Diachronic semantic shifts and distributional models

Andrey Kutuzov
University of Oslo
Language Technology Group

October 24, 2019

Contents

What is this about?

Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference "Dialogue 2019"

Moscow, May 29—June 1, 2019

TRACING CULTURAL DIACHRONIC SEMANTIC SHIFTS IN RUSSIAN USING WORD EMBEDDINGS: TEST SETS AND BASELINES

Fomin V. (wadimiusz@gmail.com),
Bakshandaeva D. (dbakshandaeva@gmail.com),
Rodina Ju. (julia.rodina97@gmail.com)

National Research University Higher School of Economics, Moscow, Russia

Kutuzov A. (andreku@ifi.uio.no)

University of Oslo, Oslo, Norway

The paper introduces manually annotated test sets for the task of tracing diachronic (temporal) semantic shifts in Russian. The two test sets are complementary in that the first one covers comparatively strong semantic changes occurring to nouns and adjectives from pre-Soviet to Soviet times, while the second one covers comparatively subtle socially and culturally de-

[Fomin et al., 2019]

What is this about?

Diachronic semantic shifts?

- Word meaning $\approx$ word contexts [Firth, 1957]
- Changes in contexts $\approx$ changes in meaning
  - a.k.a. semantic shifts.

What is this about?

Diachronic semantic shifts?

- Word meaning $\approx$ word contexts [Firth, 1957]
- Changes in contexts $\approx$ changes in meaning
  - a.k.a. semantic shifts.
- Cultural changes influence the contexts
- Studies in automatic tracing of semantic shifts require publicly available datasets and strong baselines.

Task 1: Unsupervised Lexical Semantic Change Detection

- `https://competitions.codalab.org/competitions/20948`
  1. classification task
  2. ranking task
- German, English, Swedish, Latin

Unsupervised Lexical Semantic Change Detection Challenge

September 2019 – February 2020
Major NLP conference

SemEval2020

We are participating in SemEval2020 with a task on unsupervised lexical semantic change detection for English, German, Swedish and Latin, together with Barbara McGillivray, Dominik Schlechtweg, Simon Hengchen, and Haim Dubossarsky. Come and join us!

- Trial data ready July 31, 2019
- Training data ready September 4, 2019
- Test data ready December 3, 2019
- Evaluation start January 10, 2020
- Evaluation end January 31, 2020
- Paper submission due February 23, 2020

Contents

Previous work

- Hand-picking examples [Traugott and Dasher, 2001, Daniel and Dobrushina, 2016]

Previous work

- Hand-picking examples [Traugott and Dasher, 2001, Daniel and Dobrushina, 2016]
- Distributional approaches to diachronic semantics (surveyed in [Kutuzov et al., 2018, Tang, 2018])

Previous work

- Hand-picking examples [Traugott and Dasher, 2001, Daniel and Dobrushina, 2016]
- Distributional approaches to diachronic semantics (surveyed in [Kutuzov et al., 2018, Tang, 2018])
- Various algorithms of semantic shift tracing using word embeddings:

Previous work

- Hand-picking examples [Traugott and Dasher, 2001, Daniel and Dobrushina, 2016]
- Distributional approaches to diachronic semantics (surveyed in [Kutuzov et al., 2018, Tang, 2018])
- Various algorithms of semantic shift tracing using word embeddings:
  - Training models incrementally [Kim et al., 2014]

Previous work

- Hand-picking examples [Traugott and Dasher, 2001, Daniel and Dobrushina, 2016]
- Distributional approaches to diachronic semantics (surveyed in [Kutuzov et al., 2018, Tang, 2018])
- Various algorithms of semantic shift tracing using word embeddings:
  - Training models incrementally [Kim et al., 2014]
  - Training models separately for each time bin:
    - Aligning embedding spaces [Hamilton et al., 2016]
    - Comparing distances between a given word and all others (second-rank similarity) [Yin et al., 2018]

Previous work

- Hand-picking examples [Traugott and Dasher, 2001, Daniel and Dobrushina, 2016]
- Distributional approaches to diachronic semantics (surveyed in [Kutuzov et al., 2018, Tang, 2018])
- Various algorithms of semantic shift tracing using word embeddings:
  - Training models incrementally [Kim et al., 2014]
  - Training models separately for each time bin:
    - Aligning embedding spaces [Hamilton et al., 2016]
    - Comparing distances between a given word and all others (second-rank similarity) [Yin et al., 2018]
  - Training models jointly across time bins [Bamler and Mandt, 2017, Yao et al., 2018, Rosenfeld and Erk, 2018]

Contents

Russian datasets

What we did?

- Dataset of short-term semantic shifts in Russian adjectives, based on news texts
- Re-packing a dataset of long-term semantic shifts for nouns and adjectives during the Soviet period
- Experimenting with well-established baseline algorithms for semantic shift detection, testing them on the datasets

NB: antonyms pose real problems for distributional models!

Russian datasets

'Micro' dataset

- 2000 — 2014: 15 years of Russian news texts
- 20 adjectives for each year pair (2000-2001, 2001-2002, etc...)
- selected randomly, biased towards the words chosen by the *Global Anchors* method (more details further)
- 14 year pairs $\times$ 20 words = 280 entries
- Manual annotation by 3 annotators
- 3 class labels:

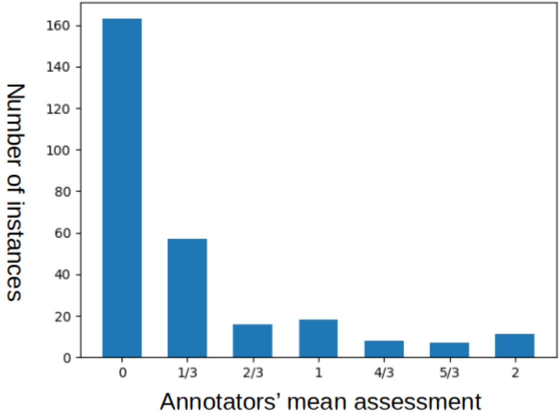
Label	**Meaning**
0	no semantic shift
1	somewhat shifted
2	significantly shifted

Russian datasets

Socio-cultural semantic shifts in adjectives in 2014, as compared to 2013 (excerpts from the 'Micro' dataset)

Class	Adjective	English translation
2	крымский	'*Crimean*'
2	приёмный	'*1) adopted; 2) something receiving*'
2	луганский	'*of Luhansk*'
1	правый	'*1) right; 2) right-wing*'
1	кипрский	'*Cyprian, Cypriot*'
0	серый	'*gray*'
0	балетный	'*of ballet*'

Russian datasets



Mean values of annotators' scores, 'Micro' dataset

Russian datasets

'Macro' dataset

- Originally from [Kutuzov and Kuzmenko, 2018]
- We publish it in a machine-readable form.
- Changes from Pre-Soviet through Soviet times
-

	Nouns	**Adjectives**
Target	38	5
Filler	152	20

- 2 class labels (no shift / shift)

Russian datasets

word	label	word	label
отделение	1	тюрьма	0
секция	1	влияние	0
богадельня	1	весна	0
особа	1	уверенность	0
уклон	1	красавица	0
молодец	1	жених	0
передовой	1	заказ	0

Table: Example entries from the ‘Macro’ dataset

Russian datasets

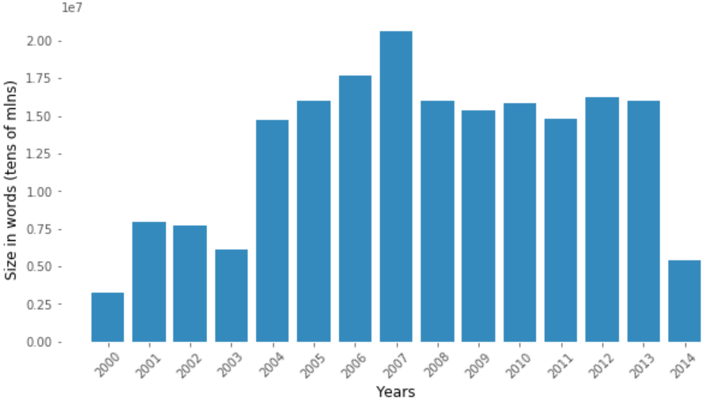
'Micro' corpus

- Newspaper subcorpus of RNC + lenta.ru
  - News texts produced in 2000,
  - News texts produced in 2001,
  - ...,
  - News texts produced in 2014,

'Macro' corpus

- Main body of RNC:
  - Texts produced before 1917 (75 millions tokens),
  - Texts produced in 1918—1990 (96 millions tokens),
  - Texts produced after 1991 (85 millions tokens)

Russian datasets



’Micro’ corpora sizes per year

Contents

Word embeddings

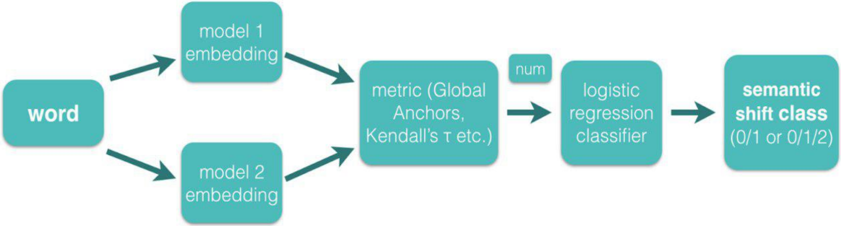
Distributional models for baselines evaluation

- 'Static' models:
  - Model trained on time bin tb_0,
  - Model trained on time bin tb_1,
  - ...
  - Model trained on time bin tb_n
- 'Incremental' models
  - Model trained on time bin tb_0,
  - Model trained on time bin tb_1, initialized with tb_0 weights,
  - ...
  - Model trained on time bin tb_n, initialized with tb_{n-1} weights.

word2vec CBOW [Mikolov et al., 2013], context window = 5, vector size 300

Contents

Baseline results



Experimental workflow

Conceptual types of methods

Local methods for semantic shift detection

Comparing words' nearest neigbors:

- Jaccard distance [Jaccard, 1901]
- Kendall's τ [Kendall, 1948]

Global methods for semantic shift detection

Comparing overall structure of semantic spaces:

- Procrustes alignment [Hamilton et al., 2016]
- Global Anchors [Yin et al., 2018]

Local methods

Jaccard distance

[Jaccard, 1901]

$$J(X, Y) = \frac{|X \cap Y|}{|X \cup Y|} \tag{1}$$

Nearest neighbors for ‘вежливый’:

X = приветливый, общительный, уравновешенный, отзывчивый, добродушный

Y = камуфляж, неравнодушный, порядочный, здравомыслящий, незнакомый

Can you guess the years for X and Y?

Local methods

Kendall's τ

Takes into account the ranking of neighbors [Kendall, 1948]

$$\frac{2}{n(n-1)} \sum_{i<j} sgn(x_i - x_j)\, sgn(y_i - y_j) \tag{2}$$

Nearest neighbors for 'луганский' ($x = 2013$, $y = 2014$):

x_1: иркутский → y_1: донецкий

… …

x_7: донецкий → y_{17}: иркутский

Global methods

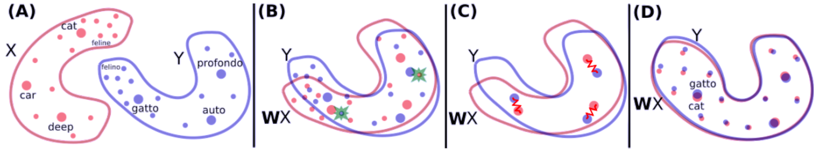
Orthogonal Procrustes Analysis

Given embedding matrices A and B, find an orthogonal matrix R that maps A to B [Hamilton et al., 2016].

$$B^T A = M$$

$$M = U \Sigma V^T$$

$$R = UV^T$$



Then simple cosine between $word^A$ and $word^B$ is calculated

Global methods

Global Anchors

[Yin et al., 2018]

Semantic shift of word *w* from year *x* to year *y*:

$$similarities_x = (x_1, ..., x_n)$$

$$similarities_y = (y_1, ..., y_n)$$

- x_i and y_i are cosine similarities between the word *w* and the i^{th} word in the intersection of *x* and *y* vocabularies.
- We compare global positions of *w* in the semantic space.
- Semantic similarity between different time periods = $cos(similarities_x, similarities_y)$

Baseline results

'Macro' dataset

Models	Glob.Anchors	Procrustes	Kendall	Jaccard	combined
Static	0.675	**0.767**	0.504	0.646	0.722
Incremental	0.598	0.681	0.475	0.576	0.617
Random choice					
≈ 0.5					

- Global methods work better
- Local methods are still applicable
- Procrustes analysis is clearly the best
- Incremental models are worse than static.

Baseline results

‘Micro’ dataset

Models	Glob.Anchors	Procrustes	Kendall	Jaccard	combined
Static	0.453	0.468	0.136	0.301	**0.503**
Incremental	0.462	0.459	0.194	0.326	0.442
Random choice					
$\approx$ 0.33					

- Global methods clearly win on granular timespans
- Local methods sometimes worse than random
- Combining methods is a good idea
- Still no (significant) profit from incremental models

Baseline results

Please re-use:

- Two manually annotated datasets with diachronic semantic shifts for Russian:
  - A short-term 'Micro' dataset, scale = years (adjectives only)
  - A long-term 'Macro' dataset, scale = centuries
- Datasets and baseline implementations:

`https://github.com/wadimiusz/diachrony_for_russian`

Contents

Recent ideas

Temporal referencing

- Time labels as tags [Dubossarsky et al., 2019]
- Each target word is replaced with a time-specific token
  - In the 1920s corpus: *computer* $\rightarrow$ *computer*$_{1920}$

Recent ideas

Temporal referencing

- Time labels as tags [Dubossarsky et al., 2019]
- Each target word is replaced with a time-specific token
  - In the 1920s corpus: *computer* $\rightarrow$ $\textit{computer}_{1920}$
- If it is a context word, it remains unchanged.

Recent ideas

Temporal referencing

- Time labels as tags [Dubossarsky et al., 2019]
- Each target word is replaced with a time-specific token
  - In the 1920s corpus: *computer* $\rightarrow$ *computer*$_{1920}$
- If it is a context word, it remains unchanged.
- One vector space is learned.
- No post-hoc alignment necessary.

Recent ideas

What else can be done?

- Semantic shifts are related to word senses

Recent ideas

What else can be done?

- Semantic shifts are related to word senses
- What about contextualized embeddings?
  - ELMo [Peters et al., 2018]
  - BERT [Devlin et al., 2019]

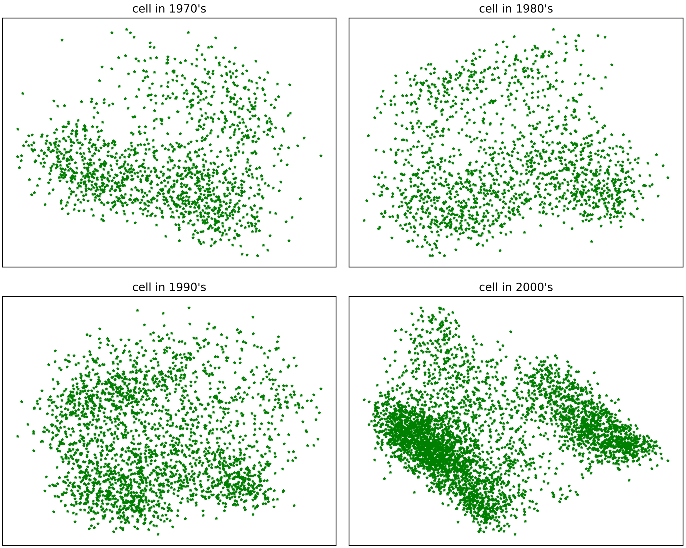
Recent ideas

What else can be done?

- Semantic shifts are related to word senses
- What about contextualized embeddings?
  - ELMo [Peters et al., 2018]
  - BERT [Devlin et al., 2019]

[Giulianelli, 2019] tries to compare clusters of BERT embeddings for word occurrences across the COHA corpus. We did it with ELMo top layer representations.

Recent ideas



ELMo representations of each occurrence of the word *'cell'* in 4 decades: actual semantic shift. Diversity significantly increased in 2000s.

Recent ideas

Prison cell

1. '...the chief turnkey on duty, for over ten years, but you wouldn't have known it from the way he processed me for the *cells*.'
2. 'It also happened to me in a jail *cell*, Peb.'
3. 'If she had been writing to somebody in the darkness of her prison *cell*, what had she done with the message?'

Biological cell

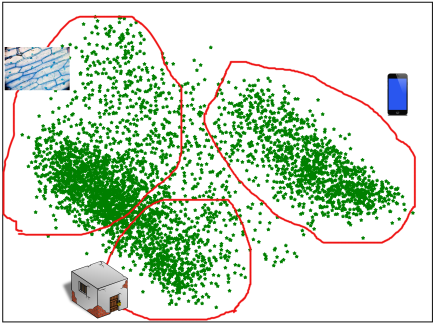
1. 'The sexual *cells* of Pyronema show this in ascomycetes.'
2. '...how a *cell* decides whether it becomes a muscle *cell* or...'
3. 'If those *cells* are found to be cancerous after being sent to a lab...'

Recent ideas

Cell phone (2000s only)

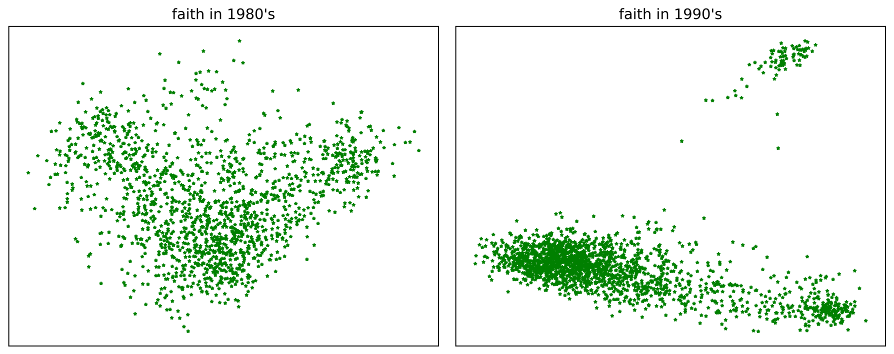
1. '...service providers fulfill that objective, and what about the other health and safety risks... that the growing use of *cell* phones raise?'
2. 'Gilles swatted Adriana on the upper arm... nearly dislodging the *cell* phone she had balanced between her chin and her left shoulder.'
3. 'You still have the same *cell* number.'

cell in 2000's



Recent ideas

But...



ELMo representations of each occurrence of the word *'faith'* in 2 decades: diversity also significantly increased. WTF?

Recent ideas

Sentences from the new cluster:

1. ‘Maybe we could - - 64 -   *FAITH* (waving down a cab) Thank you, but this is a personal matter.’
2. ‘ *FAITH* (nodding) Like a detective.’
3. ‘Perhaps you misunderstood ?   *FAITH* (trying not to panic) Are you absolutely sure he’s gone? Maybe you made a mistake.’

Recent ideas

Sentences from the new cluster:

1. ‘Maybe we could - - 64 -   *FAITH* (waving down a cab) Thank you, but this is a personal matter.’
2. ‘ *FAITH* (nodding) Like a detective.’
3. ‘Perhaps you misunderstood ?   *FAITH* (trying not to panic) Are you absolutely sure he’s gone? Maybe you made a mistake.’

- Script of the 1994 movie ‘Only You’, where ‘FAITH’ is one of the main characters!
- Often accompanied by parentheses and non-breaking space ( ).
- Contextualized representations heavily influenced by syntax and punctuation.
- False flag!

Recent ideas

Contextualized representations in semantic shifts detection

- Not entirely straightforward.

Recent ideas

Contextualized representations in semantic shifts detection

- Not entirely straightforward.
- Empirical results still do not outperform previous approaches (yet).

Recent ideas

Contextualized representations in semantic shifts detection

- Not entirely straightforward.
- Empirical results still do not outperform previous approaches (yet).
- Can we somehow filter out syntactic information?
  - learn a weighted function of layers for this task?

Recent ideas

Contextualized representations in semantic shifts detection

- Not entirely straightforward.
- Empirical results still do not outperform previous approaches (yet).
- Can we somehow filter out syntactic information?
  - learn a weighted function of layers for this task?
- Conceptual problem of determining the number of clusters.

Recent ideas

Contextualized representations in semantic shifts detection

- Not entirely straightforward.
- Empirical results still do not outperform previous approaches (yet).
- Can we somehow filter out syntactic information?
  - learn a weighted function of layers for this task?
- Conceptual problem of determining the number of clusters.
- How to align temporal models?

Recent ideas

Contextualized representations in semantic shifts detection

- Not entirely straightforward.
- Empirical results still do not outperform previous approaches (yet).
- Can we somehow filter out syntactic information?
  - learn a weighted function of layers for this task?
- Conceptual problem of determining the number of clusters.
- How to align temporal models?
- ...and lots of other interesting topics to research :-)

Recent ideas

Contextualized representations in semantic shifts detection

- Not entirely straightforward.
- Empirical results still do not outperform previous approaches (yet).
- Can we somehow filter out syntactic information?
  - learn a weighted function of layers for this task?
- Conceptual problem of determining the number of clusters.
- How to align temporal models?
- ...and lots of other interesting topics to research :-)

Thanks! Questions?

References I

Bamler, R. and Mandt, S. (2017).
Dynamic word embeddings.
In *Proceedings of the International Conference on Machine Learning*, pages 380–389, Sydney, Australia.

Daniel, M. and Dobrushina, N. (2016).
Two centuries in twenty words (in Russian).
NRU HSE.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019).
BERT: Pre-training of deep bidirectional transformers for language understanding.
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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

References II

Dubossarsky, H., Hengchen, S., Tahmasebi, N., and Schlechtweg, D. (2019).
Time-out: Temporal referencing for robust modeling of lexical semantic change.
In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 457–470, Florence, Italy. Association for Computational Linguistics.

Firth, J. (1957).
A synopsis of linguistic theory, 1930-1955.
Blackwell.

Fomin, V., Bakshandaeva, D., Rodina, J., and Kutuzov, A. (2019).
Tracing cultural diachronic semantic shifts in Russian using word embeddings: test sets and baselines.
Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Dialog conference, pages 203–218.

References III

Giulianelli, M. (2019).
Lexical semantic change analysis with contextualised word representations.
Master's thesis, University of Amsterdam.

Hamilton, W., 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*, pages 1489–1501, Berlin, Germany.

Jaccard, P. (1901).
Distribution de la Flore Alpine: dans le Bassin des dranses et dans quelques régions voisines.
Rouge.

References IV

- Kendall, M. G. (1948).
  Rank correlation methods.
  Griffin.
- Kim, Y., Chiu, Y.-I., Hanaki, K., Hegde, D., and Petrov, S. (2014).
  Temporal analysis of language through neural language models.
  In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 61–65, Baltimore, USA.
- Kutuzov, A. and Kuzmenko, E. (2018).
  Two centuries in two thousand words: Neural embedding models in detecting diachronic lexical changes.
  In *Quantitative Approaches to the Russian Language*, pages 95–112. Routledge.

References V

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. Association for Computational Linguistics.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013).
Distributed representations of words and phrases and their compositionality.
Advances in Neural Information Processing Systems, 26:3111–3119.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018).
Deep contextualized word representations.
In *Proc. of NAACL*.

References VI

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*, pages 474–484, New Orleans, Louisiana, USA.

Tang, X. (2018).
A state-of-the-art of semantic change computation.
Natural Language Engineering, 24(5):649–676.

Traugott, E. C. and Dasher, R. B. (2001).
Regularity in semantic change.
Cambridge University Press.

References VII

Yao, Z., Sun, Y., Ding, W., Rao, N., and Xiong, H. (2018).
Dynamic word embeddings for evolving semantic discovery.
In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 673–681, Marina Del Rey, CA, USA.

Yin, Z., Sachidananda, V., and Prabhakar, B. (2018).
The global anchor method for quantifying linguistic shifts and domain adaptation.
In *Advances in Neural Information Processing Systems*, pages 9433–9444.