Conjugate, decline and spell like years ago
A corpus-based morphological analyzer of 19th century Polish

Joanna Bilińska Witold Kieraś Magdalena Derwojedowa

Chair of Formal Linguistics

Institute of Polish Language
University of Warsaw

Slavic Diachronic Corpus Linguistics Conference
Tromsø, April, 21-22, 2015

The project

Automatic morphological analysis of Polish texts from 1830-1918 period with respect to evolution of inflection and spelling, grant awarded by Polish National Science Centre (DEC-2012/07/B/HS2/00570)

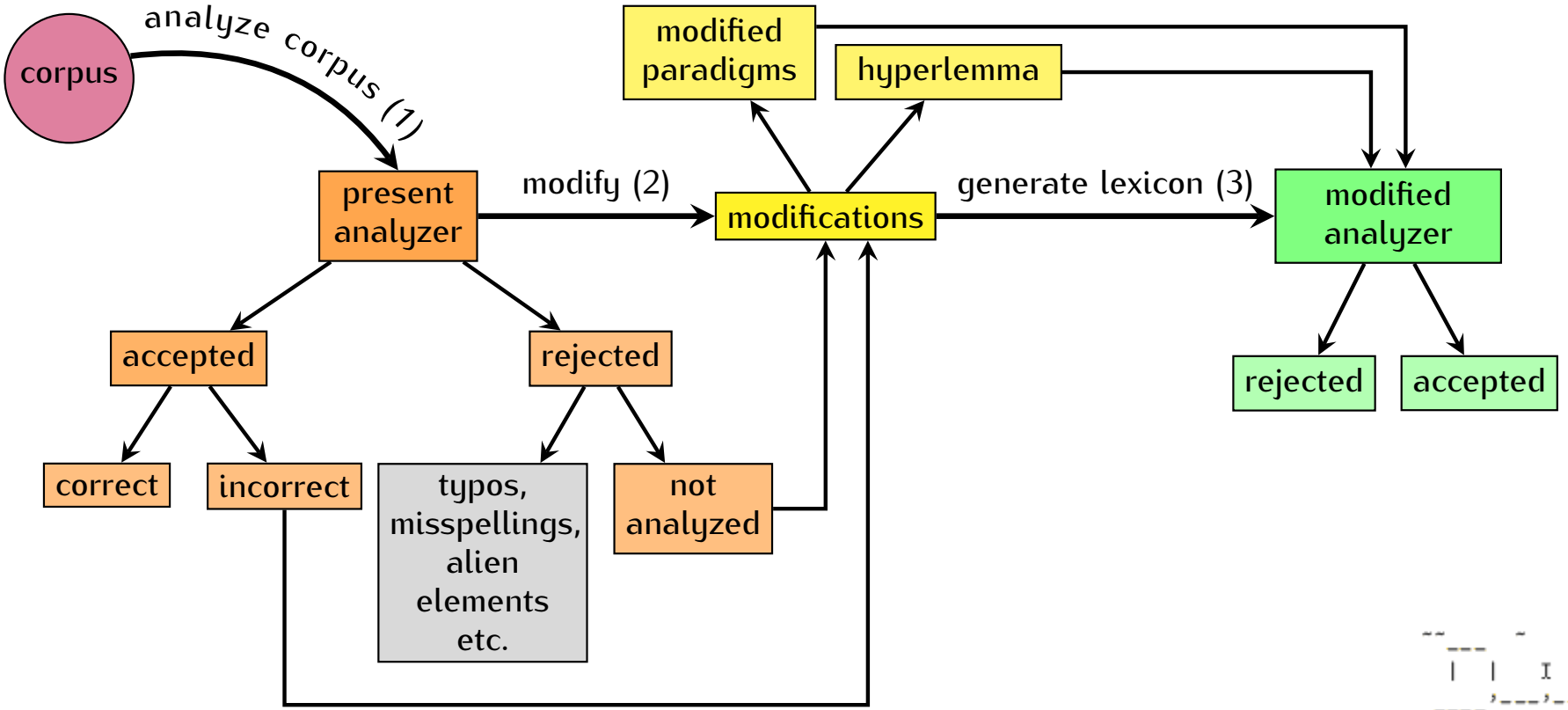
2013-2016

http://www.f19.uw.edu.pl

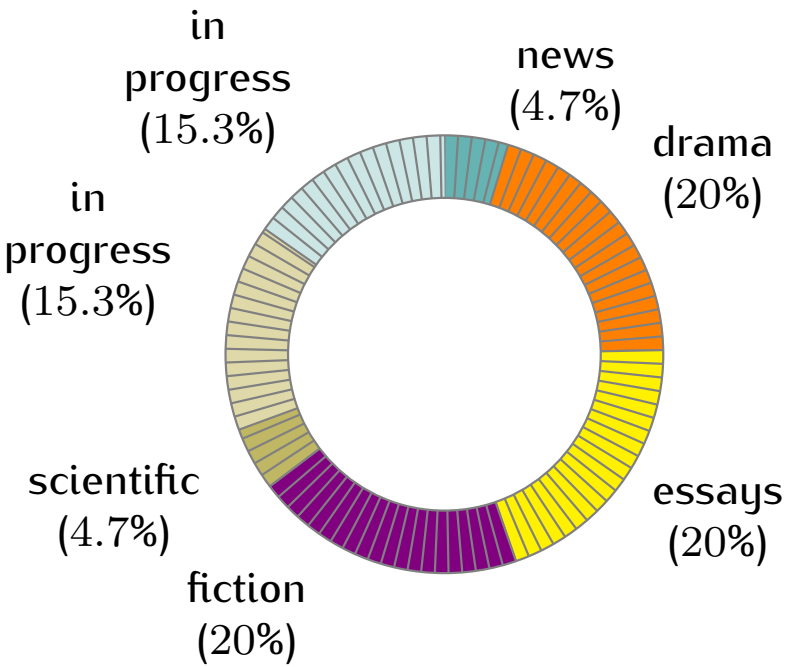
Motivation

- growing interest in NLP of old(er) Polish
- availability of adaptable tools
- extensive, detailed inflectional description of Polish
- computer-readable text from the period

A diachronic morphological analyzer in three steps



The corpus



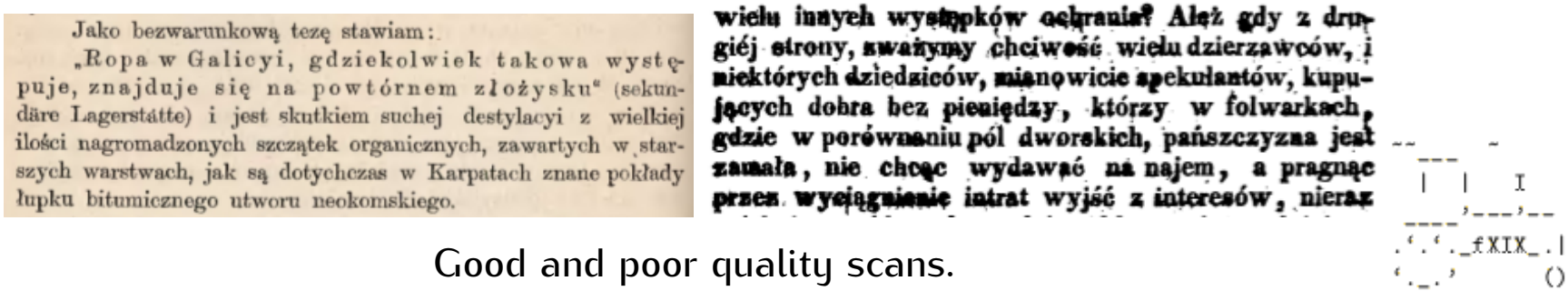
Samples collected per style

- 5 stylistic subcorpora, 200,000 segments each
- 5–13 samples/year (at least one sample of each style per year)
- more than 500 authors (inc. 119 anonymous), 480 books, 135 newspapers & magazines, 54 digital libraries
- repository ownCloud
- downloadable at

`http://www.f19.uw.edu.pl/`

The good, the bad and the ugly: excerpting samples

- stroke of luck: a text in original spelling and inflection
- time machine: aging a text from modern edition with original spelling and forms (a trick for poor quality scans)
- everyday humdrum: OCR of the digital scans (Fine Reader, tesseract)
- bad luck: typing in a text (good text, no contemporary re-editions, poor scan)


Jako bezwarunkową tezę stawiam:

„Ropa w Galicyi, gdziekolwiek takowa występuje, znajduje się na powtórnem złożysku“ (sekundäre Lagerstätte) i jest skutkiem suchej destylacyi z wielkiej ilości nagromadzonych szczątek organicznych, zawartych w starszych warstwach, jak są dotychczas w Karpatach znane pokłady łupku bitumicznego utworu neokomskiego.

wielu innych występków ochrania? Ależ gdy z drugiéj strony, zważymy chciwość wielu dzierzawców, i niektórych dziedziców, mianowicie spekulantów, kupujących dobra bez pieniędzy, którzy w folwarkach, gdzie w porównaniu pól dworskich, pańszczyzna jest zamała, nie chcąc wydawać na najem, a pragnąc przez wyciągnienie intrat wyjść z interesów, nieraz

Good and poor quality scans.

A sample and metadata

Jako bezwarunkową tezę stawiam: „Ropa w Galicyi, gdziekolwiek takowa występuje, znajduje się na powtórnem złożysku" (sekundäre Lagerstatte) i jest skutkiem suchej destylacyi z wielkiej ilości nagromadzonych szczątek organicznych, zawartych w starszych warstwach, jak są dotychczas w Karpatach znane pokłady łupku bitumicznego utworu neokomskiego.

A sample: in yellow — 19 c. forms, underlined — alien elements.

```
autor: Olszewski, Stanisław
tytuł: Przyczynek do teoryi pochodzenia
i występowania nafty w Galicyi
data wydania: 1881
miejsce wydania: Lwów
redaktor: Radziszewski, Br.
tytuł książki:
tytuł gazety, czasopisma, serii wydaw-
niczej: Kosmos. Czasopismo Polskiego
Towarzystwa Przyrodników im. Kopernika
nr: R. 6
wydawnictwo: nakł. Polskiego Towarzystwa
Przyrodników im. Kopernika
numery stron: 524-527
styl: popularnonaukowy
źródło: Śląska Biblioteka Cyfrowa
link:      http://www.sbc.org.pl/dlibra/
docmetadata?id=20199
uwagi: nie ma w HINT, chemia?
```

A metadata file.

The dictionary

podbiegunowy

Hasło	Część mowy	Rodzaj/aspekt
podbieg	subst	m3
podbiegać	v	ndk
podbiegający	pact	ndk
podbieganie	ger	ndk
podbiegły	adj	
podbiegnąć	v	dk
podbiegnięcie	subst	n2
podbiegnięcie	ger	dk
podbiegunowość	osc	f
podbiegunowy	adj	
podbielać	v	ndk
podbielający	pact	ndk
podbielanie	ger	ndk
podbielany	ppas	ndk
podbielenie	ger	dk
podbielić	v	dk
podbielony	ppas	dk
podbieracz	subst	m1
podbierać	v	ndk

Wszystkie formy | Formy bazowe

przymiotnik [SJPDor.]
P4

	l. p.					l. m.			
	m1	m2	m3	n1,n2	ż	p1	m1		pozostałe
							ndepr	depr	
M./(W.)	podbiegunowy			podbiegunowe	podbiegunowa	podbiegunowi	podbiegunowe		
D.	podbiegunowego				podbiegunowej	podbiegunowych			
C.	podbiegunowemu				podbiegunowej	podbiegunowym			
B.	podbiegunowego	podbiegunowy		podbiegunowe	podbiegunową	podbiegunowych			podbiegunowe
N.	podbiegunowym				podbiegunową	podbiegunowymi			
Ms.	podbiegunowym				podbiegunowej	podbiegunowych			
Złoż.	podbiegunowo+								

Odsyłacze

nazwa cechy	podbiegunowość
przymiotnik „zanegowany"	niepodbiegunowy

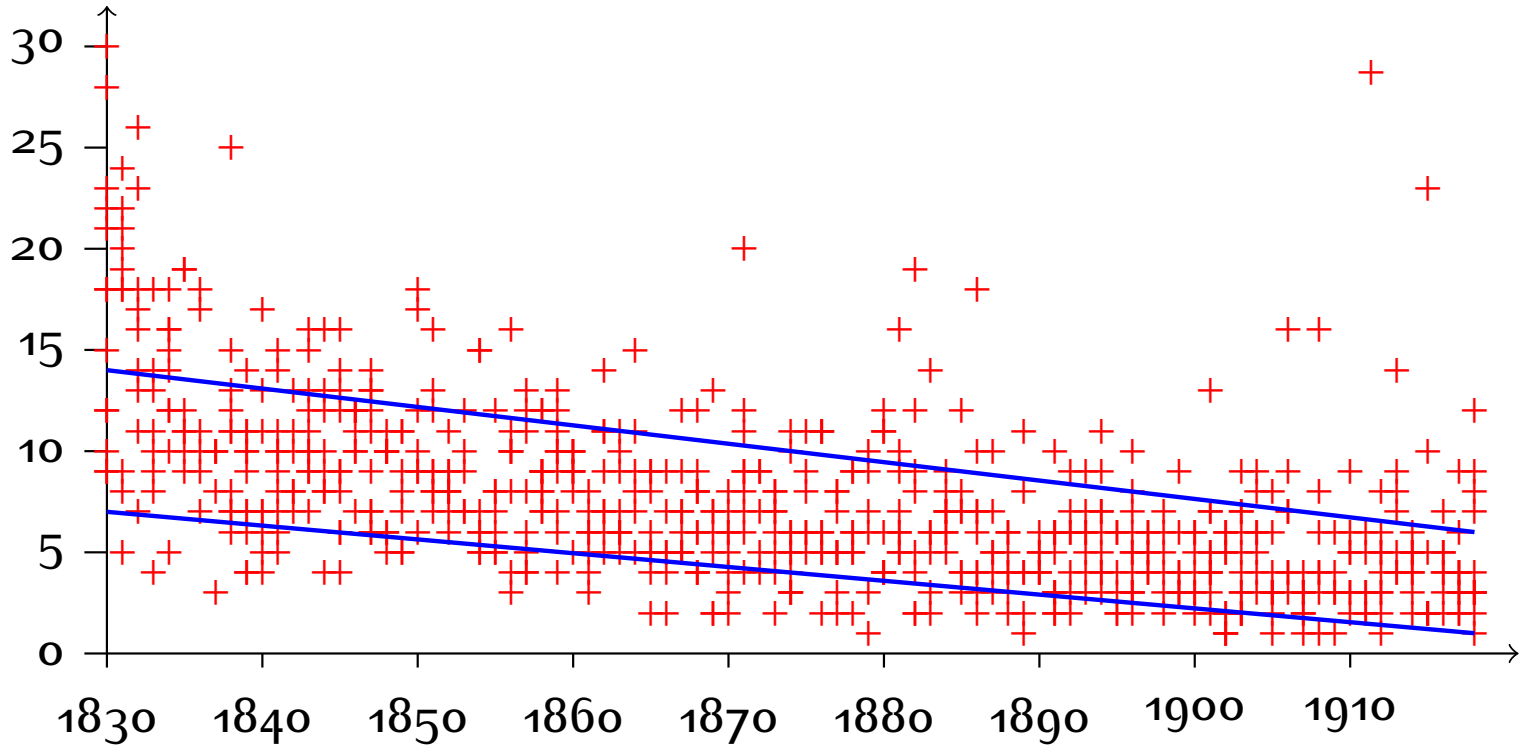
PODBIEGUNOWY ‘polar’ in *Grammatical dictionary of Polish*
(„Słownik gramatyczny języka polskiego”, Saloni et al. 2012)
http://www.sgjp.pl/kuznia/slownik/

The analyzer for modern Polish

```
Morfeusz analyzer, version: 1.9.0
Setting dictionary search path to: .
Using dictionary: sgjp (default)
Utwór iłów solnych odznacza się mniejszemi ilościami zazwyczaj cięższéj
[0,1,Utwór,utwór,subst:sg:acc:m3,pospolita,_
 0,1,Utwór,utwór,subst:sg:nom:m3,pospolita,_]
[1,2,iłów,ił:s1,subst:pl:gen:m3,pospolita,geol.
 1,2,iłów,ił:s2,subst:pl:gen:m2,pospolita,_
 1,2,iłów,ił:s2,subst:pl:gen:m3,pospolita,_]
[2,3,solnych,solny,adj:pl:acc:m1.p1:pos,_,_
 2,3,solnych,solny,adj:pl:gen:m1.m2.m3.f.n1.n2.p1.p2.p3:pos,_,_
 2,3,solnych,solny,adj:pl:loc:m1.m2.m3.f.n1.n2.p1.p2.p3:pos,_,_]
[3,4,odznacza,odznaczać,fin:sg:ter:imperf,_,_]
[4,5,się,się,qub,_,_]
[5,6,mniejszemi,mniejszemi,ign,_,_]
[6,7,ilościami,ilość,subst:pl:inst:f,pospolita,_]
[7,8,zazwyczaj,zazwyczaj,adv,_,_]
[8,9,cięższéj,cięższéj,ign,_,_]
```

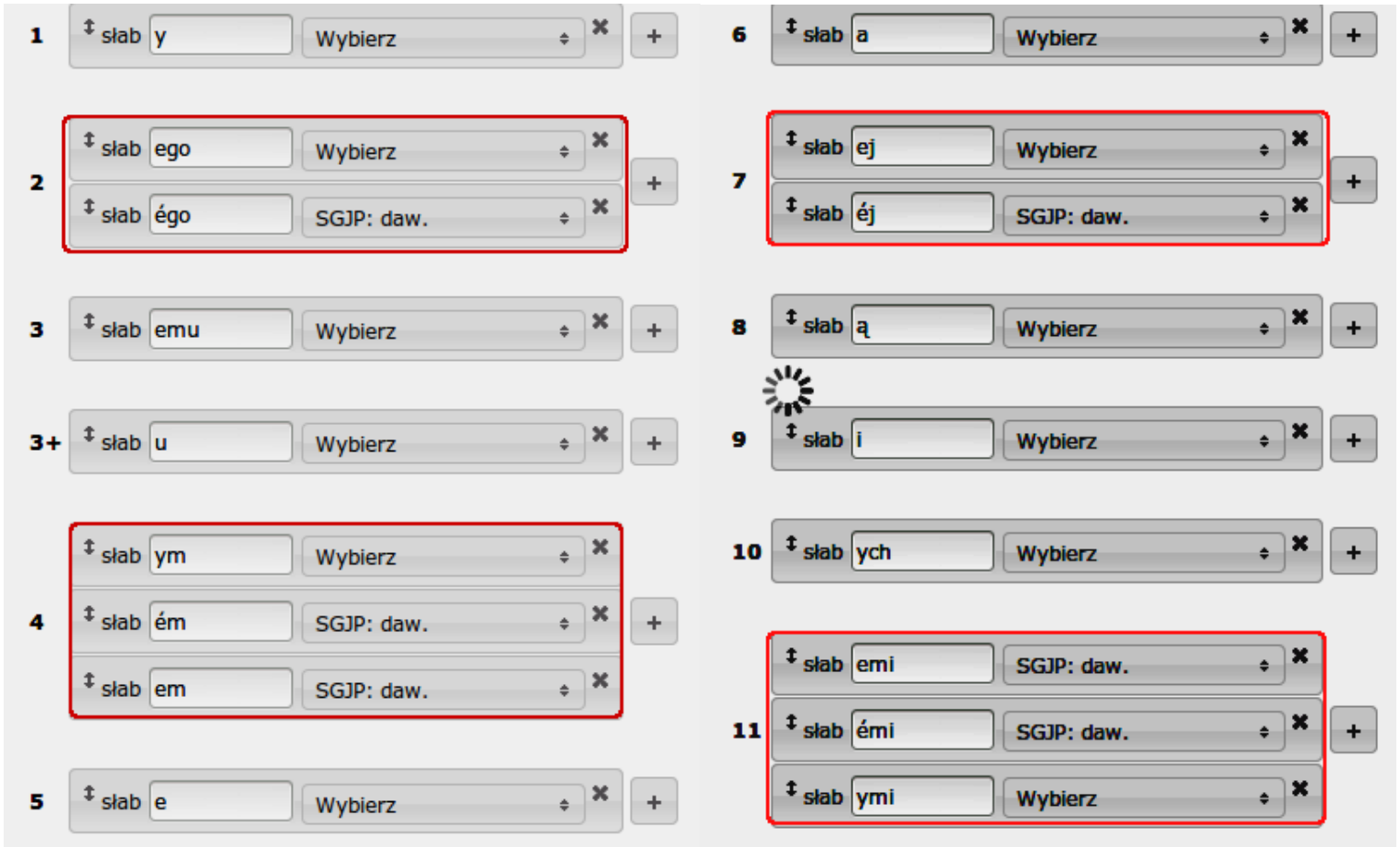
A string from a sample analyzed by *Morfeusz 2.0* (Woliński, 2014) with SGJP-based lexicon. Marked not analyzed 19c. forms.

The performance of non-modified analyzer on 1830-1918 corpus



Percent of unrecognized unique segments per file.

Modifications



Kuźnia (Smithy) editor

Days of future past: modified lexicon

1	podbiegunowy
2	podbiegunowego
3	podbiegunowemu
4	podbiegunowym
5	podbiegunowe
6	podbiegunowa
7	podbiegunowej
8	podbiegunową
9	podbiegunowi
10	podbiegunowych
11	podbiegunowymi
12	podbiegunowo+

2 G/A masc sg

4 I masc sg

7 D/C/L/V fem sg

11 I pl

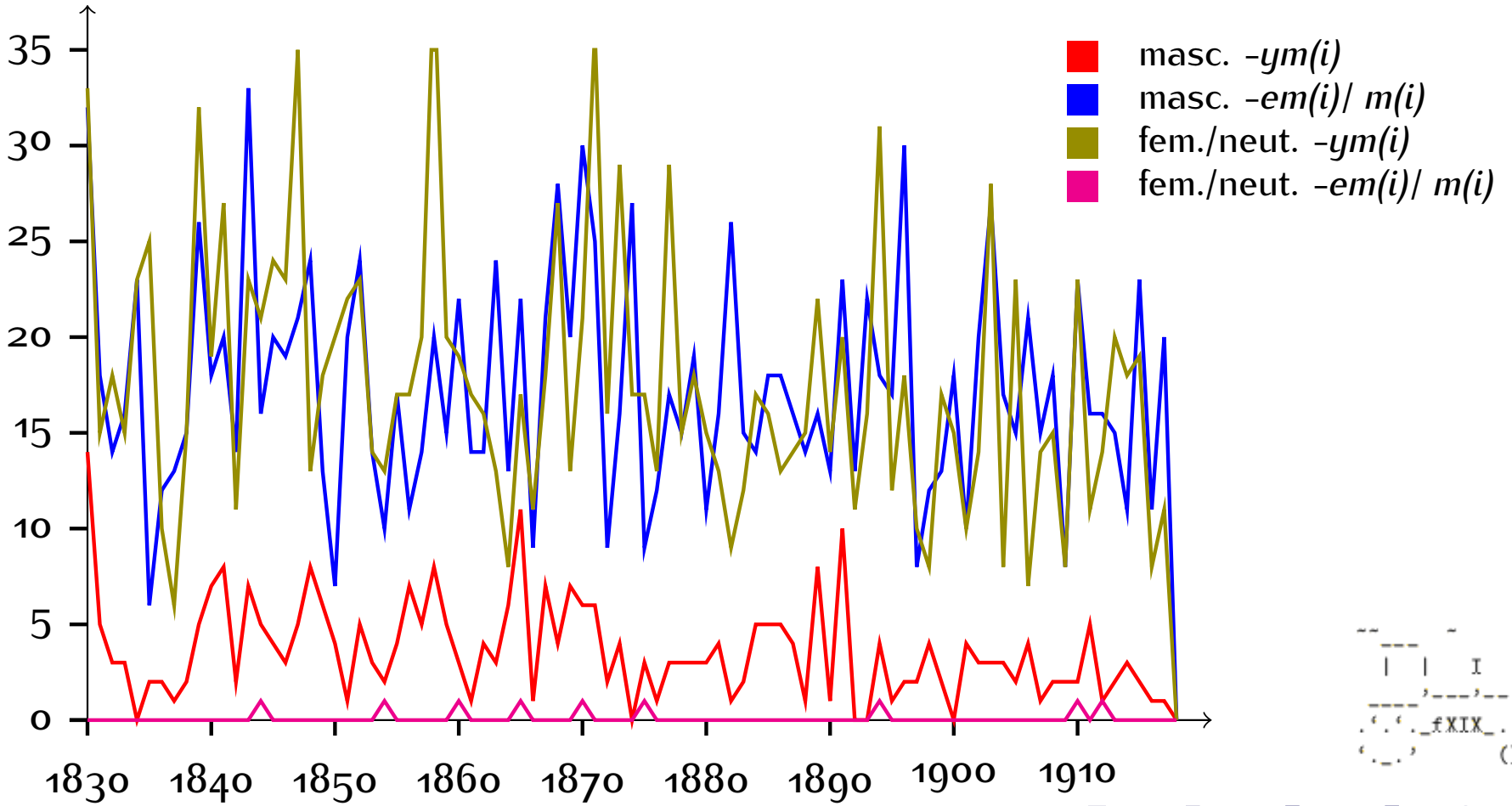
1	podbiegunowy
2	podbiegunowego podbiegunowégo *daw.*
3	podbiegunowemu
4	podbiegunowym podbiegunowém *daw.* podbiegunowem *daw.*
5	podbiegunowe
6	podbiegunowa
7	podbiegunowej podbiegunowéj *daw.*
8	podbiegunową
9	podbiegunowi
10	podbiegunowych
11	podbiegunowymi *daw.* podbiegunowemi *daw.* podbiegunowémi *daw.*
12	podbiegunowo+

Wavering instrumental

Kopczyński (1780) differentiate between instrumentalis of masculin and neuter:

	sg	*pl*
masc	ym	ymi
neut, fem	em	emi

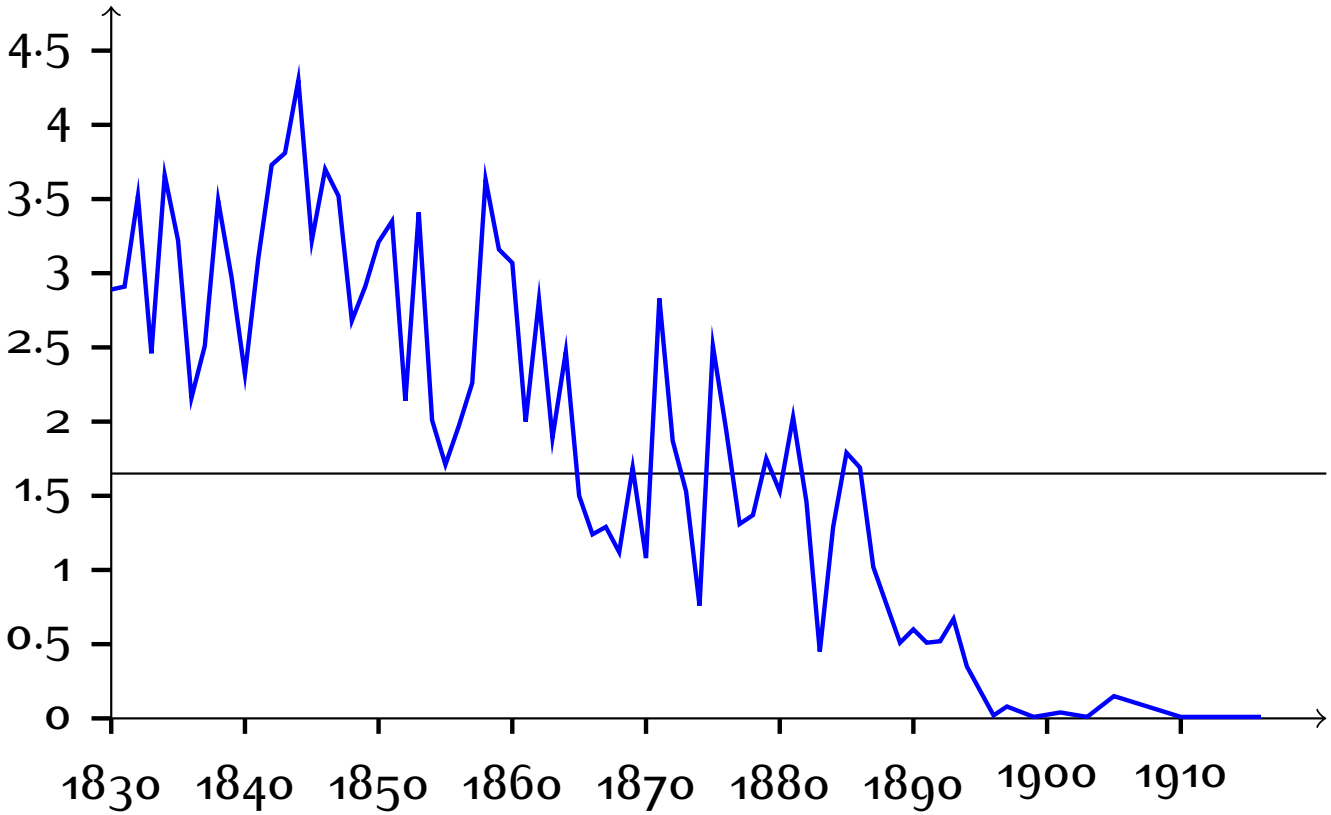
Distribution of instrumental endings across genders



Applications

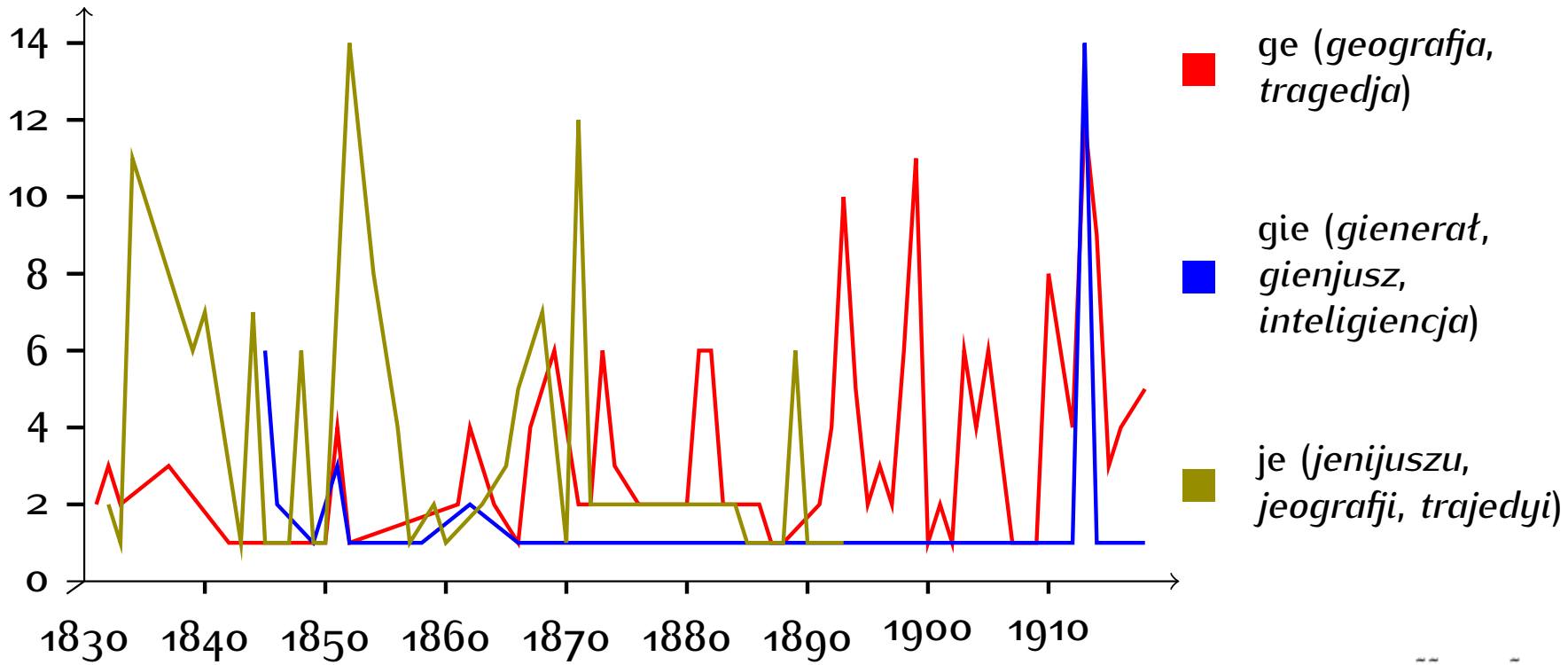
- part-of-speech tagging,
- tracing dynamics of language change (spelling and inflection)
- following adaptation stages of loanwords
- marking neologisms/disused/obsolete words with date

Example: dynamics of *é*



Percent of unique segments with per year.

Example: *ge*:*gie*:*je* alternations



Distribution of *je*:*gie*:*ge* spelling

References

Derwojedowa M., Kieraś W., Skowrońska D. i Wołosz R., *Współczesne narzędzia leksykograficzne a analiza tekstów dawniejszych*, „Polonica", t. XXXIV s. 21--27 2014a.

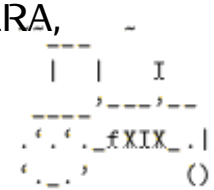
Derwojedowa M., Kieraś W., Skowrońska D. i Wołosz R., *Zasób leksykalny polszczyzny II poł. XIX wieku a możliwość automatycznej analizy morfologicznej tekstów z tego okresu*, [w:] *Leksyka języków słowiańskich w badaniach synchronicznych i diachronicznych*, red. M. Gębka-Wolak, J. Kamper-Warejko i A. Moroz, s. 183--195, 2014b.

Derwojedowa M., Skwrońska D. i Kieraś W., *Korpus polszczyzny XIX wieku – od mikrokorpusu do korpusu średniej wielkości*, „Prace Filologiczne", s. 249--254 2014c.

Kopczyński O., *Grammatyka dla szkół narodowych*, I wyd., Warszawa 1780.

Saloni Z., Woliński M., Wołosz R., Gruszczyński W. i Skowrońska D., *Słownik gramatyczny języka polskiego*, Ii wyd., Warszawa 2012, CD.

Woliński M., *Morfeusz Reloaded*, [w:] *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014*, red. N. Calzolari, K. K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk i S. Piperidis, s. 1106–1111, ELRA, Reykjavík, Iceland 2014, ISBN 978-2-9517408-8-4, URL http://www.lrec-conf.org/proceedings/lrec2014/index.html.



http://www.f19.uw.edu.pl/

funded by the Polish National Science Centre
grant DEC-2012/07/B/HS2/00570