INSTYTUT BADAŃ LITERACKICH POLSKIEJ AKADEMII NAUK
Institute of Literary Research Polish Academy of Sciences

Wrocław University of Technology

Machine Exploration of Secondary Literature with Literary Exploration Machine

Maciej Maryl
Institute of Literary Research,
Polish Academy of Sciences

Maciej Piasecki
Tomasz Walkowiak
G4.19 Research Group
Wrocław University
of Science and Technology

CLARIN-PL
Common Language Resources and Technology Infrastructure

DIGITAL HUMANITIES CENTRE

ul. Nowy Świat 72, 00-330 Warsaw, Poland
phone/fax: (22) 826 99 45, (22) 65 72 895
e-mail: sekretariat@ibl.waw.pl

Once upon a time there was a workshop

Simple Tools, … but Complicated

Simple Tools, … but Complicated

XML Text

Ciemne włosy jej były gdzieniegdzie posrebrzone siwizną, le
wyszczuplał i kształtne.

włos subst:pl:nom:m3

Simple Tools, … but Complicated


```
<chunkList> <chunk id="ch1" type="p">  <sentence id="s1">
<tok> <orth>Ciemne</orth>
        <lex disamb="1"> <base>ciemny</base>
        <ctag>adj:sg:nom:n:pos</ctag></lex>   </tok>
<tok> <orth>włosy</orth>
        <lex disamb="1"> <base>włos</base>
        <ctag>subst:pl:nom:m3</ctag></lex>   </tok>
<tok> <orth>jej</orth>
        <lex disamb="1"> <base>on</base>
       <ctag>ppron3:sg:gen:f:ter:akc:npraep</ctag> </lex></tok>
```


- **Lemmatisation**: *ciemny włos on*

LEM = Literary Exploration Machine

- Aggregating existing language tools for Polish
- Enabling interoperation with tools developed for other languages (CLUTO, Mallet)
- Simple workflow not requiring programming skills:
  - upload your corpus,
  - tweak the parameters,
  - … lay back and wait for the output
- User-driven approach – LEM is developed through case-studies according to particular research problems

http://ws.clarin-pl.eu/lem.shtml?en

LEM: User Interface

Summarize Keywords TF-IDF Inkluz TermoPL LEM MeWeX

Literary Exploration Machine

Literary Exploration Machine (LEM) provides a virtual research environment for textual scholars, allowing them to upload texts in Polish and either explore them with a suite of dedicated tools or transform them into another format (text, table, list).
The used tools include:

- Apache Tika,Morfeusz 2 with SGJP dictionary (for morphological analysis), wcrft2 (for tagging)
- WebSty

About

Instructions

Click or drag and drop files

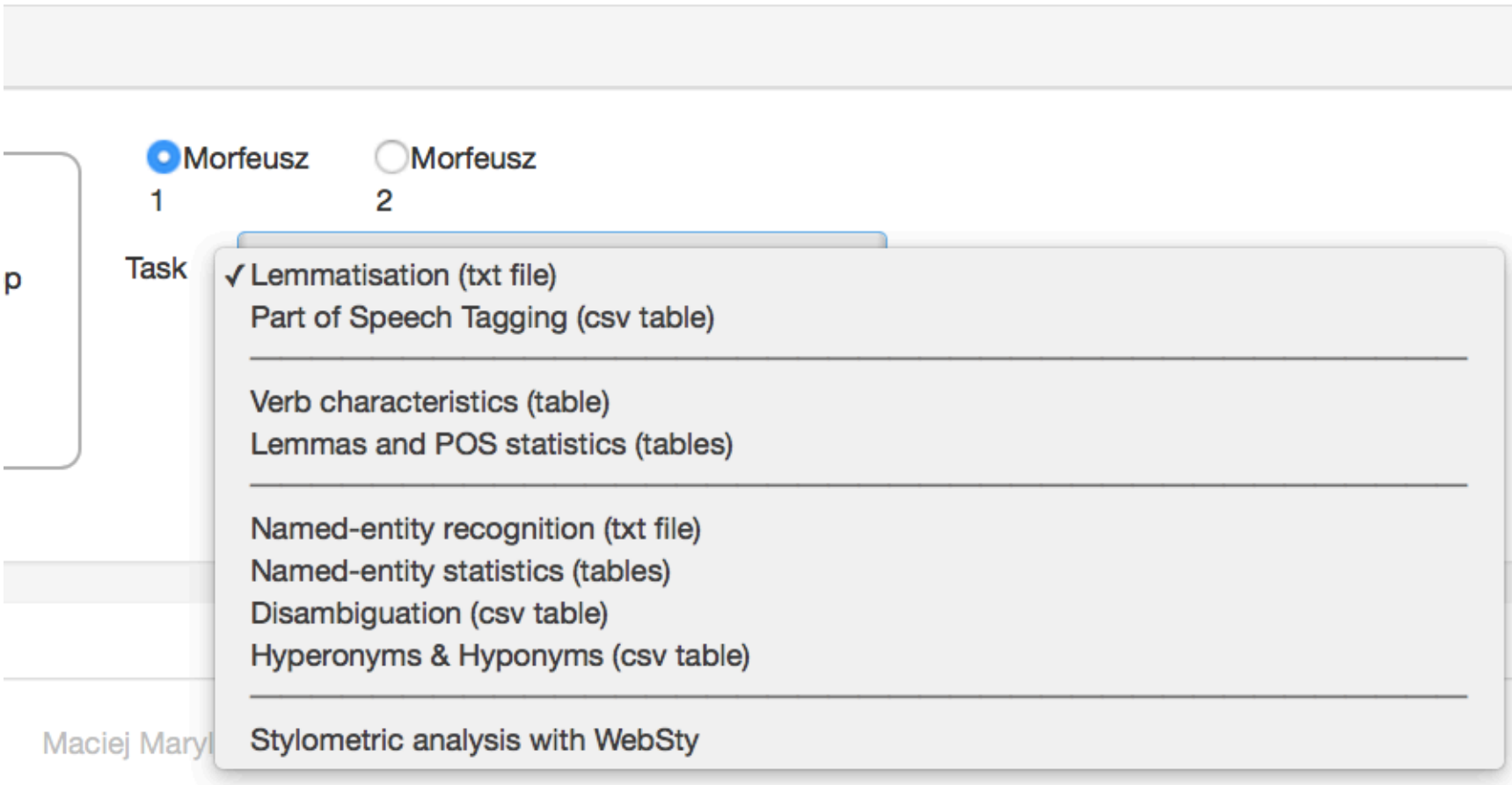
Morfeusz 1 Morfeusz 2

Task Lemmatisation (txt file)

Process

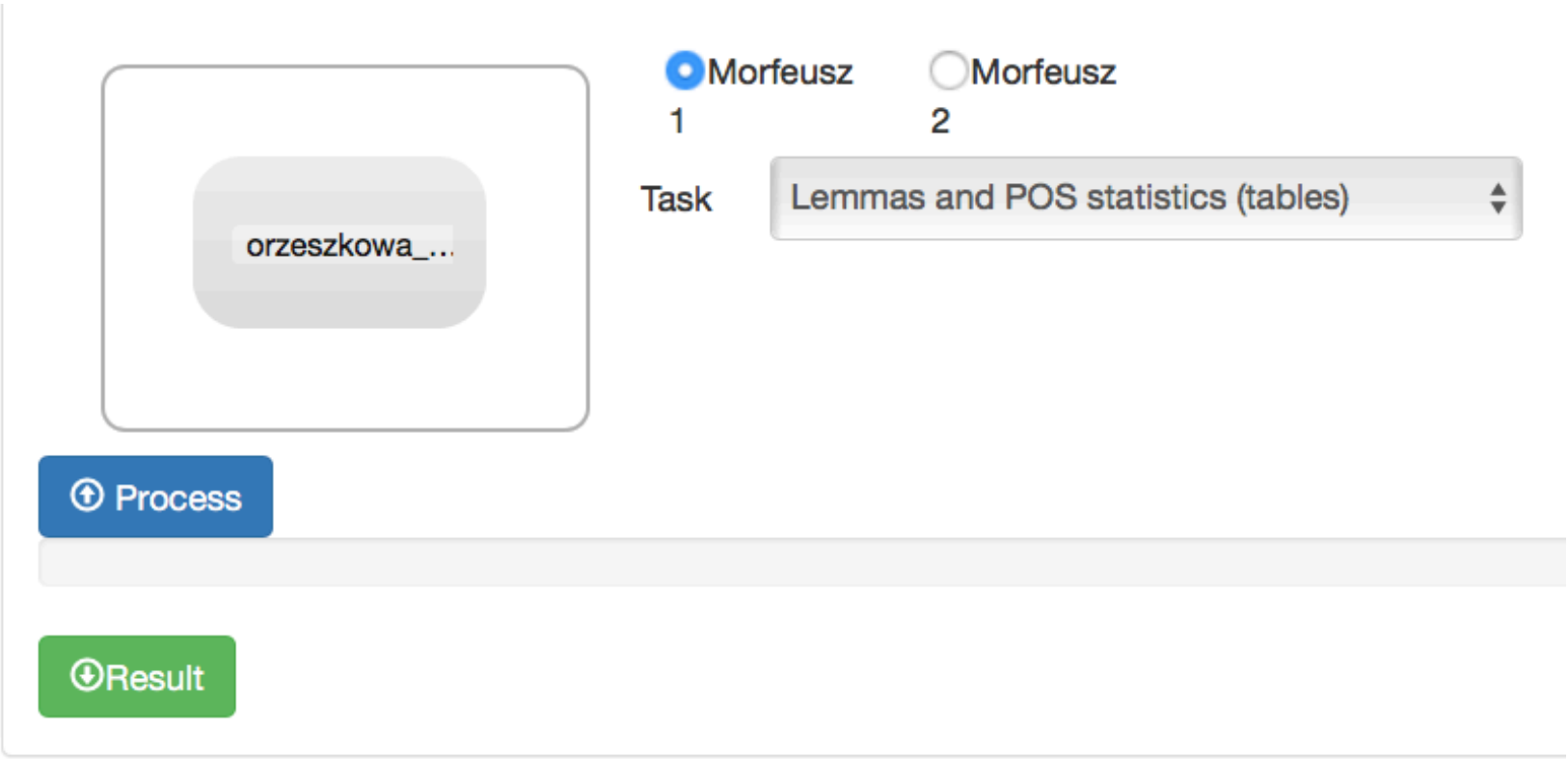
http://ws.clarin-pl.eu/lem.shtml?en

LEM: User Interface



http://ws.clarin-pl.eu/lem.shtml?en

LEM: User Interface



http://ws.clarin-pl.eu/lem.shtml?en

LEM: Functions

- Basic description
  - Lemmatisation
  - Part of Speech tags
- Combined morpho-syntactic information
  - Verb characteristics (including person and gender)
  - Lemmas and PoS
- Semantic
  - Proper Names and their frequencies
  - Disambiguation of Word Senses (WSD)
  - Description of Word Senses: Hypernyms & Hyponyms
- Stylometric analysis: predefined setting of *WebSty*

http://ws.clarin-pl.eu/websty.shtml?en

Basic Functions: Lemmatisation

- **Morphosyntactic tagging (*WCRFT2* or *MorphoDiTa-pl*)**
- **Sample text: Eliza Orzeszkowa, *Szczęśliwa***

Input (*a zip file with textual files inside*):

Wysoka, kształtna, z twarzą myślącą, zimną nieco, ale pięknie zarysowaną i bardzo świeżą, w stroju pełnym smaku i powagi, siedzi pod rozłożystemi drzewami wspaniałego parku i myśli o tem, jaki ten park jest piękny, jaki ten dzień letni jest pogodny i jaka ona sama jest szczęśliwa.

Output:

wysoki , kształtny , z twarz myśląca , zimny nieco , ale pięknie zarysować i bardzo świeży , w strój pełny smak i powaga , siedzieć pod rozłożystemi drzewo wspaniały park i myśleć o tema , jaki ten park być piękny, jaki ten dzień letni być pogodny i jaki on sam być szczęśliwy .

Basic: Part of Speech Tags

- **Output: CSV file with statistics for PoS tags (National Corpus of Polish) tagest**

WORD	LEMMA	PoS
Nie	nie	qub
była	być	praet
już	już	qub
młodą	młody	adj
,	,	interp
lecz	lecz	conj
twarz	twarz	subst
jej	on	ppron3
zachowała	zachować	praet
delikatność	delikatność	subst
rysów	rys	subst
i	i	conj
cery	cer	subst
,	,	interp
kibić	kibić	subst

Sample text:
Eliza Orzeszkowa,
Kto winien

Combined: Verb Characteristics

- Output: person and gender across verbs, expressed in the National Corpus of Polish tagset
- Excel file
  - to simplify work with UTF8 encoding
- Sample text: Eliza Orzeszkowa, *Kto winien*

		SINGULAR						PLURAL					
Tokens	Verbs	1Pers	2pers	3pers	3pers_m	3pers_f	3pers_n	1Pers	2pers	3pers	3pers_m	3pers-nm	inf
11242	1299	100	100	84	151	465	0	0	0	0	0	0	150

Combined: Lemma & PoS

- **Output: frequencies and ratios (percents) of lemmas and NCP tags**
- **Excel files**
- **Sample text: Eliza Orzeszkowa, *Kto winien***

człowiek	36
ale	34
o	34
życie	33
od	33
oko	32

subst:sg:gen:m3	122
subst:sg:nom:f	119
subst:sg:gen:n	115
subst:sg:nom:n	111
prep:gen	107
prep:gen:nwok	105

Semantic: Proper Names

- Output:
  - PN list (text file)
  - PN frequencies (Excel file)
- Tool: Liner2
- Problem
  - lemmatisation of multi-word PNs

Sample text:
Jerzy Żuławski,
Veneri et Romae

NAMED ENTITY	LEMMA	FREQ
Rzym	Rzym	19
Palatynie	Palatyn	13
Kapitolu	Kapitol	7
Forum	forum	6
Konstantyna	Konstantyn	4
Koloseum	Koloseum	3
Piotra	Piotr	3
Słońce	słońce	3
Via Sacra	via sacrum	3
Baedeker	Baedeker	2
Grecji	Grecja	2
Kastora	Kastor	2
Marka Aureljusza	Marek aureljusza	2

Semantic: Word Senses

- **Output:**
  - **frequency of plWordNet synsets**
  - **Synsets described by hypernyms and hyponyms**
- **Tool: WoSeDon**

Sample text: Jerzy Żuławski, *Veneri et Romae*

WORD	LEMMA	PoS	WORDNET SYNSET
niespokojny	niespokojny	adj	niespokojny.3(42:jak)
sen	sen	subst	spoczynek.2(23:st), sen.1(23:st)
jakiejś	jakiś	adj	jakowyś.1(42:jak); który.1(42:jak) jaki.1(42:jak); jakiś.1(42:jak) jakowy.1(42:jak); któryś.2(42:jak)
jednej	jeden	adj	pewien.1(42:jak) jeden.3(42:jak)
nocy	noc	subst	noc.2(25:czas)
jesiennej	jesienny	adj	jesienny.1(43:rel)

Stylometric analysis: WebSty

- Tool: *WebSty* run on predefined settings
- Output: visualisations and files from *WebSty*

My_Corpus....

Morfeusz 1 Morfeusz 2

Task: Stylometric analysis with WebSty

Number of groups: 2

Stylometric analysis is 'the study of measurable features of (literary) style, such as sentence length, vocabulary richness and various frequencies (of words, word lengths, word forms, etc.)'[1]. The most characteristic use of computational stylometric analysis includes the authorship attribution, which aims to propose "taxonomies of features to quantify the writing style, the so called style markers, under different labels and criteria"[2] in order to identify the author of the text if he or she is unknown or his or her authorship is disputed. In this respect stylometry has gained some publicity and the result of analysis were accepted in court cases as evidence[2][3].
LEM offers the user a simplified interface to the CLARIN-PL WebSty and returns multiple different visualization of the result of the analysis.

Process

- heatmap
- schemaball
- circle
- grouping (table)
- multidimensional scaling
- multidimensional scaling in 3D
- analiza istotności cech w grupachimportance of features

http://ws.clarin-pl.eu/websty.shtml?en

Further development

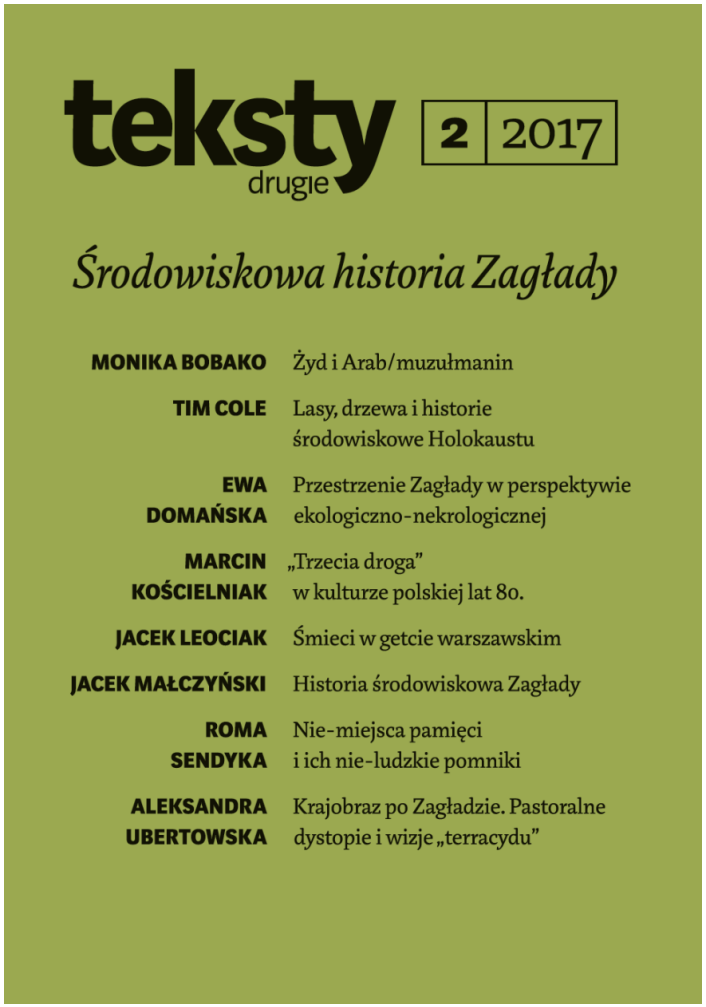
- Filtering
  - Lemmatisation stoplists: words, lemmas, tags
  - Automated based on feature importance
- Extended use of *WebSty* engine for unsupervised clustering of texts (including semantic features)
- Integrating *Mallet* for topic modelling
- *Gephi* output for relation exploration, e.g. Proper Names
- Extended use of user supplied metadata (file names and CMDI)
- Identification of characteristic features: files and groups of files
- Integration with collocation and terminology extraction
- Emotive analysis: sentiment polarity, basic emotions

Development Driven by Users' Research Tasks and Projects

1. Research on secondary literature, i.e. scientific journal articles (a comparative study of the transformation in Polish literary scholarship 1989-2014 with a matching study on historical research in that period)
2. Evolution of a collective genre on a web portal – blogs as literature
3. Working with primary sources (Radio Free Europe materials) – distant reading

- Literary studies – text similarity (Tracer) (Jan Rybicki)
- Sociology – pre-processing for quantitative analysis (Grzegorz Byrda)
- Social psychology - towards LIWC-like tool for Polish
  - studies on depression (Natalia Rohnka)
  - analysis of emotions and polarisation (Aleksandra Świderska)

Case Study: Teksty Drugie (*Second Texts*)



teksty drugie 2 2017

Środowiskowa historia Zagłady

MONIKA BOBAKO	Żyd i Arab/muzułmanin
TIM COLE	Lasy, drzewa i historie środowiskowe Holokaustu
EWA DOMAŃSKA	Przestrzenie Zagłady w perspektywie ekologiczno-nekrologicznej
MARCIN KOŚCIELNIAK	„Trzecia droga" w kulturze polskiej lat 80.
JACEK LEOCIAK	Śmieci w getcie warszawskim
JACEK MAŁCZYŃSKI	Historia środowiskowa Zagłady
ROMA SENDYKA	Nie-miejsca pamięci i ich nie-ludzkie pomniki
ALEKSANDRA UBERTOWSKA	Krajobraz po Zagładzie. Pastoralne dystopie i wizje „terracydu"

- Academic journal dedicated to literary studies
- Established in 1990
- Important for literary and cultural studies in Poland
- Special open-access editions in English
- www.tekstydrugie.pl
- 25 years:
  - 1990-2014
  - 2609 articles
  - 125 issues (21 articles/issue; 105 articles/year)

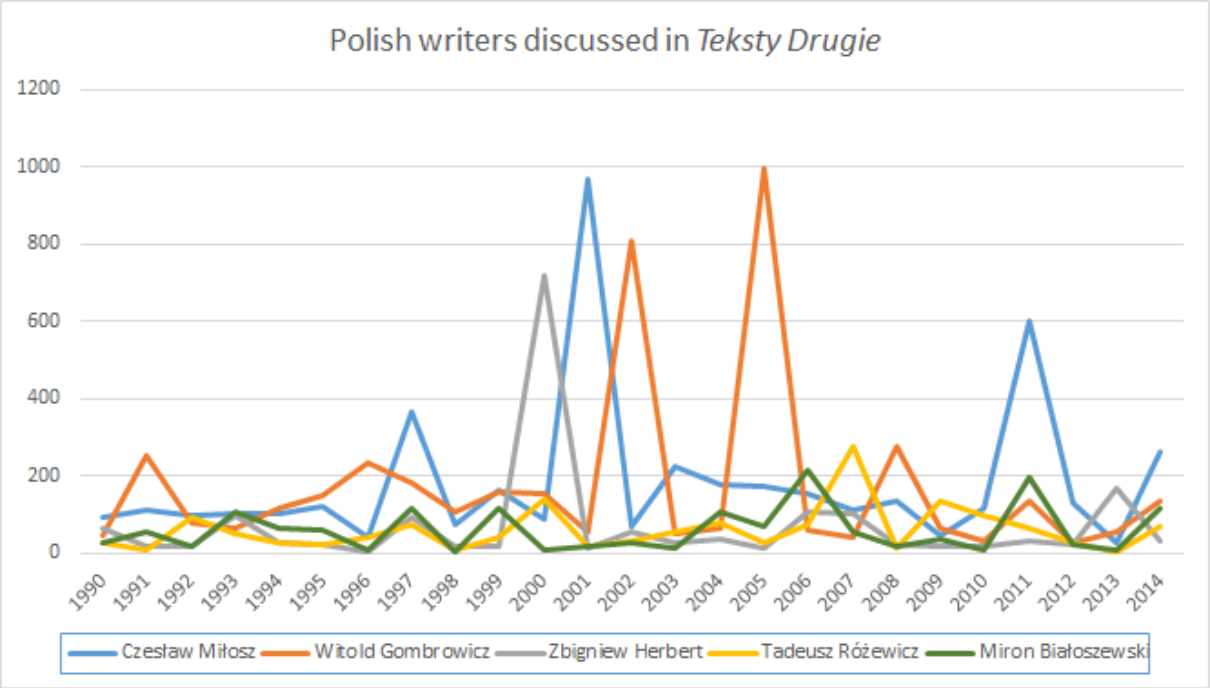
Workflow

- Phase 1. OCR and cleaning the corpus
- Phase 2. Pre-processing and manual analysis of frequencies
  - [LEM] Lemmatisation, Part of Speech tagging
  - [LEM] Generation of frequency lists
  - searching for patterns in the textual output
- Phase 3. Exploration of the word frequencies
  - Unsupervised text clustering [WebSty]
  - LEM-expanded with stoplists for filtering
  - Machine Learning for the description of classes [WebSty extended engine]

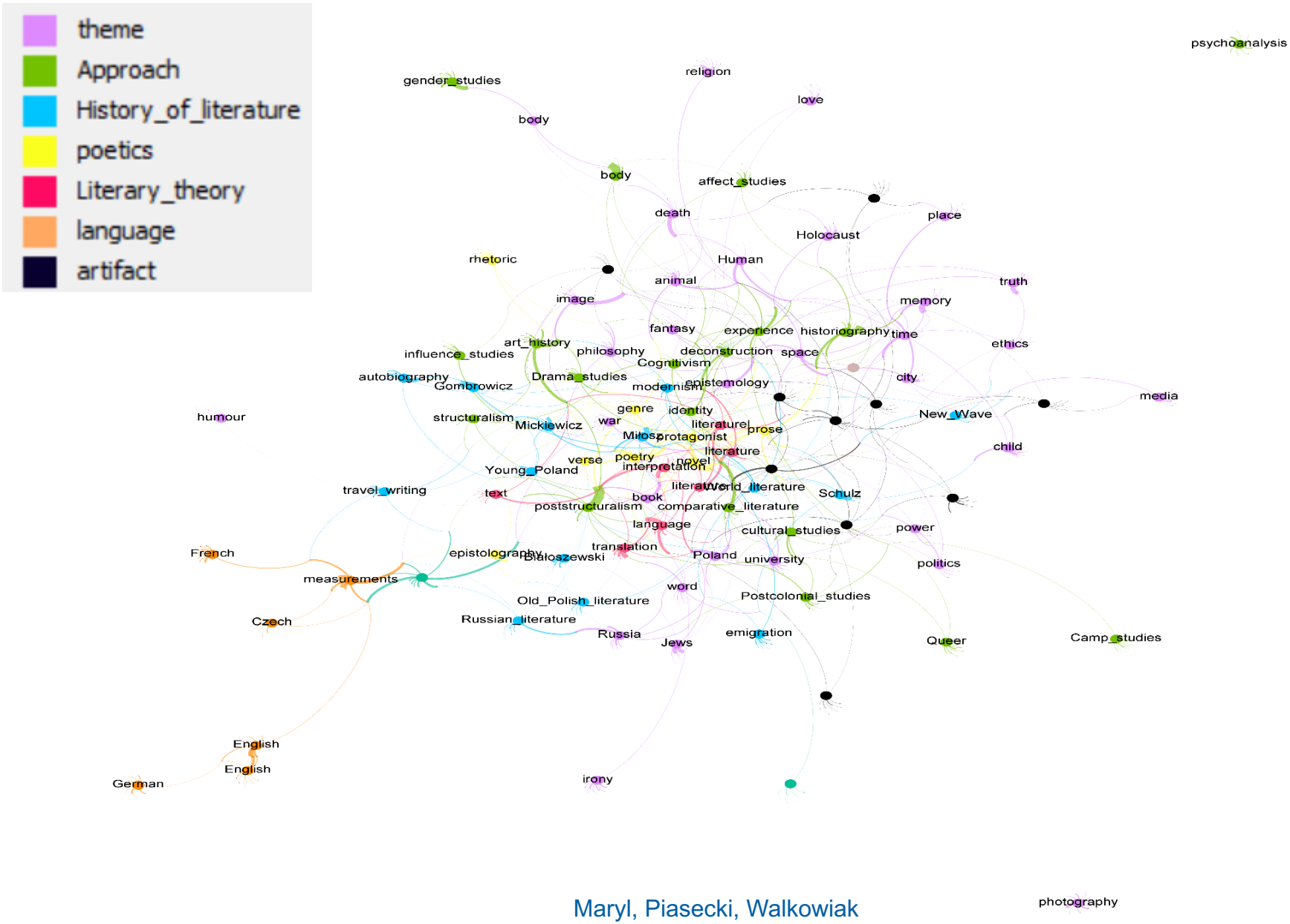
Step 2: Simple wordclouds

Step 3: Chronological patterns

- **Manual analysis of PN frequencies**

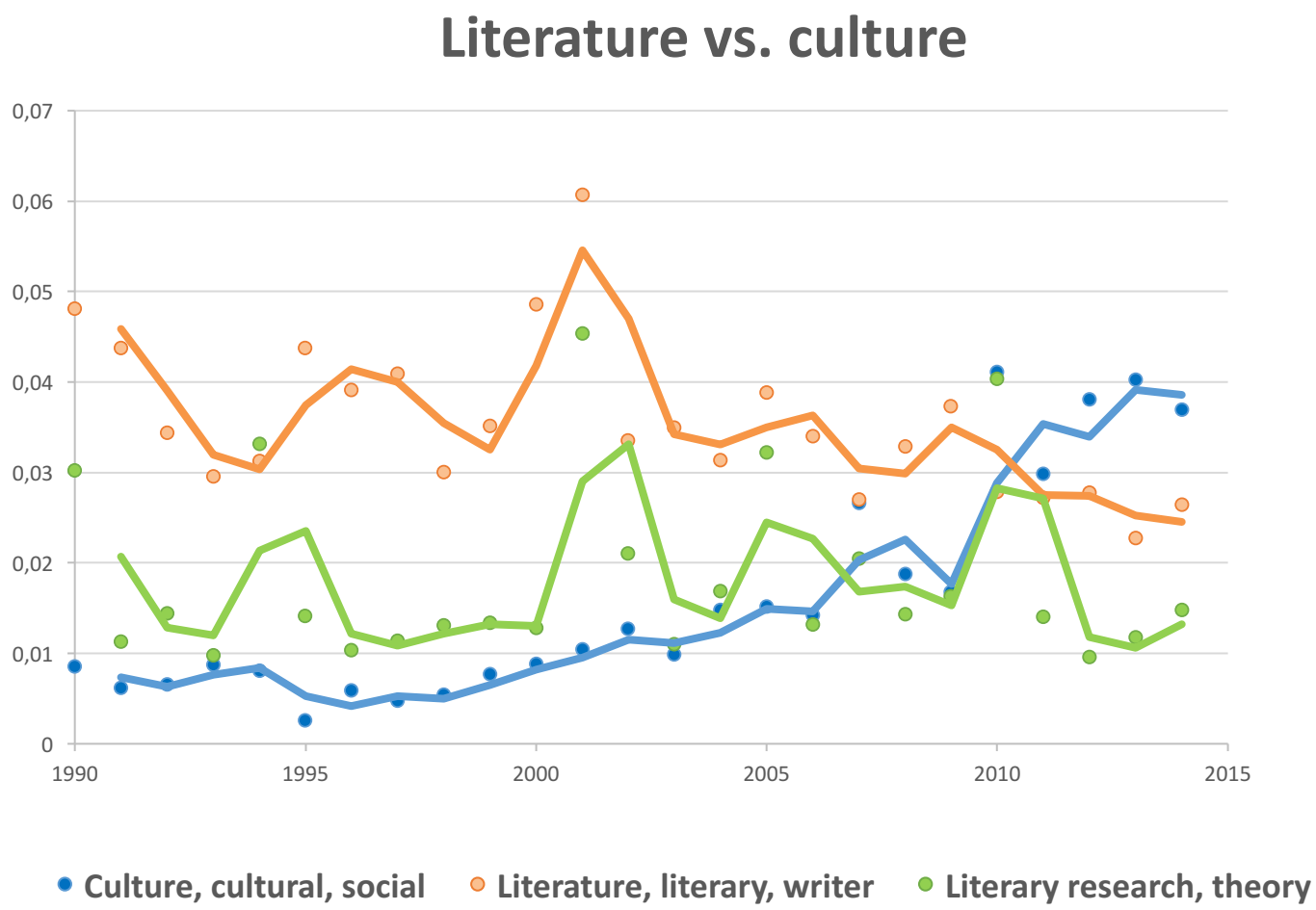


Phase 4: Topic Modelling in Mallet



Maryl, Piasecki, Walkowiak
- Machine Exploration of Literature

Phase 5: Topics overtime



To be continued …

Thank you very much for your attention!

References

- 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
- 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.
- Piasecki, M., Szpakowicz, S. & Broda, B. (2009). A Wordnet from the Ground Up. Wroclaw : Oficyna Wydawnicza Politechniki Wroclawskiej.
- 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, 2016, pp. 290-299.
- 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 [8], 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.
- Zhao, Y. and Karypis, G. (2005). Hierarchical Clustering Algorithms for Document Datasets. Data Mining and Knowledge Discovery, 10(2): 14
WebSty http://websty.clarin-pl.eu/

Maciej.Maryl@ibl.waw.pl **Maciej.Piasecki@pwr.edu.pl**