# Slavonic corpus for stylometry research

Mgr. Ján Švec

Natural Language Processing Centre, Faculty of Informatics, Masaryk University

4.12.2015

# Motivation

- Having large amounts of data remains the key to reliable results in computational stylometry.
- Data sources for stylometry research in dominant languages:
  - e-mail corpus Enron [KY04],
  - age and gender corpus of Moshe Koppel [KSAP06].
- Lack of data for under-represented languages (such as languages of Visegrád Four)
  - Stylometric corpus for Czech language containing texts written by pupils at school [ÚČNK FF UK13].
- We contribute to stylometry data sources by building *Slavonic corpus*.
- Corpus was built from web by analysis websites based on *template*.

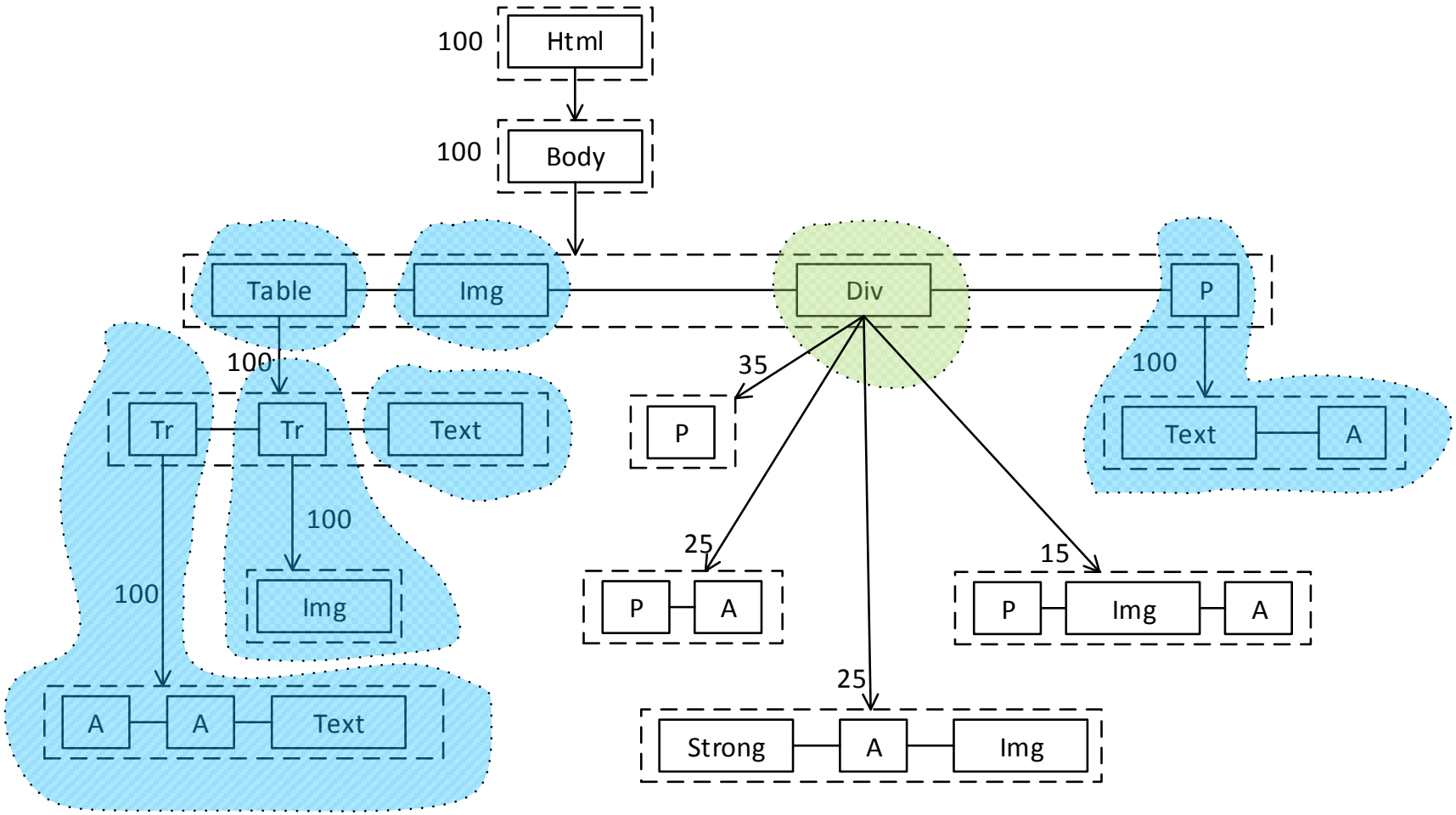
# Authorship corpora builder - system summary

- Downloading documents within one domain by modified Crawler4j[1].
  - input – URL of domain,
  - settings (number of pages, politeness, crawl depth, category limitation, threshold for cleaning),
  - name of file – hash value of file.
- Preprocessing by Jsoup[2]:
  - validation of documents,
  - removing unnecessary tags (script, style, comments),
  - file minimalization.
- Detection of structure by adapted algorithm Site Style Tree, removing the boilerplate.
- Extraction of meta-information by novel heuristic algorithms

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

[2] Jsoup: Java HTML Parser, `http://jsoup.org/`

# Sample of Site Style Tree




- SST – web domain in one structure; consists of 2 types of nodes:
  - style node (dashed) – contains number of pages of particular layout,
  - element node (solid) – corresponds with tag in DOM tree.

# Determining node importance

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

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

For particular element node $E$:

$m$ is number of all pages,

$l$ is number of child style nodes of $E$,

$p$ is a probability that web page uses *ith* child style node of $E$.

Example:

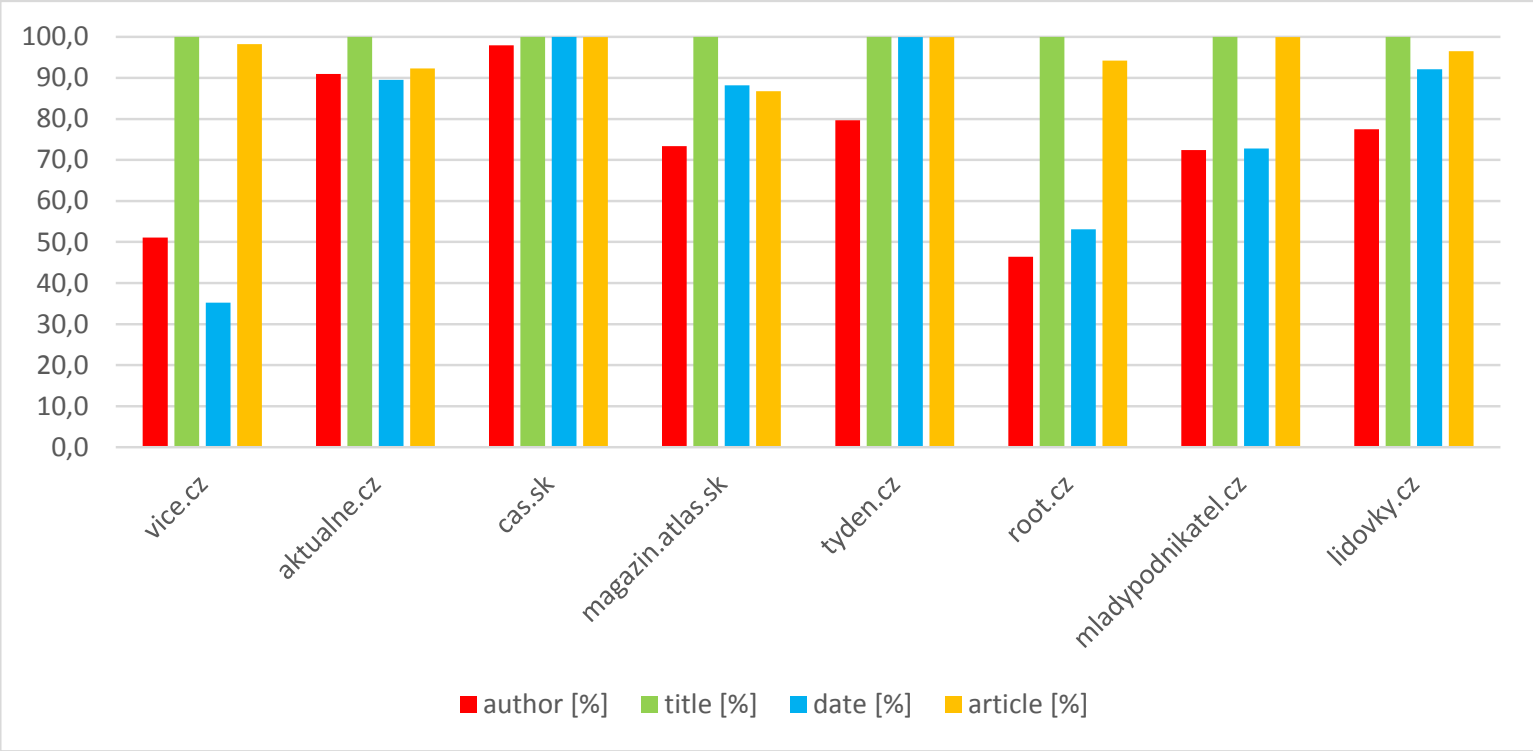
- m = 100, l = 4, p = 0.35; 0.25; 0.25; 0.15; threshold = 0.2[3]
- $-0.35 \log_{100} 0.35 - 2 * (0.25 \log_{100} 0.25) - 0.15 \log_{100} 0.15 = 0.292$
- If computed value is higher than threshold, node is important.
- $0.292 > 0.2 =$ node is important

[3] experimentally set after cleaning of small number of pages (20) with different threshold values and comparing results

# Extraction of meta-information by heuristic algorithms

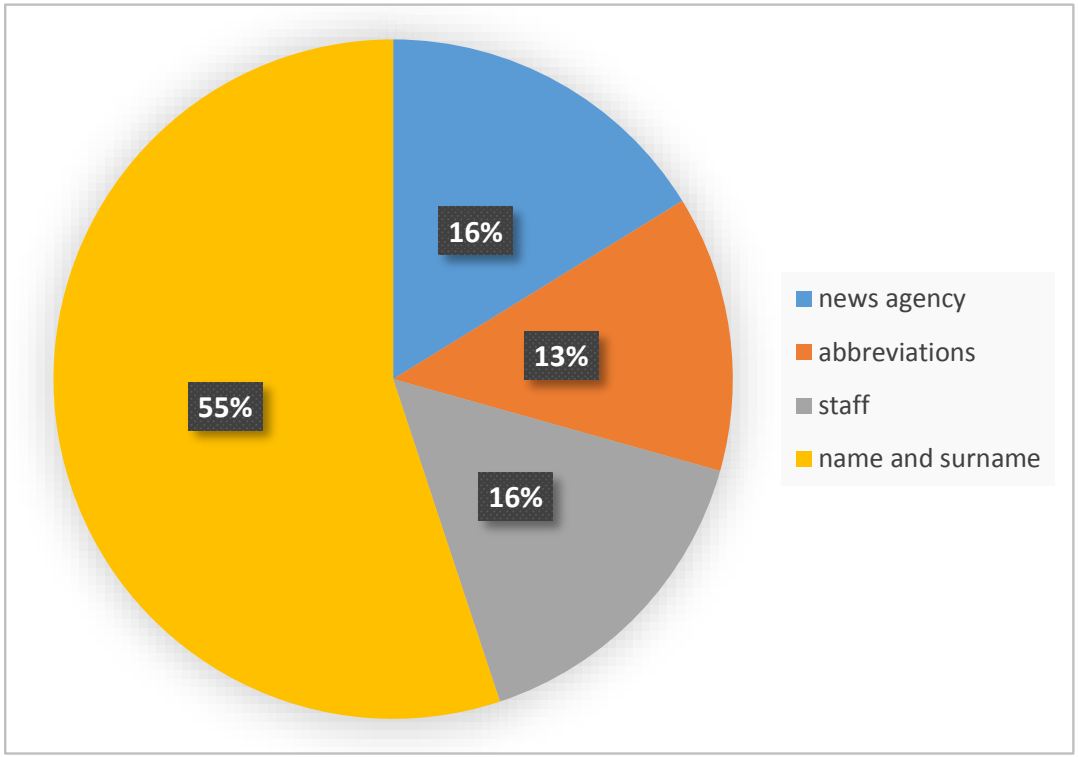
- Author – find an expression "author(s)" attributes of every tag,
  - female surnames – searching suffixes "ová",
  - search for "napísal" or "vlozil" and nearest word in SST.
- Date – regular expression in EU format (in order: day, month, year),
  - search for relative dates (today, yesterday, written before x minutes),
  - conversion to regular date format DD.MM.YYYY (based on file creation date).
- Title – tag `<h1>` or `<title>`; attribute "title" of tag `<meta/>`,
  - removal of server name in title "IDNES.cz - Novoroční projev",
  - intersection of titles within one domain (in SST).
- Article text – removing already found meta-information.

# Results




- Data is from 15 sites with articles,
  we analysed 2000 documents from each site
- Success rate of the algorithm (counting true/false) is 89,3 %.

# Author classification




- *News agency* - in our data mostly: “ČTK”, “TASR” and “SITA”,
- *abbreviations* - author name in short form (two or three letters),
- *staff* - article signed by name of web site, “staff” or “archive”,
- *name and surname* - full name of author.

# Authorship attribution example

| 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% |

Table 1: zpravy.idnes.cz

- We selected 2, 4, 8, 16 or 32 authors with at least 2 documents.
- Documents were divided to train (candidate) set and test (evaluation) set.
- Used stylometry techniques: word-length frequencies, stop-words frequencies, punctuation n-grams frequencies.
- For each data set and candidate count, accuracy was measured and baseline established (baseline is $\frac{1}{\text{number of candidate authors}}$).

# Future work

- Expand corpora by analyzing web discussions and multiple articles on one page.
- Increase the precision by adding new heuristics for searching meta-information.
- Enhancement for search on several types of websites (internet encyclopedia, social networks, etc.).
- Search more types of information (category).
- Expand corpora by extending heuristics for more languages.

# Conclusion

- Our tool *Authorship corpora builder* was used to built *Slavonic corpus*.
- The stylometry corpus contains 30 000 articles grouped by author.
- Data was used for authorship attribution experiment.
- Data is published online on:
  `https://nlp.fi.muni.cz/projekty/acb/preview`.

# End

Thank you for your attention.

# References I


- Moshe Koppel, J. Schler, Shlomo Argamon, and J. Pennebaker, *Effects of age and gender on blogging*, In AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs, 2006.
- Bryan Klimt and Yiming Yang, *Introducing the enron corpus*, CEAS 2004 - First Conference on Email and Anti-Spam, July 30-31, 2004, Mountain View, California, USA, 2004.
- ÚČNK FF UK, *SKRIPT2012: acquisition corpus of Czech written language – transcripts of the written work of pupils in primary and secondary schools in the Czech Republic*, 2013.