Computational approaches to semantic change detection
Day 1, Part I: Semantic change detection as an NLP task

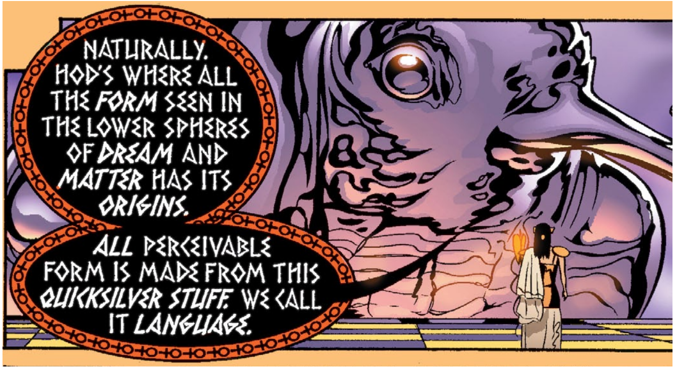
Andrey Kutuzov, Lidia Pivovarova

University of Oslo, University of Helsinki

ESSLLI'2023

1. Diachronic language change
2. Automated semantic change detection for linguists
3. NLP perspective
4. Temporal degradation of language models
5. Empirical turn

Diachronic language change


NATURALLY. HOD'S WHERE ALL THE FORM SEEN IN THE LOWER SPHERES OF DREAM AND MATTER HAS ITS ORIGINS.
ALL PERCEIVABLE FORM IS MADE FROM THIS QUICKSILVER STUFF. WE CALL IT LANGUAGE.

‘All perceivable form is made of this quicksilver stuff.
We call it language’.

Thoth, as retold by Alan Moore

Diachronic language change

- Human language is in continuous evolution [Hock and Joseph, 2009]
- New word senses arise over time
- Existing senses can change or disappear over time
- Semantic relations between words change as well.

This is a result of social and cultural dynamics or technological advances, etc.

Diachronic language change

There are many aspects and types of semantic change: a venerable field in linguistics.

Diachronic language change

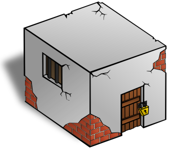
There are many aspects and types of semantic change: a venerable field in linguistics.

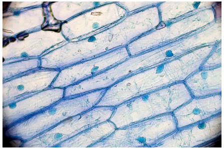
- A ‘semantic shift’ occurs when a word changes its senses: by acquiring a new sense, losing an existing one, or both. [Bloomfield, 1933]
- Sense is, roughly, an entry in a dictionary.
- Senses of a word form its ‘conventional meaning’.

Diachronic language change

There are many aspects and types of semantic change: a venerable field in linguistics.

- A 'semantic shift' occurs when a word changes its senses: by acquiring a new sense, losing an existing one, or both. [Bloomfield, 1933]
- Sense is, roughly, an entry in a dictionary.
- Senses of a word form its 'conventional meaning'.



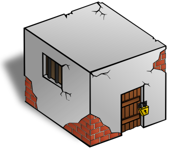


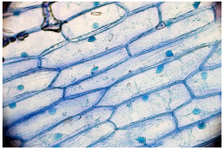
- '*Cell*' in English acquires new senses and starts to appear in new contexts: a semantic shift proper occurs.

Diachronic language change

There are many aspects and types of semantic change: a venerable field in linguistics.

- A 'semantic shift' occurs when a word changes its senses: by acquiring a new sense, losing an existing one, or both. [Bloomfield, 1933]
- Sense is, roughly, an entry in a dictionary.
- Senses of a word form its 'conventional meaning'.





- '*Cell*' in English acquires new senses and starts to appear in new contexts: a semantic shift proper occurs.
- But semantic change is not limited to discrete lexicographic senses.

Diachronic language change

Semantic proximity continuum

1. Homonymy:
   - ‘His *bark* was worse than his bite’
   - ‘He scratched the *bark* of the oak’
2. Polysemy:
   - ‘She submitted her *paper* to a journal’
   - ‘The report was printed on a piece of white *paper*’
3. Context variance:
   - ‘Careful *distancing* of blocks allow natural and controlled lighting for inner spaces’
   - ‘Self-quarantine and self-isolation are specific forms of social *distancing* in the period of COVID-19’
4. Identity:
   - ‘The *crankshaft* rotates within the engine block through use of main bearings’
   - ‘Casting is today mostly used for *crankshafts* in cheaper, lower performance engines’

[Blank, 1999] and others

Diachronic language change

Semantic proximity continuum

1. Homonymy:
   - ‘His *bark* was worse than his bite’
   - ‘He scratched the *bark* of the oak’
2. Polysemy:
   - ‘She submitted her *paper* to a journal’
   - ‘The report was printed on a piece of white *paper*’
3. Context variance:
   - ‘Careful *distancing* of blocks allow natural and controlled lighting for inner spaces’
   - ‘Self-quarantine and self-isolation are specific forms of social *distancing* in the period of COVID-19’
4. Identity:
   - ‘The *crankshaft* rotates within the engine block through use of main bearings’
   - ‘Casting is today mostly used for *crankshafts* in cheaper, lower performance engines’

[Blank, 1999] and others

- All these distinctions are gradual [Kilgarriff, 1997]
- Context variance is a semantic phenomenon as well:
  - contextual meaning can change without acquiring a new lexicographic sense
  - connotations / world knowledge / typical associations

1. Diachronic language change
2. Automated semantic change detection for linguists
3. NLP perspective
4. Temporal degradation of language models
5. Empirical turn

Automated semantic change detection for linguists

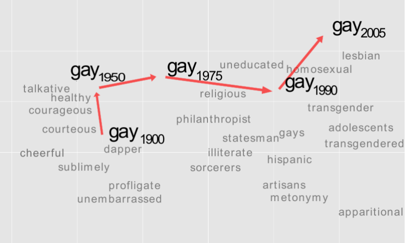
- Qualitative linguistic studies are powerful, but inherently limited:
  1. in the amount of texts to be analyzed in detail
  2. in the number of words that one can focus on simultaneously.

Automated semantic change detection for linguists

- Qualitative linguistic studies are powerful, but inherently limited:
  1. in the amount of texts to be analyzed in detail
  2. in the number of words that one can focus on simultaneously.
- With the appearance of large-scale diachronic corpora in digital format and efficient computational tools –> quantitative studies of semantic change became possible.
- One can now study semantic change across languages and different genres, over times, and in specific contexts.

Automated semantic change detection for linguists

- Qualitative linguistic studies are powerful, but inherently limited:
  1. in the amount of texts to be analyzed in detail
  2. in the number of words that one can focus on simultaneously.
- With the appearance of large-scale diachronic corpora in digital format and efficient computational tools –> quantitative studies of semantic change became possible.
- One can now study semantic change across languages and different genres, over times, and in specific contexts.



[Hamilton et al., 2016]

Automated semantic change detection for linguists

Computational data-driven methods can capture semantic change phenomena

- Historical linguists and lexicographers are naturally interested in lexical semantic change modeling...

Automated semantic change detection for linguists

Computational data-driven methods can capture semantic change phenomena

- Historical linguists and lexicographers are naturally interested in lexical semantic change modeling...
- ...or, rather in semantic change discovery:
- finding what words have changed their meaning in a given historical corpus, given the full vocabulary.

Automated semantic change detection for linguists

Computational data-driven methods can capture semantic change phenomena

- Historical linguists and lexicographers are naturally interested in lexical semantic change modeling...
- ...or, rather in semantic change discovery:
- finding what words have changed their meaning in a given historical corpus, given the full vocabulary.
- NLP still more often deals with semantic change detection (LSCD):
- analyzing a given set of target words.

Automated semantic change detection for linguists

Computational data-driven methods can capture semantic change phenomena

- Historical linguists and lexicographers are naturally interested in lexical semantic change modeling...
- ...or, rather in semantic change discovery:
- finding what words have changed their meaning in a given historical corpus, given the full vocabulary.
- NLP still more often deals with semantic change detection (LSCD):
- analyzing a given set of target words.

We will outline semantic change modeling in NLP, its methods and their relations to linguistic phenomena.
After the course, you will be able to start your own research in the area.

1. Diachronic language change
2. Automated semantic change detection for linguists
3. NLP perspective
4. Temporal degradation of language models
5. Empirical turn

NLP perspective

NLP practitioners are interested in diachronic perspective of semantics for various reasons:

NLP perspective

NLP practitioners are interested in diachronic perspective of semantics for various reasons:

- Mostly, it’s related to word senses:
  - NLP tasks like WSD, WSI, semantic similarity, etc are naturally extended to the diachronic setup

NLP perspective

NLP practitioners are interested in diachronic perspective of semantics for various reasons:

- Mostly, it’s related to word senses:
  - NLP tasks like WSD, WSI, semantic similarity, etc are naturally extended to the diachronic setup
- This includes studying various aspects of diachronic sense changes:
  - onomasiological VS semasiological
  - type of change (amelioration, metaphorization, etc)
  - source of change (technological, social, etc)

But LSCD is not limited to that.

NLP perspective

Additional topics of interest

- Discovering laws of semantic change [Hamilton et al., 2016, Dubossarsky et al., 2017]
- Constructing, testing and improving psycholinguistic and sociolinguistic theories of meaning change [Xu and Kemp, 2015, Goel et al., 2016, Noble et al., 2021]
- Surveying how the meaning of words has evolved historically [Garg et al., 2018, Kozlowski et al., 2019]
- Analyzing how meaning is currently transforming in public discourse [Azarbonyad et al., 2017, Del Tredici et al., 2019].

NLP perspective

Additional topics of interest

- Discovering laws of semantic change [Hamilton et al., 2016, Dubossarsky et al., 2017]
- Constructing, testing and improving psycholinguistic and sociolinguistic theories of meaning change [Xu and Kemp, 2015, Goel et al., 2016, Noble et al., 2021]
- Surveying how the meaning of words has evolved historically [Garg et al., 2018, Kozlowski et al., 2019]
- Analyzing how meaning is currently transforming in public discourse [Azarbonyad et al., 2017, Del Tredici et al., 2019].

And many more (some surveys: [Kutuzov et al., 2018, Tahmasebi et al., 2021])

1. Diachronic language change
2. Automated semantic change detection for linguists
3. NLP perspective
4. Temporal degradation of language models
5. Empirical turn

Temporal degradation of language models

More applied work in LSCD recently

- Large language models (LLMs) are dominating the NLP landscape now.
- But as language evolves, the LLMs training corpora become outdated and the models themselves degrade.
- How to achieve temporal generalization with LLMs is an open question [Lazaridou et al., 2021]
- But at least one needs to detect language evolution.

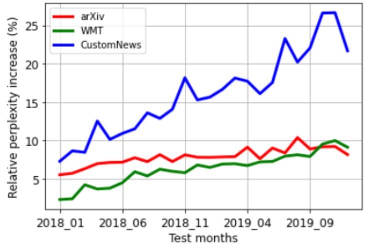


Figure 1: Relative ppl. increase of TIME-STRATIFIED over CONTROL, across test months.

Temporal degradation of language models

More applied work in LSCD recently

- Large language models (LLMs) are dominating the NLP landscape now.
- But as language evolves, the LLMs training corpora become outdated and the models themselves degrade.
- How to achieve temporal generalization with LLMs is an open question [Lazaridou et al., 2021]
- But at least one needs to detect language evolution.

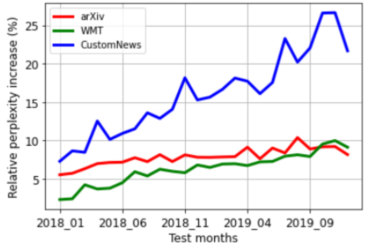


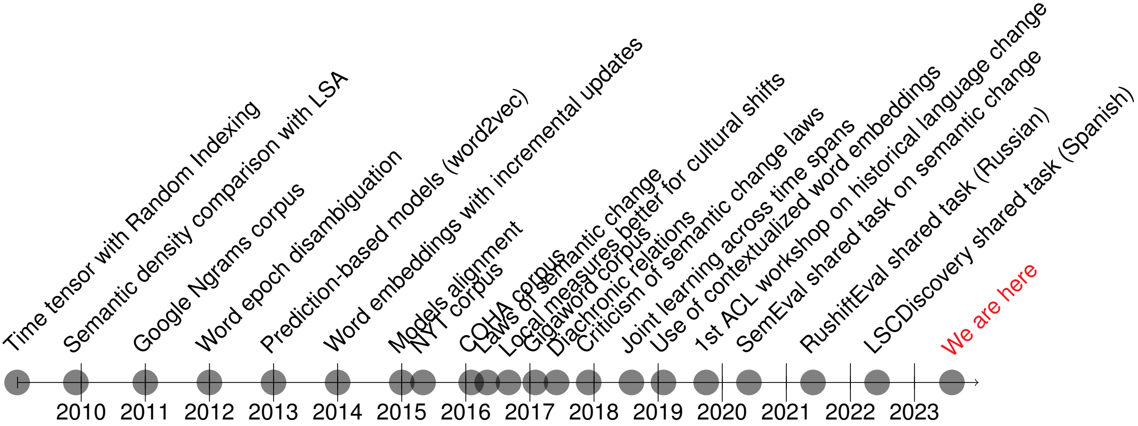
Figure 1: Relative ppl. increase of TIME-STRATIFIED over CONTROL, across test months.

See The First Workshop on Ever Evolving NLP (EvoNLP) [Barbieri et al., 2022].

1. Diachronic language change
2. Automated semantic change detection for linguists
3. NLP perspective
4. Temporal degradation of language models
5. Empirical turn

Empirical turn

(Approximate) research timeline for lexical semantic change modeling in NLP:



Time tensor with Random Indexing

Semantic density comparison with LSA

Google Ngrams corpus

Word epoch disambiguation

Prediction-based models (word2vec)

Word embeddings with incremental updates

Models alignment

NYT corpus

COHA corpus

Laws of semantic change

Local measures better for cultural shifts

Gigaword corpus

Diachronic relations

Criticism of semantic change laws

Joint learning across time spans

Use of contextualized word embeddings

1st ACL workshop on historical language change

SemEval shared task on semantic change

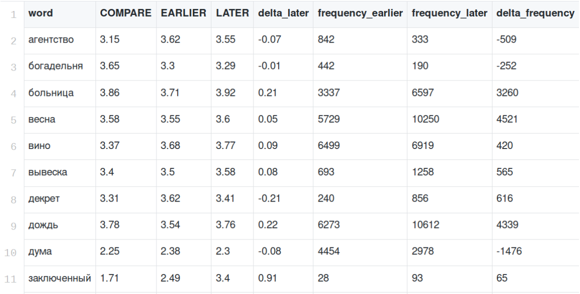
RushiftEval shared task (Russian)

LSCDiscovery shared task (Spanish)

We are here

2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021 2022 2023

Empirical turn



word	COMPARE	EARLIER	LATER	delta_later	frequency_earlier	frequency_later	delta_frequency
агентство	3.15	3.62	3.55	-0.07	842	333	-509
богадельня	3.65	3.3	3.29	-0.01	442	190	-252
больница	3.86	3.71	3.92	0.21	3337	6597	3260
весна	3.58	3.55	3.6	0.05	5729	10250	4521
вино	3.37	3.68	3.77	0.09	6499	6919	420
вывеска	3.4	3.5	3.58	0.08	693	1258	565
декрет	3.31	3.62	3.41	-0.21	240	856	616
дождь	3.78	3.54	3.76	0.22	6273	10612	4339
дума	2.25	2.38	2.3	-0.08	4454	2978	-1476
заключенный	1.71	2.49	3.4	0.91	28	93	65

[Kutuzov and Pivovarova, 2021]

- NLP is an empirical and data-driven science
  - much more empirical than (traditional) linguistics
- Hence, its reliance on well-defined datasets, benchmarks, objective system comparisons within shared tasks, etc.
- The next part of today’s lecture covers these resources.

References I

- Azarbonyad, H., Dehghani, M., Beelen, K., Arkut, A., Marx, M., and Kamps, J. (2017). Words are malleable: Computing semantic shifts in political and media discourse. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM ’17, pages 1509–1518. ACM.

- Barbieri, F., Camacho-Collados, J., Dhingra, B., Espinosa-Anke, L., Gribovskaya, E., Lazaridou, A., Loureiro, D., and Neves, L., editors (2022). *Proceedings of the The First Workshop on Ever Evolving NLP (EvoNLP)*, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

- Blank, A. (1999). Why do new meanings occur? a cognitive typology of the motivations for lexical semantic change. In *Historical Semantics and Cognition*, pages 61–90, Berlin/New York. Mouton de Gruyter.

References II

Bloomfield, L. (1933).
Language.
Allen & Unwin.

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

Dubossarsky, H., Weinshall, D., and Grossman, E. (2017).
Outta control: Laws of semantic change and inherent biases in word representation models.
In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1136–1145, Copenhagen, Denmark. Association for Computational Linguistics.

Garg, N., Schiebinger, L., Jurafsky, D., and Zou, J. (2018).
Word embeddings quantify 100 years of gender and ethnic stereotypes.
Proceedings of the National Academy of Sciences, 115(16):E3635–E3644.

References IV

- Goel, R., Soni, S., Goyal, N., Paparrizos, J., Wallach, H., Diaz, F., and Eisenstein, J. (2016).
  The social dynamics of language change in online networks.
  In *Social Informatics: 8th International Conference, SocInfo 2016, Bellevue, WA, USA, November 11-14, 2016, Proceedings, Part I 8*, pages 41–57. Springer.

- Hamilton, W. L., Leskovec, J., and Jurafsky, D. (2016).
  Diachronic word embeddings reveal statistical laws of semantic change.
  In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.

References V

- Hock, H. H. and Joseph, B. D. (2009).
  Language history, language change, and language relationship: An introduction to historical and comparative linguistics.
  Mouton de Gruyter.

- Kilgarriff, A. (1997).
  I don't believe in word senses.
  Computers and the Humanities, 31(2):91–113.

- Kozlowski, A. C., Taddy, M., and Evans, J. A. (2019).
  The geometry of culture: Analyzing the meanings of class through word embeddings.
  American Sociological Review, 84(5):905–949.

References VI

Kutuzov, A., Øvrelid, L., Szymanski, T., and Velldal, E. (2018).
Diachronic word embeddings and semantic shifts: a survey.
In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

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.

References VII

- Lazaridou, A., Kuncoro, A., Gribovskaya, E., Agrawal, D., Liska, A., Terzi, T., Gimenez, M., de Masson d'Autume, C., Kocisky, T., Ruder, S., Yogatama, D., Cao, K., Young, S., and Blunsom, P. (2021).
  Mind the gap: Assessing temporal generalization in neural language models.
  In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*, volume 34, pages 29348–29363. Curran Associates, Inc.

- Noble, B., Sayeed, A., Fernández, R., and Larsson, S. (2021).
  Semantic shift in social networks.
  In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 26–37, Online. Association for Computational Linguistics.

References VIII

- Tahmasebi, N., Borin, L., and Jatowt, A. (2021).
  Survey of computational approaches to lexical semantic change detection.
  Computational approaches to semantic change, 6:1.

- Xu, Y. and Kemp, C. (2015).
  A computational evaluation of two laws of semantic change.
  In *Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci 2015)*.