The Mysterious Letter J

Anđelka Zečević
Faculty of Mathematics
University of Belgrade
andjelkaz@matf.bg.ac.rs

Staša Vujičić Stanković
Faculty of Mathematics
University of Belgrade
stasa@matf.bg.ac.rs

Abstract

Ekavian and Ijekavian are two different variants of the contemporary standard Serbian language. The difference between them is related to the reflex of the old Slavic vowel *jat* and it influences both the speaking and writing language norms. The sensibility of existing language identification tools for both variants is of great importance for building representative corpora and development of relevant linguistics resources and tools underlying an automatic text processing. In this paper we present the results obtained after testing the three popular tools for language identification on corpora containing documents from each of the two variants. As it will be reported, the identification of Ijekavian variant is a much more difficult task since the observed tools are not adopted to it at all.

1 Introduction

The language identification is a problem of identifying the language a document is written in. It represents the fundamental step in tasks such as collecting the documents for corpora, machine translation and information retrieval. Because of its great importance, methodological approaches to the problem and submitted solutions are numerous. In the basis, the problem can be seen as a classification problem (Mitchell, 1997): if collections of known language samples represent classes, the problem of the language identification for the given document can be seen as a problem of the document assignment to the one of the classes in respect to relevant classification features.

Many sets of language features as well as classification algorithms have been tested so far. The choice of features might be linguistically motivated (diacritics and special characters) or more statistically oriented (word frequencies, n-grams of various lengths and types). The first tools were based on the analyses of character n-grams: Dunning (Dunning, 1994) introduced Markov models while Cavnar and Trenkle (Cavnar and Trenkle, 1994) worked with 1-NN classification algorithm. Nowadays the focus is on the diverse set of (dis)similarity measures (Singh, 2006) and powerful algorithms as can be read in papers discussing their performance and fields of the application (Martins and Silva, 2005).

The task of the language identification is considered much harder if the document is of modest length (for instance, e-Bay and Twitter messages or search engine queries) or the amount of available training data is limited. The same can be said for the cases when the number of considered languages is huge or languages are similar to each other. All these conditions influence the success rate as it is reported in Padró and Padró (Padró and Padró, 2004), Lui and Baldwin (Baldwin and Lui, 2010), and Milne et al. (Milne et al., 2012). We are especially interested in the latter problem since the Serbian language is closely related to the languages spoken in former Yugoslavia.

The standard Serbian language is formed on the basis of Ekavian and Ijekavian Neo-Štokavian South Slavic dialects and its form is determined by the reformer of the written language of the Serbs, Vuk Karadžić (1787-1864) (Stanojčić and Popović, 2011). In the common state of Yugoslavia this language was officially encompassed by Serbo-Croatian, a name that implied a linguistic unity with the Croats (and later with other nations whose languages were based on Neo-Štokavian dialects). In the last decade of the 20th century in Serbia the name

Proceedings of the Adaptation of Language Resources and Tools for Closely Related Languages and Language Variants, pages 40–44,
Hissar, Bulgaria, 13 September 2013.

Serbo-Croatian was replaced in general usage by the name Serbian. As mentioned above, in Serbian speaking countries two dialects coexist. The Ekavian dialect is widespread in Serbia, while the Serbian Ijekavian dialect is presented in some parts of the northern Serbia, Croatia and Montenegro as well as Bosnia and Herzegovina. The difference among the dialects is related to the old Slavic vowel called *jat* and its conflation into the vowel *e* or diphthongs *ije* and *je*. It is notable in both spoken and written language forms as Serbian has a phonologically based orthography. For instance, in respect to the Ekavian dialect the English word *flower* has forms *cv**e**t* (long **e**) in nominative singular and *cv**e**tovi* (short **e**) in nominative plural while the appropriate forms in Ijekavian dialect are *cv**ije**t* and *cv**je**tovi* respectively. Therefore, Serbian and other languages of Štokavian provenance share the Ijekavian dialect in their standard forms which makes the task of the language identification very sensitive and error-prone. From the other point of view, these languages overlapping can help in cooperative development of tools and resources necessary for an automatic language processing (Vitas et al., 2011).

In this paper we present the results obtained after testing the three popular language identification tools on the collection containing both Ecavian and Ijekavian documents. Section 2 that refers to the related work and state-of-the-art approaches is followed by two introductory sections numbered as 3 and 4 and related to tested tools and specially created test corpora. The experiment is described in Section 5 while the results are presented in Section 6. The conclusions and ambitions for future work are summarized in Section 7.

2 Related Work

There is a number of papers discussing the identification of closely related languages, varieties of polycentric languages and language dialects. All these tasks are more advanced in comparison to classical ones and require application of more subtle techniques.

In the paper (Ljubešić et al., 2007) the three-phases model for differentiating Croatian from Slovenian and Serbian is presented. In the first phase the documents written in any of these three languages are singled out by the rule of 100 most frequent words and the rule of special character elimination. In the next phase character based second-order Markov model is developed aiming to distinguish languages among themselves. In order to improve the distinction between Croatian and Serbian, in the final phase the lists of forbidden words are introduced. Those are the lists containing words that appear in one language but not in others. The model is tested on the news collection and the achieved accuracy of 0.9918 is better than any reported for this group of languages.

The case of European and Brazilian Portuguese is discussed in (Zampieri and Gebre, 2012). As the differences between these two varieties can be described at orthographic, lexical and syntactic level, the identifying algorithm analyse three groups of features: character n-grams (n varying from 2 to 6), word unigrams and word bi-grams. The language models are calculated by using the Laplace probability distribution and evaluated on the journalistic corpora containing texts from the both varieties further classified according to their length in tokens. The achieved accuracies are 0.998 for 4-grams, 0.996 for word unigrams, and 0.912 for word bi-grams.

In order to identify Spanish varieties, the authors of (Zampieri et al., 2013) compared the classical character and word n-gram model to the knowledge-rich model based on the morphosyntactic information and parts of speech. The testing was done on the newspapers corpora from four Spanish speaking countries (Spain, Argentina, Mexico and Peru) and the reported results showed the direct relationship between the performance using these two language models: for instance, the Argentina-Spain classifier performed the worst in both cases (0.843 and 0.666 in terms of accuracy) while the Argentina-Mexico classifier generated the top results with characters and words (0.999) as well as morphology and parts of speech (0.801).

3 Tested Language Identification Tools

In our experiment we have tested three tools for the language identification. A brief description of tools and the motivation for the usage is given below.

Langid.py[1] is a top-level tool developed by Lui and Baldwin (Lui and Baldwin, 2012). It is based on the multinomial naive Bayes classifier which operates on the set of features (byte level unigrams, bigrams and trigrams) selected so that their information gain represents the characteristics of the language rather than the characteristics of the training domains (Lui and Baldwin, 2011). For the training phase the corpus, which encompasses government documents, newswire, online encyclopedia, software documentations and an internet crawl in 97 languages (Lui and Baldwin, 2011) is used. In the case of Serbian, the training collection includes XML wiki dumps for the period July-August 2010 as well as the set of manually translated content strings for a number of Debian software packages[2].

CLD (Content Language Detection)[3] is a library embedded in a Google's Chromium browser able to detect a language of a web page content. Thanks to Michael McCandless, it is singled out as a separated C++/Python module and ready for use on any UTF-8 encoded content. It is not specified how many languages it can detect (at least 76[4]) and so far it does not seem that the training set can be adapted to a specific usage.

The classifier developed by Tiedemann and Ljubešić (Tiedemann and Ljubešić, 2012) (in further text **Tiedemann&Ljubešić**) aims to distinguish closely related languages such as Serbian, Croatian and Bosnian. It is in the main multinomial Naive Bayes classifier trained over a parallel collection of news from Southeast Europe known as SETimes collection[5]. The usage of the parallel training set resulted in outperforming the state-of-the-art tools significantly since the data parallelism provided the same content and the focus on the differences among the languages. The authors also reported a list of the strongest discriminators among the observed languages and for our investigation it was interesting that the list for Bosnian contains many regular Serbian words in Ijekavian pronunciation (for instance, *izvještajima*, *posjetioci*, *djelimično*).

4 Test corpus

For testing purpose, we have created a corpus which consists of documents in both Ekavian and Ijekavian variant (Table 1). Since Serbian can be written in Cyrillic or Latin script, all the documents are transliterated into Latin script.

	Size *(in number of words)*	*Size* *(in MB)*
Ekavian part	2. 078, 172	13.2
Ijekavian part	528, 749	3.2

Table 1: The structure of the corpus

The Ekavian part of the corpus includes the articles from the daily newspaper *Politika*[6] for the years 2007 and 2010, the literary works written by the local authors and the translations of many popular novels. The list of all used materials is reported in Table 2.

The Ijekavian part of the corpus includes the articles from the daily newspaper *Glas Srpske*[7] for the period January-June 2013, some columns taken from the Deutsche Welle website [8] and famous works written in the Ijekavian dialect. Table 3 depicts all the details.

5 Experiment

Due to the nature of the used tools and comparability with other reported results, we have split the corpus into lines on average 400 words long. In the next step we have randomly selected 200 lines: the first 100 lines from the Ekavian part of the corpus and the rest from the Ijekavian part of the corpus.

For the testing purpose of the **langid.py** tool each line is saved as a separate file because a redirection mode was used. The **Tiedemann&Ljubešić** tool works with a single file that contains all the texts for classification as separate

[1] https://github.com/saffsd/langid.py

[2] In the time of writing this paper, translations in both Ekavian and Iekavian variants were available.

[3] http://code.google.com/p/chromium-compact-language-detector/

[4] http://blog.mikemccandless.com/2011/10/accuracy-and-performance-of-googles.html

[5] http://www.setimes.com

[6] http://www.politika.rs

[7] http://www.glassrpske.com/

[8] http://www.dw.de

[9] titles in Serbian are *Magareće godine* and *Glava u klancu, noge na vrancu*

[10] http://www.dw.de/škljocam-i-zvocam/a-4461937

Da Vinci Code by Dan Brown
1984 by George Orwell
Around the World in Eighty Days by Jules Verne
The Little Prince by Antoine de Saint Exupéry
The Diary of Anne Frank
The Hobbit by J. R. R. Tolkien
The Lord of the Rings by J. R. R. Tolkien
Solaris by Stanisław Lem
Winnie-the-Pooh by A. A. Milne
Bridget Jones's diary by Helen Fielding
For and Against Vuk by Meša Selimović
articles from *Politika* newspaper

Table 2: Ekavian part of the corpus

Springs of Ivan Galeb by Vladan Desnica
Selected works of Petar Kočić
two novels by Branko Ćopić[9]
Dove Hole by Jovan Radulović
Rebel and Rebel Janko by Simo Matavulj
Spiders and Searching the bread by Ivo Ćipiko
The Dervish and Death by Meša Selimović
articles from *Glas Srpske* newspaper
column written by Nenad Veličković [10]

Table 3: Ijekavian part of the corpus

lines so we concatenated our test lines into the document of this form. The same was done for the testing of **CLD** Python library.

6 Results

The obtained results are summarized in Table 4.

As it can be seen, the algorithms generally can cope with the classification of the documents in Serbian Ekavian variant (an average accuracy is 74.3%). On the contrary, the classification of the documents in Serbian Ijekavian variant is a very difficult task even for the tool developed with an idea of closely related languages in mind.

During the testing of **langid.py** tool we have encountered the problem with scripts: the tool by default recognizes Serbian only if it is written in official Cyrillic alphabet even though both Latin and Cyrillic alphabets are widespread in Serbian. This certainly caused the misclassification of all tested Ijekavian documents as Croatian.

Google's **CLD** obviously favors Croatian in both cases. In all the iterations the algorithm's confident parameter is set on the true value which means it is quite sure about the final outcome. After the analysis of the wrong results referring to Ekavian tests we found that in 25 iterations the second proposed language was Slovenian, in 8 iterations Serbian, and in 5 iterations Slovak. In all the remaining iterations the algorithm was completely sure about Croatian. In the case of Ijekavian tests, in 16 iterations the second proposed language was Slovenian, in 3 iterations Slovak and in 14 iterations Serbian. There was one iteration for each of the languages: Spanish, Italian and Indonesian.

The **Tiedemann&Ljubešić** tool is very accurate in classifying the documents in Serbian Ekavian variant while it recognizes a great part of Ijekavian documents as written in Bosnian. The latter is due to the fact that the training collection contains only the news in the Ekavian variant so the rules of Serbian are strictly learnt in this manner. In 83 of 98 iterations that output Bosnian as a result, the second proposed language was Croatian, and only in 15 of them it was Serbian.

7 Conclusions and Further Work

The obtained results show that many popular tools ignore the presence of the Ijekavian variant of Serbian language. This could lead to misclassification of Serbian documents which in turn strongly influences users' experience and information needs. The next steps would be enlarging the Ijekavian part of the corpus with relevant texts diverse in topic, genre and style and testing the observed tools on the training corpora extended with this part. In our opinion, this might alleviate the problem and help language identification algorithms learn both variants equally well.

8 Acknowledgment

This research was conducted through the projects 178006 and III 47003, financed by the Serbian Ministry of Science.

References

Timothy Baldwin and Marco Lui. 2010. Language identification: The long and the short of the matter. In *Human Language Technologies: The 2010 An-*

Tool		Serbian	Bosnian	Croatian	Other
langid.py	Ekavian	100	0	0	0
	Ijekavian	0	0	100	0
CLD	Ekavian	25	0	75	0
	Ijekavian	0	0	100	0
Tiedemann&Ljubešić	Ekavian	98	2	0	0
	Ijekavian	1	98	1	0

Table 4: The results

nual Conference of the North American Chapter of the ACL, pages 229–237.

William Cavnar and John Trenkle. 1994. N-gram based text clategorization. In *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175.

Ted Dunning. 1994. Statistical identification of language. Technical report.

Nikola Ljubešić, Nives Mikelić, and Damir Boras. 2007. Language identification: How to distinguish similar languages? In *Proceedings of the 29th International Conference on Information Technology Interfaces*, pages 541–546.

Marco Lui and Timothy Baldwin. 2011. Cross-domain feature selection for language identification. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*, pages 553–561.

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 25–30.

Bruno Martins and Mário J. Silva. 2005. Language identification in web pages. In *Proceedings of the 2005 ACM symposium on Applied computing, SAC'05*, pages 764–768.

Mary Milne, Richard O'Keefe, and Andrew Trotman. 2012. A study in language identification. In *Proceedings of ADCS'12*, pages 88–95.

Tom Mitchell. 1997. *Machine Learning*. McGraw-Hill.

Muntsa Padró and Lluís Padró. 2004. Comparing methods form language identification. pages 155–162.

Anil Kumar Singh. 2006. Study of some distance measures for language and encoding identification. In *Proceedings of ACL 2006 Workshop on Linguistic Distance*.

Živojin Stanojčić and Ljubomir Popović. 2011. *Grammar of the Serbian Language*. Institute for textbook publishing and teaching aids.

Jörg Tiedemann and Nikola Ljubešić. 2012. Efficient discrimination between closely related languages. In *Proceedings of COLING 2012*, pages 2619–2634.

Duško Vitas, Ljubomir Popović, Cvetana Krstev, Mladen Stanojević, and Ivan Obradović. 2011. *Languages in the European Information Society - Serbian*. Springer, Berlin.

Marcos Zampieri and Binyam Gebrekidan Gebre. 2012. Automatic identification of language varieties: The case of Portuguese. In *Proceedings of KONVENS2012*, pages 233–237.

Marcos Zampieri, Binyam Gebrekidan Gebre, and Sascha Diwersy. 2013. N-gram language models and POS distribution for the identification of Spanish varieties. In *Proceedings of TALN2013*, pages 580–587.