Literary Exploration Machine: A New Tool for Distant Readers of Polish Literature

Maciej Piasecki
maciej.piasecki@pwr.edu.pl
Wrocław University of Science and Technology
Poland

Tomasz Walkowiak
tomasz.walkowiak@pwr.edu.pl
Wrocław University of Science and Technology
Poland

Maciej Maryl
maciej.maryl@ibl.waw.pl
Polish Academy of Sciences, Poland

Brief Summary

This paper presents an initial prototype of a web-based application for textual scholars. The goal of this project is to create a complex and stable research environment allowing scholars to upload the texts they are analysing and either explore with a suite of dedicated tools or transform them into another format (text, table, list). This latter functionality is especially important for research into Polish texts, because it allows for further processing with the tools built for the English language.

This application brings together the existing applications developed by CLARIN-PL and supplements them with new functionalities. The project is based on a close cooperation between IT professionals, linguists and literary scholars, which ensures that the tools will suit actual researchers' needs.

The main features of LEM include: lemmatization, part-of-speech tagging, text clustering, semantic text classification based on machine learning, and visualisation of its output, generating custom wordlists and lemmatized texts.

Challenge

Digital literary studies seem to be one of the most vividly developing strand of digital humanities. Different analytical systems were proposed, e.g. Mallet, PhilоLogic3 plus PhiloMine, but focused on selected techniques and mostly on English texts. Their language-processing capabilities are limited only to lemmatization and morphosyntactic tagging and they usually require from their users certain programming skills.

In order to address those challenges we have developed a prototype of a web-based system, called *Literary Exploration Machine* (LEM), which does not require installation and programming skills. LEM has a component-based architecture, remains open for expanding components, implements natural language processing on different levels and is planned to support several different paradigms of the text analysis.

Scheme of the system

Components

Word frequencies can be simply computed for English, but not for highly inflected languages such as Polish, which has more than 100 possible word forms of an adjective (however, almost-full sets of distinct forms exist only for some lemmas). In such languages, morphological forms have to be first mapped to *lemmas* by a morpho-syntactic tagger, e.g. WCRFT2 for Polish (Radziszewski, 2013). By applying different language tools, we can enrich texts with metadata revealing linguistic structures.

LEM expands WebSty - an open stylometric system, adopting the following features for text description: segmentation-based (lengths of documents, paragraphs and sentences), morphological (words, punctuations, pseudo-suffixes and lemmas), grammatical classes and categories (e.g. from the Polish National Corpus –see Przepiórkowski et al, 2012– tagset, Broda and Piasecki, 2013) and their n-grams.
This set has been additionally expanded in LEM with the following features, allowing for semantic analysis:

- semantic *Proper Name classes* – recognised by a Named Entity Recogniser Liner2 (Marcińczuk et al, 2013),
- temporal, spatial relation (Kocoń and Marcińczuk, 2015), and selected semantic binary relations (e.g. *owner of*) ,
- *lexical meanings* – synsets in plWordNet (the Polish wordnet); assigned to words and selected multiword expressions by Word Sense Disambiguation tool WoSeDon (Kędzia et al, 2015),
- generalised lexical meanings – meanings mapped to more general synsets, e.g. *an animal* instead of *a cheetah*,
- lexicographic domains from Wordnet.

Rich text description is a good basis for several processing paradigms that LEM is going to support, namely:

- *linguistic text preprocessing* - extraction of language data for further statistical analysis, i.e. computing frequencies as the initial feature values, e.g., of lemmas, tags, word senses, etc.,
- *topic modelling*,
- unsupervised *semantic text clustering* and analysis of characteristic features for clusters,
- supervised *semantic text classification* - trained on the manually annotated texts,
- stylometric analysis - performed with the help of the WebSty system.

Processing scheme

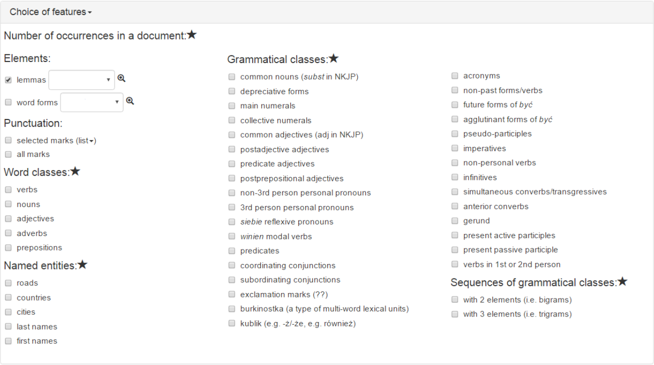
The processing paradigms share the following workflow:

- Uploading a corpus of documents together with metadata in CMDI format (Broeder et al, 2012) from the CLARIN infrastructure.
- Text extraction and cleaning.
- Choosing the features for the description of documents by users (see Fig. 1).
- Setting up the parameters for processing (users).
- Pre-processing texts with language tools.
- Calculating feature values for the pre-processed texts.
- Filtering and/or transforming the original feature values.
- Data mining.
- Presenting the results: visualisation or export of data.

To facilitate the upload, users are encouraged to deposit large text collections in the CLARI-PL dSpace repository. Users are advised to use public licences, but private research corpora can be also uploaded.

OCR-ed documents usually contain many language errors that should be corrected to some extent in the step 2. Moreover, metadata elements (e.g. page numbers, headers and footers) have to be separated during from the content and stored in a standalone annotation.

Users are not expected to have advanced knowledge of Natural Language Engineering or Data Mining. Thus, in Step 4, default settings of parameters will be provided. More advanced users will be able to tune the tool to their needs (see Fig. 1)



Choice of features▾

Number of occurrences in a document:★

Elements:

☑ lemmas ▾
□ word forms ▾

Punctuation:

□ selected marks (list▾)
□ all marks

Word classes:★

□ verbs
□ nouns
□ adjectives
□ adverbs
□ prepositions

Named entities:★

□ roads
□ countries
□ cities
□ last names
□ first names

Grammatical classes:★

□ common nouns (*subst* in NKJP)
□ depreciative forms
□ main numerals
□ collective numerals
□ common adjectives (adj in NKJP)
□ postadjective adjectives
□ predicate adjectives
□ postprepositional adjectives
□ non-3rd person personal pronouns
□ 3rd person personal pronouns
□ *siebie* reflexive pronouns
□ *winien* modal verbs
□ predicates
□ coordinating conjunctions
□ subordinating conjunctions
□ exclamation marks (??)
□ burkinostka (a type of multi-word lexical units)
□ kublik (e.g. -ż/-że, e.g. również)
□ acronyms
□ non-past forms/verbs
□ future forms of *być*
□ agglutinant forms of *być*
□ pseudo-participles
□ imperatives
□ non-personal verbs
□ infinitives
□ simultaneous converbs/transgressives
□ anterior converbs
□ gerund
□ present active participles
□ present passive participle
□ verbs in 1st or 2nd person

Sequences of grammatical classes:★

□ with 2 elements (i.e. bigrams)
□ with 3 elements (i.e. trigrams)

Figure 1. Web interface - a panel with a list of features

In Step 5 language tools are run. Each text is analysed by a part-of-speech tagger (e.g. WCRFT2) and next piped to a name entity recognizer (e.g. Liner2, Marcińczuk et al, 2013), temporal expression recognition, word sense recognition (WoSeDon, see Kędzia et al, 2015), etc.

Extraction of features encompasses counting frequencies, but also annotations matching patterns for every position in a document. In the case of wordnet-based features, meaning generalisation is done by iterating via wordnet structure.

A dedicated feature extraction module was built that is similar to Fextor (Broda et al, 2013) but much more efficient by supporting parallel processing. As a result of Step 6 every document is represented as vector of feature values and/or a sequence of language elements.

Filtering and transformation functions comes from the clustering packages or dedicated systems, e.g. SuperMatrix system (Broda and Piasecki, 2013).

Step 8 differentiates between the processing paradigms. Topic modelling, e.g. by Mallet, takes documents represented as lemma sequences. They can be also processed by corpus tools, e.g. for concordances and frequencies. Documents as feature vectors can be processed by clustering systems e.g. Cluto, or used in machine learning, e.g. Weka system.

Different processing paradigms provide varied perspectives on the data, e.g. topic modelling represents a document in terms of stochastic processes generating word occurrences from topic-related subsets in the text. Clustering reveals groups of documents based on content similarity. It is difficult to find a system that supports all paradigms.

In LEM, clustering is expanded with the extraction of features characteristic for the individual clusters. Several functions (from Weka, scikit-learn and SciPy packages), based on mathematical statistics, information theory and machine learning, are offered. The rankings of features are presented on the screen for interactive browsing and can be downloaded.

WebSty, based on elements of the same framework, can be applied to stylometric analysis.

Step 9, visualisation of clustering results (see Fig. 4), is based on Spectral Embedding (also known as Laplacian Eigenmaps). The 3D representation of the data (represented by similarity matrix) is calculated using a spectral decomposition of the graph Laplacian. Texts similar to each other are mapped close to each other in the low dimensional space, preserving local distances.

Use Case

The LEM prototype was developed by the team working with a particular textual corpus of 2553 Polish texts, published in *Teksty Drugie*, an academic journal dedicated to literary studies. The corpus consisted two parts: OCRd scans (1990-1998) and digital files (1999-2014). Given the aim of this paper (software presentation) and the shortage of space, we will treat the results only as examples of the method, without getting into too much detail.

The work on the prototype was divided into stages, conceived as a feedback loop for the developing team: on every stage a new service was added to application and the test run was performed. After the analysis of the result, the step was repeated or the team moved to the next phase.

Phase 1. Cleaning. The OCR-ed corpus has been cleaned (e.g. wordbreaks and headers were removed)

Phase 2. The corpus was lemmatized and parts of speech were tagged. Frequency lists were created what enabled the search for patterns in the textual output. For instance, Figure 2 shows the pattern of interest in particular Polish poets throughout 25 years, based on lemmatized mentions.

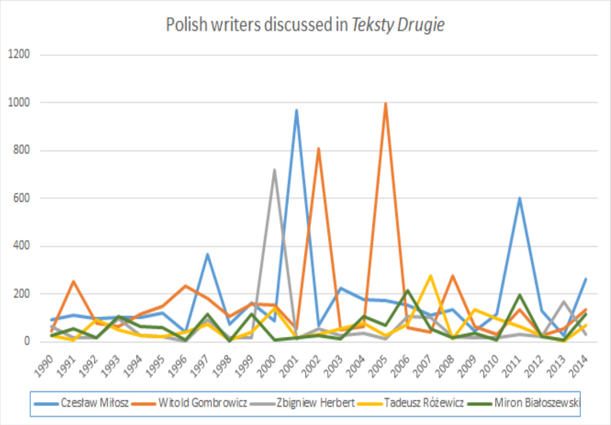


Figure 2. Pattern of interest in particular Polish writers in *Teksty Drugie* (1990-2014).

Phase 3. The analysis of the word frequencies revealed some problems with the word list, especially with numbers, years and city names, which were preserved in bibliographic references. A functionality of adopting a custom stopword list was employed. The exclusion of corpus-specific problematic words and general meaningless words (e.g. a, this, that, if) allowed for visualisation of the most frequent words in *Teksty Drugie* (Fig. 3)

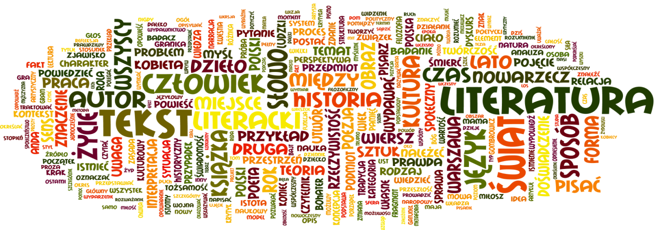


Figure 3. 300 most frequent words from *Teksty Drugie* (1990-2014) (meaningless words excluded) visualised with wordle.

Phase 4. The texts were then grouped into clusters of 20, 50 and 100 in a series of experiments. Each grouping revealed a bit different level of generalization about the texts. LEM, thanks to visualisation features (Fig. 4), allows for real-time exploration of deeper relationships between the texts.

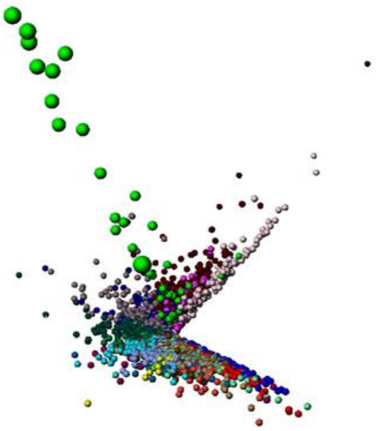


Figure 4. Visualisation of clustering results (weighting: MI-simple, similarity metric: ratio, number of clusters: 20, clustering method: agglomerative, visualization: the similarity matrix converted to distances and mapped to 3D by a spectral decomposition of the graph Laplacian - spectral embedding method).

By choosing the level of granularity (20, 50 or 100 clusters) we may analyse diverse patterns of discursive similarities between texts. Table 1 shows the differences in clustering of the same sample. The first option (20) shows the similarity between texts on a rather general level, that could be described as stylistic or genre similarity (e.g. formal vocabulary). Other options allow for more detailed exploration of general research approach (50) or particular topics analyzed in articles (100). Semantics of clusters is described by the identified characteristic features.

Number of clusters	100	50	20
Cluster size (mean)	25.33	50.66	56.65
Cluster size (median)	24	47	51,5
Smallest cluster size	13	25	2
Largest cluster size	51	91	96

Table 1. Differences between the clustering options (numbers reflect the quantity of texts assigned to particular cluster)

Researchers may explore all options and analyse the vocabulary responsible for classifying particular texts into a certain group by a virtue of being over- or under-represented in comparison to the entire sample.

The LEM is not a real time system. However, processing of the exemplar corpus (2553 documents from “Teksty Drugie”) takes less than 20 minutes. This is due to the use of a private cloud and proprietary message-oriented engine for processing texts. We plan to speed up the process, by running larger number of instances of language tools and by compressing results at each stage. Moreover, the user is able to start processing from any stage, so the processing time is shorter when the user plays with different settings.

Further Development

Currently LEM’s GUI is developed in cooperation with potential users, literary scholars working on various types of texts (fiction, journal articles, blog posts). That is also why we call this software “literary”, because further development will address the issues pertinent for literary theory, exceeding a purely linguistic perspective. Some literary-specific issues and functions will be expanded on the later stage of development, e.g. with adding language tools for Word Sense Disambiguation and partial analysis of the text structure, like anaphor resolution and discourse structure recognition. LEM’s architecture is open for such extensions. With that said, in this paper we have focused on the current stage of development.

LEM will be fully implemented and made available as a web application to the scholarly audience working on Polish. Next, it will be extended with with tools for other languages (e.g. English and German). As LEM has a modular architecture, it would require mostly linking new processing Web Services and adding converters. LEM has an open licences and we will be happy to share our tools, code and *know-how* with teams interested in doing so. Options for exporting to other formats will be added, so that researchers can easily create the output in a particular format (list, text, table) and upload it to other applications (e.g. Mallet) for further processing.

Bibliography

Broda, B., Kędzia, P., Marcińczuk, M., Radziszewski, A., Ramocki, R. and Wardyński, A. (2013). Fextor: A feature extraction framework for natural language processing: A case study in word sense disambiguation, relation recognition and anaphora resolution. *Studies in Computational Intelligence*. Berlin: Springer, vol. 458, pp. 41-62.

Broda, B. and Piasecki, M. (2013). Parallel, Massive Processing in SuperMatrix – a General Tool for Distributional Semantic Analysis of Corpora. *International Journal of Data Mining, Modelling and Management*, **5**(1):1–19.

Broeder, D., Van Uytvanck, D., Gavrilidou, M., Trippel, T., and Windhouwer, M. (2012). Standardizing a component metadata infrastructure. In: N. Calzolari (ed.), *Proceedings of LREC 2012: 8th International Conference on Language Resources and Evaluation*. European Language Resources Association (ELRA), pp. 1387-1390.

Eder, M., Kestemont, M. and Rybicki, J. (2013). Stylometry with R: a suite of tools. In: *Digital Humanities 2013: Conference Abstracts*. University of Nebraska-Lincoln, NE, pp. 487-489.

Kędzia, P., Piasecki, M. and Orlińska, M. J. (2015). Word Sense Disambiguation Based on Large Scale Polish CLARIN Heterogeneous Lexical Resources. *Cognitive Studies | Études cognitives*, (15), 269-292.

Kocoń, J. & Marcińczuk, M (2015). Recognition of Polish Temporal Expressions. In Mitkov, R., Angelova, G. & Boncheva, K. (editors), *Proceedings of the International*

Conference Recent Advances in Natural Language Processing, pages 282-290. INCOMA Ltd. Shoumen

Mallet (n.d.) http://mallet.cs.umass.edu/

Marcinczuk, M., Kocon, J. and Janicki, M. (2013). Liner2 - A Customizable Framework for Proper Names Recognition for Polish. *Studies in Computational Intelligence*. Berlin: Springer, vol. 467, pp. 231-253.

Marcińczuk, M. & Radziszewski, A (2013). WCCL Match – A Language for Text Annotation. In Kłopotek, A., M., Koronacki, Jacek, Marciniak, Małgorzata et al (editors), *Language Processing and Intelligent Information Systems*, pages 131-144. Springer Berlin Heidelberg.

PhiloLogi3 (n.d.) https://sites.google.com/site/philologic3/home

Piasecki, M.; Szpakowicz, S.; Maziarz, M. & Rudnicka, E. (2016) plWordNet 3.0 -- Almost There. In Mititelu, V. B.; Forăscu, C.; Fellbaum, C. & Vossen, P. *(Eds.)* Proceedings of the 8th Global Wordnet Conference, Bucharest, 27-30 January 2016, Global Wordnet Association, pp. 290-299.

Piasecki, M., Szpakowicz, S. & Broda, B. (2009). *A Wordnet from the Ground Up*. Wroclaw : Oficyna Wydawnicza Politechniki Wroclawskiej.

Przepiórkowski, A., Bańko, M., Górski, R. L. and Lewandowska-Tomaszczyk, B. (eds) (2012). *Narodowy Korpus Języka Polskiego*. Warszawa: PWN.

Radziszewski, A. (2013). A tiered CRF tagger for Polish, Intelligent Tools for Building a Scientific Information Platform. *Studies in Computational Intelligence*. Berlin: Springer, vol. 467, pp. 215-230.

Rygl, J. (2014) Automatic Adaptation of Author's Stylometric Features to Document Types. In Sojka, P., Horák, A., Kopeček, I. and Pala, K. (eds), *Proceedings of 17th International Conference TSD 2014*. Brno, Czech Republic, LNCS 8655, Springer.

Szałkiewicz, Ł. and Przepiórkowski, A. (2012). Anotacja morfoskładniowa. In Przepiórkowski, A., Bańko, M., Górski, R. L. and Lewandowska-Tomaszczyk, B. (eds) (2012). *Narodowy Korpus Języka Polskiego*. Warszawa: PWN., pp. 59-96.

Walkowiak, T. (2015). Web based engine for processing and clustering of Polish texts. *Proceedings of the Tenth International Conference on Dependability and Complex Systems DepCoS-RELCOMEX*. Brunów, Poland. Springer, pp. 515-522.

WebSty (n.d.) http://websty.clarin-pl.eu/

Zhao, Y. and Karypis, G. (2005). Hierarchical Clustering Algorithms for Document Datasets. *Data Mining and Knowledge Discovery*, **10**(2): 141-168.