# Language Code Switching in Web Corpora

Vladimír Benko[1,2]

[1] Slovak Academy of Sciences, Ľ. Štúr Institute of Linguistics
Panská 26, SK-81101 Bratislava, Slovakia
[2] Comenius University in Bratislava, UNESCO Chair in Plurilingual and Multicultural Communication
Šafárikovo nám. 6, SK-81499 Bratislava, Slovakia
vladimir.benko@juls.savba.sk
http://www.juls.savba.sk/~vladob

*I Love Music Párty 10. apríla 2009 v New Cage Clube v Lučenci*[3]
*I love youúúúúúú I love Americáááááááá*[4]

**Abstract.** One of the challenges in building and using web corpora is their rather high content of "noise", most notably having the form of foreign-language text fragments within otherwise monolingual text. Our paper presents an approach trying to cope with this problem by means of "exhaustive" stop-word lists provided by morphosyntactic taggers. As a side effect of the procedure, a problem of tagging text with missing diacritics is also addressed.

**Key words:** web-based corpora, language identification, morphosyntactic annotation, Aranea Project

## 1 Introduction

"Noise" in texts derived from web can be of various nature. In our work, however, we want to address just one particular type of text noise, namely language code switching, by which we also consider "code switching" in discussions, where some of the participants do not use diacritics. In the framework of our *Aranea*[5] Project [1], we are currently trying to improve the morphosyntactic annotation of the Slovak *Araneum Slovacum* corpus, so that it could be used by lexicographers as a source of lexical evidence for the multi-volume *Dictionary of the Contemporary Slovak Language*[6]. The same problem, albeit to a lesser extent, can be observed in the *Slovak National Corpus*[7], so a successful solution could be used here as well.

[3] http://www.ilovemusic.sk/clanky/i-love-music-party-10-aprila-2009-v-new-cage-clube-v-lucenci
[4] http://dolezite.sk/Raz_v_tom_ma_jasno_253.html
[5] http://aranea.juls.savba.sk/aranea_about/
[6] http://www.juls.savba.sk/sssj_6.html
[7] http://korpus.juls.savba.sk/

Aleš Horák, Pavel Rychlý, Adam Rambousek (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2017, pp. 97–105, 2017. 

| | | |
|---|---|---|
| pr-web.sk | píská a šumí, mozek se neprokrvuje... Z toho **plyne** špatná paměť, zapomínání, přidá se vysoký | ☑ |
| beo.sk | ti amici jazdit ked nie na ruskej rope a **plyne** ...na kravske prdy? ¶ _____ | ☐ |
| pluska.sk | peniaze. Ady má totiž domček v Lozorne, kde na **plyne** varí, ohrieva vodu i kúri. A každé usporené | ☐ |
| pluska.sk | je golfové ihrisko. Takže čo usporím na **plyne** , vrazím do golfu," vysvetľuje. Pritom ale | ☐ |
| euroekonom... | od roku 2001. Kupříkladu, z dat UAH MSU **plyne** ochlazovací trend mezi lednem 2001 a květnem | ☑ |
| euroekonom... | sopkami a ENSO-m) ¶ Co z vývoje Ap indexu **plyne** ? Nejspíš to, že nás čeká další rok extrémně | ☑ |
| zahori.est... | Francúzsko. K odstráneniu závislosti na ruskom **plyne** môžu prispieť aj náleziská plynu na južnom | ☑ |
| zahori.est... | troch až piatich rokov bude Slovensko na **plyne** nezávislé". ¶ Expremiér Mikuláš Dzurinda | ☐ |
| despitebor... | nariadenie nasledovala smernica o zemnom **plyne** v roku 1998 (98/30/EC). V oboch sa požadovalo | ☐ |
| despitebor... | 2003/55/EC (5) o elektrickej energii a o **plyne** predstavujú najväčšie zmeny doterajšieho | ☐ |
| despitebor... | znamenalo vyvlastnenie, čo by hlavne pri zemnom **plyne** viedlo k zvýšeniu cien pre koncových zákazníkov | ☐ |
| burjanosko... | střední školy a jeho tři zástupci. ¶ - Co z toho **plyne** : měli bychom si, jako daňoví poplatníci | ☑ |
| energia.sk | zaostávame. V Čechách mení v elektrine a zemnom **plyne** dodávateľa 15 až 20 percent domácností. | ☐ |
| energia.sk | plyn ho dokáže nahradiť, dopyt po zemnom **plyne** bude rásť a cena pôjde tiež hore. ¶ Ak sa | ☐ |
| diskusie.s... | pripravili na vojnu, a pred začiatkom klamstva o **plyne** boli v Izraeli hlavní američtí generáli | ☐ |
| sixpack.cz | telekomu atd.. Prakticky aj pri elektrike, **plyne** a vode. Tam neexistuje alternatíva, ale | ☐ |

**Fig. 1.** Czech sentences in the Slovak text ("*plyne*")

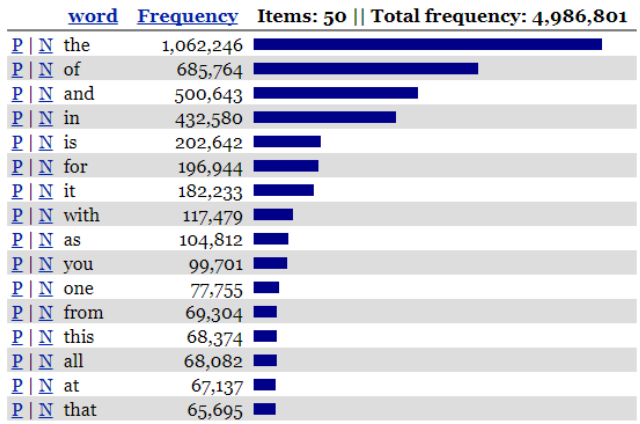
## 2 The Task

From the lexicographers' perspective, the problem of foreign language text fragments in otherwise monolingual text is mostly associated with so-called interlingual ("false") homographs, i.e., lexical items present in two languages, yet having (usually) a different meaning. This phenomenon is especially pronounced between close languages, such as Slovak and Czech. If the Slovak lexical item is rather rare and, on the other hand, the same word form in Czech is frequent, the resulting concordances may contain comparable number of occurrences in both languages. For example, the Slovak word form "*plyne*" (noun, locative case singular of "*plyn*", English "*gas*") is a form of "*plynout*" in Czech (verb, 3rd person singular, English "*flow*"). As seen in Figure 1, more than one third of the occurrences come in fact from Czech sentences (ticked in the screenshot).

Frequent foreign lexical items can cause problems even if they are not homographs, as they "spoil" the frequencies of out-of-vocabulary word forms that have to be manually checked in search for potential neologisms. If we succeed in reliable detection of foreign-language text fragments, we will be able to remove or, at least, mark them so that they would not be shown as results of corpus queries.

## 3 The Scope of the Problem

Our long-term experience with Slovak corpora shows that the most frequent foreign-language fragments in Slovak texts come from English and Czech. To estimate the rough proportion of English lexical items in the Slovak corpus we used a simple method: from a list of the most frequent word forms of our *Araneum Anglicum* English corpus, we deleted words that do exist in Slovak as



| | word | Frequency | Items: 50 \|\| Total frequency: 4,986,801 |
|---|---|---|---|
| P \| N | the | 1,062,246 | |
| P \| N | of | 685,764 | |
| P \| N | and | 500,643 | |
| P \| N | in | 432,580 | |
| P \| N | is | 202,642 | |
| P \| N | for | 196,944 | |
| P \| N | it | 182,233 | |
| P \| N | with | 117,479 | |
| P \| N | as | 104,812 | |
| P \| N | you | 99,701 | |
| P \| N | one | 77,755 | |
| P \| N | from | 69,304 | |
| P \| N | this | 68,374 | |
| P \| N | all | 68,082 | |
| P \| N | at | 67,137 | |
| P \| N | that | 65,695 | |

**Fig. 2.** English words in the Slovak corpus

well (such as "*to*", "*a*", "*on*") and used the top 50 from the resulting list to create a corpus query:

```
the|of|and|in|is|that|for|with|it|you|are
    |as|be|was|have|this|...
```

The top 16 lines of the respective frequency list are shown in Figure 2.

The total normalized frequency is 1,682.80 i.p.m., which is really quite a lot – it means, that these very 50 English words represent almost 0.17 % of all tokens of the corpus.

Using the same method, we can also estimate the proportion of the Czech lexical items. Here, however, the number of inter-lingual homographs is much larger. The situation may also be complicated by the Slovak (and also Czech) text fragments without diacritics that increase the number of the respective inter-lingual homographs. We have, however, decided not to take this phenomenon into account here. The resulting frequency distribution (first 16 lines) is shown in Figure 3.

The total normalized frequency is 1,571.0 i.p.m here, which looks pretty similar to that of English. In reality, however, the number will be much higher due to the already mentioned inter-lingual homographs.

## 4 The Method

Language identification belongs to traditional tasks in the area of Natural Language Processing, as well as in that of Information Retrieval, with state-of-the-art methods exhibiting precision well over 95 % (cf. [5]). These methods can be basically divided into two groups: (1) methods counting frequent words that are based on lists of "stop words"; (2) methods counting individual characters

| | word | Frequency | Items: 50 \|\| Total frequency: 4,655,548 |
|---|---|---|---|
| P \| N | ze | 1,530,617 | |
| P \| N | co | 837,991 | |
| P \| N | se | 482,948 | |
| P \| N | velmi | 267,340 | |
| P \| N | den | 193,408 | |
| P \| N | pro | 178,093 | |
| P \| N | jsem | 109,782 | |
| P \| N | jako | 91,904 | |
| P \| N | ve | 77,974 | |
| P \| N | jsou | 56,038 | |
| P \| N | když | 45,822 | |
| P \| N | ke | 43,205 | |
| P \| N | které | 36,798 | |
| P \| N | není | 34,947 | |
| P \| N | který | 32,108 | |
| P \| N | byl | 31,500 | |

**Fig. 3.** Czech words in the Slovak corpus

or character n-grams – "statistical methods". The main advantage of statistical methods is that they are typically able to identify language from short strings containing just several hundreds of characters, as well as that they are usually computationally "cheap". The performance of the stop-list methods usually depends on the size of the respective list. The disadvantage of both is rather low performance in distinguishing very similar languages and practical inability to cope with the texts containing code switching.

In our case, we do not require fast computation as the corpus annotation is a time-consuming process anyway. We also would like to possibly make use of existing tools – the morphological analyzers and taggers. Conceptually, our method can be put into the stop-list category, assuming that the size of the list is limited just by the size of the morphological lexicon used by the respective analyzer.

The main idea is as follows: (1) Besides the basic morphosyntactic annotation by the standard tagger (using the Slovak language model), the corpus is processed by alternative taggers (using language models for languages that we want to identify); (2) information about the result of the morphological lexicon lookup performed by the respective taggers is gathered; (3) by combing information from different taggers language of each token is estimated; (4) using summary information regarding the individual tokens of the sentence its language is stated.

The actual annotation has been carried out by *TreeTagger* [7] using our own language models for Slovak and Slovak without diacritics [2], and by the morphological component of *MorphoDiTa*[8] [9,8] using the newest Czech language model. The processing used our standard *Aranea pipeline* [3] for each

[8] As we only need information on the morphological lexicon lookup, the disambiguation phase pro-vided by the tagger in not necessary here.

| | ztag | ztag1 | ztag2 | ztag3 | Frequency | Items: 59 \|\| Total frequency: 10,000,000 |
|---|---|---|---|---|---|---|
| P \| N | 1 | 1 | 0 | 0 | 3,974,506 | |
| P \| N | 1 | 1 | 1 | 0 | 2,792,725 | |
| P \| N | 1 | 1 | 1 | 1 | 2,083,398 | |
| P \| N | 1 | 1 | 0 | 1 | 442,205 | |
| P \| N | 0 | 0 | 0 | 0 | 254,507 | |
| P \| N | 0 | 0 | 1 | 0 | 111,336 | |
| P \| N | 1 | 2 | 0 | 0 | 61,511 | |
| P \| N | 0 | 1 | 0 | 0 | 54,026 | |
| P \| N | 1 | 2 | 1 | 0 | 47,379 | |
| P \| N | 0 | 0 | 1 | 1 | 43,124 | |
| P \| N | 0 | 0 | 0 | 1 | 39,878 | |
| P \| N | 1 | 2 | 1 | 1 | 29,941 | |
| P \| N | 0 | 1 | 1 | 0 | 18,721 | |
| P \| N | 2 | 2 | 0 | 0 | 9,995 | |
| P \| N | 1 | 2 | 0 | 1 | 9,186 | |
| P \| N | 0 | 1 | 1 | 1 | 8,096 | |

**Fig. 4.** Morphological lexicon lookup results (alphabetical tokens considered only)

language, and the resulting partial verticals have been combined by the *cut* and *paste* utilities, so that the resulting vertical would contain 17 columns for corpus attributes as follows: *word*, *lemma*, *atag*, *tag*, *ztag*, *lemma1*, ... The first five attributes belonged to the original Slovak annotation, and the indexed attributes contained alterative annotations – index 1 indicated Slovak without diacritics, 2 indicated Czech, and 3 indicated English, respectively.

For our purposes, the most important in all parallel annotations is usually the respective *ztag* attribute having a non-zero value if the morphological lexicon lookup has been successful.

## 5 The Initial Experiment

The development of the language identification algorithm based on the respective parallel annotation has been started by producing a *ztag* value distribution for a random 10 million alphabetic tokens from the 200-million token test corpus. Figure 4 shows its first 16 lines.

The table is slightly complicated to read because of values in the first two columns: besides the "0" (wordfom not found in lexicon) and "1" (found) values, both Slovak language models produce also numbers larger than 1 in situations where the tagger was not able to disambiguate the respective lemma – in such cases, the number of variant lemmas is shown. It is also clear that the values in the second column cannot be smaller than those in the first one, as the "diacritics-less" language model must always yield at least the same result as the "full" model.[9] The first two rows of the table are quite expected – most wordforms have been recognized by the Slovak models, followed by both Slovak and Czech models. A surprise is the contents of the third line. What are those words that are present in Slovak, Czech and English? See Figure 5.

[9] All diacritics have been stripped from the source vertical before the diacritics-less annotation has been applied.

| | word | lemma3 | tag3 | Frequency | Items: 10,200 \|\| Total frequency: 2,083,398 |
|---|---|---|---|---|---|
| P \| N | a | a | DT | 336,754 | |
| P \| N | v | v | NN | 204,724 | |
| P \| N | na | na | TO | 202,217 | |
| P \| N | je | Je | NP | 134,035 | |
| P \| N | to | to | TO | 68,862 | |
| P \| N | o | o | NN | 59,365 | |
| P \| N | si | si | NP | 56,561 | |
| P \| N | do | do | VVP | 55,336 | |
| P \| N | z | z | SYM | 49,940 | |
| P \| N | ale | ale | NN | 33,046 | |
| P \| N | V | V | NN | 32,020 | |
| P \| N | k | k | NN | 31,113 | |
| P \| N | by | by | IN | 28,362 | |
| P \| N | od | od | NN | 27,709 | |
| P \| N | tak | Tak | NP | 27,506 | |
| P \| N | po | po | NN | 26,848 | |

**Fig. 5.** Words recognized by all language models

| | ztag | ztag1 | ztag2 | ztag3 | Frequency | Items: 58 \|\| Total frequency: 10,000,000 |
|---|---|---|---|---|---|---|
| P \| N | 1 | 1 | 0 | 0 | 5,091,887 | |
| P \| N | 1 | 1 | 1 | 0 | 3,159,066 | |
| P \| N | 1 | 1 | 1 | 1 | 621,345 | |
| P \| N | 0 | 0 | 0 | 0 | 327,822 | |
| P \| N | 1 | 1 | 0 | 1 | 264,436 | |
| P \| N | 0 | 0 | 1 | 0 | 142,349 | |
| P \| N | 1 | 2 | 0 | 0 | 74,219 | |
| P \| N | 0 | 1 | 0 | 0 | 69,113 | |

**Fig. 6.** Morphological lexicon lookup results (3 and more letters)

The beginning of the list contains lots of strange short "English" words tagged as "proper noun" ("NP"), even if they are written in lowercase letters. We therefore decided to refine our query and restrict it to words longer than 2 letters. Figure 6 shows the result.

Now the statistics looks much more "probable", and we can utilize the information that the language identification should be based on longer words.

## 6 The Algorithm

Besides the experience gathered from the introductory experiment, we also made use of information from external source: we knew that the size of the Czech morphological lexicon is almost by order of magnitude larger than that of Slovak [4], better covering not only large amounts of loanwords and rare lexical items, but also numerous proper and geographical names. This often means the Czech morphological analyzer can recognize many proper names in Slovak text, whereas the Slovak one mostly fails here.

The algorithm design has been performed in an iterative way, using the smaller (2 million token) test corpus that has been compiled by NoSketch Engine[10] [6] after each iteration to analyze the results. The values of internal

[10] `https://nlp.fi.muni.cz/trac/noske`

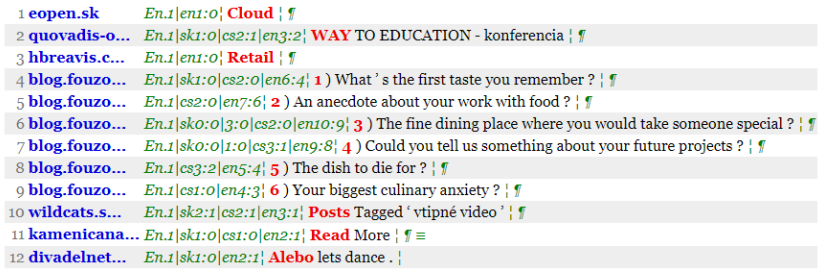



**Fig. 7.** Sentences recognized as English with the respective metadata

variables used to decide the language of the respective sentences could be conveniently encoded and displayed as attributes of the <s> structure. To display sentences according to identified language we used a CQL query like this:

```
<s lang="En.*"> []
```

Figure 7 shows several sentences resulting from that query.

The string preceding each sentence denotes the recognized language and (for each language) two values separated by a colon representing the number of "decidable" tokens and bigrams. For example, sentence 10 has been tagged as English, as it contained three decidable words and one bigram identified as English, as opposed to Slovak and Czech (two words and one bigram).

The current version of the algorithm can be described as follows:

(1) Only alphabetic tokens of a minimal length (currently 2), optionally followed by a full stop, are considered. Such tokens we call "decidable".

(2) Words marked by Slovak language model as foreign (tag "#") are not considered as Slovak.

(3) Words marked by English language model as foreign (tag "FW") or proper noun (tag "NP") are not considered as English.

(4) Words starting with a capital letter are considered in all languages only of they have been recognized by the Slovak language model.

(5) For Slovak, counts from non-diacritics language model are used.

(6) If no word in the sentence has been recognized by any language model, the sentence is marked as "undecidable" ("Xx").

(7) The language with the largest proportion of recognized words becomes the language of the sentence.

(8) If recognized words counts are equal, the greatest number of recognized bigrams is used for decision.

(9) If all bigram counts are equal, the sentence is marked as Slovak.

## 7 The Results

Although high recall has been the priority in this work, it is something that is rather difficult to evaluate in large corpora. We therefore offer data on the

**Table 1.**

| Identified as | Slovak | Czech | English | Undecidable | Total |
|---|---|---|---|---|---|
| Sentences (counts) | 10,220,517 | 97,893 | 23,423 | 177,006 | 10,518,927 |
| Sentences (%) | 97.16 | 0.93 | 0.22 | 1.68 | 100.00 |
| Tokens (counts) | 191,915,723 | 1,226,762 | 303,681 | 621,421 | 199,660,383 |
| Tokens (%) | 96.12 | 0.61 | 0.15 | 3.11 | 100.00 |

precision only. Table 1 shows some results received by the beta version of our algorithm that has been used to process the larger test corpus containing approx. 200 million tokens.

It is apparent that the procedure has identified almost 3 % of foreign-language or undecidable sentences containing approx. 4 % of all corpus tokens. We can also see that the sentences recognized as Czech and English are shorter than average, while the undecidable sentences are longer than average. A short check in the corpus reveals the cause: the long undecidables contain various lists consisting of proper names that have been excluded from our algorithm.

The precision data in Table 2 has been obtained by manual checking samples of 100 sentences from each group.

Even this rudimentary evaluation shows that identification of English text fragments within Slovak text is a relatively easy task, while distinguishing between Czech and Slovak is really a "tough" one. For evaluation of identification of Slovak text this tiny sample is naturally not sufficient. The algorithm as such, however, is already usable – the 4 % loss of data in corpus is more than acceptable. The program has been implemented in *lex* programming language and its current (9th) version is just 330 lines long, inheriting some portions of the code from other programs. We expect, however, that by using a programming language with a native utf-8 support, the program might be even simpler.

## 8 What Have We Learned

Our experiment has proved that by using existing tools this problem could be solved with minimal additional programming necessary. We have acquired

**Table 2.**

| | Identified as (by algorithm) | | | | |
|---|---|---|---|---|---|
| Identified as (manually) | Slovak | Czech | English | Undec. | Cs+En+Xx |
| Slovak | 98 | 25 | 0 | 26 | 51 |
| Czech | 0 | 43 | 0 | 2 | 45 |
| English | 0 | 3 | 89 | 6 | 98 |
| Undecidable (Xx) | 2 | 21 | 11 | 63 | 105 |
| Other language | 0 | 8 | 11 | 3 | 22 |

certain insight into the way how the respective taggers and language models work, which can be used to improve the process of morphosyntactic annotation of Slovak corpora. We have also discovered some peculiarities in the respective lexicons, such as English prepositions tagged as ("Czech") prepositions and, similarly, English articles tagged as ("Czech") adjectives in the Czech *MorphoDiTa* lexicon.

## 9 Further Work

The described method can be used for other languages as well. As English text fragments are present virtually in all corpora of the *Aranea* family (including the Chinese and Arabic ones), we would like to improve annotation of all corpora in the foreseeable future.

**Acknowledgements.** This work has been, in part, financially supported by the Slovak VEGA and KEGA Grant Agencies, Project Nos. 2/0017/17, and K-16-022-00, respectively.

## References

1. Benko, V.: Aranea: Yet another Family of (Comparable) Web Corpora. In: Text, Speech, and Dialogue. 17th International Conference, TSD 2014 Brno, Czech Republic, September 8—12. Proceedings. Eds. P. Sojka et al. Cham – Heidelberg – New York – Dordrecht – London: Springer, pp. 21—29. (2014)
2. Benko, V.: Tvorba webových korpusov a ich využitie v lexikografii. Dizertačná práca. Bratislava: Filozofická fakulty Univerzity Komenského. (2016)
3. Benko, V.: Two Years of Aranea: Increasing Counts and Tuning the Pipeline. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC). Portorož: European Language Resources Association (2016) pp. 4245–4248. (2016)
4. Petkevič, V.: Personal communication. Praha (2016)
5. Řehůřek, R. and Kolkus, M.: Language Identification on the Web: Extending the Dictionary Method In: A. Gelbukh (Ed.): CICLing 2009, LNCS 5449. Berlin – Heidelberg: Springer-Verlag, pp. 357–368. (2009)
6. Rychlý, P.: Manatee/Bonito – A Modular Corpus Manager. In: 1st Workshop on Recent Advances in Slavonic Natural Language Processing. Brno : Masaryk University, pp. 65–70. (2007)
7. Schmid, H.: Probabilistic Part-of-Speech Tagging Using Decision Trees. In: Proceedings of International Conference on New Methods in Language Processing. Manchester (1994)
8. Spoustová, D. "johanka", Hajič, J., Raab, J. and Spousta, M.: Semi-Supervised Training for the Averaged Perceptron POS Tagger. In: Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009), Athens: ACL, pp. 763–771. (2014).
9. Straková, J., Straka, M. and Hajič, J.: 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: ACL, pp. 13–18. (2014)