The SYN-series corpora of written Czech

Milena Hnátková, Michal Křen, Pavel Procházka, Hana Skoumalová

Charles University in Prague, Faculty of Arts
Nám. Jana Palacha 2, 116 38 Praha 1, Czech Republic
{milena.hnatkova, michal.kren, pavel.prochazka, hana.skoumalova}@ff.cuni.cz

Abstract

The paper overviews the SYN series of synchronic corpora of written Czech compiled within the framework of the Czech National Corpus project. It describes their design and processing with a focus on the annotation, i.e. lemmatization and morphological tagging. The paper also introduces SYN2013PUB, a new 935-million newspaper corpus of Czech published in 2013 as the most recent addition to the SYN series before planned revision of its architecture. SYN2013PUB can be seen as a completion of the series in terms of titles and publication dates of major Czech newspapers that are now covered by complete volumes in comparable proportions. All SYN-series corpora can be characterized as traditional, with emphasis on cleared copyright issues, well-defined composition, reliable metadata and high-quality data processing; their overall size currently exceeds 2.2 billion running words.

Keywords: written language, large corpus, Czech

1. Introduction

The Czech National Corpus (CNC) as a language research infrastructure strives for extensive and continuous mapping of the Czech language. This effort results in compilation, maintenance and providing access to a number of corpora that represent different language varieties. The paper concentrates on corpora of contemporary written Czech making up the SYN series and summarized in Table 1. Please note that the sizes are given in running words, i.e. tokens not including punctuation and numbers.

The series currently consists of 6 standard reference corpora. All of them are disjoint, i.e. any document can be included only into one of them. The reference corpora are invariable entities in order to ensure that identical queries always give identical results. Prefix “SYN” denotes their synchronic nature and it is followed by the year of publication (e.g. SYN2005 was published in 2005). As for their contents, there are two kinds of reference corpora: balanced and newspaper ones. The latter are composed solely of newspapers and magazines and this is denoted by their suffix “PUB”. The 3 balanced corpora cover 3 consecutive time periods and they contain a large variety of written genres (including also fiction and professional literature) subclassified into 74 categories in proportions based on language reception studies. The composition is described in detail in Křen (2013) that also discusses strong and weak points of their concept of representativeness and balance.

The reference corpora were processed with state-of-the-art versions of all tools (including tokenization, morphological tagging etc.) available at the time and remained unchanged since then. As a result, their processing may not correspond to today’s standards and it is also not comparable from one reference corpus to another. Therefore, the super-corpus SYN was introduced as a union of all the reference corpora consistently re-processed with the newest versions of available tools (Křen, 2009). Size of SYN equals to the sum of sizes of the respective original reference corpora (with a small difference caused by different tokenization) that are available as its subcorpora. SYN is a non-reference corpus intended to incorporate new corpora as the series grows over time, as well as to reflect improvements of the annotation. After publication of SYN2013PUB described in section 4, SYN version 3 is currently available with its size exceeding 2.2 billion running words.

All SYN-series corpora can be characterized as traditional (as opposed to the web-crawled corpora), with emphasis on cleared copyright issues, well-defined composition, reliable metadata and high-quality data processing outlined in the next section.

2. Text processing, standardization and clean-up

Text acquisition is based on agreements with publishers who also provide the CNC with texts in electronic format. The next step is unification of various input formats into a common intermediate format that includes also publisher-specific treatment (if necessary). This processing step is supervised by a human who selects the appropriate set of scripts, sets the thresholds and decides about possible adaptation of the individual programs. During this conversion step, all objects of non-textual nature (e.g. figures) are replaced with XML entities denoting their removal.

Manual (in case of non-periodicals) or semi-manual (periodicals) metadata annotation takes place afterwards. It includes adding both bibliographic information and evaluative text-type and genre markup. This is combined with quality control and possibility of (semi-)manual corrections of various kinds, as this is the last opportunity for human intervention. All subsequent processing is fully automatic and it includes the following procedures:

- Paragraph-level foreign languages detection.
- De-duplication (on document level).
- Detection and removal of paragraphs that contain too many numbers or punctuation symbols, too few characters with diacritics etc. The detection is based on thoroughly tested thresholds of these basic features and their combinations that indicate paragraphs with corrupted or contextless content (tables, lists etc.).

corpus	*size*	*type*	*contents*	*time span*
SYN2000	100 mil.	reference	balanced	most of the texts from 1900–1999
SYN2005	100 mil.	reference	balanced	most of the texts from 2000–2004
SYN2010	100 mil.	reference	balanced	most of the texts from 2005–2009
SYN2006PUB	300 mil.	reference	newspaper	newspapers and magazines from 1989–2004
SYN2009PUB	700 mil.	reference	newspaper	newspapers and magazines from 1995–2007
SYN2013PUB	935 mil.	reference	newspaper	newspapers and magazines from 2005–2009
SYN	2 232 mil.	non-reference	union	re-processed unification of the reference corpora

Table 1: Currently available SYN-series corpora.

It should be stressed that the texts are not corrected in any way, the whole paragraph is either left intact or deleted as a whole. The only exception are errors that result from incorrect text processing or obvious mistakes where the writer's intention was undoubtedly different from what can be found in the text, cf. subsection 3.1.

The data processing outlined in this section combines fully automatic steps with human-supervised and even manual ones. This is necessary to keep high quality standards that are not compromised despite the growing amount of the data (Křen, 2009).

3. Annotation

The SYN-series corpora are lemmatized and morphologically tagged. The lemmatization and tagging comprises of several stages: pre-processing, morphological analysis (including also segmentation and tokenization), post-morphological processing, disambiguation and final post-processing. More detailed description can be found in Hnátková et al. (2011).

3.1. Pre-processing

The pre-processing involves several operations which are very different in their nature:

- Replacing some characters or strings with XML entities; e.g. e-mail addresses and URLs are replaced with &email; and &url;, respectively.

- Joining some multi-word units together because of further morphological processing. For instance, the name *Mao Ce-tung* is inflected only in its second part as *Mao Ce-tunga*, *Mao Ce-tungovi* etc., i.e. only the nominative is used for *Mao*. However, if only the first part of the name is used, it gets inflected itself as *Maa*, *Maovi* etc. To be able to analyze the full name in its inflected forms during the later stages, both parts are joined together and the name is treated as a single unit during the morphological analysis. The full name is restored in stage 3.5.

- Treating hyphens. There is a productive forming of compounds in Czech, e.g. *česko-anglický* (Czech-English) where only the second part is inflected. As it is not reasonable to list all combinations in the dictionary, the first part is hidden, only the second part is analyzed and then the first part is restored again. Any token with a hyphen that should be treated in a different way must be explicitly marked.

  If both parts are inflected, then the token is separated, e.g. *Brně-Líšni* → *Brně - Líšni* (Brno-Líšeň, a city quarter).

  If the part after the hyphen is not an existing word, both parts are first joined together, and the hyphen is restored again after the analysis, e.g. *sci-firomán* → *sci_firomán* (sci-fi novel).

- Correcting errors caused by electronic processing of the text, either by OCR, PDF conversion, or simply wrong encoding of the texts, e.g. *AngIie* → *Anglie* (England), *pouś* → *pouť* (funfair), etc.

- Correcting other frequent mistakes which were not intended by the author, e.g. *0sm* → *Osm* (eight), *Zdeňěk* → *Zdeněk* etc.

  This includes also separating erroneously glued words, e.g. *vPraze* → *v Praze* (in Prague), and gluing erroneously separated parts of words, e.g. *fotografi e* → *fotografie* (photograph).

3.2. Morphological analysis

Segmentation (sentence boundary detection), tokenization and morphological analysis are interconnected and carried out by the tool called LanGR (Květoň, 2006). The morphological module itself is an adapted version of a tool originally developed by Jan Hajič (Hajič, 2004). It is based on a comprehensive dictionary containing ca 850 thousand lexical units. The dictionary is continually being updated and corrected, a new version of the module is published approximately two or three times a year.

The morphological analyzer assigns all possible morphological interpretations (i.e. a set of pairs consisting of morphological tag with corresponding lemma) to every token. This is done regardless of context, so that identical tokens always get the same set of lemma–tag pairs. The output of LanGR is thus a sequence of tokens divided into separate sentences, where every token is accompanied by a list of possible lemmas and tags.

For tagging the older corpora (SYN2000, SYN2005, SYN2006PUB), the original tagset provided by the morphological module was adopted with its 3 296 tags. As Czech is an inflectional language with high degree of syncretism, some of the tags contain variable values denoting

	morphology	safe rules	phrasemes	rules with heuristics
average no. of tags per token	12.81	2.71	2.7	2.17
average no. of tags per ambiguous token	19.60	5.84	5.85	5.04
percentage of unambiguous tokens	36.53 %	64.63 %	65.02 %	71.14 %

Table 2: Number of tags per token after the individual disambiguation steps.

any case, any gender, non-feminine gender etc. This version of the morphological module assigns 3.88 tags per token in the average.

For the newer corpora (SYN2009PUB, SYN2010, SYN2013PUB), the tagset was purged from the variables and only exact values for every category are used since then. The number of tags in thus adjusted current tagset is 5 711, and the average number of tags per token in a non-disambiguated text rose to 12.81.

3.3. Post-morphological processing

Before the disambiguation, some corrections to the output of the morphological analysis are made:

- Lemmas and tags are assigned to joined forms, e.g. *Mao_Ce_tungovi*, *sci_firomán* etc.
- Over-generated readings of homonymous word forms are eliminated, e.g. common word form *kořen* (root) also has a (virtually impossible) reading as passive participle derived from the verb *kořit se* (to humble oneself, to worship) which is eliminated in this stage.

3.4. Disambiguation

During the disambiguation stage, one lemma–tag pair is selected depending on the context of the particular token. To achieve this goal, a combination of rule-based and stochastic methods is used successively in the following 4 steps. The rule-based module (implemented in LanGR) performs shallow parsing of the sentence and eliminates grammatically inconsistent lemma–tag pairs (Petkevič, 2006).

3.4.1. Safe linguistic rules

The first step is very conservative and only safe rules are applied in order not to delete a correct lemma–tag pair. The system contains 3 191 safe linguistic rules; an example of such a rule is eliminating the prepositional reading of the word form *se* if it is followed by another *se*: *Nesnese se(Refl) se(Refl/Prep) sestrou.*

3.4.2. Phraseme identification module

The second step is identification of phrasemes and collocations (Hnátková, 2011). This enables disambiguation of their individual parts, which helps in further application of linguistic rules. For the identification, a list of phrasemes and collocations containing ca 32 thousand items is used. The individual items are not necessarily treated as fixed strings, they can be inflected or conjugated, their word order changed, or they can contain variables, i.e. slots that can be occupied by any word from a list, or by any word of a given part of speech.

For instance, the word form *hlavní* (main) is homonymous with both instr. sg. and gen. pl. of *hlaveň* (gun barrel). However, the phrase *hlavní nádraží* (main station) is listed and identified, so the latter reading can be eliminated.

3.4.3. Linguistic rules with heuristic

In the next step, the linguistic rules are applied once again, but this time including also heuristics to eliminate very improbable readings. For instance, vocative readings are deleted for all word forms with another, non-vocative reading. The system contains 3 509 rules with heuristics.

After this step, the average number of possible analyses per token drops down to 2.17. The total amount of work done by the linguistic rules and the phraseme module is shown in Table 2.

3.4.4. Stochastic tagger

The disambiguation is finished by stochastic tagger Morče (Votrubec, 2006). It is very robust and it does not require new training if the tagset or data change only slightly. It is also the best of all taggers tested on PDT data, with the accuracy of 95.12 % (Spoustová et al., 2007).

3.5. Post-processing

The final post-processing includes basically these different tasks:

- Restoring multi-word units joined together in the preprocessing stage.
- Adding verbal aspect to the tags. The morphological module itself does not provide this information, although it is available in the source lexicon. Therefore, a separate program is needed to add the aspect information.
- Re-classification of parts of speech in cases where the concept of the morphological module is different from what we believe is more founded (this concerns mainly the relation between adverbs and particles).
- Omitting categories that do not have good sense linguistically in a given combination, e.g. number with reflexive pronouns or tense with passive participles.

3.6. Evaluation of the tagging

The results are difficult to evaluate because there is no corpus manually tagged with the current tagset applied on the newer SYN-series corpora. An attempt at the evaluation was made in Skoumalová (2011) with reported accuracy 94.57 %. However, the results cannot be directly compared, as the evaluation had to be based on morphological data from the PDT (Hajič et al., 2006) that uses tags with variables, cf. subsection 3.2. Therefore, the reported accuracy is lower than the results of other evaluations (Spoustová et al., 2007) on the PDT tagset that report accuracy between 95.34 % and 95.68 % depending on the combination of taggers and linguistic rules. However, the PDT tagset is more coarse-grained and thus less precise.

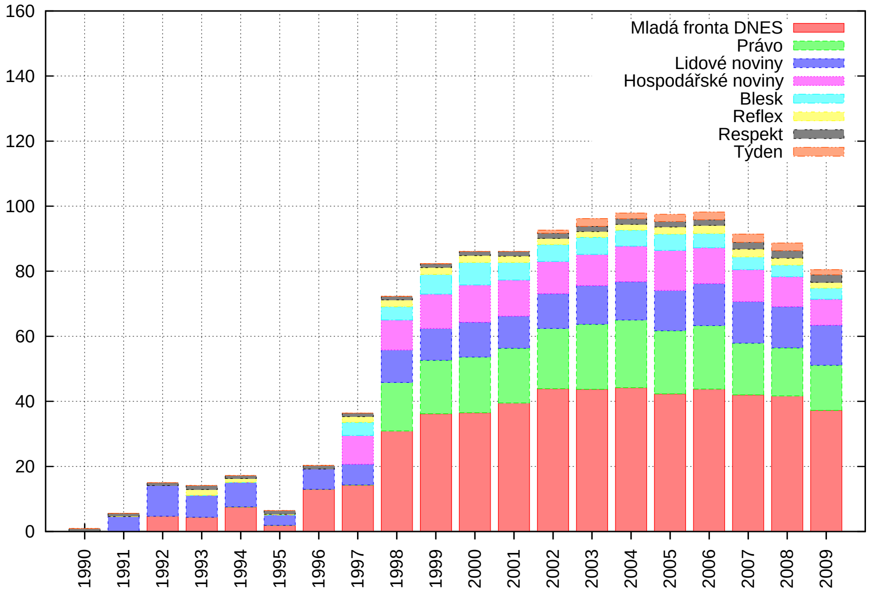

Figure 1: Composition of major periodicals in corpus SYN version 3.

In the near future, we plan to create an evaluation corpus manually tagged with the current tagset, in order to be able to measure the accuracy of the described tagging procedure.

4. SYN2013PUB

Version 2 of super-corpus SYN suffered from one drawback: under-representation of some combinations of title and time period in its newspaper part (Křen, 2013). This problem was tackled by processing a large number of respective texts that now make up corpus SYN2013PUB. It is a new corpus of written Czech newspapers and magazines published in 2013 and sized 935 million running words. It comprises 44 different newspaper and magazine titles from 2005–2009. The major part consists of the most influential national press and tabloids (Mladá fronta DNES, Právo, Hospodářské noviny, Lidové noviny, Blesk). Apart from them, popular non-specialized magazines (Reflex, Týden, Respekt) and internet periodicals (aktuálně.cz, Britské listy, centrum.cz, idnes.cz) are also included.

Regional newspapers are represented mostly by tens of the most widespread titles published by Vltava-Labe-Press and subsumed under Deníky Bohemia and Deníky Moravia. Apart from them, there are also 15 small independent local newspapers included into SYN2013PUB (e.g. Jihlavské listy, Kopřivnické noviny, Ostravská radnice, Prostějovský večerník, Roudnické noviny etc.). This shows that an effort was made to cover a large variety of different sources.

As a result of publication of SYN2013PUB, the non-reference super-corpus SYN has been updated accordingly and made available in its version 3. It contains texts of all the reference SYN-series corpora (including SYN2013PUB) re-processed with state-of-the-art versions of the tools. Size of SYN version 3 thus exceeds 2.2 billion running words, i.e. 2 685 million tokens including punctuation.

The major plus of corpus SYN is its composition, as it now covers complete volumes of all major Czech newspapers from 1998–2009 in comparable proportions (cf. Figure 1; the sizes are given in millions of running words). Therefore, SYN2013PUB can be seen as a completion of the SYN series in terms of titles and publication dates of major Czech newspapers.

5. Conclusion and future plans

Both SYN2013PUB and SYN version 3 are available to all registered users of the CNC via one of the standard interfaces (Machálek and Křen, 2013) at http://korpus.cz. Apart from that, both corpora are also available to the research community as datasets in shuffled format, i.e. randomly-ordered blocks of texts sized max. 100 tokens; this requirement results from the agreements with publishers.

In the near future, architecture of the SYN series will be revised to reflect the need for fresh data and to facilitate monitoring language change. Also the toolchain described in section 2 is currently being completely rebuilt using more standard and up-to-date tools, in order to speed up the data processing while retaining the present quality. These are the main reasons why we concentrated on processing and making available as much older data as possible in SYN2013PUB. As a result, the SYN-series corpora described in this paper will soon be followed by newly-processed enhanced corpora, including e.g. thematically-marked sections within periodicals, that will cover written Czech from 2010 onwards.

6. Acknowledgement

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.

7. References

Hajič, Jan, Panevová, Jarmila, Hajičová, Eva, Sgall, Petr, Pajas, Petr, Štěpánek, Jan, Havelka, Jiří, Mikulová, Marie, Žabokrtský, Zdeněk, and Ševčíková-Razímová, Magda. (2006). Prague dependency treebank 2.0. Linguistic Data Consortium, Philadelphia.

Hajič, Jan. (2004). *Disambiguation of Rich Inflection (Computational Morphology of Czech)*. Karolinum Charles University Press, Prague.

Hnátková, Milena, Petkevič, Vladimír, and Skoumalová, Hana. (2011). Linguistic Annotation of Corpora in the Czech National Corpus. In Труды международной конференции "Корпусная лингвистика – 2011", pages 15–20. St.-Petersburg State University, Institute of Linguistic Studies (RAS), Russian State Herzen Pedagogical University.

Hnátková, Milena. (2011). Výsledky automatického vyhledávání frazémů v autorských korpusech. In Petkevič, Vladimír and Rosen, Alexandr, editors, *Korpusová lingvistika Praha 2011 - 3 Gramatika a značkování korpusů*, pages 171–185, Prague. NLN.

Květoň, Pavel. (2006). *Rule-based Morphological Disambiguation*. Ph.D. thesis, MFF UK, Prague.

Křen, Michal. (2009). The SYN Concept: Towards One-Billion Corpus of Czech. In *Proceedings of the Corpus Linguistics Conference*, Liverpool.

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

Machálek, Tomáš and Křen, Michal. (2013). Query interface for diverse corpus types. In Gajdošová, Katarína and Žáková, Adriána, editors, *Natural Language Processing, Corpus Linguistics, E-learning*, pages 166–173. RAM Verlag, Lüdenscheid.

Petkevič, Vladimír. (2006). Reliable Morphological Disambiguation of Czech: Rule-Based Approach is Necessary. In Šimková, Mária, editor, *Insight into the Slovak and Czech Corpus Linguistics*, pages 26–44, Bratislava. Veda.

Skoumalová, Hana. (2011). Porovnání úspěšnosti tagování korpusu. In Petkevič, Vladimír and Rosen, Alexandr, editors, *Korpusová lingvistika Praha 2011 - 3 Gramatika a značkování korpusů*, pages 199–207, Prague. NLN.

Spoustová, Drahomíra, Hajič, Jan, Votrubec, Jan, Krbec, Pavel, and Květoň, Pavel. (2007). The Best of Two Worlds: Cooperation of Statistical and Rule-Based Taggers for Czech. In *ACL '07 Proceedings of the Workshop on Balto-Slavonic Natural Language Processing: Information Extraction and Enabling Technologies*, pages 67–74. ACL.

Votrubec, Jan. (2006). Morphological Tagging Based on Average Perceptron. In *WDS'06 Proceedings of Contributed Papers*, Prague. MFF UK.