Maciej Eder & Rafał L. Górski
Institute of Polish Language, Polish Academy Sciences
{maciejeder; rafalg}@ijp-pan.krakow.pl

The use of stylometry in historical linguistics: a case study in recent diachronic change in Polish

Question 1: Do the texts cluster according to the date of creation?
Question 2: Can we automatically identify a turning point between two (sub)periods in language?

1918 commonly accepted beginning of Late Modern Polish – a claim based on extralinguistic factors:

- Polish regains the status of an official language as well as the language of instruction
- Poles living in three different countries became citizens of one state
- WW I is a turning point in European history and culture

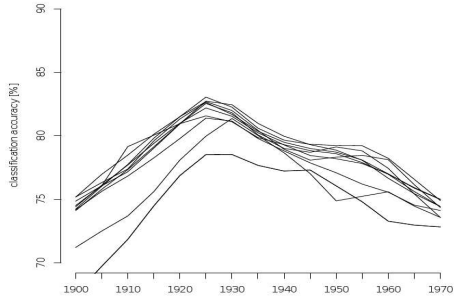
Hypothesis: these external factors influence the system of the language

A **corpus** of 76 novels ranging from 1828 to 2010. The texts were lemmatized and POS tagged

No more than two novels by one author

- only fiction
- as far as possible an author wasn't represented by early or late works in order to prevent the texts to cluster according to the author or genre

1. NSC for bigrams of POS tags



In several **experiments** we checked if the texts tend to cluster according to the time of their creation. Discriminators:

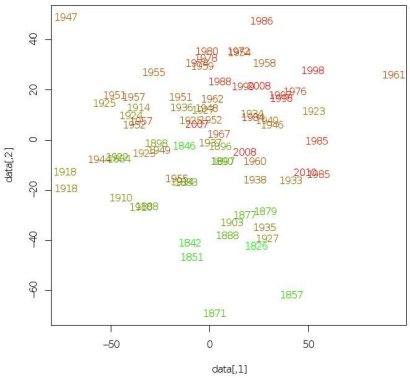
- words
- lemmata
- grammatical tags, which are an approximation of syntactic structure

all as uni-, bi- and trigrams, (one feature at a time)

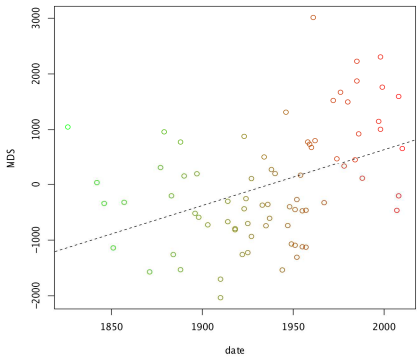
Clustering techniques: 1. Multidimensional scaling (MDS),
2. Nearest Shrunken Centroids (NSC).

Assumption: the corpus covers two subperiodes of Modern Polish. Where is the borderline between them? We move the borderline between 1900 and 1970 iteratively by 5 years. Texts are (via NSC) assigned to the „earlier" or the „recent" period. Which of the assesed borderlines gives best results of classification?

2. MDS for the texts of the corpus



3. MDS against the timeline (trigrams of lemmata)



Conclusion
The fig. 1 exhibits a peak about 1925. It can be interpreted as an effect of the changes which started in 1918
The results support claims of the historians of language that both 1918 and 1939/1945 were turning points in modern Polish
Some authors continue to write in an „old fashoned style" after 1918 (fig. 2 and 3)

Open questions
What is the character of the two shifts: systemic or stylistic?
Do other text types show similar shifts?

The research is supported by the National Science Centre grant DEC-2013/11/B/HS2/02795