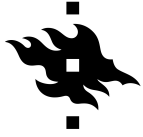
Computational approaches to semantic change detection

Day 5

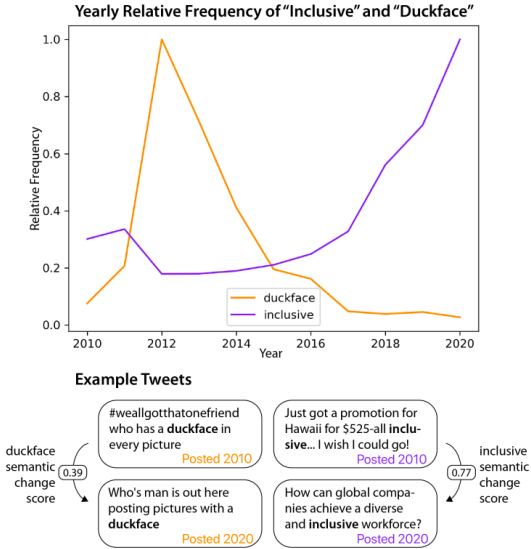
Part I: Applications

Andrey Kutuzov, Lidia Pivovarova

University of Oslo, University of Helsinki



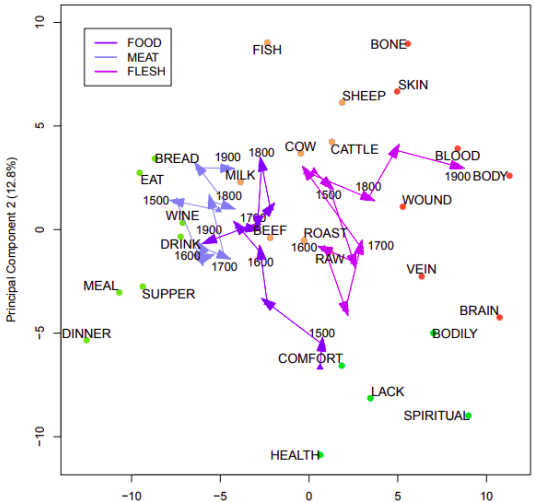
Language Studies



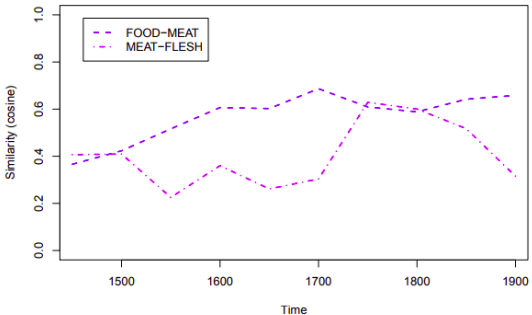
[Keidar et al., 2022]:

- Study of differences between slang and standard language
- Slang words tends to preserve their meaning over time but may have large variety in frequency

Language Studies

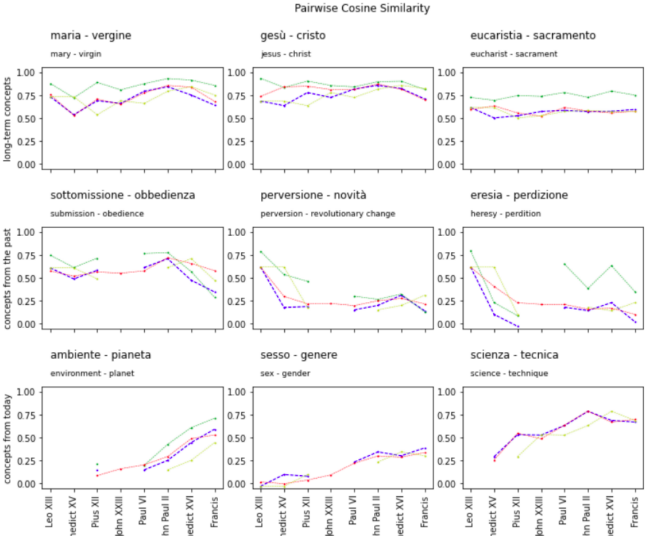


[Zimmermann, 2019]:
a chain of semantic change
FOOD -> MEAT -> FLESH



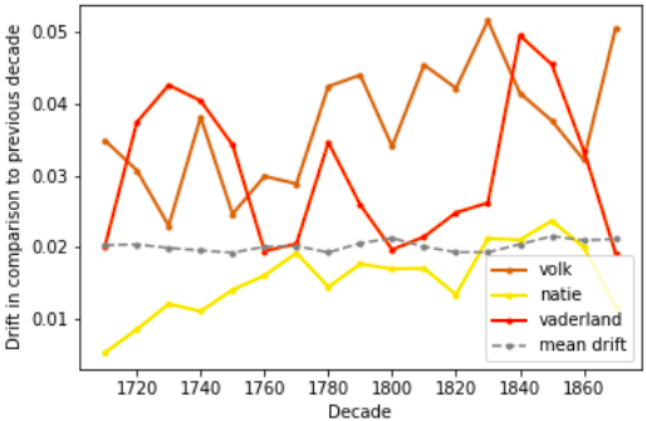
Historical Studies

[Castano et al., 2022]: Vatican publications study



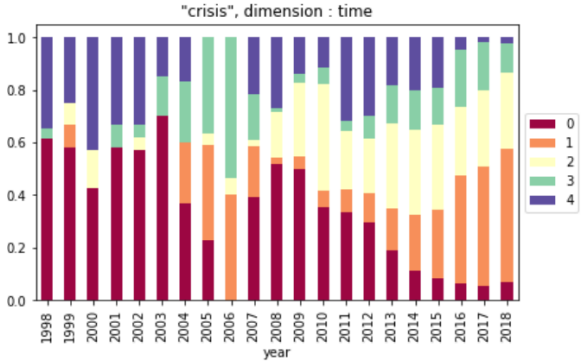
Historical Studies

[Timmermans et al., 2022]: nation building as presented in the 18th century Dutch fiction



Media Montiroting

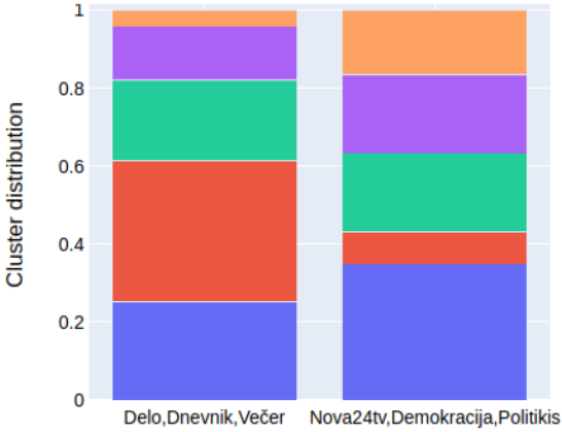
[Montariol et al., 2020]: tracking word usage change in financial domain



N^o	Keyword examples - Word = *crisis*
0	liquidity, funding, contingency, cash, collateral, outflows
1	marketing, business, management, design, advertising, media
2	european, debt, credit, sovereign, countries, eurozone, banks
3	financial, accident, capital, regulatory, loss, liquidity, funding
4	credit, financial, global, markets, debt, european, recession

Media Montiroting

[Martinc et al., 2021]: same techniques can be used to track difference across other dimensions, e.g. conservative vs. liberal media



#	Keywords
	partnership, homosexual relationship, polygamous relationship, intimate relationship, problem
	family relationship, end work relationship, woman, field, less
	power in the relationship, relationship between people, percentage, man and woman, story
	family relationship, marriage, children, Ljubljana, coalition for children
	woman, love relationship, work relationship, film, meet

Summary

- Computational approaches to semantic shift detection have a large potential for applications in
  - Linguistics
  - Digital humanities
  - Computational social science
  - Media monitoring

Summary

- Computational approaches to semantic shift detection have a large potential for applications in
  - Linguistics
  - Digital humanities
  - Computational social science
  - Media monitoring
- Interpretable methods are strogly preferred in downstream applications

Summary

- Computational approaches to semantic shift detection have a large potential for applications in
  - Linguistics
  - Digital humanities
  - Computational social science
  - Media monitoring
- Interpretable methods are strogly preferred in downstream applications
- Currently, most of this research are rather at the proof-of-concept stage

Laws of Semantic Change

- In addition to classifying *individual words* as either changing or non-changing it would be interesting to find *general regularities* of word meaning change
  - i.e. laws of semantic change

Laws of Semantic Change

- In addition to classifying *individual words* as either changing or non-changing it would be interesting to find *general regularities* of word meaning change
  - i.e. laws of semantic change
- This is a much more complex task, as explaining language phenomena is more difficult than studying specific use cases

Laws of Semantic Change

- In addition to classifying *individual words* as either changing or non-changing it would be interesting to find *general regularities* of word meaning change
  - i.e. laws of semantic change
- This is a much more complex task, as explaining language phenomena is more difficult than studying specific use cases
  - for example, usually earlier corpora are much smaller in size that ones collected in later periods, which skews many standard measures for change detection

Laws of Semantic Change

- In addition to classifying *individual words* as either changing or non-changing it would be interesting to find *general regularities* of word meaning change
  - i.e. laws of semantic change
- This is a much more complex task, as explaining language phenomena is more difficult than studying specific use cases
  - for example, usually earlier corpora are much smaller in size that ones collected in later periods, which skews many standard measures for change detection
- We discuss the **latest** papers on this topic, since it learns from previous studies and correct their drawbacks

Laws of Semantic Change

- In addition to classifying *individual words* as either changing or non-changing it would be interesting to find *general regularities* of word meaning change
  - i.e. laws of semantic change
- This is a much more complex task, as explaining language phenomena is more difficult than studying specific use cases
  - for example, usually earlier corpora are much smaller in size that ones collected in later periods, which skews many standard measures for change detection
- We discuss the **latest** papers on this topic, since it learns from previous studies and correct their drawbacks
  - If you interested, read the paper and then look at the previous work it is referring to

[Kawasaki et al., 2022]

Research Question

- Whether a tendency to change meaning correlates with
  - frequency,
  - length,
  - polysemousness?

[Kawasaki et al., 2022]

Research Question

- Whether a tendency to change meaning correlates with frequency, length and polysemousness?

Cognate Study

- Spanish, French, and Italian cognates, 794 in total

[Kawasaki et al., 2022]

Research Question

- Whether a tendency to change meaning correlates with frequency, length and polysemousness?

Cognate Study

- Spanish, French, and Italian cognates, 794 in total
  - Cognates *by definition* originated from the same (Latin) word
  - If cognates have different meanings at least one of them experienced a semantic change
    e.g. Latin LONGU -> French *long* (long) vs. Spansh *luengo* (erudite wording)
  - If cognates still share the same meaning, they likely remain unchanged

[Kawasaki et al., 2022]

Research Question

- Whether a tendency to change meaning correlates with frequency, length and polysemousness?

Cognate Study

- Spanish, French, and Italian cognates, 794 in total
  - Cognates *by definition* originated from the same (Latin) word
  - If cognates have different meanings at least one of them experienced a semantic change
    e.g. Latin LONGU -> French *long* (long) vs. Spansh *luengo* (erudite wording)
  - If cognates still share the same meaning, they likely remain unchanged

Main idea

- Regression analysis to predict semantic distance between cognates in a pair of languages from their frequency, length, polysemy, etc.

[Kawasaki et al., 2022]

Research Question

- Whether a tendency to change meaning correlates with frequency, length and polysemousness?

Main idea

- Regression analysis to predict semantic distance between cognates in a pair of languages from their frequency, length, polysemy, etc.
  - Semantic distance is measured as a **cosine similarity** between vectors that represent cognates in an **aligned** embedding space

[Kawasaki et al., 2022]

Research Question

- Whether a tendency to change meaning correlates with frequency, length and polysemousness?

Results

35% of variety in data can be explained by

- the polysemy of Latin etymon (*POLY*$_{lat}$),
- the length of Latin etymon (*LEN*$_{lat}$),
- an averaged frequency in Romance languages (*FREQ*$_{rom}$),
- and an averaged edit distance between Latin etymon and Romance cognates (*EDIT*).

	Coef.	SE	t	$p > \|t\|$
Intercept	0.00	0.03	0.00	1.00
FREQ_{lat}	-0.08	0.04	-1.82	0.07
POLY_{lat}	**0.10**	**0.04**	**2.28**	**0.02**
LEN_{lat}	**-0.21**	**0.03**	**-6.29**	**0.00**
FREQ_{rom}	**-0.54**	**0.03**	**-18.40**	**0.00**
NORM_{rom}	–	–	–	–
EDIT	**0.13**	**0.03**	**4.07**	**0.00**

[Kawasaki et al., 2022]

Research Question

- Whether a tendency to change meaning correlates with frequency, length and polysemousness?

Results

The negative correlation with the frequency of a Latin etymon (*FREQ$_{lat}$*) is not significant

	Coef.	SE	t	$p > \|t\|$
Intercept	0.00	0.03	0.00	1.00
FREQ$_{lat}$	-0.08	0.04	-1.82	0.07
POLY$_{lat}$	**0.10**	**0.04**	**2.28**	**0.02**
LEN$_{lat}$	**-0.21**	**0.03**	**-6.29**	**0.00**
FREQ$_{rom}$	**-0.54**	**0.03**	**-18.40**	**0.00**
NORM$_{rom}$	–	–	–	–
EDIT	**0.13**	**0.03**	**4.07**	**0.00**

[Kawasaki et al., 2022]

Research Question

- Whether a tendency to change meaning correlates with frequency, length and polysemousness?

Results

The negative correlation with the frequency of a Latin etymon (*FREQ*$_{lat}$) is not significant

- Most probably because length, polysemy and frequency are interdependent

	Coef.	SE	t	$p > \|t\|$
Intercept	0.00	0.03	0.00	1.00
FREQ$_{lat}$	-0.08	0.04	-1.82	0.07
POLY$_{lat}$	**0.10**	**0.04**	**2.28**	**0.02**
LEN$_{lat}$	**-0.21**	**0.03**	**-6.29**	**0.00**
FREQ$_{rom}$	**-0.54**	**0.03**	**-18.40**	**0.00**
NORM$_{rom}$	–	–	–	–
EDIT	**0.13**	**0.03**	**4.07**	**0.00**

[Kawasaki et al., 2022]

Research Question

- Whether a tendency to change meaning correlates with frequency, length and polysemousness?

Results

Edit distance indicates the difference between words that were in the language from the very beginning

CAPUT (head) -> *chef* (fr), *jefe* (es)

from those that were borrowed from the Medieval Latin

ANIMAL (animal) -> *animal* (fr), *animal* (es)

	Coef.	SE	t	$p > \|t\|$
Intercept	0.00	0.03	0.00	1.00
FREQ$_{lat}$	-0.08	0.04	-1.82	0.07
POLY$_{lat}$	**0.10**	**0.04**	**2.28**	**0.02**
LEN$_{lat}$	**-0.21**	**0.03**	**-6.29**	**0.00**
FREQ$_{rom}$	**-0.54**	**0.03**	**-18.40**	**0.00**
NORM$_{rom}$	–	–	–	–
EDIT	**0.13**	**0.03**	**4.07**	**0.00**

[Kawasaki et al., 2022]

Research Question

- Whether a tendency to change meaning correlates with frequency, length and polysemousness?

Results

Is it possible to predict meaning change from the state of word in the origin language only?

[Kawasaki et al., 2022]

Research Question

- Whether a tendency to change meaning correlates with frequency, length and polysemousness?

Results

Is it possible to predict meaning change from the state of word in the origin language only?

- A negative correlation with the frequency and length of a Latin etymon is statistically significant
- However, only 6% of variance is explained, which means other factors play a more important role

	Coef.	SE	t	$p > \|t\|$
Intercept	0.00	0.03	0.00	1.00
FREQ_{lat}	**-0.10**	**0.04**	**-2.74**	**0.01**
POLY_{lat}	–	–	–	–
LEN_{lat}	**-0.27**	**0.04**	**-7.10**	**0.00**

[Kawasaki et al., 2022]

Summary

- **frequency** is negatively correlated with semantic change

[Kawasaki et al., 2022]

Summary

- **frequency** is negatively correlated with semantic change
- **polysemy** is positively correlated with semantic change

[Kawasaki et al., 2022]

Summary

- **frequency** is negatively correlated with semantic change
- **polysemy** is positively correlated with semantic change
- **word length** is negatively correlated with semantic change

[Kawasaki et al., 2022]

Summary

- **frequency** is negatively correlated with semantic change
- **polysemy** is positively correlated with semantic change
- **word length** is negatively correlated with semantic change
- the **longer** word exist in a language, the more it is prone to change

[Kawasaki et al., 2022]

Summary

- **frequency** is negatively correlated with semantic change
- **polysemy** is positively correlated with semantic change
- **word length** is negatively correlated with semantic change
- the **longer** word exist in a language, the more it is prone to change
- effect is **stronger for verbs than for nouns** and adjectives:
  - **nouns** are more susceptible to extra-linguistic factors such as socio-cultural circumstances, technological advances, language contacts

[Kawasaki et al., 2022]

Summary

- **frequency** is negatively correlated with semantic change
- **polysemy** is positively correlated with semantic change
- **word length** is negatively correlated with semantic change
- the **longer** word exist in a language, the more it is prone to change
- effect is **stronger for verbs than for nouns** and adjectives:
  - **nouns** are more susceptible to extra-linguistic factors such as socio-cultural circumstances, technological advances, language contacts

- All this corresponds to previous studies, common sense and linguistic theory.

[Kawasaki et al., 2022]

Summary

- **frequency** is negatively correlated with semantic change
- **polysemy** is positively correlated with semantic change
- **word length** is negatively correlated with semantic change
- the **longer** word exist in a language, the more it is prone to change
- effect is **stronger for verbs than for nouns** and adjectives:
  - **nouns** are more susceptible to extra-linguistic factors such as socio-cultural circumstances, technological advances, language contacts

- All this corresponds to previous studies, common sense and linguistic theory.
- However, this is just one study on limited material

[Kawasaki et al., 2022]

Summary

- **frequency** is negatively correlated with semantic change
- **polysemy** is positively correlated with semantic change
- **word length** is negatively correlated with semantic change
- the **longer** word exist in a language, the more it is prone to change
- effect is **stronger for verbs than for nouns** and adjectives:
  - **nouns** are more susceptible to extra-linguistic factors such as socio-cultural circumstances, technological advances, language contacts

- All this corresponds to previous studies, common sense and linguistic theory.
- However, this is just one study on limited material
  ? Why did not they use Latin embeddings?

[Kawasaki et al., 2022]

Summary

- **frequency** is negatively correlated with semantic change
- **polysemy** is positively correlated with semantic change
- **word length** is negatively correlated with semantic change
- the **longer** word exist in a language, the more it is prone to change
- effect is **stronger for verbs than for nouns** and adjectives:
  - **nouns** are more susceptible to extra-linguistic factors such as socio-cultural circumstances, technological advances, language contacts

- All this corresponds to previous studies, common sense and linguistic theory.
- However, this is just one study on limited material
  ? Why did not they use Latin embeddings?
- Many factors remain unexplained: a lot of work for all of you!

References I

- Castano, S., Ferrara, A., Montanelli, S., Periti, F., et al. (2022).
  Semantic shift detection in Vatican publications: a case study from Leo XIII to Francis.
  In *CEUR WORKSHOP PROCEEDINGS*, volume 3194, pages 231–243. CEUR-WS.
- Kawasaki, Y., Salingre, M., Karpinska, M., Takamura, H., and Nagata, R. (2022).
  Revisiting statistical laws of semantic shift in Romance cognates.
  In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 141–151, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

References II

- Keidar, D., Opedal, A., Jin, Z., and Sachan, M. (2022).
  Slangvolution: A causal analysis of semantic change and frequency dynamics in slang.
  In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1422–1442, Dublin, Ireland. Association for Computational Linguistics.

- Martinc, M., Perger, N., Pelicon, A., Ulčar, M., Vezovnik, A., and Pollak, S. (2021).
  EMBEDDIA hackathon report: Automatic sentiment and viewpoint analysis of Slovenian news corpus on the topic of LGBTIQ+.
  In *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*, pages 121–126, Online. Association for Computational Linguistics.

References III

- Montariol, S., Allauzen, A., and Kitamoto, A. (2020).
  Variations in word usage for the financial domain.
  In *Proceedings of the Second Workshop on Financial Technology and Natural Language Processing*, pages 8–14, Kyoto, Japan. -.
- Timmermans, M., Vanmassenhove, E., and Shterionov, D. (2022).
  “vaderland”, “volk” and “natie”: Semantic change related to nationalism in Dutch literature between 1700 and 1880 captured with dynamic Bernoulli word embeddings.
  In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 125–130, Dublin, Ireland. Association for Computational Linguistics.

References IV

Zimmermann, R. (2019).
Studying semantic chain shifts with Word2Vec: FOOD>MEAT>FLESH.
In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 23–28, Florence, Italy. Association for Computational Linguistics.