# Slavonic Corpus for Stylometry Research

Ján Švec and Jan Rygl

Natural Language Processing Centre
Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
{svec,rygl}@fi.muni.cz

**Abstract.** Stylometry techniques such as authorship recognition, machine translation detection and pedophile identification are daily used in applications for the most widely used languages. But under-represented languages lack data sources usable for stylometry research. In this paper, we propose an algorithm to build corpora containing meta-information required for stylometry experiments (author information, publication time, document heading, document borders) and introduce our tool *Authorship Corpora Builder* (ACB). We modify crawling and data-cleaning techniques for purposes of stylometry field and add heuristic layer to detect and extract meta-information.

The system was used on Czech and Slovak web domains to build a *Slavonic corpus* for stylometry research. Collected data have been published and we are planning to build collections for other languages and gradually extend existing ones.

**Keywords:** stylometry, slavonic corpus, web structure detection, corpora building

## 1 Introduction

Internet users are daily confronted with hundreds of situations in which they could use stylometry techniques. These situations include authorship detection (anonymous threats, false product reviews [1,2]); age and gender recognition (pedophile detection [3]); and spam and machine translation classification.

For dominant languages, there are many valuable data sources which can be used for a stylometry research (e-mail corpus Enron [4], age and gender corpus of Moshe Koppel [5]). Existence of these data sources enables fast implementation and comparison of the best techniques and facilitates further research.

But classic corpora (e.g. for Czech [6] and Slovak [7]) are unsuitable for several reasons:

- meta-information is missing (we do not know genres and publication times of texts; authors' identities, ages and genders);
- document borders are unclear;

Aleš Horák, Pavel Rychlý, Adam Rambousek (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2015, pp. 11–21, 2015. © Tribun EU 2015

- formatting is omitted (stylometry can use typography which cannot be witnessed in vertical corpus format)[1].

The lack of data for under-represented languages (such as languages of Visegrád Four) slows down development of stylometry tools for these languages. Therefore, we should contribute to building stylometry data sources before conducting further stylometry experiments in minor languages.

We propose a novel approach to building internet stylometry corpora and a modular system for collecting document with meta-information suitable for stylometry research is described. There are many systems for document crawling and text extraction but they are predominantly used for general corpora building (deduplication, boiler-plate removal, . . . ). Selecting the most suitable algorithms and adding a layer of heuristic enable fully automated data acquisition. Our data are automatically annotated using information from web site. Documents without meta-information are omitted.

Having data with meta-information is the key to reliable results in computational stylometry. In this paper, we present a system, which was successfully used to build the Slavonic stylometry corpus[2], a freely available Czech and Slovak corpus that can be used for stylometry research and many other applications. We are also planning to build collections for other minor Slavonic languages and publish them.

## 2 Building a Slavonic corpus

Building a stylometry web corpora based on internet articles consists of downloading data from web (predominant crawlers can be used); detecting the structure of the web page (classic algorithms are optimized for boiler-plate removal; we need to modify them); text extraction (we modify text processing to keep valuable information for stylometry); and novel heuristic evaluation of extracted data.

Stylometry corpora differ from classic corpora by giving more emphasis on relationship between author and his documents. It is focused on documents, which are associated with certain author, so we can simply get various information from it. We can determine what vocabulary is the author using, how is he constructing sentences etc.[8]

### 2.1 Downloading data from web

A web crawler is used to gather all pages needed for further extraction process. We have implemented our crawler using slightly modified *Crawler4j*[3] java library. In our approach, we modify crawler to work with specific category on selected domain. Our work is focused on web domains built on template, which

[1] word-per-line (WPL) text, as defined at the University of Stuttgart in the 1990s

[2] Data are available at `http://nlp.fi.muni.cz/projekty/acb/preview`

[3] Crawler4j, `https://github.com/yasserg/crawler4j`

can be analyzed. We use a filter on URL for category, to ensure that the template on downloaded pages within one domain is the same.

## 2.2 Site Style Tree algorithm

Site style tree can be used for cleaning web pages. It is based on analysis of templates of HTML pages. It assumes that the important information on the page differs in content, size and shape opposite to non-important parts which have the same structure among many pages on the same domain.

It uses a structure called *Style Tree*. Pages are first parsed into DOM tree and then transferred to *Style Tree* that consists of two types of nodes, namely, *style nodes* and *element nodes*.

*Style node* represents a layout or presentation style and it consists of two components. First is a sequence of element nodes, second is number of pages that has this particular style. *Element node* is similar to node in DOM tree, but differs in his pointer to child nodes – which in element node is set on sequence of style nodes. Interconnection of *style node* and *element node* creates *Style tree* [9]. Example of Site Style Tree can be seen on fig. 1.

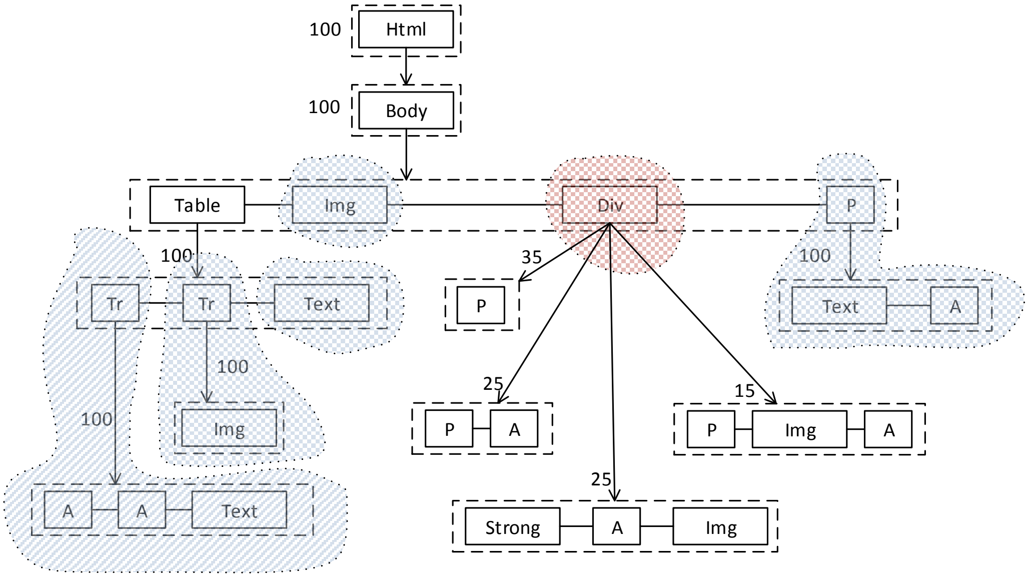

**Fig. 1.** Site Style Tree build from 100 pages.

In the SST we determine the important nodes, which have informational value like this: The more different child nodes *element node* has, the more important it is and, vice versa. We use weight for this attribute, which value can be between 0 and 1.

We used a metric, which measures the importance of element node [10]:

$$NodeImp(E) = \begin{cases} -\sum_{i=1}^{l} p_i \cdot log_m \cdot p_i & if\ m > 1 \\ 1 & if\ m = 1 \end{cases} \quad (1)$$

For particular element node $E$:

$m$ is number of pages containing $E$,
$l$ is number of child style nodes of $E$,
$p$ is number of pages of particular child style node of $E$.

From the equation we can see that when $E$ contains little child nodes, the value of NodeImp(E) will be small, and vice versa. For example – we can count NodeImp for Table (from fig. 1), like this:
$-0.35 \log_{100} 0.35 - 2 * (0.25 \log_{100} 0.25) - 0.15 \log_{100} 0.15 = 0.292 > 0.$

Computed value is afterwards compared with threshold. We can set the threshold in settings of our application. When the value is lower than threshold, element node is marked as non–informative, and vice versa.

### 2.3 Modifying SST for Stylometry Research

For our purpose we modified the SST algorithm in according way. If the input data contains a portal, which is unique between the pages, due to the *NodeImp* equation it receives value of one. Because we need to distinguish pages based on the same template and remove the boilerplate, portal pages and various pages which are very different are not the subject of our interest. We decided to remove unique pages from our analysis. For our needs we marked these pages as non–important. For increasing the efficiency of our algorithm, we explicitly increased the *NodeImp* value to nodes which may include the title, author or date, ensuring that these elements are kept. We use this approach to remove boilerplate from each web page. Afterwards it contains main text with meta data of the article - author, date and title.

### 2.4 Finding the article text

From modified SST we obtain only main text with meta-data. Text can contain boilerplate text – parts such as *a share button*, *a rate the article button*, *a link to comments*, etc. We can safely remove common text parts which are shorter than 20 characters[4] and doing so, boilerplate is removed.

Last step is to remove meta-data within the main text – author, date and title. When the required tags are found, we delete these tags from the main content of the page[5]. After the deletion we have only clean text of the article, which is stored.

[4] We have experimentally set the size limit to 20 after analysis of 5 different data sources

[5] More detailed description is in next subsections

### 2.5 Finding the author

*ACB* tries to find an expression "author(s)" in attributes of every tag in Slovak, Czech and English language (also tag `<author>` is checked). The score of each candidate tag is counted to determine the author (content must in most of cases abide name rules – regular expression preferring two words starting with capital letter). Because our current work aimed on small Slavic languages, we can detect female surnames by searching suffixes "ová". If tag with attribute "author" is not found, the algorithm tries to find the author in whole page. For example, we can look for name context such as *written by, published by, author*, etc. It is also important that the author is found near article text node in the SST, so the closer it is the higher score it has.

### 2.6 Finding the title

Title in most cases can be found in one of three different places:

1. tag `<h1></h1>` (or less probably in `<h2>`, `<h3>`, ...;
2. tag `<title></title>`;
3. attribute "title" of tag `<meta/>`.

A variant with the most diverse content (most of documents should have different titles) and abiding size limits (experimentally set min and max text length limits) is selected. If a found title contains boilerplate, e.g. *Server name: title name*, common phrase is automatically removed from the beginning of text.

### 2.7 Finding the date

In all examined languages, the algorithm searches an expression "date" between attributes of HTML tag. If we find this expression between attributes of a HTML tag, we check the content of this tag by regular expression recognizing date format (we also check relative times such as *today*, *yesterday*, ...

If there are multiple dates in one article, algorithm prefers the tag, which contains the attribute "date" and its contents corresponds to regular expression. It also prefers when date is found near article text in SST. If algorithm cannot find the expressions such as "date", it looks for all texts corresponding to date regular expression. If the matching text is found, its value is saved.

## 3 Results

We built a *Slavonic corpus* consisting of 15 Czech and Slovak web sites, from each we extracted 2000 articles. Therefore created corpus contains 30 000 documents and we plan to increase that number gradually. Collected data is published on `https://nlp.fi.muni.cz/projekty/acb/`[6].

[6] Data can be found in results section as version 2.

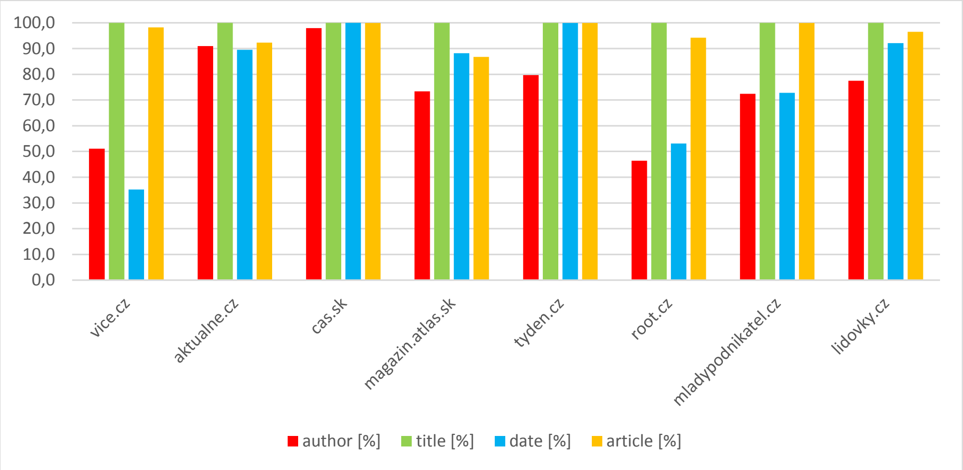

**Fig. 2.** Success rate of the first 8 sites

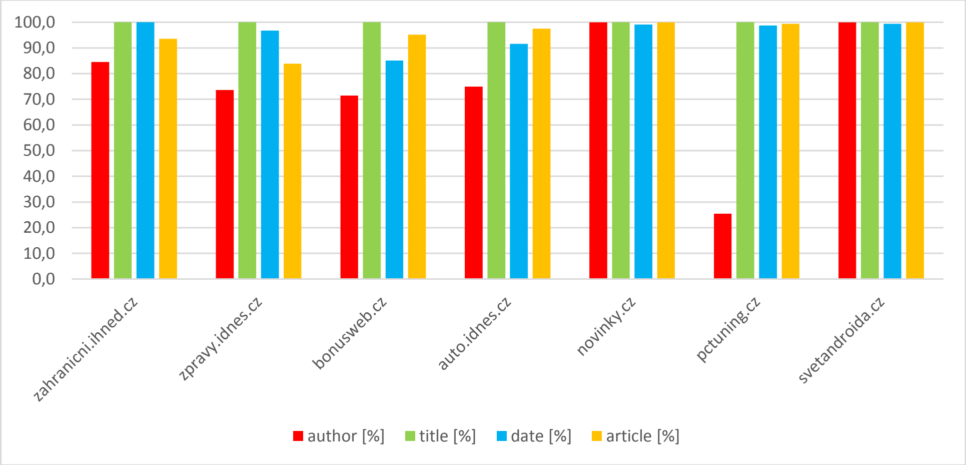

**Fig. 3.** Success rate of next 7 sites

We decided to use accuracy metric to determine the quality of acquired data (correctly found value is success, other cases are considered to be failure). Outside of scope of evaluation, we want to omit pages without found meta-information – portals, help pages, . . . .

We randomly selected 2000 documents and manually annotated them. Four categories were observed: author of the article, title, date and main text. If a correct author was found on 1800 pages from 2000, authors' accuracy was 90%; date was found on 1900 pages from 2000, accuracy is 95%, etc. We count accuracy for each category and we also compute average for each category and for whole algorithm. Success rate of the algorithm can be seen on fig. 2 and fig. 3.

Fig. 2 and 3 imply that the algorithm is most accurate in finding document title. For other categories the results are varying. It means that in some cases

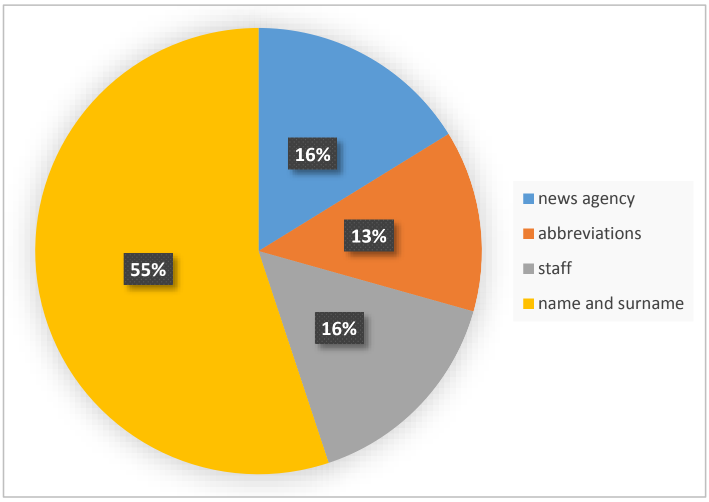

**Fig. 4.** Author data distribution

the important data was found only partly due to irregular structure of web pages. For example when it evaluated the page *vice.cz*, success rate of finding the date was 35.2%, because on most pages it uses different data format than we defined. It could be improved by extending the heuristic for finding date by adding more date formats.

Success rate for finding author is 74.6% , title is found in every case - 100%, date in 86,8% and article text in 95,8%. Average success rate of whole algorithm is 89.3%. Results indicate that algorithm described in this paper is suitable for building a *Slavonic corpus* for stylometry.

### 3.1 Authorship attribution

The **authorship attribution** problem can defined as follows: *Given a particular sample of text known to be by one of a set of authors, determine which one* [2, p. 238]. We selected authorship attribution for evaluation of our reference corpus because it is one of the most researched problems using a stylometry. The first deterministic stylometry technique was used to detect an authorship of documentsby T. C. Mendenhall in 1887 ([11]) and since then, authorship attribution is examined by security agencies, forensic experts, web authentication services and in many other tasks.

We used our *Slavonic corpus* as a training data source for authorship attribution. For our research the most important meta-information is author of the article, therefore if the algorithm don't find the author, the article is discarded – the corpora will always contain relevant data.

Correctly found authors (with their articles) are divided into 4 categories:

1. *news agency* - in our data mostly: "ČTK", "TASR" and "SITA",
2. *abbreviations* - author name in short form (two or three letters),
3. *staff* - article signed by name of web site, "staff" or "archive",

4. *name and surname* - full name of author.

Percentage of each category of our data is shown on fig. 4. For the problem of authorship attribution we focus only on data with authors' *name and surname* and omit other three categories.

In stylometry analysis, we want meta-information representing authors to correspond to one real person. Therefore, the 4th author category (full name of author is present) was used. Also, to define problem as having *n* candidates and trying to assign an author to an anonymous document, we want to document of each author to be present in tested data set to evaluate the task fairly. Therefore, we selected only authors with at least two documents, one document was used in a candidate set, once document was used in a test set of documents with examined authorship.

For each downloaded data set, we conducted following experiments:

1. We selected 2, 4, 8, 16 or 32 authors with at least 2 documents. If given set there wasn't enough documents, we took the highest possible count (e.g. in set with 10 documents, 2, 4, 8 and 10 candidates would be tested).
2. We divided documents to train (candidate) set and test (evaluation) set.
3. Documents were processed and analyzed using following stylometry techniques: word-length frequencies, stop-words frequencies, punctuation n-grams frequencies.
4. For each data set and candidate count, accuracy was measured and baseline established (baseline is $\frac{1}{\text{number of candidate authors}}$). Results are shown in Tables 1–7.

**Table 1.** atlas.sk

| Author count | True | False | Accuracy | Baseline |
|---|---:|---:|---:|---:|
| 2 | 25 | 1 | 96.15% | 50.00% |
| 4 | 26 | 4 | 86.67% | 25.00% |
| 8 | 38 | 40 | 48.72% | 12.50% |
| 14 | 53 | 61 | 46.49% | 7.14% |

**Table 2.** cas.sk

| Author count | True | False | Accuracy | Baseline |
|---|---:|---:|---:|---:|
| 2 | 4 | 0 | 100.00% | 50.00% |
| 4 | 5 | 3 | 62.50% | 25.00% |
| 8 | 9 | 11 | 45.00% | 12.50% |
| 16 | 11 | 28 | 28.21% | 6.25% |
| 32 | 12 | 74 | 13.95% | 3.12% |

**Table 3.** root.cz

| Author count | True | False | Accuracy | Baseline |
|---|---|---|---|---|
| 2 | 2 | 1 | 66.67% | 50.00% |
| 4 | 3 | 3 | 50.00% | 25.00% |
| 8 | 3 | 9 | 25.00% | 12.50% |
| 13 | 12 | 11 | 52.17% | 7.69% |

**Table 4.** svetandroida.cz

| Author count | True | False | Accuracy | Baseline |
|---|---|---|---|---|
| 2 | 5 | 1 | 83.33% | 50.00% |
| 4 | 13 | 7 | 65.00% | 25.00% |
| 8 | 20 | 23 | 46.51% | 12.50% |
| 16 | 21 | 58 | 26.58% | 6.25% |
| 19 | 29 | 65 | 30.85% | 5.26% |

**Table 5.** tyden.cz

| Author count | True | False | Accuracy | Baseline |
|---|---|---|---|---|
| 2 | 8 | 0 | 100.00% | 50.00% |
| 4 | 7 | 5 | 58.33% | 25.00% |
| 8 | 14 | 11 | 56.00% | 12.50% |
| 16 | 13 | 26 | 33.33% | 6.25% |
| 20 | 15 | 33 | 31.25% | 5.00% |

**Table 6.** vice.cz

| Author count | True | False | Accuracy | Baseline |
|---|---|---|---|---|
| 2 | 3 | 3 | 50.00% | 50.00% |
| 4 | 7 | 5 | 58.33% | 25.00% |
| 7 | 10 | 11 | 47.62% | 14.29% |

**Table 7.** zpravy.idnes.cz

| Author count | True | False | Accuracy | Baseline |
|---|---|---|---|---|
| 2 | 3 | 1 | 75.00% | 50.00% |
| 4 | 6 | 1 | 85.71% | 25.00% |
| 8 | 7 | 11 | 38.89% | 12.50% |
| 16 | 19 | 15 | 55.88% | 6.25% |
| 32 | 22 | 53 | 29.33% | 3.12% |

## 4 Conclusions

In this paper we described an application *ACB* which was used to build *Slavonic corpus* for stylometry research. The corpus contains contains articles grouped by author for each web domain. Every article contains also reference to its source, title and creation date. Main strength of the application is that building a corpus is fully automatic process – user is only required to insert start URL, and optionally adjust the settings.

Conducted authorship attribution experiments show that for low number of candidates, authorship can be attributed very reliably using basic stylometric features. For these scenarios, experimental results can be used as a baseline for other authorship attribution algorithms. If other scientist compares their data on the same data set, the methods become comparable.

Simple stylometric analysis is not good enough for higher number of candidates, but authors of other methods are welcome to establish more accurate baseline for these data using their techniques. We also plan to prepare download and attribute archives in which documents will be already divided into candidates authors' examples and examined instances. Each set will be named and highest achieved accuracy for the set will be dynamically updated in archive meta data.

## 5 Future work

We are currently working on full support of collecting of web discussions and very short documents – sites with more than one article per HTML page. We also plan to increase the precision by adding new heuristics for searching meta-information. We could also easily add an module for searching new types of meta-information. We are now working on expanding our *Slavonic corpus* and extending heuristics for other languages. Our goal is to build the biggest collection of stylometry data for under-represented European languages.

**Acknowledgements** This work has been partly supported by the Ministry of Education of CR within the LINDAT-Clarin project LM2010013.

## References

1. Chaski, C.E.: Who wrote it? steps toward a science of authorship identification. National Institute of Justice Journal (1997) 15–22
2. Joula, Patrick: Authorship Attribution. Foundations and Trends in Information Retrieval. (2008)
3. Peersman, C., Vaassen, F., Asch, V.V., Daelemans, W.: Conversation level constraints on pedophile detection in chat rooms. In: CLEF 2012 Evaluation Labs and Workshop, Online Working Notes, Rome, Italy, September 17-20, 2012. (2012)
4. Klimt, B., Yang, Y.: Introducing the enron corpus. In: CEAS 2004 - First Conference on Email and Anti-Spam, July 30-31, 2004, Mountain View, California, USA. (2004)
5. Koppel, M., Schler, J., Argamon, S., Pennebaker, J.: Effects of age and gender on blogging. In: In AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs. (2006)
6. Suchomel, V.: Recent czech web corpora. In Aleš Horák, P.R., ed.: 6th Workshop on Recent Advances in Slavonic Natural Language Processing, Brno, Tribun EU (2012) 77–83
7. Medved', M., Jakubíček, M., Kovář, V., Němčík, V.: Adaptation of czech parsers for slovak. In Aleš Horák, P.R., ed.: RASLAN 2012 Recent Advances in Slavonic Natural Language Processing, Brno, Czech Republic, Tribun EU (2012) 23–30

8. Koppel, M., Argamon, S., Shimoni, A.: Automatically categorizing written texts by author gender (2003)
9. Deepa, R., Nirmala, D.R.: Noisy elimination for web mining based on style tree approach. International Journal of Engineering Technology and Computer Research (IJETCR) **3** (2013) 23–26
10. Yi, L., Liu, B., Li, X.: Eliminating noisy information in web pages for data mining. In: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '03, New York, NY, USA, ACM (2003) 296–305
11. Mendenhall, T.C.: The characteristic curves of composition. The Popular Science **11** (1887) 237–246