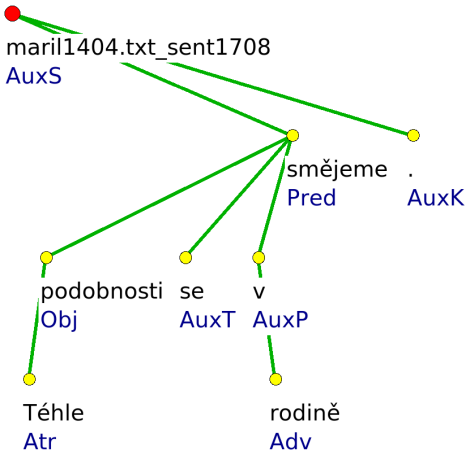

Figure 4: Example of a dependency structure.

The automatic syntactic annotation (parsing) was performed by TurboParser (Martins et al., 2013) trained on the PDT data. The TurboParser uses a fast and reliable parsing algorithm and achieves an accuracy of 87.23 % unlabelled attachment score (UAS), 81.38% labelled attachment score (LAS). To further increase parsing precision, we devised a data simplification procedure (Jelínek, 2014). It is based on a simple observation: parsers have to use lemmas or word forms to achieve high accuracy, they cannot rely on morphological tags only. In the training data, however, only a limited number of lemmas appear frequently enough for a reliable language modelling, and many words in new texts are out-of-vocabulary. On the other hand, there are many categories of words (such as numerals or several groups of proper names) with identical syntactic behaviour. We identify such categories and replace their members by proxies, reducing the variability of lemmas by approx. 20 %. For example, all feminine given names such as *Mary*, *Jane*, *Lucy* are replaced with just one proxy name, e.g. *Alice*. This procedure is applied both on the training data (the parser is trained on simplified data) and on the data to be parsed; in the latter case, the original forms and lemmas are preserved in a backup file. TurboParser with the text simplification method achieves an accuracy of 88.48/82.46 % UAS/LAS on Czech.

Most of the existing tools for full-fledged treebank querying, such as PML-TQ (Pajas and Štěpánek, 2009), were designed for relatively small treebanks and they hardly scale up to a 100-million token corpus. Nevertheless, syntactic annotation can be very useful to search for sentences matching specified syntactic properties even in simple "flat" architectures. Therefore, a set of (morpho)syntactic attributes has been devised for KonText and assigned to every token next to the morphological tag and lemma; these attributes can be used to query some syntactic properties of the tokens. Some syntactic attributes are related to the token itself, other attributes are related to its parent (governing node).

The (morpho)syntactic attributes related to the token itself are: *afun* (syntactic function as defined in PDT, e.g. *Obj*, *Sb*, *Atr*), *parent* (a number defining the relative position of the parent) and *prep* (lemma of a preposition; applies to syntactic nouns in PPs formally dependent on a preposition). Attributes related to the token's parent are: *p_pos*, *p_tag*, *p_form*, *p_lemma*, *p_afun*. For example, the word form "podobnosti" – 'similarity' in the example structure in Fig. 4 has the following attributes: afun="Obj", parent="+4" (the governing nodes position is 4 to the right), p_pos="V" (the governing node is a verb) etc. The attributes have been designed and tested to make querying the syntactic features as easy as possible, although KonText interface does not support user-friendly querying of syntactic structures yet.

7. Availability

SYN2015 has been released within the framework of the Czech National Corpus. It is accessible to all its registered users via the standard corpus query interface KonText[4] and it is also available to the research community as a dataset in shuffled format, i.e. randomly-ordered blocks of texts sized max. 100 tokens;[5] this requirement results from the agreements with publishers.

8. Acknowledgements

This paper resulted from the implementation of the Czech National Corpus project (LM2015044) funded by the Ministry of Education, Youth and Sports of the Czech Republic within the framework of Large Research, Development and Innovation Infrastructures.

3 http://universaldependencies.org/

4 http://kontext.korpus.cz/

5 LINDAT/CLARIN digital library at http://hdl.handle.net/11234/1-1593

9. References

Biber, D. (1993). Representativeness in Corpus Design. *Literary and Linguistic Computing* 8(4), pp. 243–257.

Cavnar, W. B., Trenkle, J. M. (1994). N-Gram-Based Text Categorization. In *Proceedings of Third Annual Symposium on Document Analysis and Information Retrieval*. Las Vegas, USA. Available online at http://www.info.unicaen.fr/~giguet/sigdat95/cavnar_trenkle_ngram.ps

Cvrček, V., Čermáková, A., Křen, M. (in print): Nová koncepce synchronních korpusů psané češtiny. *Slovo a slovesnost*.

Hnátková, M., Křen, M., Procházka, P., Skoumalová, H. (2014). The SYN-series corpora of written Czech. In *Proceedings of LREC 2014*. Reykjavík: ELRA, pp. 160–164. Available online at http://www.lrec-conf.org/proceedings/lrec2014/pdf/294_Paper.pdf

Jelínek, T. (2014). Improvements to Dependency Parsing Using Automatic Simplification of Data. In *Proceedings of LREC 2014*. Reykjavík: ELRA, pp. 73–77. Available online at http://www.lrec-conf.org/proceedings/lrec2014/pdf/228_Paper.pdf

Králík, J., Šulc, M. (2005). The Representativeness of Czech Corpora. *International Journal of Corpus Linguistics* 10(3), pp. 357–366.

Křen, M. (2013). *Odraz jazykových změn v synchronních korpusech*. Prague: NLN.

Machálek, T., Křen, M. (2013). Query interface for diverse corpus types. In *Natural Language Processing, Corpus Linguistics, E-learning*. Lüdenscheid: RAM Verlag, pp. 166–173.

Martins, A. F. T., Almeida, M. B., Smith, N. A. (2013). Turning on the Turbo: Fast Third-Order Non-Projective Turbo Parsers. In *Proceedings of ACL 2013*. Sofia.

Pajas, P., Štěpánek, J. (2009). System for Querying Syntactically Annotated Corpora. In *Proceedings of the ACL-IJCNLP 2009*. Singapore: Suntec, pp. 33–36.

Pomikálek, J. (2011). *Removing Boilerplate and Duplicate Content from Web Corpora*. PhD thesis. Brno: FI MU. Available online at http://is.muni.cz/th/45523/fi_d/phdthesis.pdf

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

10. Language Resource References

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., Zikánová, Š. (2013). *Prague Dependency Treebank 3.0*. Available online at LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague, http://hdl.handle.net/11858/00-097C-0000-0023-1AAF-3.