UDK 811.162.4’322
811.162.4’373.46:34
Izvorni znanstveni rad
Rukopis primljen 10. XII. 2020.
Prihvaćen za tisak 16. XI. 2021.
doi.org/10.31724/rihjj.48.1.2

Radovan Garabík
Jana Levická
Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences
Panská 26, SK-81101 Bratislava
orcid.org/0000-0003-1691-3157
radovan.garabik@kassiopeia.juls.savba.sk
orcid.org/0000-0001-6027-604X
jana.levicka@korpus.juls.savba.sk

NAÏVE TERMINOLOGICAL ANNOTATION OF LEGAL TEXTS IN SLOVAK – CAN IT BE USEFUL?

Correct automatic terminological annotation of texts in a corpus can be sometimes a challenging task, especially for moderately or heavily inflected languages with relatively free word order. We explore the possibility of simple annotation based on sequence matching of lemmatized texts to annotate Slovak language corpus with IATE terminological entries. The accuracy of annotating legal language is very good when annotating multiword terms, while accuracy of single-word terms can be increased by applying simple filters based on word lengths and blacklisting most frequent false positives.

1. Introduction

By terminological annotation we understand assigning terminological concepts to individual terms contained within the texts (as opposed to annotating the whole documents, a different task that constitutes a document classification problem). On one end of the annotation stands an approach taken in a well-known manually annotated corpus of medical articles CRAFT (Bada 2010), annotating words and multiword expressions by their terminological identifier (according to a given terminology database). On the other end, by extending the annotation to its logical conclusion, we get the Semantic web, annotating many items within full blown semantic ontology (Hitzler 2021). Wikipedia, at least in its original conception, stands aside, because it directly marks units (words) only by their relation to a relevant article, not the semantic value of the unit by itself; nevertheless, this is enough to be successfully used for automatic semantic annotation of texts, including (almost as a side effect) also terminological annotation (Brank et al. 2017).

Correct automatic terminological annotation of texts in a corpus can be sometimes a challenging task, taking into account the process of determinologisation of terms in real language, reflecting all kinds of morphological and syntactic variants terms can take, especially considering inflected languages with relatively free word order. Although contemporary advances in Natural Language Processing evoke cautious optimism in dealing with hitherto intractable problems (for example, it is demonstrated by the ongoing revolution in NLP caused by unreasonable effectiveness of transformers, see Vaswani et al. 2017), applications of such advanced methods often hit barriers either in the form of non-existing terminological or linguistic resources for non-major languages or an inappropriate investment in time, effort and research activity to bring such an annotation to acceptable levels of quality. As a result, general terminological annotation in language corpora is either not performed at all, or only with a limited scope. The goal of this article is to present a way of simple annotation based on sequence matching of lemmatized texts and discrimination on surface-level word attributes based solely on the term length, and to evaluate the suitability of annotating a Slovak language corpus of legal texts with IATE terminological entries. The annotation is aimed at legal language and is developed and tested on a corpus of laws of the Slovak Republic (Váradi et al. 2020).

2. Theoretical Background

2.1. Slovak Language

The Slovak language belongs to the West Slavic group of Slavic languages. It is the official and dominant language in Slovakia.

As a "typical" Slavic language it can be characterized as a medium-level inflected language with three or four genders, six or seven cases, two grammatical numbers, three tenses and two verbal aspects interacting in a rather complex way. Adjectives are inflected for gender, number and case and have to agree with the noun in these categories. The language can be characterized as a generally head-initial and subject–verb–object language with a relatively free word order. These general grammatical features of the language play an important role in the setup of our work described in this article.

2.2. Terminological Annotation

As a result of terminological annotation as defined in the introduction we obtain a word or a multiword expression marked by an identifier denoting its terminological status, provided the word or expression is a terminological unit (out of a given set of terminological units). Such a terminologically annotated corpus provides richer possibilities for research by allowing queries to be combined with queries for terminological units or general queries to be specified within a defined terminological area, and to obtain statistical information about various aspects of terminological usage.

Common problems of automatic terminological annotation (relevant for Slovak, but valid universally in languages of equivalent complexity) are:

lexical or semantic homonymy – a word can be a terminological unit but it can also have an unrelated meaning

overlapping terms – most likely some constituent or a sequence of constituents of a multiword term is a terminological unit by itself

inflections – terms in terminology databases tend to be in their "base form" i.e. singletons are lemmatized though there are exceptions; noun phrases have their head in the base form but the remaining words must agree with their syntactical and valency positions, adjectives are in gender agreement (while the base form of an adjective is masculine) with following nouns, non-concordant attributes, etc.

word order – if the word order is not absolutely rigid, multiword expressions might have their constituents shuffled, or other modifiers inserted inside them other variations: abbreviations, ellipses, typographical variations, insertions etc.

In general, making correct automatic terminological annotation belongs to a class of word level classification and disambiguation problems, such as lemmatization, morphological description or syntactic parsing. While considered to be an almost solved problem, successful approaches (considering medium to high inflected languages, since languages with simple and predictable morphology are easier to work with because they could be lemmatized and POS tagged algorithmically, archetypal example of such an algorithm is the Porter Stemming Algorithm (Porter 1980)) are based on statistical methods solving classification problems, necessarily using reasonably large quantities of manually annotated training data. For comparison, see Coman et al. 2019 about terminological annotation in Romanian – the authors do not consider lemmatization, but instead devise a method of "compressing" word orthographic representation (by removing certain vowels and sequences of letters) and show a good accuracy when matching the compressed forms. This is feasible because Romanian is a low inflected language and the compression method efficiently takes care of what inflections remain in the language, but also of some word derivations.

2.3. Interactive Terminology for Europe – IATE

Interactive Terminology for Europe – IATE[1] 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, though there is a disparity in coverage and quality between the "original" EU languages and the languages of the "new" countries, including Slovakia. IATE is not static, though – new terms are being added, errors are fixed, wrong terms deleted.

The public data of IATE are considered not copyright protected and are available for download, the database structure use (for personal, non-commercial or commercial purposes) is explicitly authorised by the European Union. This makes the use of IATE attractive in corpus linguistics and research environ-

[1] https://iate.europa.eu/.

ments where one of the main problems is licensing and legal availability of various language resources.

In the article, we are using the Slovak IATE database, version from January 2020, without the *multiple languages* and *Latin* entries (such entries are generally used translingually in several domains). Slovak IATE entries are classified into 21 domains according to Eurovoc descriptors, the most numerous of them being the domain of the European Union, social questions, agriculture, forestry and fisheries, finance, law, and industry.

Though IATE represents a terminology database, its entries do not feature only genuine terms that "designate a general concept in a particular language" (User's Handbook, p. 24). For translation purposes, the database includes four other types of entries: abbreviations (including acronyms, initialisms, contractions or truncations), phrases (without terminological status but occurring repeatedly and having a "standard translation"), formulae (chemical formulae, mathematical and other scientific expressions) and short forms (e. g. the common name of an agreement or the short, unofficial name of a country). Strictly speaking, we could remove such entries from the annotation process; however, this would have no bearing on the validity of the automatic annotation process, neither it introduces significant changes in the accuracy of the annotation (with the exception of short abbreviations and acronyms that form a significant part of the set of single-word terms).

2.4. Terminology of our Article

Since the length of expressions (both in characters and words) is one of the key components in our work, we feel the need to define several terms we are going to use throughout the article:

token is a basic unit of text, typically corresponds to one word, numeral or a punctuation character

word-length is the number of tokens the term consists of (i.e. words or numerals – we already discard punctuation characters from the terms)

single-word term is a terminological unit consisting of only one word

character-length is the length of the term in characters. Usually, space separating words is counted as one character, however, we apply the character-length only to single-word terms, therefore the status of space is irrelevant; we consider the character sequences “ch”, “dz”, “dž” to be two characters long each[2].

3. Linguistic Preprocessing of the Corpus and Terminology Database

3.1. Lemmatization and Preprocessing

Lemmatization, i.e. finding the base form of the word is an indispensable step in compiling any reasonable corpora of moderately inflected languages. Given close relations between parts of speech, grammatical categories and lemmas in Slavic languages, the lemmatization process is often closely coupled with part of speech marking and full morphological/grammatical description (MSD). Although logically a separate step, disambiguation (selecting the correct, or the most probable, variant out of several possibilities) on the space of lemmas and/or MSD tags is often included in the lemmatization. This is also the case of the major Slovak corpora (Benko 2014; Slovenský národný korpus 2020), where there are two attributes automatically assigned to each of the tokens in the corpus – lemma and MSD tag, and we accept (approximately) 95 % accuracy of the process.

Because our terminological annotation is performed on the space of lemmas, the prerequisite of our terminological annotation is the lemmatization of the corpus (i.e. the texts to be terminologically annotated). As the next step, we also lemmatize the entries in the IATE database, and perform IATE term matching by comparing the lemmatized strings. This means that no semantic disambiguation of IATE terms has been carried out.

Since the punctuation in typical Slovak texts can be somewhat flexible (compared to the formal rigidity of prescribed rules of punctuation), we compare the terms discarding differences in punctuation characters, which is achieved by

[2] These digraphs are considered one letter/characters each in Slovak. In order to avoid confusion, we are not following this custom.

stripping the punctuation from the lemmatized IATE database and corpus texts in the process of matching.

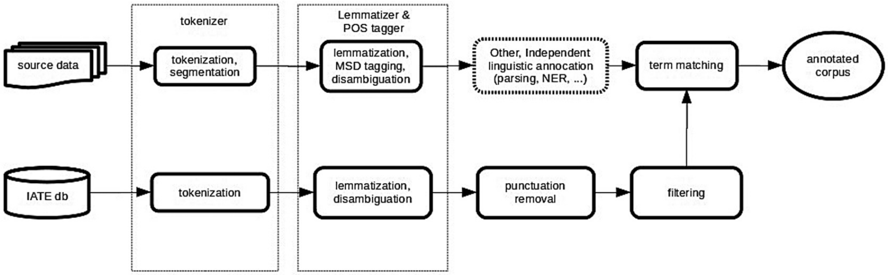
tokenizer
Lemmatizer & POS tagger
source data
tokenization, segmentation
lemmatization, MSD tagging, disambiguation
Other, Independent linguistic annocation (parsing, NER, ...)
term matching
annotated corpus
IATE db
tokenization
lemmatization, disambiguation
punctuation removal
filtering

Figure 1: Flowchart demonstrating the corpus processing pipeline

We started with an unaltered Slovak part of the IATE database containing 46 399 terms (single or multiword ones). After lemmatization, the number of unique entries is reduced to 46 219 terms. Many of the duplicates are caused by inclusion of both the singular and plural versions of the term, and some limited number of them by differences in capitalization (common vs. proper nouns, if the lemmatization process failed to recognize the proper noun). Some of the examples are shown in Table 1.

Table 1: Examples of duplicates appearing after lemmatization

lemmatized term	unlemmatized first occurrence	IATE ID	unlemmatized second occurrence	IATE ID
kompletný krmivo	kompletné krmivo	IATE-756110	kompletné krmivá	IATE-756110
hospodársky zviera	hospodárske zvieratá	IATE-3567039	hospodárske zviera	IATE-756184
pektín	pektín	IATE-757931	pektíny	IATE-828369
účtovný záznam	účtovný záznam	IATE-759239	účtovné záznamy	IATE-1078023 IATE-2142004 IATE-3574468
kritérium	kritériá	IATE-3574505	kritérium	IATE-760097
dlhový nástroj	dlhový nástroj	IATE-761216	dlhové nástroje	IATE-3570705
rabat	rabat	IATE-764721	Rabat	IATE-1891482
technický rezerva	technická rezerva	IATE-1070688	technické rezervy	IATE-765015 IATE-3545012 IATE-3563359

We call our annotation “naïve”, because it is perhaps the most simple way of annotating the corpus if we have an external (to the corpus) terminological dictionary (or a database) – the annotation consists of matching sequences of lemmatized tokens in the corpus to the lemmatized terminology database and annotating matching sequences by their IATE IDs. Matching is performed only within one sentence, though there are some entries in the database consisting of several sentences. Such entries will not be matched in the text – an examination of them revealed that these are longer descriptions, not typical terminology entries (we do not expect them to appear verbatim in the annotated texts anyway). An example of such a sentence-spanning entry is IATE-3521330: *“Kyanoakrylát. Nebezpečenstvo. V priebehu niekoľkých sekúnd zlepí pokožku a oči. Uchovávajte mimo dosahu detí.”*

While matching, we ignore punctuation characters in both the annotated text and the terminology database, and we always consider the longest (by word-length, not character-length) possible match. The reasoning is that longer terms are more probably genuine ones, because the probability of randomly occurring sequences of words identical to those appearing in the longer terms is low.

After punctuation removal, the number of unique entries is reduced from previous 46 219 terms to 46 181 terms (not many such duplicates are present, relatively speaking), many of them having the same IATE ID. Table 2 shows several duplicates, where the terms originally differed only in punctuation. Note that having the same IATE ID is expected (the entries are then just variations of the term).

Table 2: Examples of duplicates appearing after punctuation removal

terms	IATE IDs
double - no - touch opcia double no touch opcia	IATE-3565103 IATE-3570082
medziregionálna skupina Regióny Pobaltia medziregionálna skupina „Regióny Pobaltia“	IATE-3519021 IATE-3519021
medziregionálna skupina Víno medziregionálna skupina „Víno“	IATE-3519027 IATE-3519027
zásada „znečisťovateľ platí“ zásada “znečisťovateľ platí”	IATE-764076 IATE-3533381

3.2. Analysis of Term Lengths

We reason that examining term length can bring valuable insights into the text analysis and can improve our goal of automatic terminology annotation, therefore as the first step, we take a closer look into the distribution of term length after punctuation removal (lemmatization does not affect word-length[3]) in the IATE database.

The distribution of word-lengths of the terms is shown in Table 3 (just terms of "reasonable" length[4]) and Figure 2 (the complete set of terms). Mean word-length of the terms is 3.46, median and mode are 2.

Table 3: Number of terms of given word-length

word-length	number of terms
1	7607
2	16247
3	8207
4	4652
5	3017
6	1967
7	1268
8	813
9	534
10	386
11	308
12	229

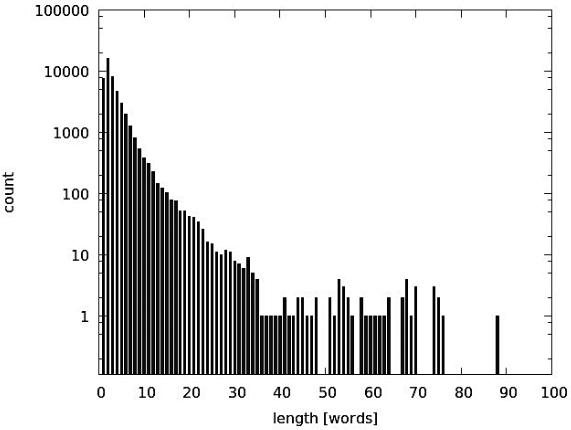

Figure 2: Distribution of word-lengths of terms in the IATE database (note the y-axis is logarithmic)

[3] Although there are some agglutinated word forms in Slovak, they are lemmatized by agglutinated lemmas as well, keeping the bijection between a word form and a lemma – following the principles laid out in the predominant lemmatization used in big Slovak language corpora (Garabík, Bobeková 2021; Garabík, Šimková 2012).

[4] The longest entry, in fact a phrase in IATE terms, is 121 tokens long (!) IATE-930909 and we replicate it verbatim: "*Zmluva medzi Belgickým kráľovstvom, Dánskym kráľovstvom, Spolkovou republikou Nemecko, Helénskou republikou, Španielskym kráľovstvom, Francúzskou republikou, Írskom, Talianskou republikou, Luxemburským veľkovojvodstvom, Holandským kráľovstvom, Rakúskou republikou, Portugalskou republikou, Fínskou republikou, Švédskym kráľovstvom, Spojeným kráľovstvom Veľkej Británie a Severného Írska (členskými štátmi Európskej únie) a Českou republikou, Estónskou republikou, Cyperskou republikou, Lotyšskou republikou, Litovskou republikou, Maďarskou republikou, Maltskou republikou, Poľskou republikou, Slovinskou republikou, Slovenskou republikou o pristúpení Českej republiky, Estónskej republiky, Cyperskej republiky, Lotyšskej republiky, Litovskej republiky, Maďarskej republiky, Maltskej republiky, Poľskej republiky, Slovinskej republiky a Slovenskej republiky k Európskej únii*".

3.3. Single-word Terms

One potentially very problematic issue is the existence of single-word terms; often they are homonymous with short function words (commonly a preposition or a conjunction) in Slovak.

There are two distinct issues present here: one is the lemmatization of acronyms (written in all capitals), sometimes homonymous with a (potentially inflected) “normal” Slovak word, and subsequently sometimes (depending on context) lemmatized into a lowercase base form of the word. This is less of a concern when dealing with multiword terms, since we can safely assume the same lemmatization will be applied to the text in the corpus and the lemmatized term will match the lemmatized text (even if the lemmatization “behind the scenes” is incorrect).

However, in case of single-word terms, such a lemmatization will more often than not produce an incorrect base form, while the corresponding occurrence in the text will be (again, thanks to contextual information) lemmatized into the correct (unchanged) word form.

The same argument can be made for proper names – single-word terms in title-case (starting with a capital letter and continuing in minuscule) could be erroneously lemmatized as words in minuscule, while in the text, the lemmas will be correctly capitalized. Therefore we decided not to lemmatize single-word terms written in capital letters, because we can expect single-word terms to be already in their base form[5].

[5] ^{m}However, we found an exception – the only one three-letter term that is not in the base form is the word *dal* (IATE-3507619, lemmatized as an L-participle of the verb *dať*).

Table 4: Single-word terms, distribution of character-lengths

length	count
1	9
2	97
3	923
4	940
5	757
6	636
7	671
8	720
9	691
10	619

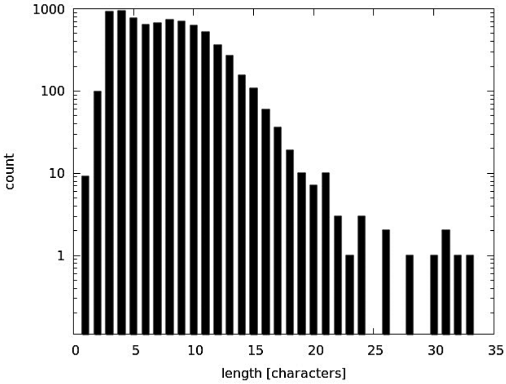

Figure 3: Distribution of character-lengths of single-word terms (the y-axis is logarithmic)

Table 5: Single-word uppercase terms, distribution of character-lengths

character-length	count
1	8
2	91
3	846
4	756
5	368
6	119
7	45
8	18
9	3
10	2

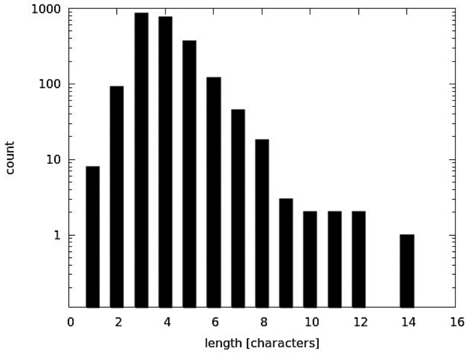

Figure 4: Single-word uppercase terms, distribution of character-lengths (the y-axis is logarithmic)

Table 6: Single-word titlecase terms, distribution of character-lengths

character-length	count
1	8
2	4
3	5
4	22
5	38
6	51
7	53
8	47
9	38
10	24

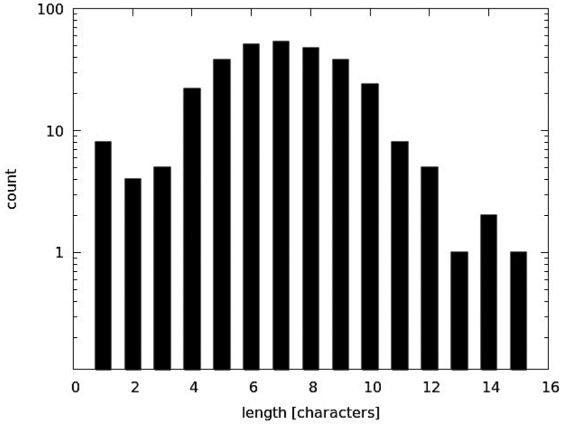

Figure 5: Single-word titlecase terms, distribution of character-lengths (the y-axis is logarithmic)

There are also a handful of one- or two-character long single-word terms that are homonymous with very frequent prepositions or conjunctions, the most misleading are *K* (lemmatized as the preposition *k*), *A* (lemmatized as the conjunction *a*), *SO*[6] (lemmatized as the preposition *s*[7]).

The only one one-letter term that is lowercase is *t* (for metric ton), and fortunately it is not homonymous with anything but the letter *t* itself, however, it is used for several unrelated purposes (time value; part of the abbreviation *t. j.*; vehicle category). Cursory examination of the sample of the corpus (100 occurrences of the lemma *t*) revealed that only 4 percent of them is the metric ton usage[8].

The complete list of single-word terms of character-length equal to one is:

A B C D E F K t W

The accuracy of matching the correct term based on the sample of 20 random occurrences each is shown in Table 7.

[6] Acronym for *sprostredkovateľský orgán* (intermediate body).

[7] *so* is the vocalized form of the preposition; vocalized forms are customarily lemmatized as their bare unvocalized counterparts.

[8] 95% Clopper-Pearson confidence interval is (0.011, 0.099), assuming binomial distribution of the term in the sample.

Table 7: Complete list of one-character long single-word terms, with the number of occurrences (in samples of 20 concordances) where the word was really the expected terminological unit

term	*longer name (in English)*	*IATE ID*	*# of correct matches*
A	*Directorate A*	IATE-3547016	0
B	*byte* *Directorate B*	IATE-1327726 IATE-3535468	0
C	*Directorate C*	IATE-3520864	0
D	*Department D* *credit (“dal”)*	IATE-3547458 IATE- 3507619	0
E	*Directorate E*	IATE-3524082	0
F	*Directorate F*	IATE-3544617	0
K	*kelvin*	IATE-1097346	0
t	*tonne*	IATE-1428563	0
W	*Wobbe index* *Watt*	IATE-1076785 IATE-788614	3

Table 8: Several capitalization classes of single-word terms of character-length equal to 2, with the number of occurrences (in samples of 20 concordances) where the word was really the expected terminological unit

capitalization	some examples	# of correct matches
[A-Z][A-Z]	*SD* (*súdny dvor* or *sadzba dane*)	0
[A-Z][a-z]	*La* (part of the term *Ro-La*, IATE-799398 or IATE-1876237 (*rolling road*)	0 (only one occurrence of the lemma *la* in the corpus, but in a different meaning – an abbreviation of *legislatívny akt*)
[a-z][a-z]	*hm, íľ, úľ*	20
[a-z][A-Z]	no hits in the corpus	N/A

In the Table 8, [A-Z] stands for an(y) uppercase letter, [a-z] for a lowercase one[9]. We examine the accuracy of case sensitive lemma matching. The only two two-letter long terms that are not acronyms are *úľ* and *íl* (and they are already present

[9] Including those characters with diacritics. The regular expression-like sequence is just a shorthand familiar for the reader.

in their base form). The only one lowercase acronym is *hm*, for "*hmotnostný zlomok*" (mass fraction)[10].

Therefore, the annotation rules we chose for single-word terms can be succinctly written as:

- ignore those with character-length equal to one
- if the character-length is equal to two, ignore terms with capital letters (anywhere in them, i.e. either uppercase or titlecase)
- if the character-length is greater or equal to three, do not lemmatize uppercase or titlecase terms (i.e. consider lemma to be identical with the word, but do not ignore the term)

4. Distribution of Terms in the Corpus

With the rules for term matching formalized, we can explore the coverage of the corpus by the IATE terms. We look at the distribution of terms by their word-length (Table 9, Figure 6) and the distribution of single-word terms by their character lengths (Table 10, Figure 7); all the numbers are in instances per million. We disregard overlapping terms (we consider only the first of them). The corpus contains texts of laws and other legally binding documents (decrees, government resolutions etc.) of the Slovak Republic, as published in the official Collection of Laws of the Slovak Republic[11]. The size of the corpus is 22 252 043 tokens; we selected only texts published in or after the year 1993. The texts are deduplicated on the paragraph level, using default *onion*[12] parameters, as described in (Benko 2013). Somewhat unsurprisingly, the most frequent terms are single-word ones and the distribution of term word-lengths is roughly analogous to Zipf's law (but we do not attempt to read too much into this).[13] Accuracy in following tables is based on manually verifying samples of 20 entries randomly selected from the corpus.

[10] Homonymous with an interjection rather frequent in spoken Slovak; though this interjection is not likely to appear in formal texts (apart from direct speech).

[11] Courtesy of the SLOV-LEX portal, https://slov-lex.sk.

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

[13] The longest entry, or rather a phrase, in the corpus is 36 token long IATE-769101, *Dohovor o prepustení zdravotníckeho, chirurgického a laboratórneho vybavenia do režimu dočasného použitia pre nemocnice a iné zdravotnícke zariadenia na diagnostické a terapeutické účely s úplným oslobodením od dovozného cla, daní a iných platieb vyberaných pri dovoze.*

Table 9: Distribution of word-lengths of terms in the corpus and the accuracy of annotation

word-length	i.p.m.	accuracy
1	112703	0.5
2	31029	0.9
3	4074	1
4	1156	1
5	204	1
6	70	1
7	26	1
8	17	1
9	4	1
10	2	1

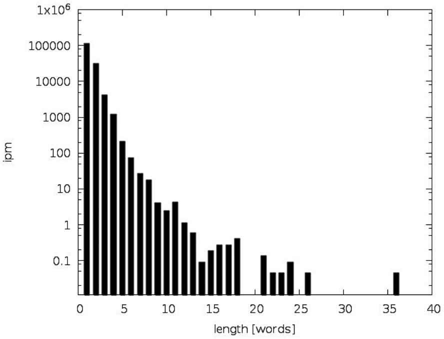

Figure 6: Distribution of word-lengths of matched terms in the corpus (the y-axis – logarithmic – shows the number of terms of given word-lengths in instances per million)

Distribution of character-length of single-word terms is more normal (though with a significant tail[14]), which allows us to get meaningful values for several statistical parameters, giving an intuitive overview of the distribution: mean 7.74, median 7, mode 6, standard deviation 2.31.

For terms two characters long, the main reason for the rather low accuracy (0.3) was the incorrect tagging of the word *dá* (3rd person singular indicative of the very frequent[15] word *dať* "to give"), matched by the IATE-3507619 *dal* (an accounting term, see footnote 5). Blacklisting this one term improved the accuracy to 0.95. Overall accuracy is uneven, there is no obvious correlation with term lengths, and an examination of various lengths shows that often a frequent word incorrectly matched against IATE is responsible for most of the errors, and blacklisting such words can improve the accuracy further – however, we did not want to follow this direction, in order to keep the annotating system simple.

Many of the errors are caused by the sparsity of IATE – often there is only a field specific narrow definition in IATE, while outside of the field, the term is used in broader meaning (e.g. the word *kandidát* is used in many different contexts, but

[14] The longest single-word term is 23 characters long IATE-3568059 *metyléndioxypyrovalerón*.

[15] Not that frequent in the corpus of the body of law (i.p.m. 45), but very frequent in general language (e.g. in the corpus *Araneum Slovacum V Maximum* (Benko 2014) i.p.m. 1082).

there is only one single-word entry IATE-3526213, in the air and space transport domain, a candidate for a Pilot or Inspector license; all the other meanings of the word are present only in multiword terms and therefore not matched by our algorithm). This is something that cannot be overcome without either extending/supplementing IATE or introducing semantic disambiguation in the corpus.

Table 10: Distribution of character-lengths of single-word terms in the corpus

character-length	i.p.m.	accuracy
1	N/A	N/A
2	22	0.30/0.95+
3	945	0.85
4	4887	0.45
5	12235	0.85
6	20420	0.75
7	18811	0.65
8	16328	0.20
9	14558	0.60
10	9752	0.55

+After blacklisting the word *dal* from matching.

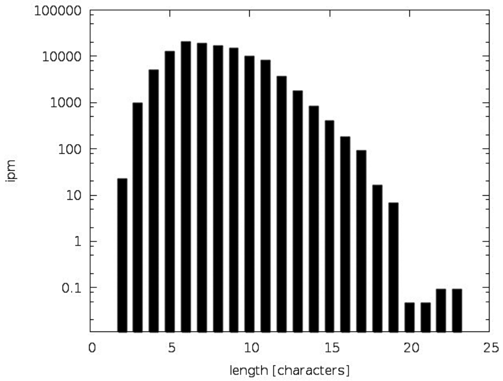

Figure 7: Distribution of character-lengths of single-word matched terms in the corpus (the y-axis – logarithmic – shows the number of terms of given character-lengths in instances per million)

5. Conclusion

We created a (deliberately) simple terminological annotation system based on existing NLP tools (lemmatization and disambiguation) to annotate Slovak texts by IATE terms. By investigating the accuracy of the shortest terms we show that by applying simple filters, we evade many systematic errors in the annotation. The annotation by matching sequences of lemmas in the IATE database gives sufficiently good results. Such an approach is relevant especially in a situation where we need a terminologically annotated corpus maximizing recall, we have basic NLP tools at our disposal, but (as is often the situation) we lack the resources to invest into more complex systems, e.g. a manually terminologically annotated corpus of considerable size. The accuracy of tagging multiword terms is very good, and while the accuracy of single-word terms is significantly lower, it can easily be increased by applying simple filters based on word lengths and blacklisting frequent false positives, keeping the effort invested into the annotation reasonably low.

Acknowledgements

This work was supported by the European Commission in the CEF Telecom Programme, Action no. 2017-EU-IA-0136 "Multilingual Resources for CEF.AT in the legal domain" (MARCELL).

References

Bada, Michael et al. 2010. An overview of the CRAFT concept annotation guidelines. *Proceedings of the Fourth Linguistic Annotation Workshop*. Ed. Xue, Nianwen; Poesio, Massimo. Association for Computational Linguistics. Uppsala. 207–211.

Benko, Vladimír. 2013. Data Deduplication in Slovak Corpora. *Slovko 2013: Natural Language Processing, Corpus Linguistics, E-learning*. Ed. Gajdošová, Katarína; Žáková, Adriána. RAM-Verlag. Lüdenscheid. 27–39.

Benko, Vladimír. 2014. Aranea: Yet Another Family of (Comparable) Web Corpora. *Text, Speech and Dialogue. 17th International Conference, TSD 2014*. Ed. Sojka, Petr et al. Springer International Publishing Switzerland. Brno. 257–264. doi.org/10.1007/978-3-319-10816-2_31.

Brank, Janez; Leban, Gregor; Grobelnik, Marko. 2017. Annotating Documents with Relevant Wikipedia Concepts. *Proceedings of the Slovenian Conference on Data Mining and Data Warehouses (SiKDD 2017).*

Coman, Andrei; Mitrofan, Maria; Tufiș, Dan. 2019. Automatic identification and classification of legal terms in Romanian law texts. *Proceedings of the International Conference on Linguistic Resources and Tools for Natural Language Processing (ConsILR 2019).* Ed. Onofrei, Mihaela et al. Faculty of Computer Science "Alexandru Ioan Cuza", University of Iași. Cluj-Napoca. 3–12.

Garabík, Radovan; Bobeková, Kristína. 2021. Lematizácia, morfologická anotácia a dezambiguácia slovenského textu – webové rozhranie. *Slovenská reč* 86/1. 104–109.

Garabík, Radovan; Šimková, Mária. 2012. Slovak Morphosyntactic Tagset. *Journal of Language Modelling* 0/1. 41–63. doi.org/10.15398/jlm.v0i1.35.

Hitzler, Pascal. 2021. A Review of the Semantic Web Field. *Communications of the ACM* 64/2. 76–83.

IATE Data fields explained. https://iate.europa.eu/fields-explained (accessed 11 November 2020).

IATE (= European Union Terminology) – User's Handbook. 2021. https://iate.europa.eu/assets/IATE_Handbook_public.pdf (accessed 26 July 2021).

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*.

Porter, Martin F. 1980. An algorithm for suffix stripping. *Program* 14/3. 130−137.

Slovenský národný korpus – prim-9.0. Jazykovedný ústav Ľ. Štúra Slovenskej akadémie vied. Bratislava. http://korpus.juls.savba.sk.

Váradi, Tamás et al. 2020. The MARCELL Legislative Corpus. *Proceedings of the 12th Language Resources and Evaluation Conference* (*LREC 2020)*. Ed. Calzolari, Nicoletta et al.

Vaswani, Ashish et al. 2017. Attention is all you need. *Advances in Neural Information Processing Systems 30 (NIPS 2017).* 5998–6008.

Naivna terminološka anotacija zakonskih tekstova u slovačkom – može li biti korisna?

Sažetak

Ispravna automatska terminološka anotacija tekstova u korpusu ponekad može biti izazovan zadatak, posebno za iznimno flektivne jezike s razmjerno slobodnim redoslijedom riječi. U članku istražujemo mogućnost jednostavne anotacije na temelju podudarnosti lematiziranih tekstova kako bi korpus slovačkoga jezika bio anotiran terminološkim zapisima IATE. Točnost anotacije višerječnih termina vrlo je dobra, dok se točnost jednorječnih termina može povisiti primjenom jednostavnih filtara na temelju duljine riječi i stavljanja na crnu listu najčešćih lažnih pozitivnih rezultata.

Keywords: terminology, corpus, Slovak language, corpus annotation, IATE
Ključne riječi: terminologija, korpus, slovački jezik, anotacija korpusa, IATE