Detecting contact-induced semantic shifts: what can embedding-based methods do in practice?

Filip Miletic

Under the supervision of
Anne Przewozny-Desriaux and Ludovic Tanguy

25 October 2021

Outline

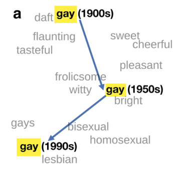
Outline

Background

- Growing interest in semantic change detection methods, but their descriptive contributions remain unclear

English (diachronic, 90y) Top-100 words identified by our method cover all the words attested as real semantic shift in Hamilton et al. (2016b)'s top-10 except the word 'wanting'. Specifically, three attested words, 'gay', 'major' and 'check' are present in our top-10, which also has more interesting words not present in Hamilton et al. (2016b)'s top-10 (1900 vs. 1990): **van** (captain vs. vehicle), **press** (printing vs. places), **oxford** (location vs. university). In addition, interesting words that came up in the top-30 list are the following: **headed** (body part vs. move in a direction), **mystery** (difficulty in understanding vs. book genre).

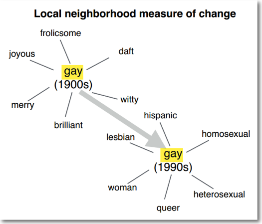
Gonen et al. (2020)



Hamilton et al. (2016b)

We compare the cosine similarity of the word vectors for same words in different years to identify words that have moved significantly in the vector space during that time period. Our model identifies words such as *cell* and *gay* as having changed between 1900–2009. The model additionally identifies words whose change is more subtle. We also analyze the yearly movement of words across the vector space to identify the specific periods during which they changed. The trained word vectors are publicly available.[1]

Kim et al. (2014)



Hamilton et al. (2016a)

In diachronic semantics, especially lexical semantic change, the interaction between use and meaning (crucial for distributional semantics) has traditionally been the focus of interest for theoretical linguists (Deo 2015, Traugott & Dasher 2001). For instance, the contexts of use of *gay* reflect a gradual change in meaning during the twentieth century: from a meaning similar to 'cheerful' to its current predominant use as 'homosexual.' This can be seen in the contrast between the sen-

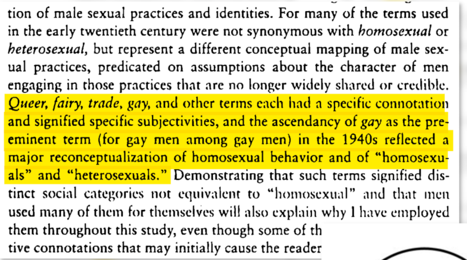
Boleda (2020)

Word	Change year	Description
aeroplane	1919-1920	First use as weapon of war and commercial flights
cinema	1900	movie theatre
computer	1940	digital computer
cool	1964	a way of being
flight	1918	after WWI commercial aviation grows rapidly
gay	1985	recommended for use instead of homosexual
memory	1960	digital memory
mouse	1965	the computer mouse was introduced

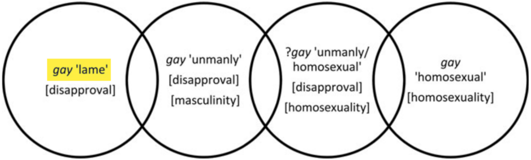
Dubossarsky et al. (2019)

Background

- Growing interest in semantic change detection methods, but their descriptive contributions remain unclear


tion of male sexual practices and identities. For many of the terms used in the early twentieth century were not synonymous with *homosexual* or *heterosexual*, but represent a different conceptual mapping of male sexual practices, predicated on assumptions about the character of men engaging in those practices that are no longer widely shared or credible. *Queer, fairy, trade, gay,* and other terms each had a specific connotation and signified specific subjectivities, and the ascendancy of *gay* as the preeminent term (for gay men among gay men) in the 1940s reflected a major reconceptualization of homosexual behavior and of "homosexuals" and "heterosexuals." Demonstrating that such terms signified distinct social categories not equivalent to "homosexual" and that men used many of them for themselves will also explain why I have employed them throughout this study, even though some of th
tive connotations that may initially cause the reader

Chauncey (1994)



The 'chain' of semantic change from *gay* 'homosexual' to *gay* 'lame'

Robinson (2012)

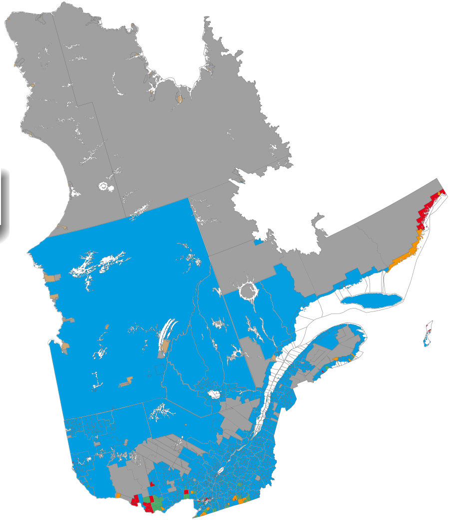
Research question

How useful are these methods
when applied to a precisely defined
descriptive issue?

Contact-induced semantic shifts
in Quebec English

Toronto ‘dishonesty’
Stop misleading. We need truth, not **deception** from our leaders.

Montreal ‘déception’
Big **deception**... you were not present in the Pride Parade in Montreal today. [...] I keep waiting for a breakthrough but Conservatives keep disappointing.



Majority language communities in Quebec
•• English-speaking •• French-speaking
Source: Statistics Canada (2006)

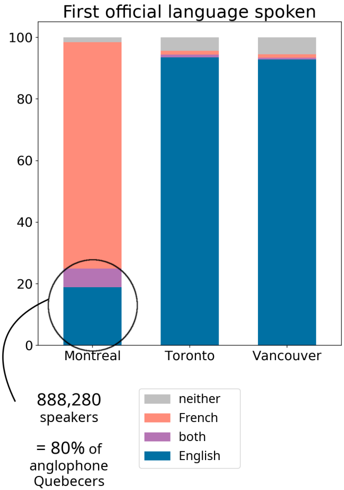
Evaluating semantic change detection

- Quantitative evaluation of top semantic change candidates
  - Synthetic corpora (Shoemark et al., 2019)
  - Lexicographic information (Basile and McGillivray, 2018)
- Quantitative evaluation on binary classification or ranking tasks (e.g. Basile et al., 2020; Schlechtweg et al., 2019, 2020)
- Qualitative evaluation of hand-picked examples (e.g. Hamilton et al., 2016b; Rodda et al., 2017)

- We aim to evaluate the practical usability of these methods and the qualitative nature of the results they output on empirically occurring data

Outline

Corpus



Source: Statistics Canada, 2016 Census.

Corpus design criteria

- A large amount of regional data
- Basic sociolinguistic information
- Limited noise

Corpus creation

- Identification of relevant Twitter users
- Crawl of their entire timelines
- Detailed data filtering

Subcorpus	Users	Tweets	Tokens
Montreal	55 k	11 m	193 m
Toronto	51 k	13 m	223 m
Vancouver	48 k	11 m	213 m
Total	154 k	35 m	629 m

Test set for binary classification

- 40 semantic shifts based on the sociolinguistic literature
  (e.g. Boberg, 2012; Fee, 1991, 2008; Rouaud, 2019)
- 40 stable words of Anglo-Saxon origin
  $\Rightarrow$ limited formal similarity with French
- The classes are balanced for POS and frequency,
  and the presence of target uses is checked in the corpus

Sem. shift	Fr. meaning	Freq.	Stable word
formidable	‘terrific’	1.48	damp
circulation	‘traffic’	2.12	campfire
deceive	‘disappoint’	2.98	cram
souvenir	‘memory’	3.11	hassle
resume	‘summarize’	4.91	arise

Outline

Intro: type-level word embeddings

- Looks like I'll be eating **poutine** in Montreal tonight.
- If you're not using real cheese curds, it's not real **poutine**.
- What type of **poutine** doesn't have gravy?

	eat	cheese	gravy	developer	engineer	system	...
poutine	52	16	5	-	-	-	
fries	24	24	10	-	-	-	
software	4	-	-	129	64	24	
design	-	-	-	6	26	97	
...							

Intro: type-level word embeddings

- Count-based distributional models:
  $\rightarrow$ computation of co-occurrence frequencies

- A more recent solution: predictive models (word embeddings)
- Different methods, including word2vec (Mikolov et al. 2013)
- Different algorithms:
  - **CBOW** (continuous bag of words)
    predicts the probability that a target word is used given a word that appears in its context
  - **SGNS** (skip-gram with negative sampling)
    given a target word, predicts the probability that another word appears in its context

- One vector per word, all senses taken together
- Efficient models good at capturing general trends

Experimental setup

- We use the general method previously shown to be the most robust (Basile et al., 2020; Schlecthweg et al., 2020)
- We experiment with several parameters given the specifics of our setup

Model type	word2vec SGNS
Window size	2, 5, 10
Vector dims	100, 300
Alignment	OP, SR

OP = Orthogonal Procrustes (Hamilton et al., 2016)
SR = Spatial Referencing (Dubossarsky et al., 2019)

- We compute a variation score for each word in the shared vocabulary

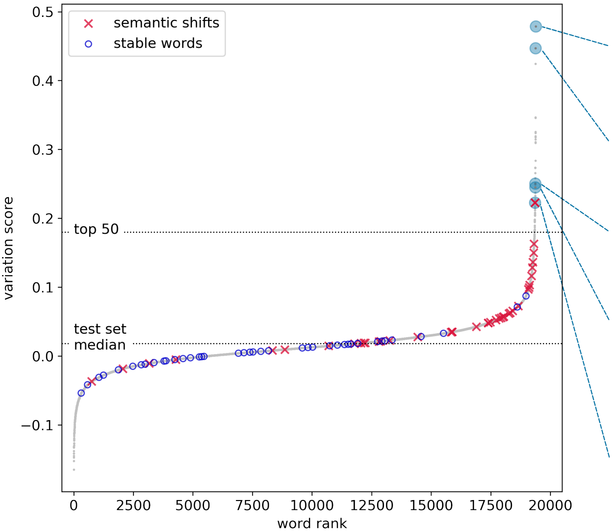
$$var(w) = \frac{CD(w_m, w_t) + CD(w_m, w_v)}{2} - CD(w_t, w_v)$$

Finding the best-performing model

- Classification based on the median score ⇒ accuracy
- Max = 0.8; min = 0.625

Parameters		Accuracy		
		mean	min	max
Dim	100	**0.738**	**0.700**	**0.800**
	300	0.675	0.625	0.750
Win	2	**0.713**	**0.675**	0.750
	5	**0.713**	0.650	**0.800**
	10	0.694	0.625	0.775
Type	OP	0.700	**0.650**	**0.800**
	SR	**0.713**	0.625	0.775

Deploying the model



⇒ precision = 0.02

False positives

pour ‘for’
French homographs in codeswitched tweets

plateau ‘Plateau-Mont-Royal’
local proper noun denoting a Montreal neighborhood

trough ‘through’
misspelling indicative of L2 English

detached ‘separate (house)’
topical variation driven by Toronto and Vancouver

True positive

exposition ‘art exhibition’
previously described case, included in our test set

Outline

Intro: token-level word embeddings

- BERT — `bert-base-uncased`, 12 layers, 768 dims (Devlin et al., 2019)

- Stop misleading. We need truth, not **deception** from our leaders. How can anyone trust you if you won't tell the truth ?

12	0.960	0.113	0.221	-0.027	0.210	-0.058	0.116	0.043	0.110	-0.147	-0.005	-0.083	-0.295	0.716	-0.520	...
11	1.325	-0.124	0.138	0.040	0.325	0.089	0.239	-0.144	0.087	-0.166	-0.269	-0.027	-0.393	1.066	-0.706	...
10	1.157	-0.319	0.189	0.170	0.181	0.197	0.331	-0.153	-0.001	-0.559	-0.083	0.196	-0.443	1.015	-0.854	...
9	1.105	0.087	0.104	0.473	0.234	0.166	0.126	-0.383	0.122	-0.348	0.090	0.437	-0.560	0.915	-0.787	...
avg	1.137	-0.061	0.163	0.164	0.238	0.099	0.203	-0.159	0.080	-0.305	-0.067	0.131	-0.423	0.928	-0.717	...

Intro: token-level word embeddings

- BERT — `bert-base-uncased`, 12 layers, 768 dims (Devlin et al., 2019)

- Stop misleading. We need truth, not **deception** from our leaders. How can anyone trust you if you won't tell the truth ?

1.137	-0.061	0.163	0.164	0.238	0.099	0.203	-0.159	0.080	-0.305	-0.067	0.131	-0.423	0.928	-0.717	...

- Big **deception**... you were not present in the Pride Parade in Montreal today. I keep waiting for a breakthrough but Conservatives keep disappointing.

0.836	-0.279	0.192	0.271	0.740	0.933	0.294	-0.087	-1.019	-0.585	0.364	-0.574	0.380	0.259	-0.767	...

Experimental setup

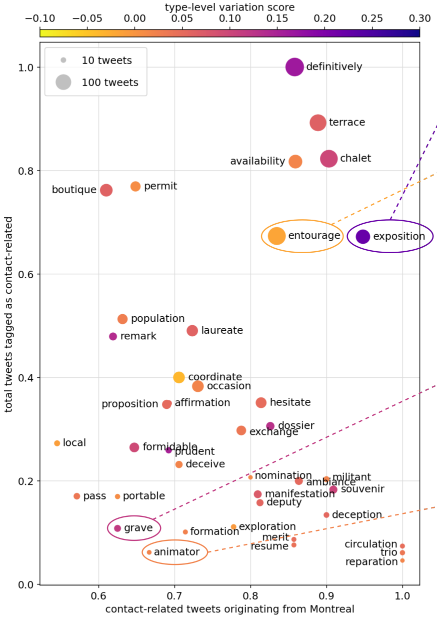
For each of the 40 semantic shifts in the test set:

- Extract all of the word's occurrences (up to 1,000)
- Feed each tweet into BERT
- Get an embedding for each occurrence by averaging over the last 4 hidden layers for the target word
- Cluster the embeddings using affinity propagation
- Retain the clusters with >50% tweets from Montreal
- Manually annotate each cluster (up to 10 per word) for the presence or absence of contact influence

Exploring the clusters

of Janet Werner's upcoming	**exposition**	at the museum . Starting November
Come to admire Laura Granata's	**exposition**	at #CLDV
Such a beautiful	**exposition**	!!! #mbam #art #montrealmuseum
turned into a citizens' area with	**exposition**	space and multipurpose room .
STATION - It's now part of the	**exposition**	events space in Montreal . Its
are showcasing their work at an	**exposition**	hall in Trois-Rivière .
	exposition	d'aquarelles , exhibition of my
	Exposition	du World Press Photo 2016 #photo
	Exposition	en cours - Galerie d'art Stewart Hall

Patterns of semantic change



Contact use is dominant and regionally specific

exposition ‘art exhibition’

prevalent and clear contact use

entourage ‘group of friends’

evident contact use, but related to referential knowledge rather than contextual differences

Contact use is limited and of varying regional specificity

grave ‘serious’

most occurrences are false positives with the French homograph, as in *ce n’est pas grave* ‘it doesn’t matter’

animator ‘group leader’

contact use is present, but rare due to topical effects (thriving animation industry in Montreal)

Outline

Summary

- Diachronic word embedding methods applied to contact-induced semantic shifts in synchrony
  $\Rightarrow$ SOTA-like results on standard quantitative evaluation
- Low precision on the discovery task
  $\Rightarrow$ limited practical value for new semantic shifts
- Token-level embeddings to isolate regionally specific occurrences
  $\Rightarrow$ faster manual analysis, clearer patterns in the data
- New 80-item test set for semantic shifts in Quebec English
- The first quantitative, corpus-based study of this issue

Discussion

- The choice of evaluation is crucial,
  especially when establishing practical usability
- Going back to corpus data remains necessary,
  even when extensive filtering is applied
- Multiple dimensions of variation appear to be at play
  $\Rightarrow$ different types of semantic change should be identified

And now what?

- Almost 500 000 ppl showed up at the Montreal walk for the environment (manifestation). Not only is this walk the biggest for environment in Quebec's history. This walk is the biggest **manifestation** for this week.

- *Ok, so this – I'm gonna say 3 – is awkward, it's an awkward way to say it. But it's super common for me because that is the only way that my partner says "protest".*

References

- Basile, P., Caputo, A., Caselli, T., Cassotti, P., & Varvara, R. (2020). DIACR-Ita @ EVALITA2020: Overview of the EVALITA2020 Diachronic Lexical Semantics (DIACR-Ita) Task. *Proceedings of EVALITA*.
- Basile, P., & McGillivray, B. (2018). Exploiting the Web for Semantic Change Detection. In L. Soldatova, J. Vanschoren, G. Papadopoulos, & M. Ceci (Eds.), *Discovery Science* (Vol. 11198, pp. 194–208).
- Boberg, C. (2012). English as a minority language in Quebec. *World Englishes*, 31(4), 493–502.
- Boleda, G. (2020). Distributional semantics and linguistic theory. *Annual Review of Linguistics*, 6, 213–234.
- Chauncey, G. (1994). *Gay New York: Gender, urban culture, and the makings of the gay male world, 1890-1940.* Basic Books.
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional Transformers for language understanding. *Proceedings of NAACL-HLT*, 4171–4186.
- Dubossarsky, H., Hengchen, S., Tahmasebi, N., & Schlechtweg, D. (2019). Time-Out: Temporal referencing for robust modeling of lexical semantic change. *Proceedings of ACL*, 457–470.

References

- Fee, M. (2008). French borrowing in Quebec English. *Anglistik: International Journal of English Studies*, 19(2), 173–188.
- Fee, M. (1991). Frenglish in Quebec English newspapers. *Papers of the Fifteenth Annual Meeting of the APLA*, 12–23.
- Gonen, H., Jawahar, G., Seddah, D., & Goldberg, Y. (2020). Simple, interpretable and stable method for detecting words with usage change across corpora. *Proceedings of ACL*, 538–555.
- Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016a). Cultural shift or linguistic drift? Comparing two computational measures of semantic change. *Proceedings of EMNLP*, 2116–2121.
- Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016b). Diachronic word embeddings reveal statistical laws of semantic change. *Proceedings of ACL*, 1489–1501.
- Kim, Y., Chiu, Y.-I., Hanaki, K., Hegde, D., & Petrov, S. (2014). Temporal analysis of language through neural language models. *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, 61–65.

References

- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *Proceedings of ICLR*.
- Robinson, J. A. (2012). A *gay* paper: Why should sociolinguistics bother with semantics? *English Today*, 28(4), 38–54.
- Rodda, M. A., Lenci, A., & Senaldi, M. S. G. (2017). Panta rei: Tracking semantic change with distributional semantics in Ancient Greek. *Italian Journal of Computational Linguistics*, 3(1), 11–24.
- Rouaud, J. (2019). *Lexical and phonological integration of French loanwords into varieties of Canadian English since the seventeenth century*. Doctoral dissertation, Université Toulouse - Jean Jaurès.
- Schlechtweg, D., Hätty, A., Del Tredici, M., & Schulte im Walde, S. (2019). A wind of change: Detecting and evaluating lexical semantic change across times and domains. *Proceedings of ACL*, 732–746.
- Schlechtweg, D., McGillivray, B., Hengchen, S., Dubossarsky, H., & Tahmasebi, N. (2020). SemEval-2020 task 1: Unsupervised lexical semantic change detection. *Proceedings of SemEval*, 1–23.

References

- Shoemark, P., Liza, F. F., Nguyen, D., Hale, S., & McGillivray, B. (2019). Room to Glo: A systematic comparison of semantic change detection approaches with word embeddings. *Proceedings of EMNLP-IJCNLP*, 66–76.