

Computational approaches to semantic change detection

Day 1, Part II: Annotation schemas and datasets

Andrey Kutuzov, Lidia Pivovarova

University of Oslo, University of Helsinki



- 1 Manual annotation
- 2 Synthetic datasets
- 3 Regularities in semantic change

Manual annotation

- ▶ A variety of diachronic corpora is available for research;

Manual annotation

- ▶ A variety of diachronic corpora is available for research;

Popular English corpora for diachronic research

Corpus	Size, words	Reference
Helsinki Corpus	10^6	[Rissanen et al., 1993]
New-York Times Annotated Corpus	$\approx 2 \times 10^9$	[Sandhaus, 2008]
Google Books Ngrams	$\approx 100 \times 10^9$	[Michel et al., 2011]
English Gigaword	$\approx 4 \times 10^9$	[Parker et al., 2011]
Corpus of Historical American English (COHA)	400×10^6	[Davies, 2012]
Amazon Movie Reviews	$\approx 9 \times 10^8$	[McAuley and Leskovec, 2013]
Twitter (also in other languages)	$\approx 14 \times 10^9$	[Banda et al., 2023]

Manual annotation

- ▶ A variety of diachronic corpora is available for research;
- ▶ To develop computational methods we need at least testing data, **annotated** for the task at hand;

Manual annotation

- ▶ A variety of diachronic corpora is available for research;
- ▶ To develop computational methods we need at least testing data, **annotated** for the task at hand;
- ▶ **Meaning** is a property of a **word as a whole** and to detect word meaning change we need annotation on a level of **lexicon**;

Manual annotation

- ▶ A variety of diachronic corpora is available for research;
- ▶ To develop computational methods we need at least testing data, **annotated** for the task at hand;
- ▶ **Meaning** is a property of a **word as a whole** and to detect word meaning change we need annotation on a level of **lexicon**;
- ▶ More specifically, we need a list of words, there each word is marked as either changed or not (binary score) or associated with degree of change;

Manual annotation

- ▶ A variety of diachronic corpora is available for research;
- ▶ To develop computational methods we need at least testing data, **annotated** for the task at hand;
- ▶ **Meaning** is a property of a **word as a whole** and to detect word meaning change we need annotation on a level of **lexicon**;
- ▶ More specifically, we need a list of words, there each word is marked as either changed or not (binary score) or associated with degree of change;
- ▶ This is different from annotation for many other NLP tasks where specific **sentences** or **tokens** are annotated.

Manual annotation

- ▶ A variety of diachronic corpora is available for research;
- ▶ To develop computational methods we need at least testing data, **annotated** for the task at hand;
- ▶ **Meaning** is a property of a **word as a whole** and to detect word meaning change we need annotation on a level of **lexicon**;
- ▶ More specifically, we need a list of words, there each word is marked as either changed or not (binary score) or associated with degree of change;
- ▶ This is different from annotation for many other NLP tasks where specific **sentences** or **tokens** are annotated.

How to obtain such annotation?

Word-level annotations

[Gulordava and Baroni, 2011]

- ▶ 100 words annotated by 5 participants for semantic change between the 1960s and 1990s
- ▶ Four point annotation scale from *0: no change* to *3: significant change*, average score used as gold standard (GS)
- ▶ Inter-rater agreement = 0.51

Rank	Word	Rank	Word
1	program	100	sleep
2	domain	99	baseball
3	virtual	98	orange
4	address	97	street
5	disk	96	doubt

Word-level annotations

[Gulordava and Baroni, 2011]

- ▶ 100 words annotated by 5 participants for semantic change between the 1960s and 1990s
- ▶ Four point annotation scale from *0: no change* to *3: significant change*, average score used as gold standard (GS)
- ▶ Inter-rater agreement = 0.51

[Del Tredici et al., 2019]

- ▶ 100 words from the *r/LiverpoolFC* subreddit posts from 2011-2013 and 2017
- ▶ Annotated by forum participants, binary decision
- ▶ 26 participants, 8.8 annotations per word
- ▶ Participants saw few randomly chosen examples for each word
- ▶ Inter-rater agreement = 0.58

Sentence-level annotation

DURel: [Schlechtweg et al., 2018]

- Five native speakers of German annotated **use pairs** on the 4-point scale

Sentence-level annotation

DURel: [Schlechtweg et al., 2018]

- Five native speakers of German annotated **use pairs** on the 4-point scale

	A	B	C
1	target sentence 1	...	target sentence 2
303	Bemerkungen. Ein Donnerwetter in Paris ist mit so vielen Verdrieslichkeiten verknüpft, daß ichs hier anführen muß. Wir hatten heute Abends eins von 6. Uhr bis halb 11. Uhr des Nachts.		Der andre observirte schärfer mit dem Ausruf: „ Donnerwetter , sollte ich mich irren! Sie changirt nicht Farbe, und doch zuckte sie zusammen, als die Lupinus ihr was ins Ohr sagte.“

Sentence-level annotation

DURel: [Schlechtweg et al., 2018]

- ▶ Five native speakers of German annotated **use pairs** on the 4-point scale
- ▶ Some pairs were from the same time period; annotators did not know from which time period they are taken

Sentence-level annotation

DURel: [Schlechtweg et al., 2018]

- ▶ Five native speakers of German annotated **use pairs** on the 4-point scale
- ▶ Some pairs were from the same time period; annotators did not know from which time period they are taken
- ▶ 60 use pairs per word, 1320 use pairs for 22 target words in total:
 - ▶ 20 pairs in each group: *EARLIER* (1750-1800), *LATER* (1850-1900) and *COMPARE*

Sentence-level annotation

DURel: [Schlechtweg et al., 2018]

- ▶ Five native speakers of German annotated **use pairs** on the 4-point scale
- ▶ Some pairs were from the same time period; annotators did not know from which time period they are taken
- ▶ 60 use pairs per word, 1320 use pairs for 22 target words in total:
 - ▶ 20 pairs in each group: *EARLIER* (1750-1800), *LATER* (1850-1900) and *COMPARE*
- ▶ Degree of change:

Sentence-level annotation

DURel: [Schlechtweg et al., 2018]

- ▶ Five native speakers of German annotated **use pairs** on the 4-point scale
- ▶ Some pairs were from the same time period; annotators did not know from which time period they are taken
- ▶ 60 use pairs per word, 1320 use pairs for 22 target words in total:
 - ▶ 20 pairs in each group: *EARLIER* (1750-1800), *LATER* (1850-1900) and *COMPARE*
- ▶ Degree of change:
 - ▶ $\Delta LATER(w) = Mean_{EARLIER} - Mean_{LATER}$

Sentence-level annotation

DURel: [Schlechtweg et al., 2018]

- ▶ Five native speakers of German annotated **use pairs** on the 4-point scale
- ▶ Some pairs were from the same time period; annotators did not know from which time period they are taken
- ▶ 60 use pairs per word, 1320 use pairs for 22 target words in total:
 - ▶ 20 pairs in each group: *EARLIER* (1750-1800), *LATER* (1850-1900) and *COMPARE*
- ▶ Degree of change:
 - ▶ $\Delta LATER(w) = Mean_{EARLIER} - Mean_{LATER}$
—positive vs. negative values indicate innovative vs. reductive meaning change.

Sentence-level annotation

DURel: [Schlechtweg et al., 2018]

- ▶ Five native speakers of German annotated **use pairs** on the 4-point scale
- ▶ Some pairs were from the same time period; annotators did not know from which time period they are taken
- ▶ 60 use pairs per word, 1320 use pairs for 22 target words in total:
 - ▶ 20 pairs in each group: *EARLIER* (1750-1800), *LATER* (1850-1900) and *COMPARE*
- ▶ Degree of change:
 - ▶ $\Delta LATER(w) = Mean_{EARLIER} - Mean_{LATER}$
 - ▶ $COMPARE(w) = Mean_{COMPARE}(w)$

Sentence-level annotation

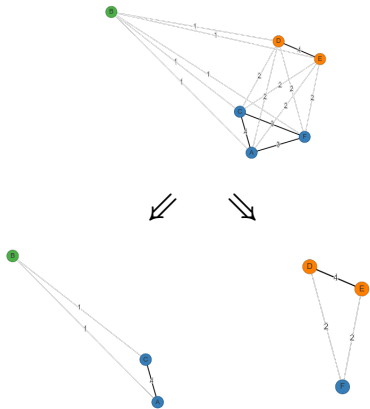
DURel: [Schlechtweg et al., 2018]

- ▶ Five native speakers of German annotated **use pairs** on the 4-point scale
- ▶ Some pairs were from the same time period; annotators did not know from which time period they are taken
- ▶ 60 use pairs per word, 1320 use pairs for 22 target words in total:
 - ▶ 20 pairs in each group: *EARLIER* (1750-1800), *LATER* (1850-1900) and *COMPARE*
- ▶ Degree of change:
 - ▶ $\Delta\text{LATER}(w) = \text{Mean}_{\text{EARLIER}} - \text{Mean}_{\text{LATER}}$
 - ▶ $\text{COMPARE}(w) = \text{Mean}_{\text{COMPARE}}(w)$

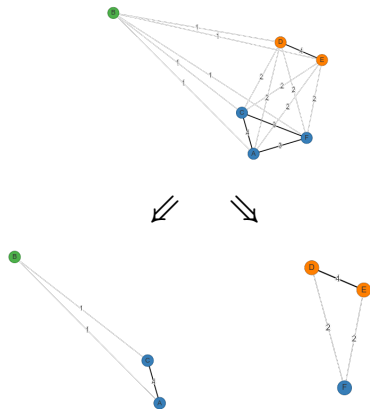
Let's try it!

<https://durel.ims.uni-stuttgart.de/>

Diachronic Word Usage Graphs



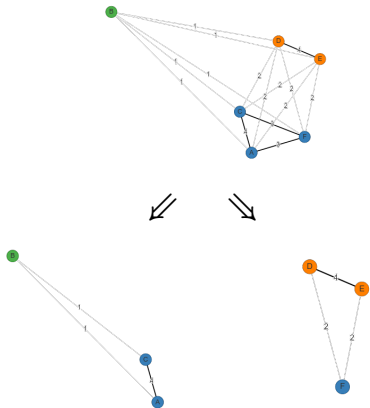
Diachronic Word Usage Graphs



[Schlechtweg et al., 2021]

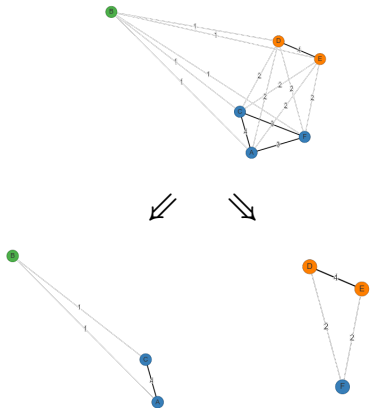
1. Randomly sample pairs of sentences for annotation
2. Annotate on the 1-4 scale
3. Cluster the graph
4. Sample additional edges to connect smaller cluster to bigger
5. Repeat 2-4 until the point each cluster has been compared to each other cluster
6. Degree of change:
 - ▶ Binary: appearing/disappearing cluster
 - ▶ Graded: difference between cluster distributions

Diachronic Word Usage Graphs



- ▶ This procedure was used in the SemEval 2020 Task 1 Shared Task, which standardized and established benchmarks for semantic change detection [Schlechtweg et al., 2020]
- ▶ Later used in annotation for other datasets [Zamora-Reina et al., 2022, Kutuzov et al., 2022]

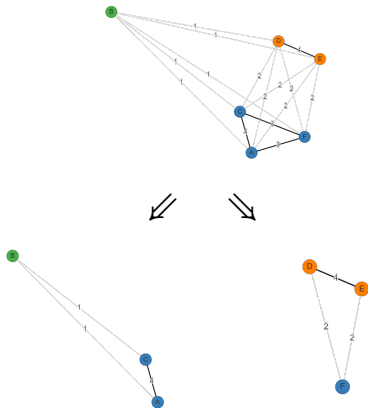
Diachronic Word Usage Graphs



But

- ▶ Requires additional efforts in annotating pairs from the same time period
- ▶ Sampling and clustering hyperparameters may affect the results
- ▶ Many datasets only use annotations of pairs from different time periods (*COMPARE* score) to reduce annotation efforts

Diachronic Word Usage Graphs



But2

- ▶ In some cases it is hard for annotators to judge whether a word used in the same sense in a pair of sentences
- ▶ Connection between sentence clusters and word **senses** is not straightforward
- ▶ Instead of using sentence pairs, some datasets annotated by explicitly linking sentences to senses from dictionaries [Schlechtweg et al., 2020] (for Latin), [Basile et al., 2020]

(Almost) all existing datasets

Data

	EN	DE	LA	SW	IT	NO-1	NO-2	RU-1	RU-2	RU-3	ES	SI
Period 1	1810-1860	1800-1899	-200-0	1790-1830	1945-1970	1929-1965	1980-1990	1700-1916	1918-1990	1700-1916	1810-1906	1997
Period 2	1960-2010	1946-1990	0-2000	1895-1903	1990-2014	1970-2013	2012-2019	1918-1990	1992-2016	1992-2016	1994-2020	2018
Tokens (mln)	7+7	70+72	2+9	71+110	52+197	57+175	43+649	93+122	122+107	93+107	13+22	70+80
Targets	37	48	40	32	18	80	80	99	99	99	80	104
Graded	✓	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓
Binary	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	✓	✗
Reference	[Schlechtweg et al., 2020]				[Basile et al., 2020]	[Kutuzov et al., 2022]		[Kutuzov and Pivovarova, 2021]			[Zamora-Reina et al., 2022]	[Martinc et al., 2022]

(Almost) all existing datasets

Data

	EN	DE	LA	SW	IT	NO-1	NO-2	RU-1	RU-2	RU-3	ES	SI
Period 1	1810-1860	1800-1899	-200-0	1790-1830	1945-1970	1929-1965	1980-1990	1700-1916	1918-1990	1700-1916	1810-1906	1997
Period 2	1960-2010	1946-1990	0-2000	1895-1903	1990-2014	1970-2013	2012-2019	1918-1990	1992-2016	1992-2016	1994-2020	2018
Tokens (mln)	7+7	70+72	2+9	71+110	52+197	57+175	43+649	93+122	122+107	93+107	13+22	70+80
Targets	37	48	40	32	18	80	80	99	99	99	80	104
Graded	✓	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓
Binary	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	✓	✗
Reference	[Schlechtweg et al., 2020]				[Basile et al., 2020]	[Kutuzov et al., 2022]		[Kutuzov and Pivovarova, 2021]			[Zamora-Reina et al., 2022]	[Martinc et al., 2022]

- A variety of datasets to develop **generalizable** methods for semantic change detection

(Almost) all existing datasets

Data

	EN	DE	LA	SW	IT	NO-1	NO-2	RU-1	RU-2	RU-3	ES	SI
Period 1	1810-1860	1800-1899	-200-0	1790-1830	1945-1970	1929-1965	1980-1990	1700-1916	1918-1990	1700-1916	1810-1906	1997
Period 2	1960-2010	1946-1990	0-2000	1895-1903	1990-2014	1970-2013	2012-2019	1918-1990	1992-2016	1992-2016	1994-2020	2018
Tokens (mln)	7+7	70+72	2+9	71+110	52+197	57+175	43+649	93+122	122+107	93+107	13+22	70+80
Targets	37	48	40	32	18	80	80	99	99	99	80	104
Graded	✓	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓
Binary	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	✓	✗
Reference	[Schlechtweg et al., 2020]				[Basile et al., 2020]	[Kutuzov et al., 2022]		[Kutuzov and Pivovarova, 2021]			[Zamora-Reina et al., 2022]	[Martinc et al., 2022]

- ▶ A variety of datasets to develop **generalizable** methods for semantic change detection
- ▶ BUT:
 - ▶ Only Indo-European languages from three language families
 - ▶ "Only" few dozen words for each language, not all phenomena presented
 - ▶ Large variety among corpora that eclipses variety among languages
 - ▶ Only two time periods for most languages

(Almost) all existing datasets

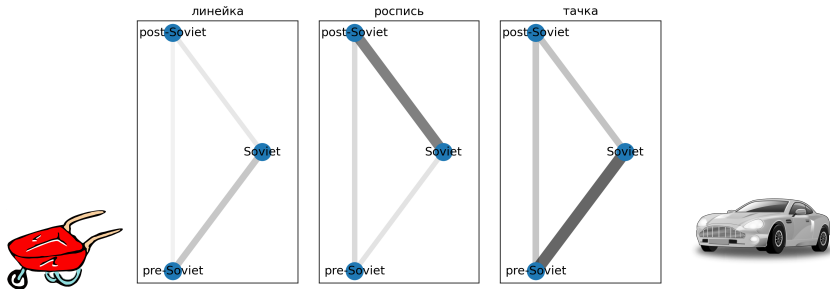
Data

	EN	DE	LA	SW	IT	NO-1	NO-2	RU-1	RU-2	RU-3	ES	SI
Period 1	1810-1860	1800-1899	-200-0	1790-1830	1945-1970	1929-1965	1980-1990	1700-1916	1918-1990	1700-1916	1810–1906	1997
Period 2	1960-2010	1946-1990	0-2000	1895-1903	1990-2014	1970-2013	2012-2019	1918-1990	1992-2016	1992-2016	1994–2020	2018
Tokens (mln)	7+7	70+72	2+9	71+110	52+197	57+175	43+649	93+122	122+107	93+107	13+22	70+80
Targets	37	48	40	32	18	80	80	99	99	99	80	104
Graded	✓	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓
Binary	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	✓	✗
Reference	[Schlechtweg et al., 2020]				[Basile et al., 2020]	[Kutuzov et al., 2022]		[Kutuzov and Pivovarova, 2021]			[Zamora-Reina et al., 2022]	[Martinc et al., 2022]

- ▶ A variety of datasets to develop **generalizable** methods for semantic change detection
- ▶ BUT:
 - ▶ Only Indo-European languages from three language families
 - ▶ "Only" few dozen words for each language, not all phenomena presented
 - ▶ Large variety among corpora that eclipses variety among languages
 - ▶ Only two time periods for most languages

Diachronic trajectory types in RuShiftEval

1. **changes in every period pair**, all relatedness scores are low: линейка ('carriage/ruler/series of goods')
2. **change in the Soviet period VS the pre-Soviet period**: роспись ('list/painting')
3. **change in the post-Soviet period VS the Soviet period**: тачка ('wheelbarrow/car')
4. (trivial) **no changes**: all three relatedness scores are high.
5. (not found) **change in the Soviet period then coming back to the original meaning**



Time relatedness graphs. **Nodes**: time periods; **edge width**: relatedness scores.

- 1 Manual annotation
- 2 Synthetic datasets**
- 3 Regularities in semantic change

Synthetic datasets

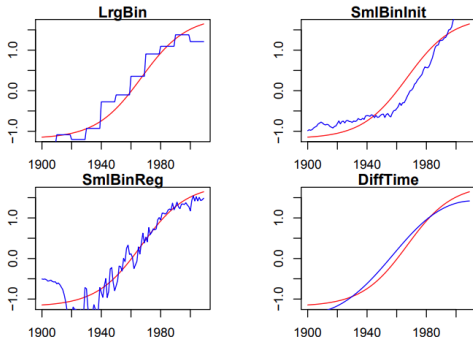
Pseudo-words

- ▶ [Rosenfeld and Erk, 2018]: Mentions of *word1* and *word2* are replaced with synthetic word *word1* ◦ *word2*
 - ▶ lobster ◦ banana
 - ▶ contexts for *word1* and *word2* are over- or under-sampled according to *sigmoid* change rule

Synthetic datasets

Pseudo-words

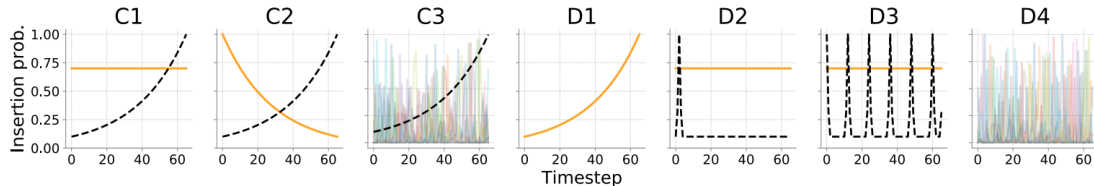
- ▶ [Rosenfeld and Erk, 2018]: Mentions of *word1* and *word2* are replaced with synthetic word *word1* ◦ *word2*
 - ▶ lobster ◦ banana
- ▶ contexts for *word1* and *word2* are over- or under-sampled according to *sigmoid* change rule



Synthetic datasets

Pseudo-words

- ▶ [Rosenfeld and Erk, 2018]: Mentions of *word1* and *word2* are replaced with synthetic word *word1* \circ *word2*
 - ▶ lobster \circ banana
 - ▶ contexts for *word1* and *word2* are over- or under-sampled according to *sigmoid* change rule
- ▶ [Shoemark et al., 2019]: A larger variety of sampling patterns to emulate semantic change (C), and distinguish it from mere frequency change (D)



Synthetic datasets

Pseudo-words

- ▶ [Rosenfeld and Erk, 2018]: Mentions of *word1* and *word2* are replaced with synthetic word *word1* ◦ *word2*
 - ▶ lobster ◦ banana
 - ▶ contexts for *word1* and *word2* are over- or under-sampled according to *sigmoid* change rule
- ▶ [Shoemark et al., 2019]: A larger variety of sampling patterns to emulate semantic change, and distinguish it from mere frequency change

Sense-annotated corpus

- ▶ Instead of merging unrelated words, it is possible to use sense-annotated corpus for under/oversampling of certain senses *for the same word*
- ▶ This approach has been used in monolingual [Schlechtweg and Im Walde, 2020] and multilingual [Montariol and Allauzen, 2021] setting.

Synthetic datasets

- ▶ Synthetic datasets allows to evaluate LSCD in controlled setting, without any annotation efforts

Synthetic datasets

- ▶ Synthetic datasets allows to evaluate LSCD in controlled setting, without any annotation efforts
- ▶ **BUT:**
 - ▶ Artificial is artificial: much variety from data is removed
 - ▶ E.g. when all 'time periods' are sampled from a modern corpus, only target words are changing their meaning, while context words remain unchanged

Synthetic datasets

- ▶ Synthetic datasets allows to evaluate LSCD in controlled setting, without any annotation efforts
- ▶ **BUT:**
 - ▶ Artificial is artificial: much variety from data is removed
 - ▶ E.g. when all 'time periods' are sampled from a modern corpus, only target words are changing their meaning, while context words remain unchanged
 - ▶ It is necessary to retrain all models in multiple copies of the corpus
 - ▶ In the era of LLMs and huge datasets this may be unfeasible

Synthetic datasets

- ▶ Synthetic datasets allows to evaluate LSCD in controlled setting, without any annotation efforts
- ▶ **BUT:**
 - ▶ Artificial is artificial: much variety from data is removed
 - ▶ E.g. when all 'time periods' are sampled from a modern corpus, only target words are changing their meaning, while context words remain unchanged
 - ▶ It is necessary to retrain all models in multiple copies of the corpus
 - ▶ In the era of LLMs and huge datasets this may be unfeasible
 - ▶ Resources necessary for more subtle strategies, i.e. sense-annotated dictionaries or WordNets, exist only for major languages

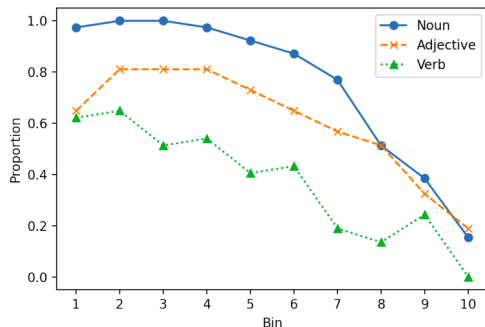
- 1 Manual annotation
- 2 Synthetic datasets
- 3 Regularities in semantic change**

Regularities in semantic change

- ▶ It is possible to exploit existing linguistic knowledge: historical dictionaries, known cases of semantic shift, established regularities
- ▶ Cognate studies [Uban et al., 2021, Kawasaki et al., 2022]
 - ▶ Cognates are *by definition* words that originated from the same ancestor
 - ▶ Cognates could be either *translations* of each other or *false friends*, in which case at least one of them undergone semantic change

Regularities in semantic change

- ▶ It is possible to exploit existing linguistic knowledge: historical dictionaries, known cases of semantic shift, established regularities
- ▶ Cognate studies [Uban et al., 2021, Kawasaki et al., 2022]
 - ▶ Cognates are *by definition* words that originated from the same ancestor
 - ▶ Cognates could be either *translations* of each other or *false friends*, in which case at least one of them undergone semantic change
 - ▶ This can be used as a sanity check




[Kawasaki et al., 2022]: A proportion of translations among different bins of automatically determined semantic similarity among Spanish-French cognates (1 - most semantically similar, 10 - most dissimilar)

What's next


- ▶ During the next three lectures we will go through various methods for semantic change detection
- ▶ This will include **hands on** exercises in a form of **Jupyter Notebooks**
- ▶ To participate in hand-on sessions you should:
 1. clone our repository https://github.com/lmphcs/semshift_esslli2023
 2. install all required packages and make sure you can run Python notebooks
- ▶ It's ok to skip exercises and listen for the lecture part only
- ▶ In any case, we won't have time to setup the environment in the class

References I

 Banda, J. M., Tekumalla, R., Wang, G., Yu, J., Liu, T., Ding, Y., Artemova, K., Tutubalina, E., and Chowell, G. (2023).

A large-scale COVID-19 Twitter chatter dataset for open scientific research - an international collaboration.

This dataset will be updated bi-weekly at least with additional tweets, look at the github repo for these updates. Release: We have standardized the name of the resource to match our pre-print manuscript and to not have to update it every week.

 Basile, P., Caputo, A., Caselli, T., Cassotti, P., and Varvara, R. (2020).

Diacr-ita@ evalita2020: Overview of the evalita2020 diachronic lexical semantics (diacr-ita) task.

Evaluation Campaign of Natural Language Processing and Speech Tools for Italian.

References II



Davies, M. (2012).

Expanding horizons in historical linguistics with the 400-million word corpus of historical american english.

Corpora, 7(2):121–157.



Del Tredici, M., Fernández, R., and Boleda, G. (2019).

Short-term meaning shift: A distributional exploration.


In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2069–2075, Minneapolis, Minnesota. Association for Computational Linguistics.

References III

 Gulordava, K. and Baroni, M. (2011).

A distributional similarity approach to the detection of semantic change in the Google Books ngram corpus.




In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 67–71, Edinburgh, UK. Association for Computational Linguistics.

 Kawasaki, Y., Salinger, 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 IV


-  Kutuzov, A. and Pivovarova, L. (2021).
Three-part diachronic semantic change dataset for Russian.
In *Proceedings of the 2nd International Workshop on Computational Approaches to Historical Language Change 2021*, pages 7–13, Online. Association for Computational Linguistics.
-  Kutuzov, A., Touileb, S., Mæhlum, P., Enstad, T., and Wittemann, A. (2022).
NorDiaChange: Diachronic semantic change dataset for Norwegian.
In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2563–2572, Marseille, France. European Language Resources Association.
-  Martinc, M., Dobrovoljc, K., and Pollak, S. (2022).
Semantic change detection datasets for Slovenian 1.0.
Slovenian language resource repository CLARIN.SI.

References V

 McAuley, J. J. and Leskovec, J. (2013).

From amateurs to connoisseurs: Modeling the evolution of user expertise through online reviews.

In *Proceedings of the 22Nd International Conference on World Wide Web, WWW '13*, pages 897–908. ACM.

 Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. A., and Aiden, E. L. (2011).

Quantitative analysis of culture using millions of digitized books.


Science, 331(6014):176–182.

References VI

 Montariol, S. and Allauzen, A. (2021).


Measure and evaluation of semantic divergence across two languages.

In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1247–1258, Online. Association for Computational Linguistics.

 Parker, R., Graff, D., Kong, J., Chen, K., and Maeda, K. (2011).

English Gigaword Fifth Edition LDC2011T07.

Technical report, Technical Report. Linguistic Data Consortium, Philadelphia.

 Rissanen, M. et al. (1993).

The Helsinki corpus of English texts.

Kyttö et. al, pages 73–81.

References VII

 Rosenfeld, A. and Erk, K. (2018).


Deep neural models of semantic shift.

In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 474–484, New Orleans, Louisiana. Association for Computational Linguistics.

 Sandhaus, E. (2008).

The New York Times annotated corpus overview.



Linguistic Data Consortium, Philadelphia, 6(12):e26752.

 Schlechtweg, D. and Im Walde, S. S. (2020).

Simulating lexical semantic change from sense-annotated data.

In *The Evolution of Language: Proceedings of the 13th International Conference (Evolang13)*, page 393.

References VIII

-  Schlechtweg, D., McGillivray, B., Hengchen, S., Dubossarsky, H., and Tahmasebi, N. (2020).
SemEval-2020 task 1: Unsupervised lexical semantic change detection.
In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1–23, Barcelona (online). International Committee for Computational Linguistics.
-  Schlechtweg, D., Schulte im Walde, S., and Eckmann, S. (2018).
Diachronic usage relatedness (DURel): A framework for the annotation of lexical semantic change.
In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 169–174, New Orleans, Louisiana. Association for Computational Linguistics.

References IX



Schlechtweg, D., Tahmasebi, N., Hengchen, S., Dubossarsky, H., and McGillivray, B. (2021).

DWUG: A large resource of diachronic word usage graphs in four languages.

In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7079–7091, Online and Punta Cana, Dominican Republic.

Association for Computational Linguistics.





Shoemark, P., Liza, F. F., Nguyen, D., Hale, S., and McGillivray, B. (2019).

Room to Glo: A systematic comparison of semantic change detection approaches with word embeddings.

In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 66–76, Hong Kong, China.

Association for Computational Linguistics.

References X

-  Uban, A. S., Cristea, A. M., Dinu, A., Dinu, L. P., Georgescu, S., and Zoicas, L. (2021). Tracking semantic change in cognate sets for English and Romance languages. In *Proceedings of the 2nd International Workshop on Computational Approaches to Historical Language Change 2021*, pages 64–74, Online. Association for Computational Linguistics.
-  Zamora-Reina, F. D., Bravo-Marquez, F., and Schlechtweg, D. (2022). LSCDiscovery: A shared task on semantic change discovery and detection in Spanish. In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 149–164, Dublin, Ireland. Association for Computational Linguistics.