sciendo
DOI 10.2478/jazcas-2023-0004

CORPUS OF SLOVAK LEGISLATIVE DOCUMENTS

RADOVAN GARABÍK
Jazykovedný ústav Ľudovíta Štúra SAV, v. v. i., Bratislava

GARABÍK, Radovan: Corpus of Slovak legislative documents. Jazykovedný časopis (Journal of Linguistics), 2022, Vol. 73, No 2, pp. 175–189.

Abstract: The article describes the construction of the corpus of Slovak legislative documents. By analyzing several statistical values of the source metadata and documents, we efficiently improve corpus quality. We describe the methods used to clean up small variations in metadata, length based discrimination of document and examine the effectiveness of several strategies of deduplication. The corpus is a part of a comparable corpus of legislative documents of seven languages, created in the *Multilingual Resources for CEF.AT in the Legal Domain (MARCELL)* project.

Key words: corpus, Slovak language, body of law, legislation

1. INTRODUCTION

Text corpora have become an indisputable and irreplaceable source of research data in linguistics during the last few decades. In Slovak linguistics, huge, representative corpora are available at least through their query interface, though there are situations where access to full texts is desirable, or outright necessary for either research or as training data. There is source of text data that is usually exempt from copyright protection (unlike web based texts) – legal acts, which are often specifically excluded from copyright laws (see Law No. 185/2015).

Presented article describes the analysis of existing Slovak body of law sources and a construction of linguistic corpus based on these data, with the aim to be used by linguists in a "traditional" linguistic research, as a source of free data for other researchers, but also as a useful domain specific text corpus for (e.g.) terminology research/application. The article is intended to describe contemporary processes and analysis applied in building Slovak language corpora.

The corpus is a part of seven comparable corpora of national legislative documents of seven countries – Bulgaria, Croatia, Hungary, Poland, Romania, Slovakia and Slovenia. The corpora have been developed within the Connecting Europe Facility (CEF) Telecom Action *Multilingual Resources for CEF.AT in the Legal Domain (MARCELL)*[1] that aims to enhance the eTranslation system[2] developed by the European Commission (see Váradi et al. 2020).

[1] http://marcell-project.eu

[2] https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/eTranslation

The corpus introduces several hitherto unused annotation rules and practices into Slovak corpora – it is the first (to our knowledge) publicly available Slovak corpus with a dependency parsing annotation, with named entities annotation, and the first corpus with a new, improved morphological description and lemmatization, the first corpus with an included terminological annotation.

2. OTHER LEGAL LANGUAGE CORPORA

The first Slovak corpus containing legal language was the *legal-1.0* corpus containing the then-current body of law of Slovak Republic, compiled at the Slovak National Corpus department of the Ľ. Štúr Institute of Linguistics. The first version has been available (for querying through a corpus manager interface, not for download) since 2014; somewhat unusually named, the version *legal-1.1* is not a revision of the corpus, but was built in parallel and contains only deduplicated (unique) texts.

The corpus *od-justice-1.0* has been compiled in 2018 (also by the Slovak National Corpus department of the Ľ. Štúr Institute of Linguistics) and contains the texts of court proceedings of the Slovak Republic, based on the Open Data Initiative releases.[3]

These corpora were, however, opportunistically created – not much effort has been put into analysis of end users' needs and requirements, and in any case, these corpora were one-time only, reflecting the body of law at that time, and they were not expanded by new texts.

3. SOURCE DESCRIPTION

The source archives have been provided by the Slov-Lex portal[4], the official Legislative and Information Portal of the Ministry of Justice of the Slovak Republic, providing access to the body of law of Slovak Republic. The source archive (updated monthly) contains also working versions of laws; however, in our corpus, we deal only with the final, published (and therefore in effect) versions. The size of the archive is 7.3GiB of zipped data in XML format. The data are organized in a tree structure, first level is the type of documents, second level the year of the document. The overall number of legislative documents (without appendices and attachments) is 61 627; the number of final version documents is 23 070. Because the data is changing with each new release of the archive, we analyze and describe here one specific snapshot, from 1st April 2019.

3.1 Source Metadata Analysis

In the source, we analyzed following metadata keys as relevant for the corpus (there are several other metadata items, but only these are reliable and relevant for corpus annotation):

[3] https://obcan.justice.sk/opendata

[4] https://www.slov-lex.sk

- *predpisOznacenie* – (unique, within one type) identifier of the document, as published in the official and legally binding Collection of Law/Official Journal
- *predpisTyp* – type of the document (law, resolution, declaration etc…)
- *predpisDatum* – date of the document, in human readable form, in the genitive, e.g. "z 2. decembra 2011," (including the comma)
- *predpisNadpis* – second (!) part of the name of the document, can be empty
- *predpisPodnadpis* – first (!) part of the name of the document, can be empty

Each of these data can be empty sometimes, however, if the *predpisOznacenie* key is empty, it is a sign of conversion error in the source data; similarly, empty value of both *predpisNadpis* and *predpisPodnadpis* signals an error in the source data.

We construct the (human readable) title as a concatenation of strings *predpisTyp* + *predpisDatum* + *predpisPodnadpis* + *predpisNadpis* (separated by spaces). Title constructed this way forms a readable, grammatically correct description, including the date of publication. It is entirely possible such a title was originally in the database or annotation of scanned documents and the metadata in the archive have been constructed by splitting the original title.

The metadata is not uniform – there is some amount of typos, spelling and formatting errors, sometimes there are fields with obviously swapped values, and the word order use of capitalization is inconsistent. All this strongly suggests there was a lot of manual work involved, perhaps in transforming existing pre-electronic data.

We also noticed there is a non negligible amount of documents in Czech in the laws and regulations from before 1990. These can be often detected by looking at the *predpisTyp* value – if the description is in Czech, the text data will be in Czech as well.

3.2 Initial Metadata Cleanup

The key *predpisTyp* is the most important in initial cleaning up of the database, because it determines the classification of the document and its usefulness for inclusion into the corpus.

In the first stage, we replaced erroneous *predpisTyp* values by their corrected forms, thus eliminating easily detectable typos and eliminating incorrectly converted and annotated documents (with no reasonable way of correcting them). In the first step, we converted the value into uppercase (uppercase is the canonical form of the value, everything else is to be considered a typo or a mistake). After this case normalization, there are 314 unique values for this key, with 237 of them appearing only once. The next step consists of replacing alternate values with the canonical form (see Table 1 for the most frequent replacements). To replace the values, it is sufficient if the alternate value (first column of the table) is a substring of *predpisTyp*. Special value *blacklist* is reserved for unrecognized values and values obviously in Czech are not processed further. After these cleanups, the numbers of documents of given types are in Table 2.

original	replacement
VLÁDNE NARIADENIE	NARIADENIE VLÁDY
USNESENÍ*)	blacklist
ZÁKONNÉ OPATRENIE	OPATRENIE
NAŘÍZENÍ VLÁDY*)	blacklist
VYHLÁŠKA MINISTRA ZAHRANIČNÝCH VECÍ	VYHLÁŠKA
VLÁDNA VYHLÁŠKA	VYHLÁŠKA
ZÁKONNÉ OPATŘENÍ*)	blacklist
UZNESENIE VLÁDY	UZNESENIE
OPATRENIE1)	OPATRENIE
ZÁKONNÍK PRÁCE	ZÁKON
OPATŘENÍ*)	blacklist

Table 1. List of replacements of the predpisTyp values. Entries marked with *) are in Czech and they are excluded from further processing.

nr of documents	document type (*predpisTyp*)
5089	VYHLÁŠKA (decree)
4775	OZNÁMENIE (announcement)
4243	ZÁKON (law)
1805	NARIADENIE VLÁDY (government regulation)
936	blacklist
597	OPATRENIE
392	REDAKČNÉ OZNÁMENIE
205	UZNESENIE
189	ROZHODNUTIE
140	NÁLEZ
103	ÚSTAVNÝ ZÁKON
42	VÝNOS

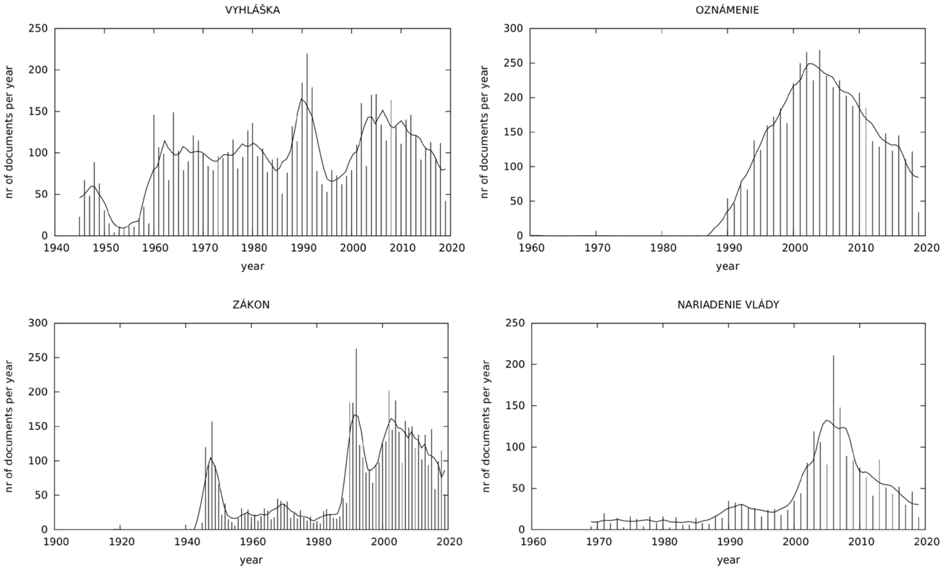
Table 2. Number of documents of given type.

First four types are important, since they are legally binding with their legal power equivalent to that of a law. Also ÚSTAVNÝ ZÁKON (constitutional law) has been merged with ZÁKON (law). REDAKČNÉ OZNÁMENIE is a notification about the publishing process – mostly corrections of errors in published texts (but not typos). The language is very repetitive and it quotes sentences from other documents. Although often legally binding and important from a legal point of view, it is quite unusable for NLP training, and strictly speaking, these documents do not consist of full continuous texts, so we exclude them from further processing.

4. STATISTICAL DATA

4.1 Temporal Distribution of Documents

We look into temporal distribution of the documents by the year they went into effect (i.e. the date of their publication in the official journal). The distribution is depicted on Figures 1–4, we show both the raw per-year count of documents and a smoothed moving average, with a window size of 5 years. The year 1969 marks the establishment of the Czechoslovak Federation, and the conversion of Slovak National Council into a parliament of the republic with legislative powers (clearly observable in the temporal distribution of documents with the *predpisTyp* value "NARIADENIE VLÁDY").



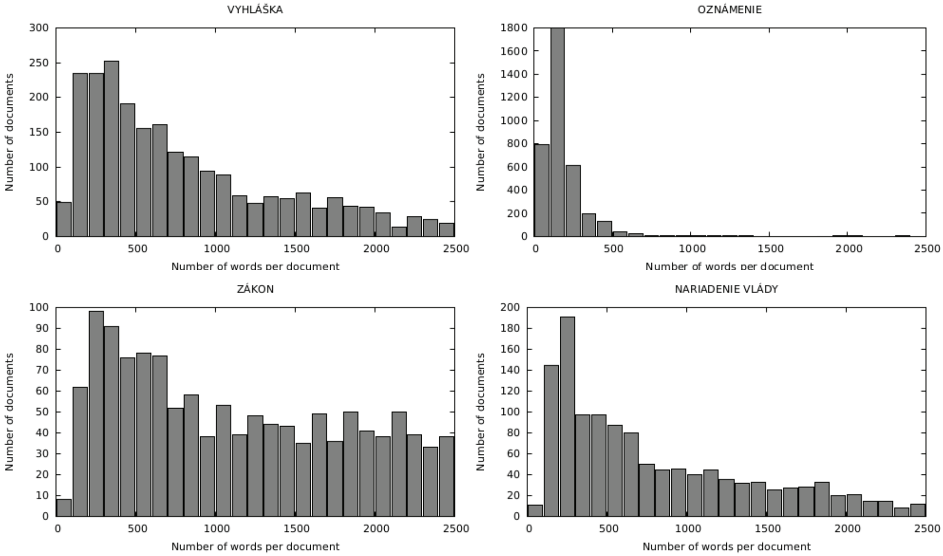
Figures 1–4. Number of documents per year. Smooth line marks values smoothed by moving average, window size of 5 years.

From a corpus linguistic point of view, users prefer (in synchronic corpora) to consider only texts containing "sufficiently contemporary" language, and the sociolinguistic situation comparable to the current one. We considered several possibilities for such a cut off date – the end of 1989, which coincides with profound changes in the Slovak society (Velvet Revolution, the transformation from socialism into modern European society); or the year 1993, marking the independence of Slovakia (conveniently not long after a minor orthography reform in 1991); for NLP training purposes a good cut off seems be at the local minima around 1997 (which also marked significant changes in political situation in Slovakia – the culmination of an authoritarian regime and illiberal democracy). For the purposes of the *MARCELL* project, we have chosen 1993 as the starting point for the released data (while the documents from the previous years are still kept in the corpus) for the CEF AT. For linguistic research, the version of the corpus available at https://www.juls.savba.sk/legalcorp.html and via the NoSketchEngine interface contains texts starting with the year 1955, the date when the last significant orthography reform went into effect being January 1st 1954, and assuming it took some time for the new orthography to stabilize (this is the same approach as adopted in the Slovak National Corpus).

4.2 Length Distribution of the Documents

Many of the documents are short amendments to existing laws, containing repetitive text and sentence fragments, and as such not really suitable for corpus linguistic purposes. Since these amendments are often rather short (compared to "regular" laws) and the legislative documents vary considerably in their lengths, we hope to get an insight into the status of the texts by looking at their length distribution. We focused on the time period where the majority of the documents is concentrated, i.e. from the year 2000 and later.

Following histograms (Figures 5–8) depict a number of documents per their length in tokens. For the sake of clarity, we display only documents up to 2500 words long, there is otherwise a long shallow-sloped tail towards extremely long documents.



Figures 5–8. Histogram of document length, in tokens. Only the beginning of the length axis is shown.

For the laws (*predpisTyp* ZÁKON) we hoped to find a region in the lower lengths where law amendments are concentrated. Unfortunately, this turns out not to be so sharply defined and we had to resort to a more sophisticated classification, described in the next section.

4.3 Document Length Threshold – Manual Annotation

We randomly selected a set of documents within certain length limits (in tokens) and manually annotated the documents for their suitability to be included in the corpus. The general rule of the annotation was if there was enough of "new" text in the document. If there is roughly more than half of new text (added paragraphs etc.), the document is annotated as suitable. The results of the annotation for laws are succinctly depicted at Figure 9.

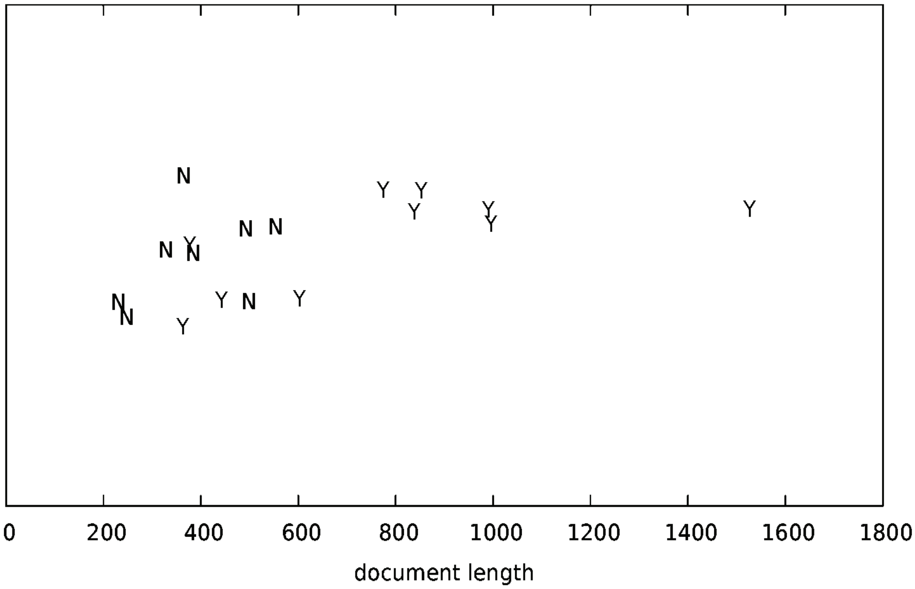


Figure 9. Manually annotated documents (*predpisTyp* ZÁKON). 'Y' means the document should be included in the corpus, 'N' it should not be included. Vertical positions are randomized for better visualization.

We used a logistic regression model to train a classifier on the manually annotated data, assuming equal weights for the documents and labels. The length threshold under which the documents would not be included in the corpus is 388 tokens. Of course, logistic regression for such a simple problem can be replaced by other methods; however, the advantage is a straightforward interpretation of the parameters of the regression and simple selection of probability cutoff and its eventual modification, if desired.

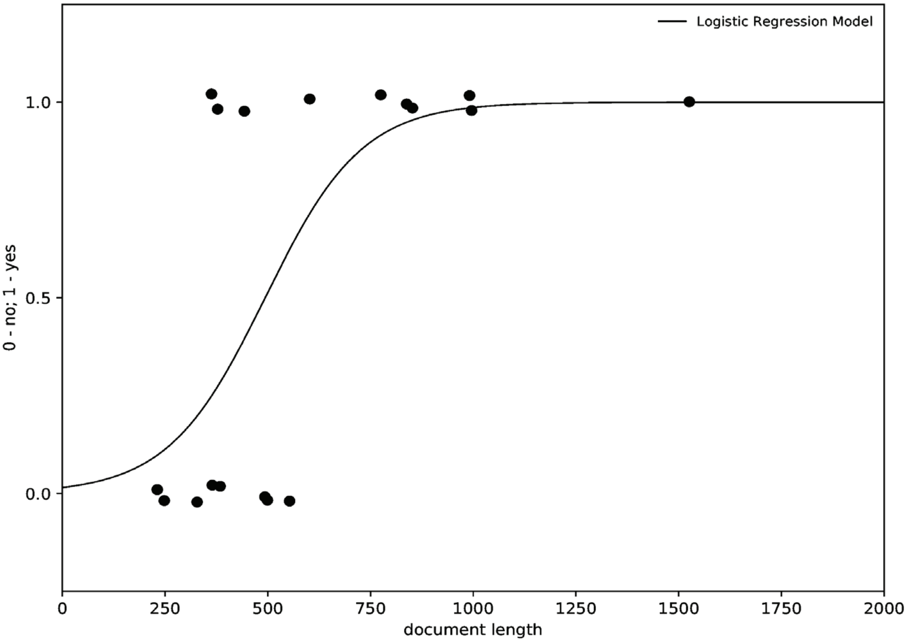


Figure 10. Manually annotated documents (*predpisTyp* ZÁKON) for their inclusion into the corpus; value 0 means they should not be included, value 1 they should be included, vertical positions are slightly randomized for better visualization. The curve shows a logistic regression model.

For comparison, governmental regulations (NARIADENIE VLÁDY) exhibit much sharper length dependency (see Figure 11), with a document length threshold of 855 tokens. However, other document types do not show any reasonable dependency between their lengths and their suitability.

4.4 Deduplication

Legal language is well known for a huge amount of fixed phrases, repetitions and mutual similarity between different documents. This is even more pronounced in the case of legislative texts, and especially with amendments, that often either cite previous versions, or change some sentences only very slightly. For correct language model training, removing duplicities is therefore paramount.

We are using the tools *onion*[5] for deduplication (Pomikálek 2011). *onion* expects vertical text, each line is one token or an XML-like tag, with two special tags <doc> to delimitate documents and <p> paragraphs. Deduplication process can

[5] http://corpus.tools/wiki/Onion

be fine tuned to specific requirements of input texts, the most important parameters are the n-gram length to be compared and the similarity threshold, beyond which the documents are flagged as duplicities.

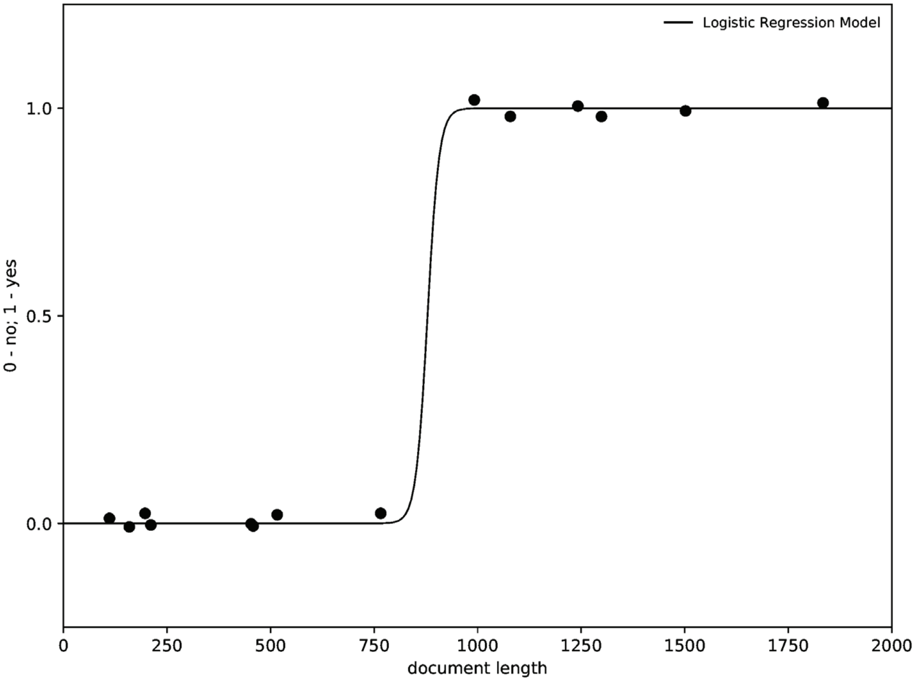


Figure 11. Manually annotated documents (NARIADENIE VLÁDY) for their inclusion into the corpus; value 0 means they should not be included, value 1 they should be included, vertical positions are slightly randomized for better visualization. The curve shows a logistic regression model.

We evaluated four different deduplication strategies:

- deduplicate on the paragraph level (i.e. remove duplicate paragraphs)
- deduplicate on the sentence level (remove duplicate sentences)
- deduplicate on the paragraph level, but consider all numbers (numerals written in digits) to be equal (technically achieved by replacing all consecutive digits with 0); in other words, we replaced all the numbers with a placeholder (this approach is marked as “no digits” in following table and figure)
- deduplicate on the sentence level, numbers replaced with a placeholder (marked as “no digits”)

Unifying numbers in legal language texts is justified by numerous repeating sentences that differ only in the (numeric) law ID or the year they mention, extremely common is the fixed phrase “as specified in the law NUMBER/YEAR, section NUMBER”. In the following analysis, we are working on texts from the year 1990 and later.

threshold	strategy			
	paragraphs	paragraphs, no digits	sentences	sentences, no digits
0	5774338	4245649	6240452	4706050
0.2	7443710	6106281	7539086	6189892
0.4	10062575	8556360	9998727	8480724
0.6	12736750	11056295	12580439	10886920
0.8	15638292	13720368	15393495	13454778
0.9	17286887	15189806	16982277	14863244
0.95	18424580	16205575	18078741	15834553
0.975	19212892	16904831	18825453	16494295
0.999	19783367	17360954	19286206	16847338
1	28388789			

Table 3. Size of deduplicated corpus (in tokens) as a function of the threshold parameter, paragraph and sentence levels, with or without placeholder numbers. Threshold=1 means the corpus is not deduplicated. Comparing 7-gram contexts.

Comparison of different strategies and varying the threshold is recorded in Table 3. There is a notable difference when the numbers were replaced with a placeholder, whereas selecting sentences as opposed to paragraphs does not change the size of the corpus significantly (noticeable only in the extremities, if using very low threshold, where the more coarse-grained segmentation into paragraphs means the whole paragraphs got deleted while not being really similar).

We conclude that there is no obvious quality/size cutoff point for the deduplication threshold, we should choose it according to the requirements of the task the corpus is used for. But the deduplication as such is useful to remove at least those completely identical pieces of text (the limit at the graph above where the threshold→1^{-}), and for many real life situations, a lower threshold should be applied. In our corpus, we use context size of 7 tokens and the threshold 0.5, to remain compatible with other Slovak language corpora of the ARANEA family (Benko 2014).

We also investigated a dependency on n-gram deduplication context. For the sake of brevity, we do not include the results here, but in general, it reproduces the outcome described in (Pomikálek 2011), with trigrams giving unsuitable performance, tetragrams exhibiting some level of improper deduplication, and longer contexts giving comparably reasonable quality, with no explicitly objectively best selection of parameters.

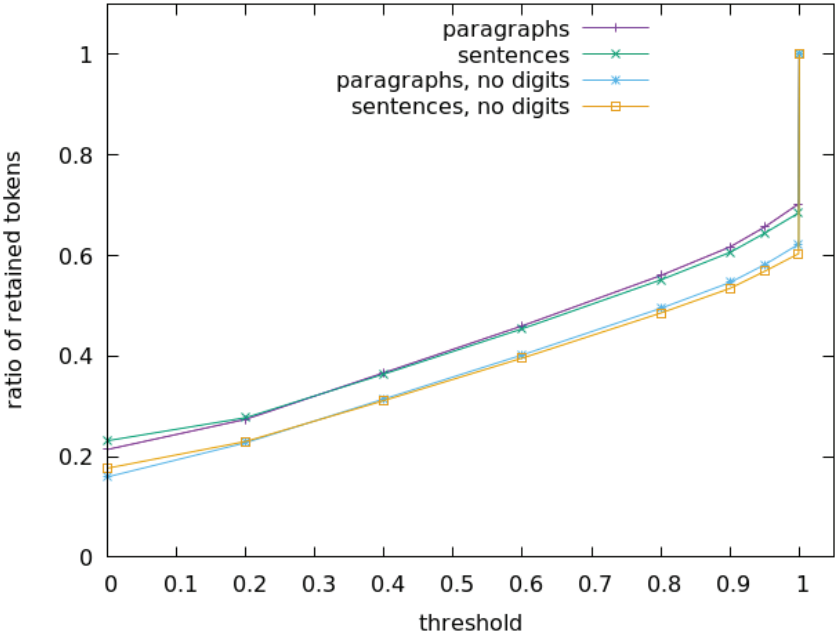


Figure 11. Ratio of tokens retained after deduplication, for different deduplication strategies. Comparing 7-gram contexts.

Since the desirability of deduplicated texts varies depending on the intended use of the corpus, we do not remove duplicate texts, we just annotate them (at the paragraph level by a XML attribute *dup* of the tag <p>), thus leaving the choice of using the information in the hands of the users of the corpus.

5. CORPUS ANNOTATION

5.1 Metadata

Corpus metadata structure is based on the source metadata – the documents have the following attributes: *number* is the identifier of the document (law), as given in the Collection; *docid* is a unique identifier of the document in the corpus (derived from the number), *entype*="announcement" is an English name of the document type (one of: *decree*, *announcement*, *law*, *regulation*, *measure*, *resolution*, *decision*, *finding*, *act*), *type* is the Slovak original name of the type, *nadpis* and *podnadpis* are the second and the first parts of the original document title, *title* is a human readable extended title of the document, *date* is the year when the law went into effect, *tokcount* is the number of tokens in the document, *tokcountdd* is the number of tokens after deduplication.

The paragraph has only one attribute, *dup* (set to “1” if the paragraph is a duplicate of a previous one, “0” otherwise) that can be used to filter search results or to create a subcorpus.

5.2 Lemmatization & MSD tagging

Since the Slovak language belongs to the group of moderately inflected languages, corpus based linguistic research and a reasonably large part of NLP applications directly depend on correct lemmatization and MSD tagging in Slovak corpora. State-of-the-art lemmatization and tagging for Slovak is performed by the MorphoDiTa tagger (Straková et al. 2014) that we trained on manually lemmatized and MSD annotated Slovak corpus *r-mak-6.0* of 1.2 million tokens.[6] The accuracy of lemmatization on general texts is 98.2%; the accuracy of MSD tagging 94.2%; the accuracy of POS tagging 98.1% and the combined accuracy of lemmatization+MSD tagging is 93.5% (Garabík – Mitana 2022). Unfortunately, we cannot easily estimate the accuracy on the texts from the legal language domain; nevertheless, since the training corpus *r-mak-6.0* contains one manually lemmatized and MSD annotated legal text, although strictly speaking not a law (*Programové vyhlásenie vlády*[7]), we can obtain at least a rough estimate by training a separate tagger model on the rest of the corpus and calculating the accuracy on the one legal text. For the lemmatization, we get 99.4%, for MSD tags 97.1%, POS 99.5%, and the combined lemmatization+MSD tagging 96.8%. This accuracy is significantly better and the improvement can be ascribed mostly to lower amount of out of vocabulary (OOV) words (these are lemmatized and MSD tagged using a combination of statistical and heuristic algorithms) – the ratio of OOV words in the *r-mak-6.0* corpus is 3.33%, in the whole corpus of Slovak Legislative Documents it is 1.35%, but in the *Programové vyhlásenie vlády* OOV words comprise 1.18% of the tokens.

5.3 Dependency Annotation

Dependency annotation is performed by the UDPipe (Straka – Straková 2017), Slovak model version 2.4, trained on a subset of Slovak Dependency Treebank (Šimková, Garabík 2006). We replace the lemmatization and MSD annotation provided by UDPipe by the output of the MorphoDiTa tagger trained on Slovak corpus, as described in the previous section. The annotation is accessible in two NoSketch Engine attributes, *head* (head of the current token within the sentence) and *deprel* (universal dependency relation to the head), as used in the Universal Dependencies platform (Zeman 2017). However, only queries related to the

[6] https://korpus.sk/r-mak/

[7] Policy Statement of the Government

dependency relation are straightforwardly supported by the NoSketch Engine CQL syntax, queries related to the head quickly become rather unwieldy. The dependency tree visualization can be accessed via a clickable URL in the *s.tree* attribute of the NoSketch Engine interface, using a simple purpose-built CGI wrapper around the *conllu-viewer* software.[8]

5.4 Named Entities Recognition

Named Entities Recognition is performed by the NameTag 1 package.[9] We used transfer learning to train the NameTag on the machine-translated[10] Czech Named Entity Corpus 2.0. The corpus has been filtered by removing sentences with named entities containing Czech specific letters (ě,ř,ů), sentences containing named entities that could not be automatically identified in the Slovak translation (by comparing their stems using a simple stemming algorithm, taking into account predictable morphological and orthographic changes between Czech and Slovak), sentences where the entities were not in the same order. After filtering, the Slovak corpus contains 6735 sentences (75% of the original amount), 13173 named entities (37% of the original amount). The F-measure of this model is 31.9%; the reason for such a low score is the excessive filtering which removes a substantial number of named entities from the train data. A new manually annotated Slovak named entity corpus is being prepared at the Slovak National Corpus department of the Ľ. Štúr Institute of Linguistics by the time of writing this article, preliminary estimates give the F-measure of 75.8%, compared to the Czech 77.8% (Straková et al. 2014).

5.5 IATE Terminology Annotation

Interactive Terminology for Europe – IATE1 is the terminology database of the European Union. It includes all of the previously existing EU terminology databases (Johnson, Macphail 2000) in all official European Union languages. The corpus is annotated by the Slovak IATE database, version from January 2020, without the *multiple languages* and *Latin* entries. The annotation is achieved by simple pattern matching of the tokens by assigning them their corresponding IATE identifiers, if the token is a term or a part of a multi-word term (Garabík – Levická 2022). Since there is no disambiguation of multiple-meaning terms, the annotation is preferably used in offline settings, for e.g. automatic domain classification or in cross-referencing law texts across several languages in the comparable corpora of the MARCELL project. The annotation is accessible as a NoSketch Engine attribute and can be used e.g. to retrieve or filter concordancies by IATE identifiers.

[8] https://github.com/rug-compling/conllu-viewer

[9] https://ufal.mff.cuni.cz/nametag/1

[10] Using a well-known commercial machine translation system: https://translate.google.com

6. SUMMARY

The corpus of Slovak body of law is available (as a part of a comparable corpus of seven-languages) via the ELRC-SHARE platform.[11] For interactive use, it is indexed in a NoSketch Engine corpus manager, accessible[12] via the webpage of the Ľ. Štúr Institute of Linguistics (no registration needed), providing also an alternate download location for the data, where the original texts are exempted from copyright projection, and our additional annotation and corpus data are (to the extent we might hold copyright and database rights to them) released under the creative commons CC0 license, i.e. with as little restrictions as possible.

Bibliography

BENKO, Vladimír (2013): Data Deduplication in Slovak Corpora. In: K. Gajdošová – A. Žáková (eds.): *Slovko 2013: Natural Language Processing, Corpus Linguistics, E-learning*. Lüdenscheid: RAM-Verlag, pp. 27–39.

BENKO, Vladimír (2014): Aranea: Yet Another Family of (Comparable) Web Corpora. In: P. Sojka – A. Horák – I. Kopeček – K. Pala (eds.): *Text, Speech and Dialogue. 17th International Conference, TSD 2014*, Brno, Czech Republic, September 8–12, 2014. Proceedings. LNCS 8655. Springer International Publishing Switzerland, pp. 257–264.

GARABÍK, Radovan – BOBEKOVÁ, Kristína (2021): Lematizácia, morfologická anotácia a dezambiguácia slovenského textu – webové rozhranie. In: *Slovenská reč*, Vol. 86, No. 1, pp. 104–109.

GARABÍK, Radovan – LEVICKÁ, Jana (2022): Naïve Terminological Annotation of Legal Texts in Slovak – Can it Be Useful?. In: *Rasprave: Časopis Instituta za hrvatski jezik i jezikoslovlje*. Vol. 48, No. 1, pp.2022, pp. 27–44.

GARABÍK, Radovan – MITANA, Denis (2022): Accuracy of Slovak Language Lemmatization and MSD Tagging – MorphoDiTa and SpaCy. In: *LLOD Approaches for Language Data Research and Management*, Abstract Book, Mykolo Romerio universitetas, Vilnius, pp. 93–95.

GARABÍK, Radovan – ŠIMKOVÁ, Mária (2012): Slovak Morphosyntactic Tagset. In: *Journal of Language Modelling*, No. 1, pp. 41–63.

JOHNSON, Ian – MACPHAIL, Alastair (2000): *IATE-Inter-Agency Terminology Exchange: development of a single central terminology database for the institutions and agencies of the European Union*. Workshop on Terminology resources and computation.

Law No. 185/2015 col. (copyright law of the Slovak Republic)

POMIKÁLEK, Jan (2011): *Removing Boilerplate and Duplicate Content from Web Corpora*. Ph.D. Thesis, Faculty of Informatics, Masaryk University in Brno.

ŠIMKOVÁ, Mária – GARABÍK, Radovan (2006): Синтаксическая разметка в Словацком национальном корпусе. In: *Труды международной конференции Корпусная лингвистика – 2006*. Sankt-Petersburg: St. Petersburg University Press, pp. 389–394.

[11] https://www.elrc-share.eu

[12] https://www.juls.savba.sk/legalcorp.html

STRAKOVÁ, Jana – STRAKA, Milan – HAJIČ, Jan (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.* Baltimore, Maryland, June 2014. Association for Computational Linguistics, pp. 13–18.

STRAKA, Milan – STRAKOVÁ, Jana (2017): Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In: *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Vancouver, Canada, August 2017.

VÁRADI, Tamás – KOEVA, Svetla – YALAMOV, Martin – TADIĆ, Marko – SASS, Bálint – NITOŃ, Bartłomiej – OGRODNICZUK, Maciej – PĘZIK, Piotr – BARBU MITITELU, Verginica – ION, Radu – IRIMIA, Elena – MITROFAN, Maria – PĂIȘ, Vasile – TUFIȘ, Dan – GARABÍK, Radovan – KREK, Simon – REPAR, Andraž – RIHTAR, Matjaž. (2020): The MARCELL Legislative Corpus. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference.* Marseille, France. May 2020. European Language Resources Association, pp. 3761–3768.

ZEMAN, Daniel (2017): Slovak Dependency Treebank in Universal Dependencies. In: *Jazykovedný časopis*, Vol. 68, No. 2, pp. 385–395.