

Exploring Code-switching and Borrowing Using Word Vectors

Steven Coats

English Philology, University of Oulu, Finland
steven.coats@oulu.fi

14th ESSE Conference, Brno
September 1st, 2018

Outline

1. Code-switching and borrowing German-English in a Twitter data set
2. Data collection
3. Word vectors and embeddings
4. Identification of code-switches and borrowings
5. Tracing changes in meaning and visualizing borrowings

Slides for the presentation are on my homepage at <https://cc.oulu.fi/~scoats>

Code-switching and borrowing German-English

Code-switching: Use of two or more languages in an utterance/turn

Lexical borrowing: Use of a second-language lexical item as a lone element within an L1 matrix (Myers-Scotton 1997)

Code-switching and borrowing German-English

Code-switching: Use of two or more languages in an utterance/turn

Lexical borrowing: Use of a second-language lexical item as a lone element within an L1 matrix (Myers-Scotton 1997)

German-English code-switched tweet

- Der Dauerregen ist zuviel für unseren Platz, daher müssen wir das heutige Spiel absagen. - Land under, the rain won today. So it's a rainout. (The unrelenting rain is too much for our pitch, so we have to cancel today's game...)

Code-switching and borrowing German-English

Code-switching: Use of two or more languages in an utterance/turn

Lexical borrowing: Use of a second-language lexical item as a lone element within an L1 matrix (Myers-Scotton 1997)

German-English code-switched tweet

- Der Dauerregen ist zuviel für unseren Platz, daher müssen wir das heutige Spiel absagen. - Land under, the rain won today. So it's a rainout. (The unrelenting rain is too much for our pitch, so we have to cancel today's game...)

German tweets with English borrowings

- @user Gibt es schon einen Ort, also location für das Treffen? (@user *Is there already a place, that is, a location for the meet-up?*)

Code-switching and borrowing German-English

Code-switching: Use of two or more languages in an utterance/turn

Lexical borrowing: Use of a second-language lexical item as a lone element within an L1 matrix (Myers-Scotton 1997)

German-English code-switched tweet

- Der Dauerregen ist zuviel für unseren Platz, daher müssen wir das heutige Spiel absagen. - Land under, the rain won today. So it's a rainout. (The unrelenting rain is too much for our pitch, so we have to cancel today's game...)

German tweets with English borrowings

- @user Gibt es schon einen Ort, also location für das Treffen? (@user Is there already a place, that is, a location for the meet-up?)
- Bahhhh draußen scheint voll die Sonne. EKELEHAFT. Erinnert mich daran, dass ich noch immer nicht in shape für den Sommer bin. 🤢 (Bahhhh the sun is totally shining outside. DISGUSTING. Reminds me that I'm still not in shape for summer. 🤢)

Code-switching and borrowing German-English in a Twitter data set

- Lexical borrowings can undergo semantic shift: many Anglicisms in German have meanings incommensurate with their meanings in English (Onysko 2007)

Code-switching and borrowing German-English in a Twitter data set

- Lexical borrowings can undergo semantic shift: many Anglicisms in German have meanings incommensurate with their meanings in English (Onysko 2007)

@user Danke für den like ♥ (@user Thanks for the like! ♥)

Code-switching and borrowing German-English in a Twitter data set

- Lexical borrowings can undergo semantic shift: many Anglicisms in German have meanings incommensurate with their meanings in English (Onysko 2007)

@user Danke für den like 💜 (@user Thanks for the like! 💜)

@user yo danke für deinen follow bro! ich weiß das zu schätzen. #RealHipHop
(@user yo thanks for your follow bro! I appreciate it. #RealHipHop)

Code-switching and borrowing German-English in a Twitter data set

- Lexical borrowings can undergo semantic shift: many Anglicisms in German have meanings incommensurate with their meanings in English (Onysko 2007)

@user Danke für den like 💜 (@user Thanks for the like! 💜)

@user yo danke für deinen follow bro! ich weiß das zu schätzen. #RealHipHop
(@user yo thanks for your follow bro! I appreciate it. #RealHipHop)

- How can we trace the semantic shift of English borrowings in German?

Code-switching and borrowing German-English in a Twitter data set

- Lexical borrowings can undergo semantic shift: many Anglicisms in German have meanings incommensurate with their meanings in English (Onysko 2007)

@user Danke für den like 💜 (@user Thanks for the like! 💜)

@user yo danke für deinen follow bro! ich weiß das zu schätzen. #RealHipHop

(@user yo thanks for your follow bro! I appreciate it. #RealHipHop)

- How can we trace the semantic shift of English borrowings in German?
- By using word embeddings from large corpora that contain **borrowings**

Data collection

- 653,457,659 tweets with *place* metadata collected globally from the Twitter Streaming API from November 2016 until June 2017

Data collection

- 653,457,659 tweets with *place* metadata collected globally from the Twitter Streaming API from November 2016 until June 2017
- 60,683 authors of at least one German-language tweet with place metadata from Germany, Austria or Switzerland identified and all of their tweets/most recent 3,250 tweets (whichever was larger) downloaded from REST API in April 2018

Data collection

- 653,457,659 tweets with *place* metadata collected globally from the Twitter Streaming API from November 2016 until June 2017
- 60,683 authors of at least one German-language tweet with place metadata from Germany, Austria or Switzerland identified and all of their tweets/most recent 3,250 tweets (whichever was larger) downloaded from REST API in April 2018
- Retain tweets in German according to Twitter's metadata

Data collection

- 653,457,659 tweets with *place* metadata collected globally from the Twitter Streaming API from November 2016 until June 2017
- 60,683 authors of at least one German-language tweet with place metadata from Germany, Austria or Switzerland identified and all of their tweets/most recent 3,250 tweets (whichever was larger) downloaded from REST API in April 2018
- Retain tweets in German according to Twitter's metadata
- 36,240,530 (59.3%) of tweets in German = 534,211,366 tokens

Identifying sentences with code-switching/borrowing

- Tokenize all tweets, remove punctuation, URLs, user names, hashtags, emoji

Identifying sentences with code-switching/borrowing

- Tokenize all tweets, remove punctuation, URLs, user names, hashtags, emoji
- Match each word in each message with large German and English word lists

Identifying sentences with code-switching/borrowing

- Tokenize all tweets, remove punctuation, URLs, user names, hashtags, emoji
- Match each word in each message with large German and English word lists

Anyone in Oberwart, der am Abend nach Wien fährt & mir was abholen und mitbringen könnte? Biete Aufwandsentschädigung & ewige Dankbarkeit! *(Anyone in Oberwart who is driving this evening to Vienna and can pick up something and bring it to me? I offer reimbursement for the effort & eternal thanks!)*

- 1 English word of 19: Borrowing

Identifying sentences with code-switching/borrowing

- Tokenize all tweets, remove punctuation, URLs, user names, hashtags, emoji
- Match each word in each message with large German and English word lists

Anyone in Oberwart, der am Abend nach Wien fährt & mir was abholen und mitbringen könnte? Biete Aufwandsentschädigung & ewige Dankbarkeit! *(Anyone in Oberwart who is driving this evening to Vienna and can pick up something and bring it to me? I offer reimbursement for the effort & eternal thanks!)*

- 1 English word of 19: Borrowing

Seit Anfang dieses Jahres habe ich soooo oft heißhunger auf Asiatisches Essen. This year I have so often cravings for asian #food. 🍱🍜🍝🍲

- 9 English words of 21: Code-switching

Identifying sentences with code-switching/borrowing

- Tokenize all tweets, remove punctuation, URLs, user names, hashtags, emoji
- Match each word in each message with large German and English word lists

Anyone in Oberwart, der am Abend nach Wien fährt & mir was abholen und mitbringen könnte? Biete Aufwandsentschädigung & ewige Dankbarkeit! *(Anyone in Oberwart who is driving this evening to Vienna and can pick up something and bring it to me? I offer reimbursement for the effort & eternal thanks!)*

- 1 English word of 19: Borrowing

Seit Anfang dieses Jahres habe ich soooo oft heißhunger auf Asiatisches Essen. This year I have so often cravings for asian #food. 🇩🇪🇸🇯🇵🇰🇷🇻🇳

- 9 English words of 21: Code-switching
- We can distinguish code-switching from borrowing on the basis of counts of English and German types

Lexica

English words:

- 236,736 English words from NLTK (Bird et al. 2009)

Lexica

English words:

- 236,736 English words from NLTK (Bird et al. 2009)

German words:

- 50k most frequent German words (Dave 2017, Lison & Tiedemann 2016)

Lexica

English words:

- 236,736 English words from NLTK (Bird et al. 2009)

German words:

- 50k most frequent German words (Dave 2017, Lison & Tiedemann 2016)

Corpus for word embeddings

- Tweets with least 8 tokens, of which one or two are English words from the list
- 2,488,673 of 36m tweets have borrowings (many more have code-switches!)

Word embeddings

- Distributional hypothesis (Harris 1968): Word meanings correspond to their aggregate contexts of use

Word embeddings

- Distributional hypothesis (Harris 1968): Word meanings correspond to their aggregate contexts of use
- Collocational information can be represented with vectors of co-occurrence probabilities within a word span

Word embeddings

- Distributional hypothesis (Harris 1968): Word meanings correspond to their aggregate contexts of use
- Collocational information can be represented with vectors of co-occurrence probabilities within a word span
- Similarity of collocational context (and thus meaning) for any two types in a data set (corpus) can then be quantified

Word embeddings

- Distributional hypothesis (Harris 1968): Word meanings correspond to their aggregate contexts of use
- Collocational information can be represented with vectors of co-occurrence probabilities within a word span
- Similarity of collocational context (and thus meaning) for any two types in a data set (corpus) can then be quantified
- Word2Vec algorithm (Mikolov et al. 2013) in Gensim (Řehůřek and Sojka 2010), 5-token co-occurrence span, minimum of 20 occurrences, 200-dimensional vectors

Word embeddings

- Distributional hypothesis (Harris 1968): Word meanings correspond to their aggregate contexts of use
- Collocational information can be represented with vectors of co-occurrence probabilities within a word span
- Similarity of collocational context (and thus meaning) for any two types in a data set (corpus) can then be quantified
- Word2Vec algorithm (Mikolov et al. 2013) in Gensim (Řehůřek and Sojka 2010), 5-token co-occurrence span, minimum of 20 occurrences, 200-dimensional vectors
- Vectors for 51,336 types (mostly German words, but many English words as well)

Cosine similarity

- For word types a and b , corresponding to vectors $\mathbf{a}$ and $\mathbf{b}$:

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\| \|\mathbf{b}\|} = \frac{\sum_{i=1}^n a_i b_i}{\sqrt{\sum_{i=1}^n a_i^2} \sqrt{\sum_{i=1}^n b_i^2}}$$

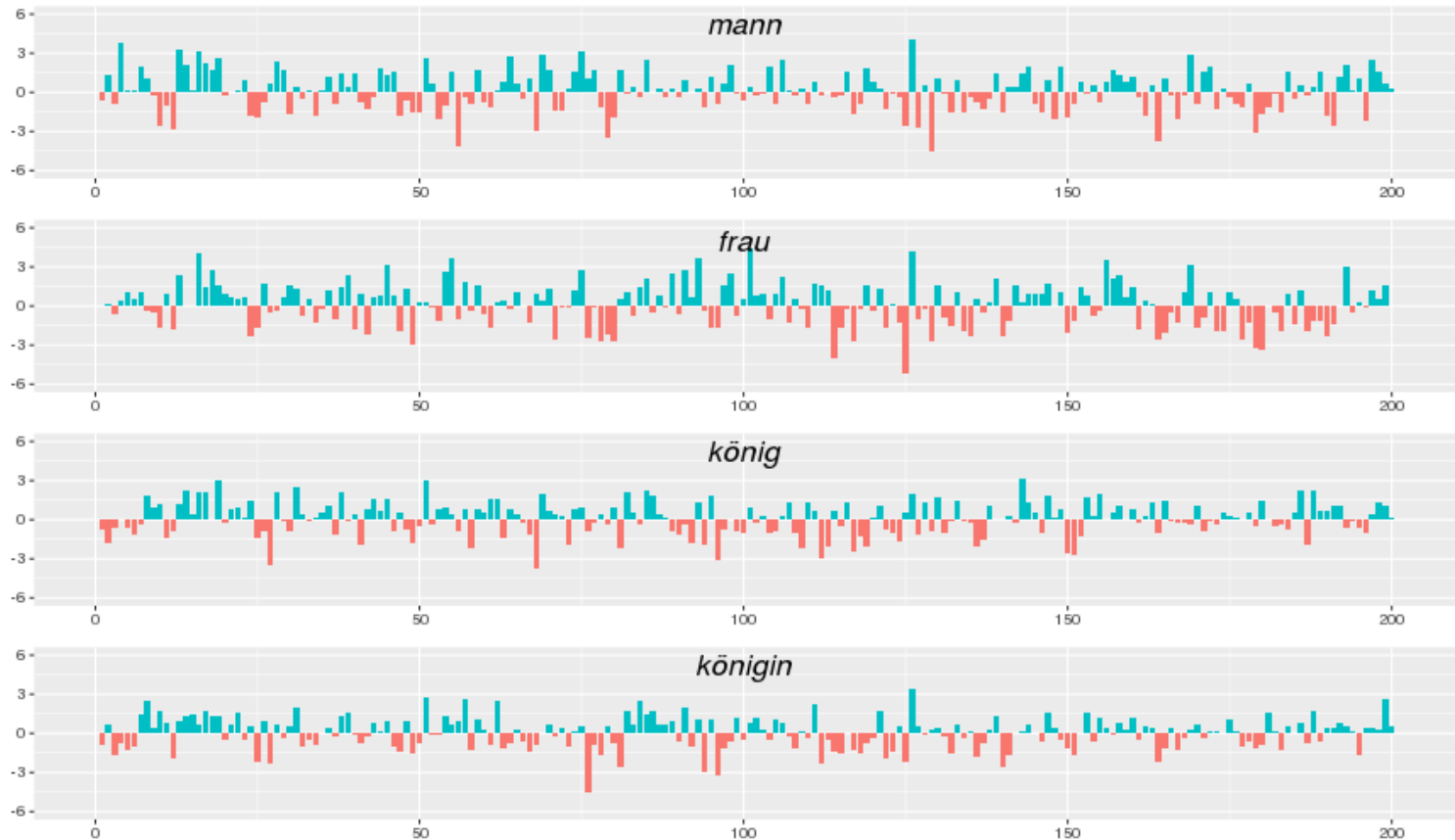
Cosine similarity

- For word types a and b , corresponding to vectors $\mathbf{a}$ and $\mathbf{b}$:

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\| \|\mathbf{b}\|} = \frac{\sum_{i=1}^n a_i b_i}{\sqrt{\sum_{i=1}^n a_i^2} \sqrt{\sum_{i=1}^n b_i^2}}$$

- Value can range from -1 (types never occur in same context, meanings probably very different) to 1 (types occur in exact same contexts, meanings probably very similar)

Vectors for *mann*, *frau*, *könig*, and *königin*



Cosine similarity

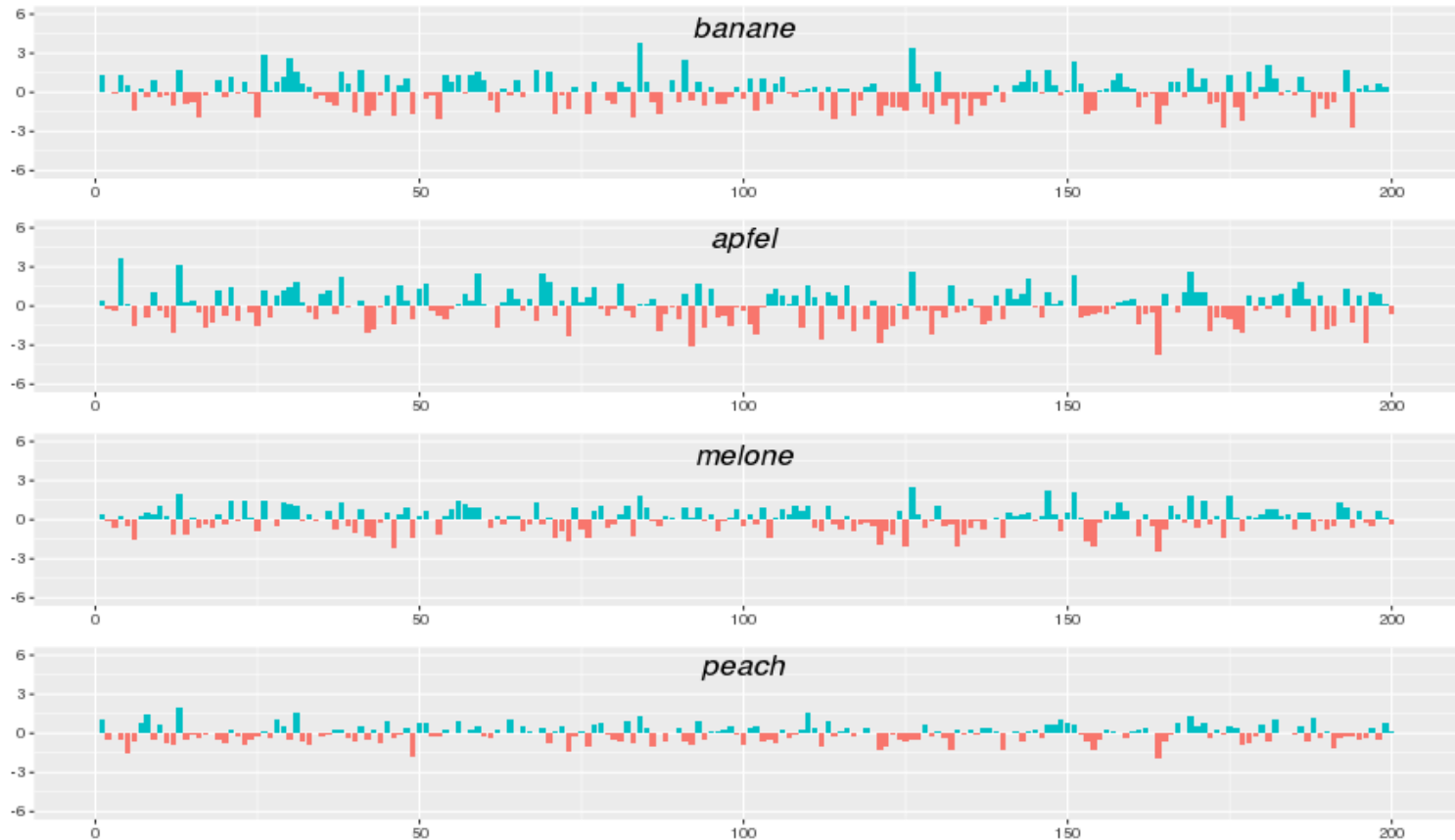
	frau	könig	königin	mann
frau	1.000000	0.154443	0.489897	0.576267
könig	0.154443	1.000000	0.570677	0.341579
königin	0.489897	0.570677	1.000000	0.276945
mann	0.576267	0.341579	0.276945	1.000000

- Cosine similarity preserves semantic relations

Cosine similarity

	apps	frau	könig	königin	mann
apps	1.000000	-0.015972	-0.203364	-0.199718	-0.055016
frau	-0.015972	1.000000	0.154443	0.489897	0.576267
könig	-0.203364	0.154443	1.000000	0.570677	0.341579
königin	-0.199718	0.489897	0.570677	1.000000	0.276945
mann	-0.055016	0.576267	0.341579	0.276945	1.000000

Vectors for *banane*, *apfel*, *melone*, and *peach*



Cosine similarity

	apfel	banane	melone	peach
apfel	1.000000	0.588266	0.527162	0.305676
banane	0.588266	1.000000	0.783086	0.324065
melone	0.527162	0.783086	1.000000	0.431262
peach	0.305676	0.324065	0.431262	1.000000

Cosine similarity

	apfel	banane	melone	peach
apfel	1.000000	0.588266	0.527162	0.305676
banane	0.588266	1.000000	0.783086	0.324065
melone	0.527162	0.783086	1.000000	0.431262
peach	0.305676	0.324065	0.431262	1.000000

- If using data that contains many borrowings, some semantic relations are preserved cross-linguistically

Research questions

- Which English borrowings have meanings closest to or furthest from their German translations (i.e. have undergone semantic shift)
- What role do frequency effects play?
- How can we visualize the meanings of the Anglicisms in the German lexicon?

Change in meaning of borrowed words

- Identify all English words that occur >300 times in the 2.5m borrowing tweets
- Translate the words to German using Google Translate API (507 types)
- Measure cosine similarity of English borrowings to their German-language lexical equivalents
- Regress cosine similarity with log odds ratio of English word frequency to German word frequency

Change in meaning of borrowed words

Show entries Search:

	eng	ger	cos_sim	freq_eng	freq_ger	log_or
1	monitoring	überwachung	0.345	2131	327	1.874
2	module	modul	0.599	694	330	0.743
3	active	aktiv	0.178	892	1765	-0.682
4	mavericks	einzelgänger	-0.075	443	8	4.014
5	released	freigegeben	0.522	648	190	1.227
6	icons	symbole	0.527	926	89	2.342
7	phone	telefon	0.449	4282	1313	1.182

Showing 1 to 507 of 507 entries

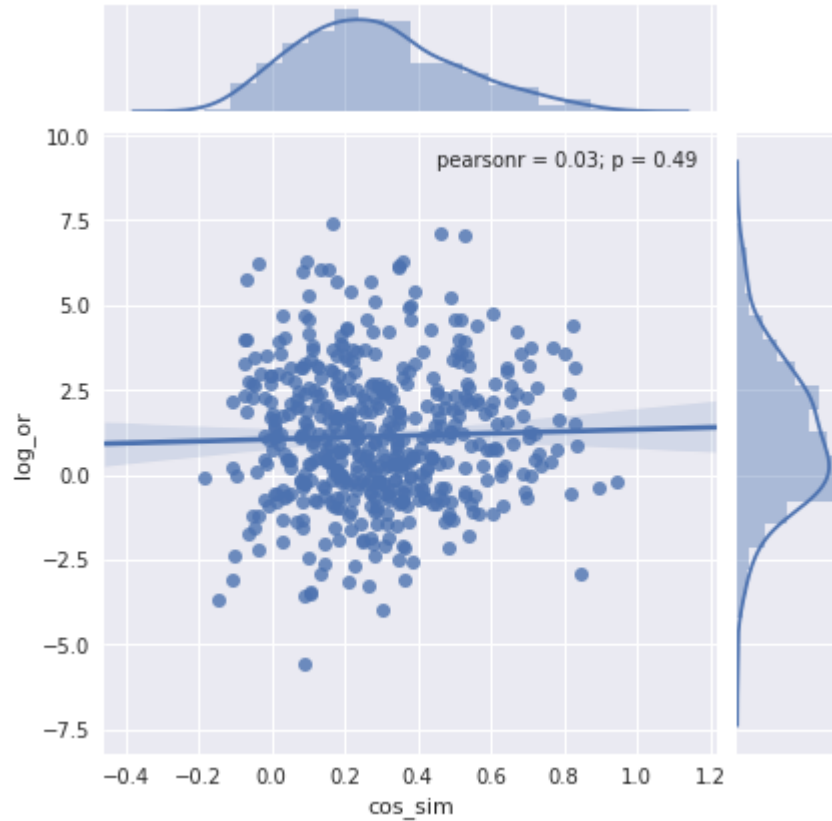
Previous

1

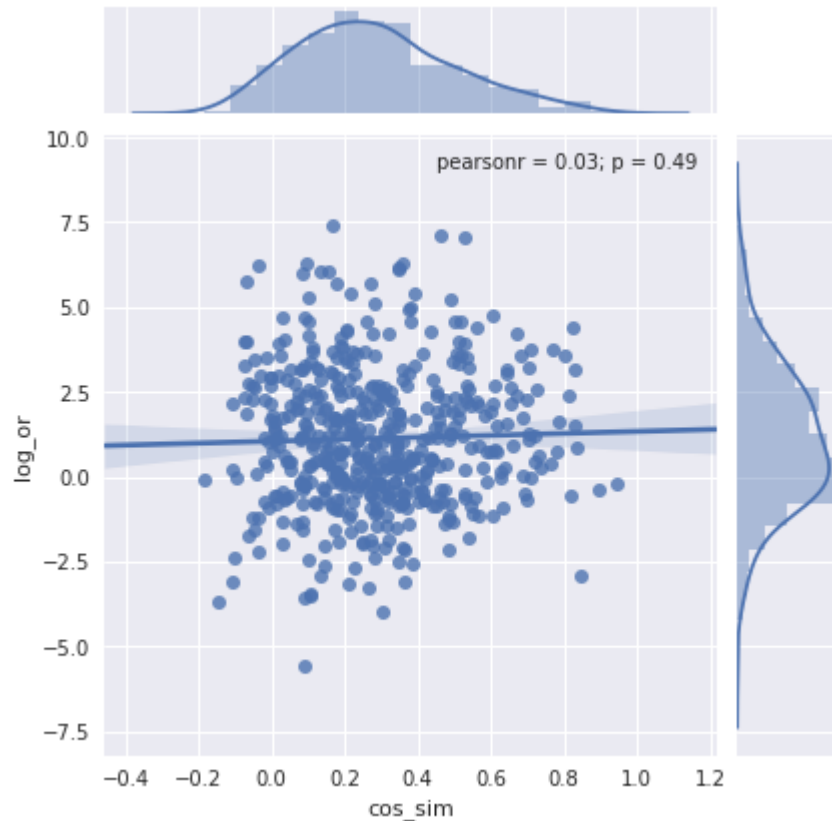
Next

- English-German lexical pairs with high cosine similarity values: English borrowing means more or less the same as the German lexical item
- English-German pairs with low cosine similarity values: Borrowing has undergone semantic shift

Change in meaning of borrowed words

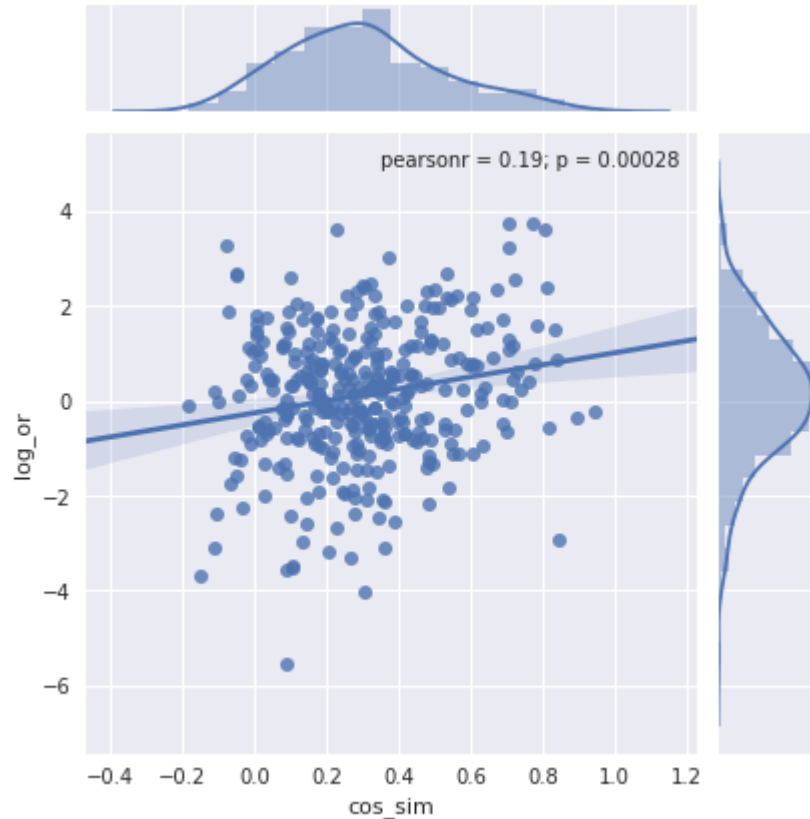


Change in meaning of borrowed words



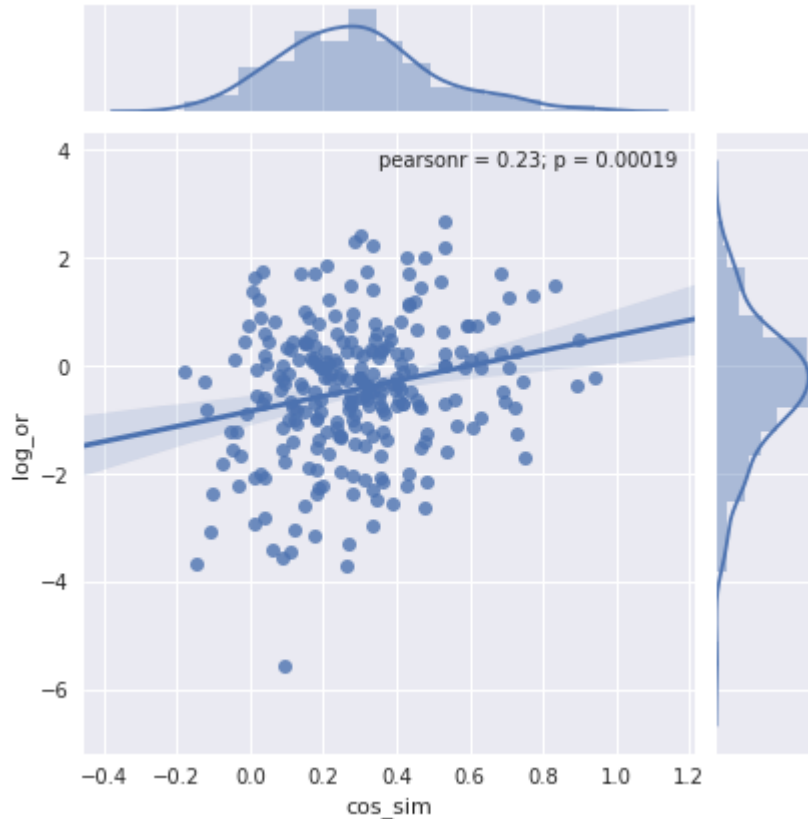
- Very weak correlation between frequency of English-German log odds and cosine similarity... **BUT**

Change in meaning of borrowed words



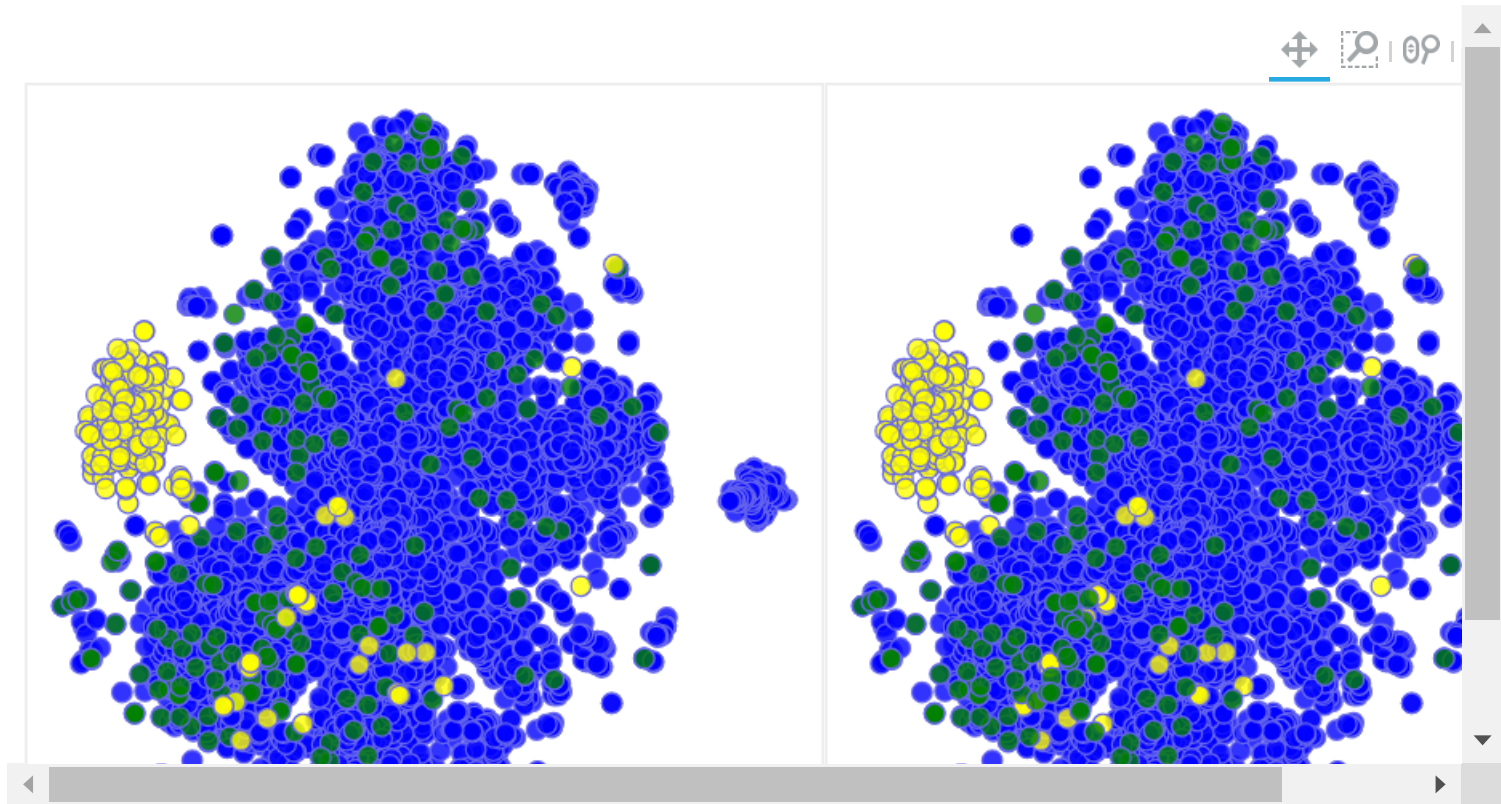
- Only types for which the German lexemes occur at least 100 times

Change in meaning of borrowed words



- Only types for which the English and German lexemes occur at least 300 times

3,845 common German words and 227 common English borrowings



200 dimensions reduced to 2 with t-SNE (van der Maaten & Hinton 2008)

Blue: German word types. Yellow: English word types. Green: German lexical equivalents of the English types

Summary

- Large corpora of messages containing borrowings or code-switches can be created from social media (Twitter)
- Word list methods can be used to identify messages with borrowings or code-switches
- Word embeddings may be able to shed light on semantic shift of lexical borrowings

Issues to address

- "Contaminated" word lists
 - Manually correct word lists, use stemmers
- Inaccurate translations (polysemy, case syncretism)

Issues to address

- "Contaminated" word lists
 - Manually correct word lists, use stemmers
- Inaccurate translations (polysemy, case syncretism)
- Use "artificial codeswitching" sentences to train the embedding models (Gouws and Søgaard 2015, Wick et al. 2016)
- Use WordNet and German WordNet to more accurately compare semantic fields
- Consider English borrowings in more than one language (cf. Görlach 2001)

Thank you! Dekuji! Danke!

References I

- Bird, S., Loper, E. and Klein, E. 2009. *Natural Language Processing with Python*. Newton, MA: O'Reilly.
- Dave, H. (2016). *FrequencyWords*.
- van der Maaten, L. and Hinton, G. 2008. Visualizing high-dimensional data using t-SNE. *Journal of Machine Learning Research* 9: 2579–2605.
- Faruqui, M., and Padó, S. 2010. Training and evaluating a German named entity recognizer with semantic generalization. In *Proceedings of Konvens 2010*. Saarbrücken, Germany.
- Görlach, M. (ed.). 2001. *A Dictionary of European Anglicisms*. Cambridge, UK: Cambridge University Press.
- Gouws, S. and Søgaard, A. 2015. Simple task-specific bilingual word embeddings. *Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL*, 1386–1390.
- Harris, Z. 1968. *Mathematical Structures of Language*. New York: Interscience.
- Lison, P. and Tiedemann, J. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, 923–929.
- Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J. and McClosky, D. 2014. The Stanford CoreNLP natural language processing toolkit. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 55–60.
- Mikolov, T., Yih, W. and Zweig, G. 2013. Linguistic regularities in continuous space word representations. *Proceedings of HLT-NAACL 13*, 746–751.

References II

- Myers-Scotton, C. 1997. Code-switching. In *The Handbook of Sociolinguistics*, ed. Florian Coulmas, 212–237. Oxford: Blackwell.
- Onysko, A. 2007. *Anglicisms in German: Borrowing, lexical productivity, and written codeswitching*. Berlin/New York: De Gruyter.
- Řehůřek, R. and Sojka, P. 2010. Software framework for topic modelling with large corpora. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50.
- Roesslein, J. 2015. [Tweepy](#).
- Schmid, H., Fitschen, A. and Heid, U. (2004). SMOR: A German computational morphology covering derivation, composition, and inflection. *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, 1263–1266.
- Schreiber, J. 2017. [A German word list for GNU Aspell](#).
- Wick, M., Kanani, P. and Pocock, A. 2016. Minimally-constrained multilingual embeddings via artificial code-switching. *Proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI-16)*, 2849–2855.