Distributional semantics for lexical semantic variation and change

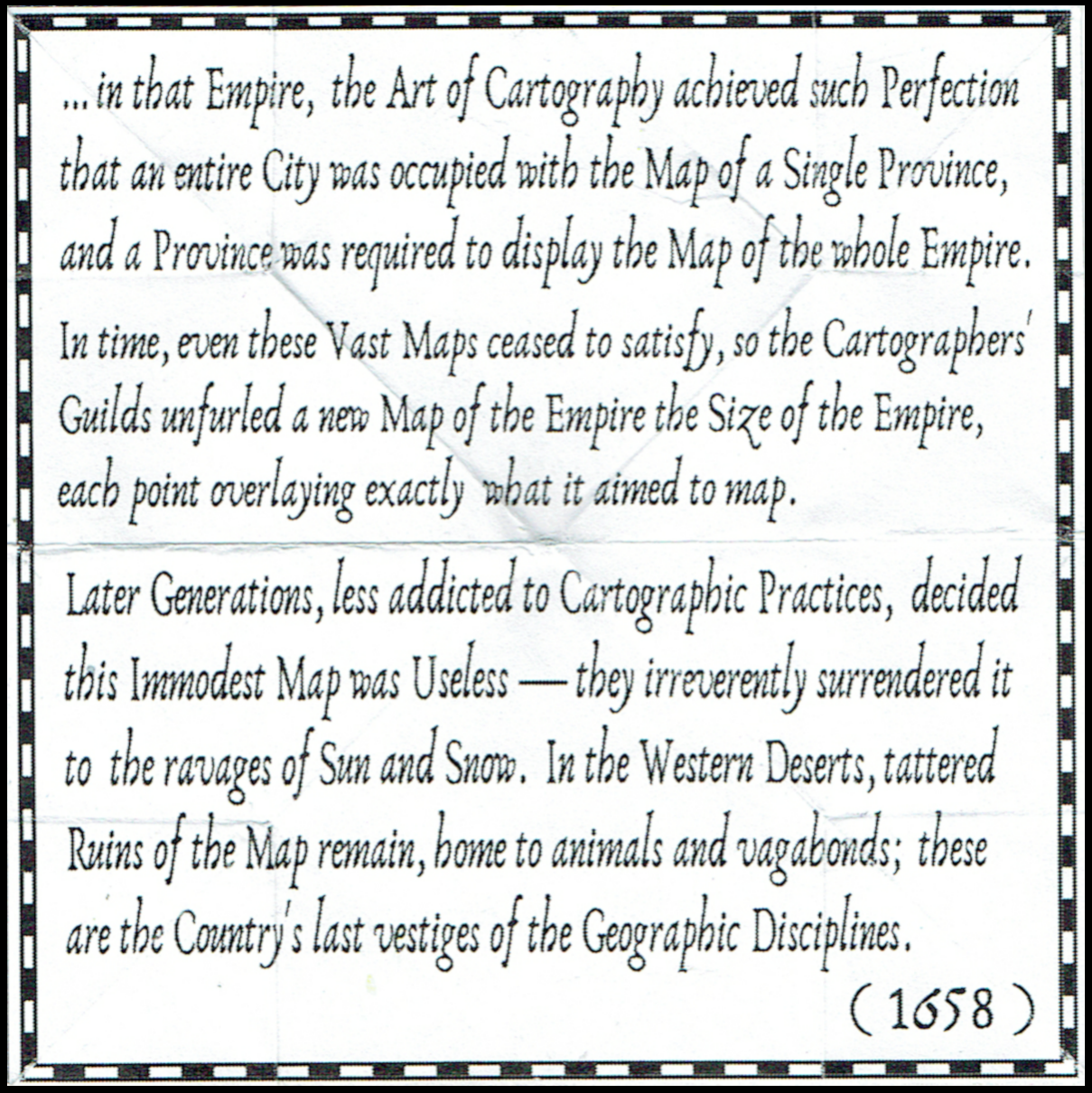
Bill Noble

January 27, 2021

CLASP Seminar

University of Gothenburg

All models are wrong...



...in that Empire, the Art of Cartography achieved such Perfection that an entire City was occupied with the Map of a Single Province, and a Province was required to display the Map of the whole Empire.

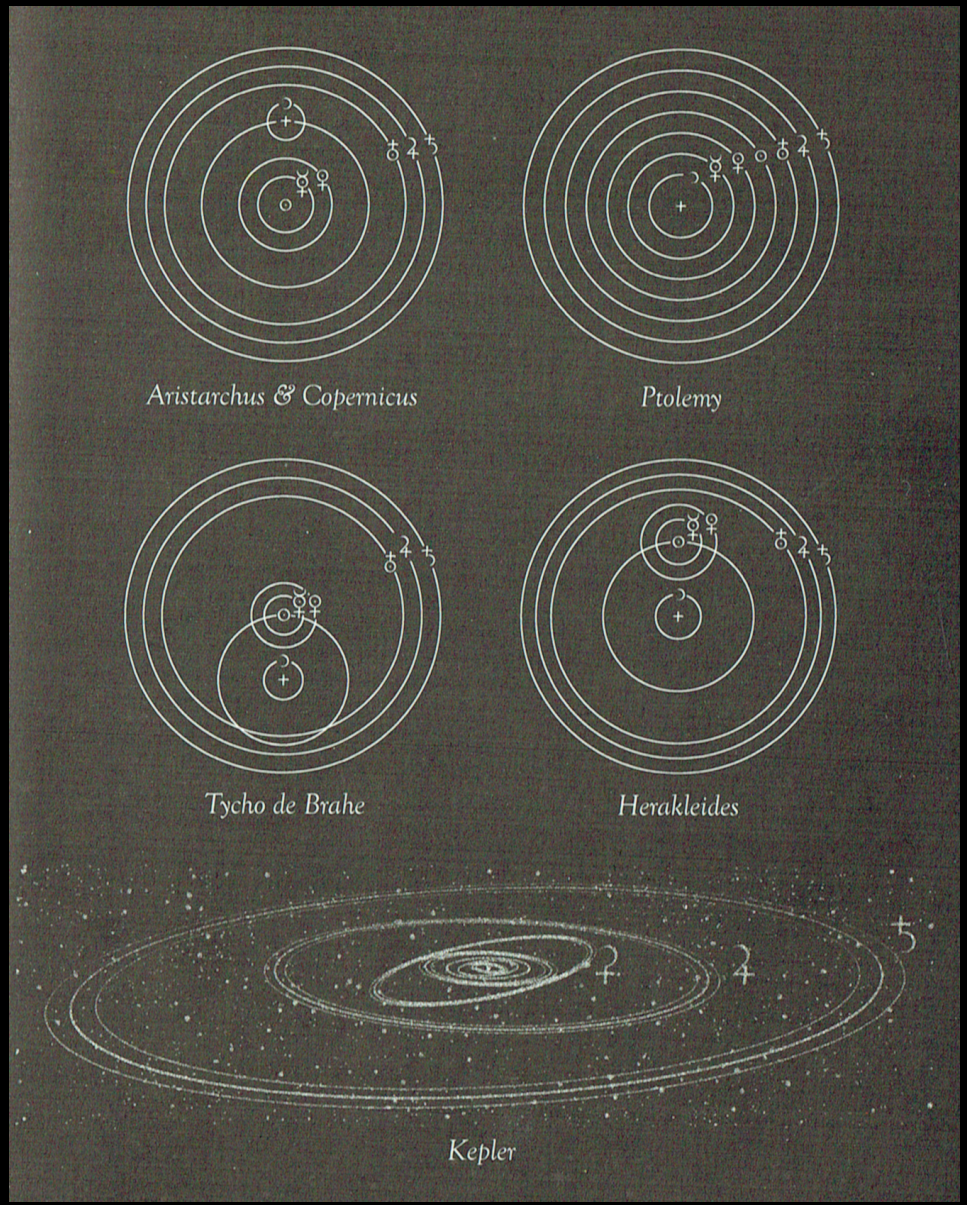
In time, even these Vast Maps ceased to satisfy, so the Cartographers' Guilds unfurled a new Map of the Empire the Size of the Empire, each point overlaying exactly what it aimed to map.

Later Generations, less addicted to Cartographic Practices, decided this Immodest Map was Useless — they irreverently surrendered it to the ravages of Sun and Snow. In the Western Deserts, tattered Ruins of the Map remain, home to animals and vagabonds; these are the Country's last vestiges of the Geographic Disciplines.

(1658)

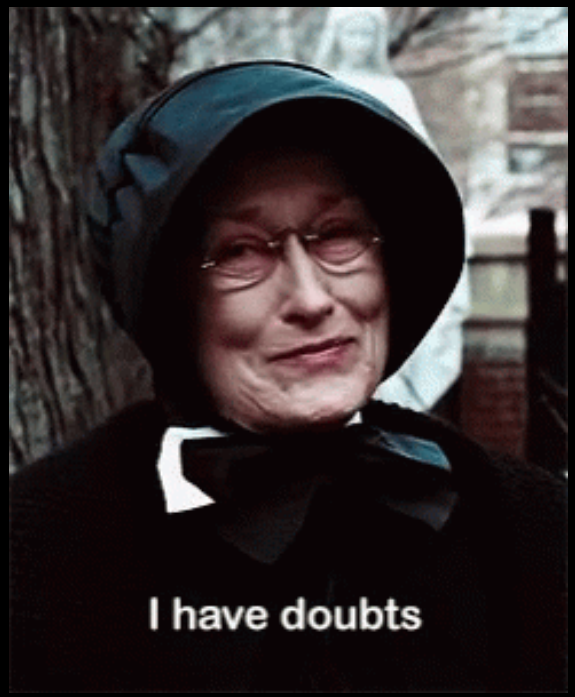
From: *On Exactitude in Science* by Jorge Luis Borges (trans. Noah K. Mease)

...but some are useful



Source: Martineau (2002)

Such doubts!



Variation and change

Distributional semantics

Two studies

- Change: Diachronic skip-gram model
- Variation: Conditional language models

Comparing models

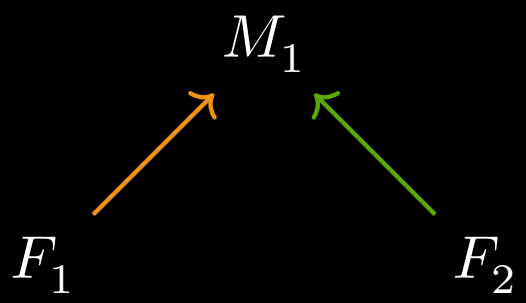
Variation and change

Variation

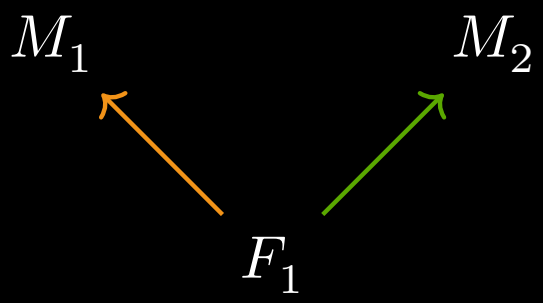
Linguistic variation: systematic differences across speech communities.

- If language is part of the common ground (Clark, 1996) of a community, variation is what changes between communities
- Contrast with Linguistic style (individual differences)
- We usually consider variation with in the “same” language
- Variationist sociolinguistics (e.g., Labov, 1963; Milroy, 2000; Eckert, 2000; Podesva, 2007; Campbell-Kibler, 2010)

Two kinds of variation



(a) **Type 1**: C_1 and C_2 use different forms for M_1.



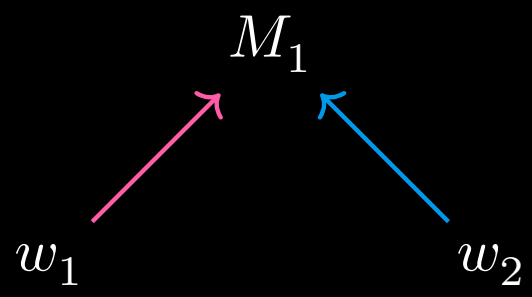
(b) **Type 2**: C_1 and C_2 interpret F_1 differently.

Figure 1: Sociolinguists typically only consider the first kind of variation (Anttila, 2004).

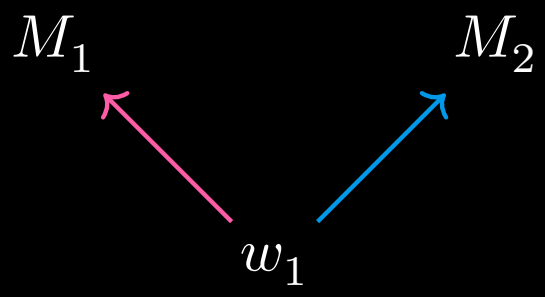
Semantic variation isn't studied much in sociolinguistics

- Semantics is at the top of the classical linguistic hierarchy
- It's difficult to establish that different forms are being used to say the same thing
- Exception: Hasan (2009) (using functional linguistics)

Change is just variation over time?



(a) **Type 1**: w_1 previously meant M_1, but now w_2 is used (too).



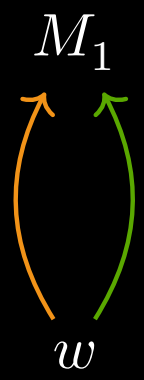
(b) **Type 2**: w_1 previously meant M_1, but now it means M_2 (too).

Figure 2: Semantic change typically only refers to type 2 variation.

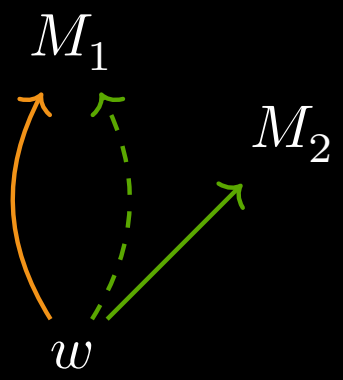
Type 2 semantic change has been studied extensively:

- in historical linguistics (e.g., Paul, 1886; Bloomfield, 1933; Traugott and Dasher, 2001)
- with distributional semantics (recent surveys: Tang, 2018; Kutuzov et al., 2018; Tahmasebi et al., 2018)

Change results in variation



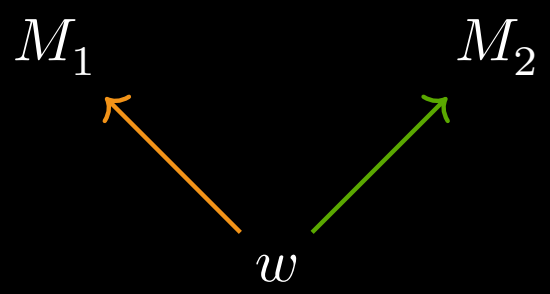
(a) C_1 and C_2 have the same meaning for w.



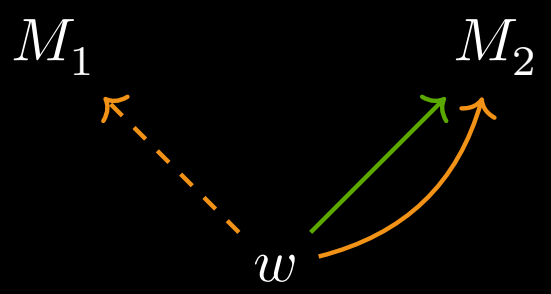
(b) C_2 now has a different (additional) meaning for w.

Figure 3: When the meaning of a word changes in one community but not another, semantic variation is the result.

Variation leads to change



(a) C_1 and C_2 have different meanings for w.



(b) C_1 has has adopted the C_2 meaning for w.

Figure 4: One source of change within a community is adopting semantic variants from other communities.

Distributional semantics

The distributional hypothesis (Harris, 1954)

If we consider *oculist* and *eye-doctor* we find that, as our corpus of actually-occurring utterances grows, these two occur in almost the same environments, except for such sentences as *An oculist is just an eye-doctor under a fancier name*, or *I told him Burns was an oculist, but since he didn't know the professional titles, he didn't realize that he could go to him to have his eyes examined*.

[...]

If A and B have almost identical environments except chiefly for sentences which contain both, we say they are synonyms: oculist and eye-doctor. If A and B have some environments in common and some not (e.g. oculist and lawyer) we say that they have different meanings, the amount of meaning difference corresponding roughly to the amount of difference in their environments.

Reasons to doubt

Simply compare (a representation of) the environments (contexts) of w in t_1 (or C_1) with that of w in t_2 (or C_2).

The amount of change (or variation) in word meaning should correspond to the amount of difference in the two contexts.

- Distributional representations of meaning are ungrounded (e.g., Bender and Koller, 2020; Bisk et al., 2020)?
- Differences in context can reflect differences in topic distribution.
  - *Tomato* in a gardening forum vs. *tomato* in a cooking forum (context: *dirt* vs. context: *sauté*).
  - *Trump* in 2014 (TV performer) vs. *Trump* in 2016 (politician).
- Contexts are *at best* a noisy approximation of meaning

Meaning potential

A useful distinction (Norén and Linell, 2007; de Saussure et al., 2011):

- meaning potential (parole)
- situated use (langue)

We are interested in changes in *meaning potential*, but we approximate them with changes in *situated use*. Is that at all justified?

Reasons to doubt your doubts i

[...] the whole theory of language change can be reduced to one question: what is the relationship between prevailing usage and the speech activity of an individual? How is the speech of an individual determined by prevailing usage in the community, and how in turn does the individual's speech affect prevailing usage?

Hermann Paul, *Principles of the History of Language* (1886)
(trans. Herbert A. Strong, 1891)

Reasons to doubt your doubts ii

Each word tastes of the context and contexts in which it has lived its socially charged life: all words and forms are populated by intentions.

[...]

Prior to this moment of appropriation, the word does not exist in a neutral and impersonal language it is not, after all, out of a dictionary that the speaker gets his words!), but rather it exists in other people's mouths, in other people's contexts, serving other people's intentions: it is from there that one must take the word, and make it one's own.

M.M. Bakhtin, *Discourse in the Novel* (1941)
(trans. Caryl Emerson and Michael Holquist 1981)

Two studies

Semantic shift in social networks

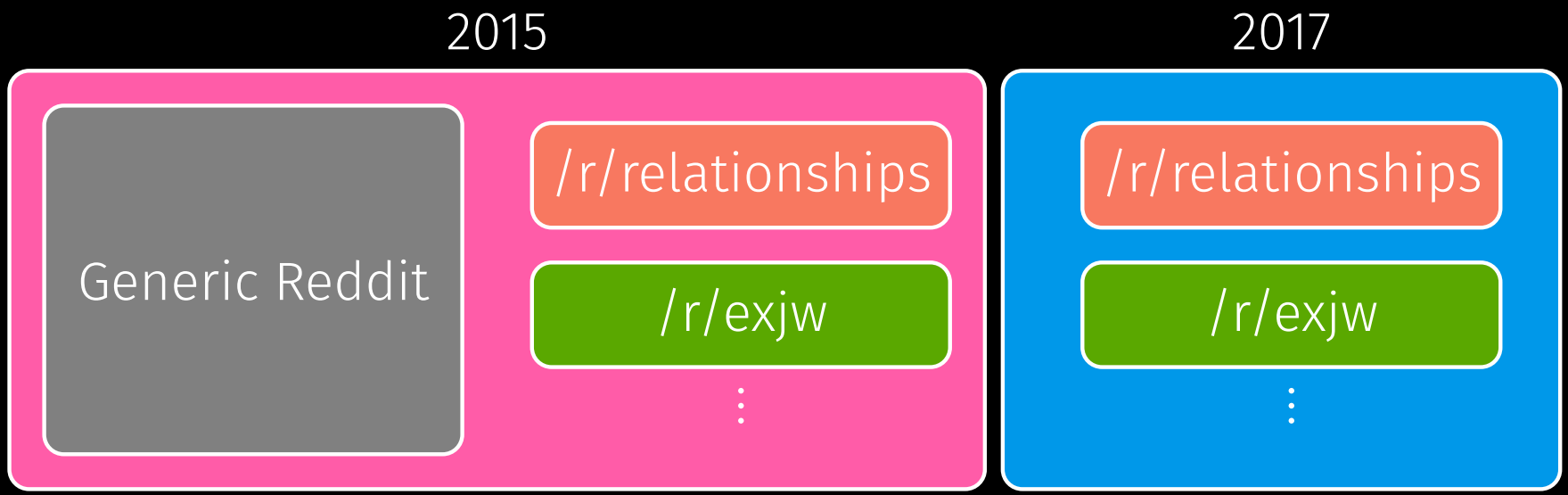
- Model: Diachronic skip-gram with negative sampling
- Question: How does social network structure affect semantic change?
- People: Asad Sayeed, Staffan Larsson, Raquel Fernández

Community-conditioned language models

- Model: Community-conditioned language models
- Question: What (if any) linguistically distinguishing community features do the LMs encode?
- People: Jean-Philippe Bernardy

Data: Reddit comments

- Social media comments
  - threaded replies
  - authorship identified by username
- Two time periods: 2015 and 2017 (one year gap)
- 46 randomly selected communities (avg. 282K comments per community)
- A larger "generic" 2015 corpus of comments randomly selected from all of Reddit (55M comments)



Two studies

Change: Diachronic skip-gram model

Diachronic skip-gram (Kim et al., 2014)

- Skip-gram with Negative Sampling (SGNS) tries to guess, for a given word, whether another word was drawn from its context window or not (i.e. if it is a negative sample)
- The diachronic skipgram procedure we followed is as follows (adapted from Del Tredici et al. (2019)):
    1. Train a base model, M_{15}, on the generic Reddit 2015 corpus.
    2. For each subreddit c:
        - 2.1 Initialize with M_{15} and train a community-specific 2015 model, M_{15}^{c}.
        - 2.2 Initialize with M_{15}^{c} and train a community-specific 2017 model M_{17}^{c}.

Cosine change

For a community c and word w, cosine change is defined as the angular distance between the corresponding vectors for the two time periods:

$$\Delta_c^{\cos}(w) = \text{angular distance}(\vec{w}_{t_0}^c, \vec{w}_{t_1}^c) \tag{1}$$

Research questions

- How does social network connectivity (clustering coefficient) of a community affect the pace of semantic change?
- How does it interact with other community features?
  - Size (number of active members)
  - Stability (membership overlap between years)
  - Mean posts per member

Clustering coefficient

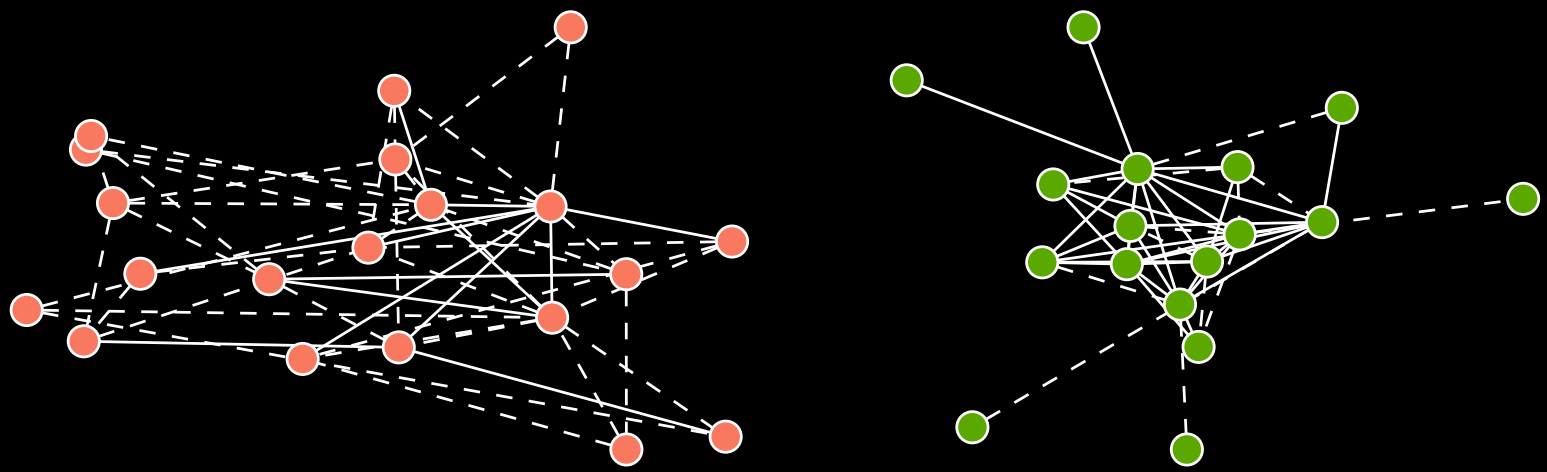
For an individual, i, the clustering coefficient C^i is defined as the proportion of possible connections that exist between individuals connected to i:

$$C_G^i = \frac{|\{\{j,k\} \in G \mid j,k \in N(i)|\}}{|N(i)|(|N(i)|-1)} \tag{2}$$

where $N(i) = \{j \in U \mid \{i,j\} \in G\}$ is the *neighborhood* of i. The clustering coefficient for the community as a whole is the mean clustering coefficient of its members:

$$C_G = \frac{\sum_{i \in U} C_G^i}{|U|} \tag{3}$$

Strong and weak ties



(a) /r/relationships

(b) /r/exjw

Figure 5: Sub-graphs of two communities with different weak and strong tie pattern. Weak ties are shown with dashed lines. Top: $C_s = 0.04$, $C_w = 0.47$, $C = 0.51$. Bottom: $C_s = 0.42$, $C_w = 0$, $C = 0.53$.

Exploratory analysis

- Multi-stage regressions and model selection by backwards elimination
- First, we try to explain change with word-level features (e.g., frequency)
- Then we use community-level features to predict the residuals of that regression model.
- Clustering coefficient has a negative correlation with cosine change (more clustered communities experienced less change).
- This is especially true for large, unstable communities.

Two studies

Variation: Conditional language models

Community-conditioned language model (CCLM)

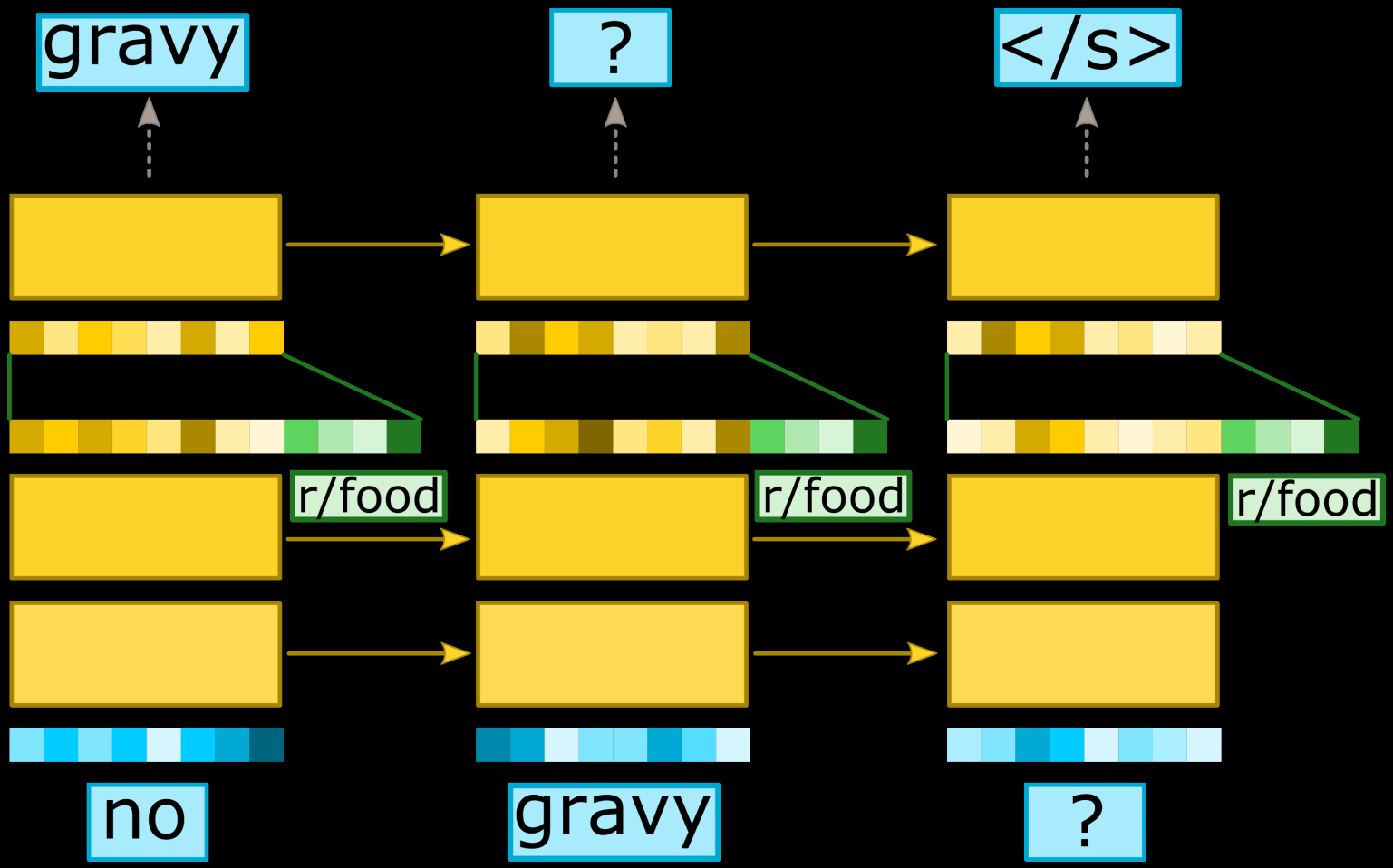
Language model:

$$P(w_1, ...w_n) = \prod_{i=1}^{n} P(w_i \mid w_1, ...w_{i-1}) \tag{4}$$

Conditional language model:

$$P(w_1, ...w_n \mid c) = \prod_{i=1}^{n} P(w_i \mid w_1, ...w_{i-1}; c) \tag{5}$$

Neural CCLM



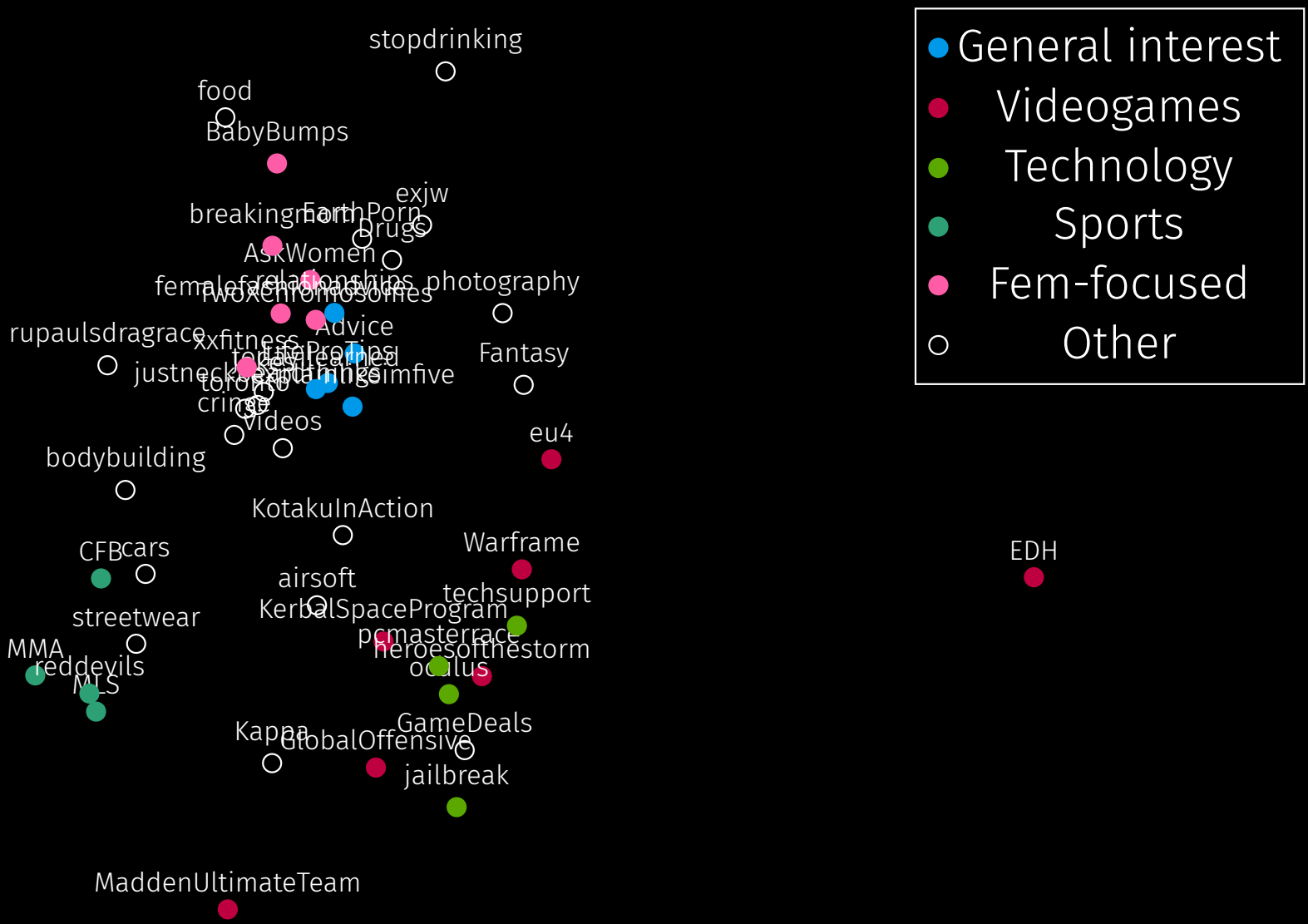
Research questions

- Do the CCLM community embeddings correlate with non-linguistic community features?
- How do communities differ in language model perplexity?
- Can we use the trained CCLM to classify comments by community (i.e., as a LMCC)?
- How does the layer at which the community embedding is concatenated affect LM performance?

Conditioning on community improves LM performance

	l_c	test epoch	Perplexity	Info. gain
LSTM	-	21	51.99	-
	0	17	50.83	1.023
	1	34	49.66	1.047
	2	11	50.23	1.035
	3	16	**49.60**	**1.048**
Transformer	-	20	61.43	-
	0	7	58.71	1.046
	1	12	61.69	0.992
	2	7	78.76	0.780
	3	10	**52.28**	**1.054**

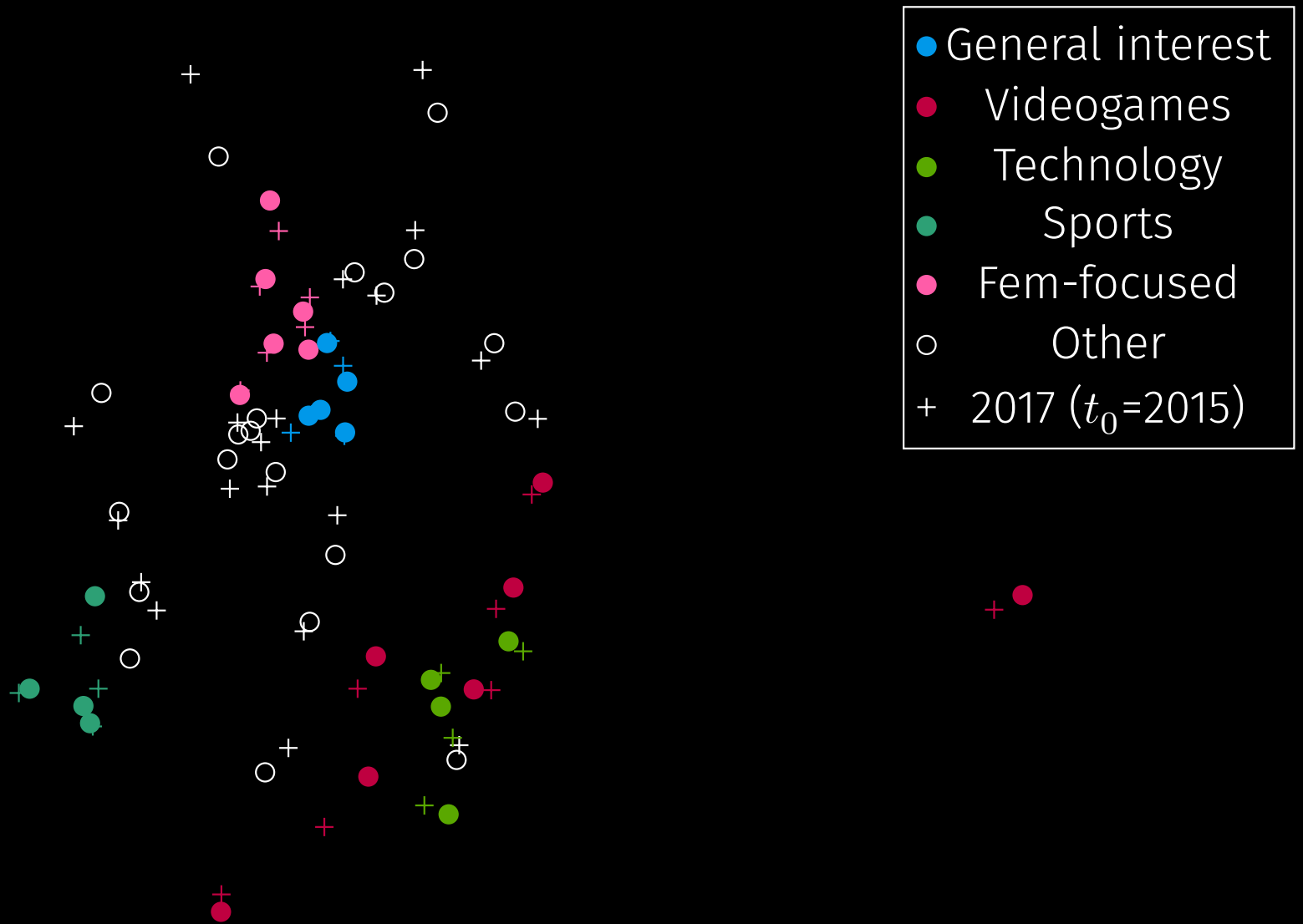
The community embedding (PCA)



Diachronic community-conditioned models: Naive approach

- Idea: Use an embedding for each community $\times$ time period
  - With 46 communities and 2 time periods (2015, 2017) we now have 92 conditional vectors.
- Concatenate the community embedding at layer 0 (i.e., directly to the word embedding)

Diachronic community embedding



Comparing models

SGNS vs CCLM

SGNS

- Highly sensitive to word frequency confound.
- Must be careful about vector space drift and managing vocabularies.
- Extensively tested on long-term (and a little bit on short-term) semantic change detection.
- Skip-gram training scheme

CCLM

- Doesn't seem to have as much of a problem with word frequency (in preliminary tests)
- No worries about maintaining the same vector space.
- "Bonus" community/time period embedding.
- May not be able to pick out specific word-level changes.
- Customizable language model training scheme.

References

Arto Anttila. 2004. Variation and Phonological Theory. In *The Handbook of Language Variation and Change*, chapter 8, pages 206–243. John Wiley & Sons, Ltd.

M. M. Bakhtin. 1981. *The Dialogic Imagination: Four Essays*. Number no. 1 in University of Texas Press Slavic Series. University of Texas Press, Austin.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In *ACL 2020*.

References ii

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. 2020. Experience Grounds Language. *arXiv:2004.10151 [cs]*.

Leonard Bloomfield. 1933. *Language*, 14. impr edition. Allen & Unwin, London.

Kathryn Campbell-Kibler. 2010. The sociolinguistic variant as a carrier of social meaning. *Language Variation and Change*, 22(3):423–441.

Herbert H. Clark. 1996. *Using Language*. Cambridge University Press.

References iii

Marco Del Tredici, Raquel Fernández, and Gemma Boleda. 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*, volume 1 (Long and Short Papers), pages 2069–2075, Minneapolis, Minnesota. Association for Computational Linguistics.

Penelope Eckert. 2000. *Linguistic Variation as Social Practice: The Linguistic Construction of Identity in Belten High*. Number 27 in Language in Society. Blackwell Publishers, Malden, Mass.

References iv

Zellig S. Harris. 1954. Distributional Structure. *Word*, 10(2-3):146–162.

Ruqaiya Hasan. 2009. *Collected Works of Ruqaiya Hasan. Vol. 2, Semantic Variation: Meaning in Society and in Sociolinguistics*. Equinox, London ; Oakville.

Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. Temporal Analysis of Language through Neural Language Models. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 61–65, Baltimore, MD, USA. Association for Computational Linguistics.

References v

Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 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.

William Labov. 1963. The Social Motivation of a Sound Change. *WORD*, 19(3):273–309.

John Martineau. 2002. *A Little Book of Coincidence: In the Solar System*. Bloomsbury Publishing USA.

References vi

Lesley Milroy. 2000. Social Network Analysis and Language Change: Introduction. *European Journal of English Studies*, 4(3):217–223.

Kerstin Norén and Per Linell. 2007. Meaning potentials and the interaction between lexis and contexts: An empirical substantiation. *Pragmatics*, 17(3):387–416.

Hermann Paul. 1886. *Prinzipien der Sprachgeschichte*. Max Niemeyer.

Hermann Paul. 1891. *Principles of the History of Language*. London ; New York : Longmans, Green.

References vii

Robert J. Podesva. 2007. Phonation type as a stylistic variable: The use of falsetto in constructing a persona1. *Journal of Sociolinguistics*, 11(4):478–504.

Ferdinand de Saussure, Wade Baskin, Perry Meisel, and Haun Saussy. 2011. *Course in General Linguistics*. Columbia University Press, New York.

Nina Tahmasebi, Lars Borin, and Adam Jatowt. 2018. Survey of Computational Approaches to Diachronic Conceptual Change. *arXiv:1811.06278 [cs]*.

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

References viii

Elizabeth Closs Traugott and Richard B. Dasher. 2001. *Regularity in Semantic Change*. Cambridge Studies in Linguistics. Cambridge University Press, Cambridge.