

Florent Perek

Distributional semantic plots

**A data-driven approach to recent change
in syntactic productivity**

Syntactic productivity

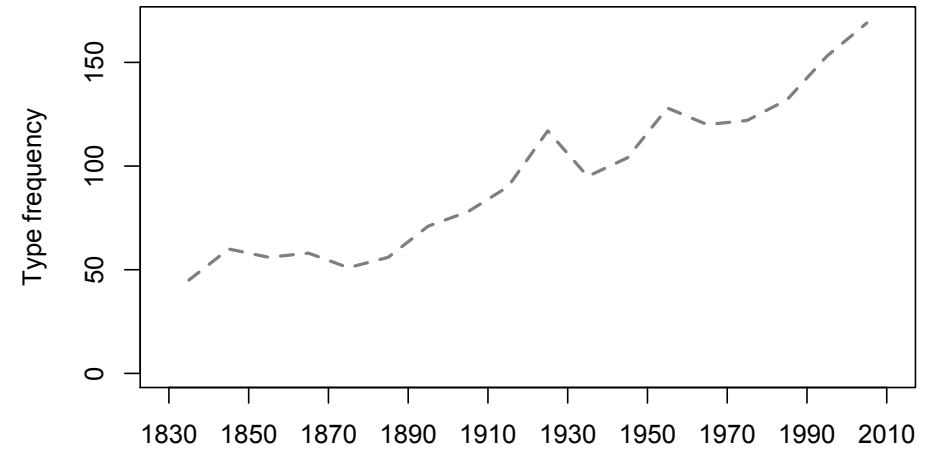
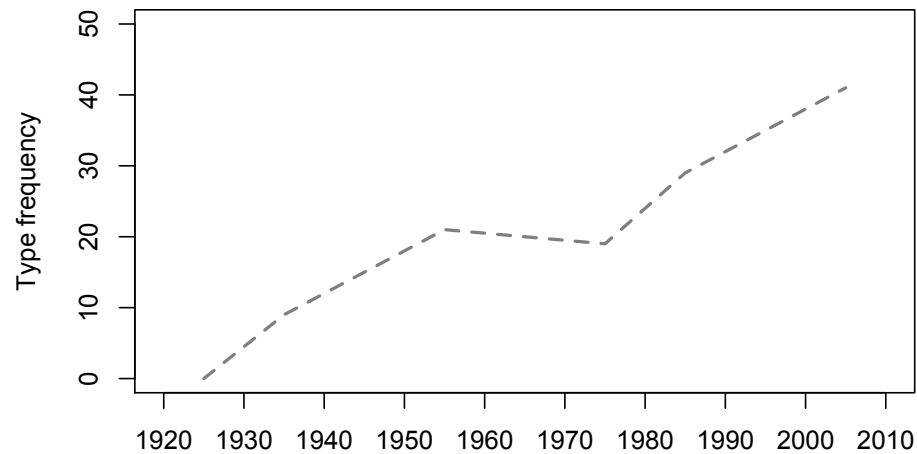
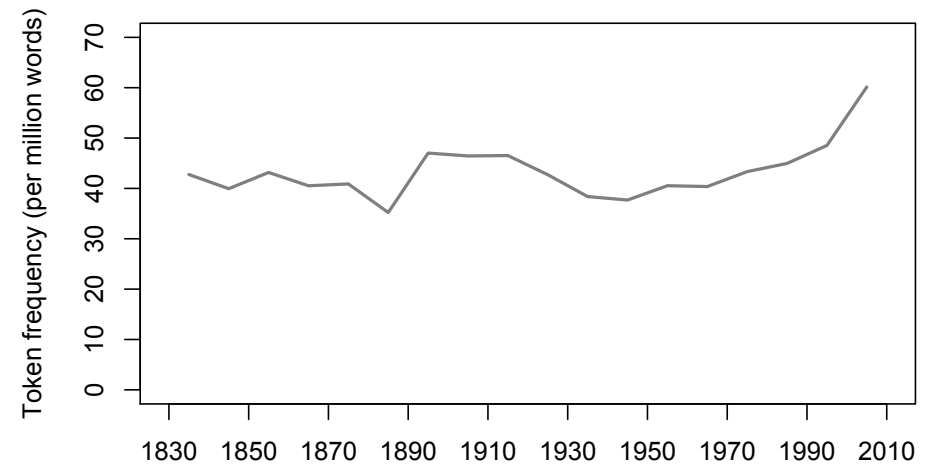
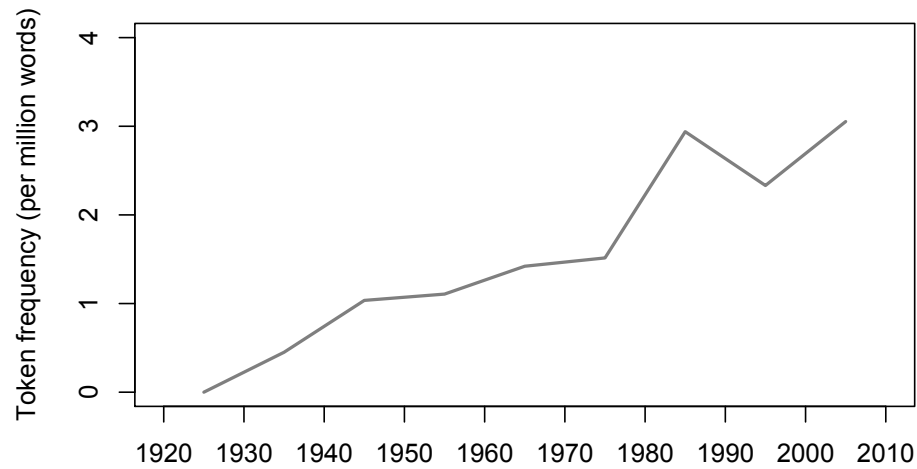
- Morphological productivity
 - Property of a word formation process to coin new words
 - E.g., *nouniness*: *noun* + *-y* + *-ness* (Ross 1973)
- Syntactic productivity
 - Syntactic constructions are similarly able to combine words in creative ways
 - E.g., *He sneezed the napkin off the table* (Goldberg 1995)

Syntactic productivity in diachrony

- The lexical distribution of syntactic constructions may vary over time
- For instance, the *way*-construction (Israel 1996)
 - Verbs of physical actions attested from the 16th century
They hacked their way through the jungle.
 - Abstract means of reaching a goal only appear in the 19th century
She typed her way to a promotion.

Token and type frequencies

- Token frequency: how often a construction is used?
- Type frequency: with how many different lexical items?
- Example: verbs in the *hell*-construction and the *way*-construction
 - The *hell*-construction (Perek 2014, to appear)
[V *the hell out of* NP]
You scared the hell out of me.
I enjoyed the hell out of that show!
 - The *way*-construction (Goldberg 1995, Israel 1996)
[V poss way PP]
Their hacked they way through the jungle.
She typed her way to a promotion.



hell-construction

[V *the hell out of* NP]

e.g., *I enjoyed the hell out of that show!*

way-construction

[V poss *way* PP]

e.g., *Their hacked they way through the jungle.*



Type frequency

- Type frequency reflects the lexical range of a construction
- But it is a purely quantitative measure of lexical diversity
 - No account of how *different* items are
 - Coarse indication of productivity
 - Must take into account semantic diversity
- Questions:
 - What kinds of verbs joined the distribution?
 - Did it become more semantically diverse?
 - Are there particular semantic domains favored by the construction?

How to operationalize semantic similarity?

- Introspection
 - Subjective and time-consuming
 - Does not lend itself to quantification
- Semantic norming (Bybee & Eddington 2006)
 - Similarity judgments provided by a group of speakers
 - Also time-consuming and constraining
 - Limited in terms of the number of lexical items considered
- Proposal: using distributional semantics to measure semantic similarity

Distributional semantics

“You shall know a word by the company it keeps.” (Firth 1957: 11)

- Words that occur in similar contexts tend to have related meanings (Miller & Charles 1991)
- Therefore, a way to characterize the meaning of words is through their distribution in large corpora
- Semantic similarity is quantified by similarity in distribution

Distributional semantic model

- “Bag of word” approach
 - Extraction of lexical collocates of each verb in a 5-word window from a large corpus
 - Each verb is assigned an array of numerical values (a vector) derived from co-occurrence frequencies
 - Vectors interpreted as dimensions in a high-dimensional space
- Semantic similarity measured by similarity between vectors
- The more frequent collocates are shared by two words, the more similar they will be considered

Visualization

- Output: pairwise distances between verbs
- Define a semantic space that can be plotted for visualization
 - By means of *t*-Distributed Stochastic Neighbor Embedding algorithm (*t*-SNE) (Van der Maaten & Hinton 2008)
 - Places objects in a 2-dimensional space such that the between-object distances are preserved as well as possible
 - Superior to multidimensional scaling (MDS) for dense spaces with many dimensions
 - Distance matrix converted to a set of coordinates for each verb
- Semantic domain of the construction plotted for different time periods

Example 1: the *hell*-construction

- Verb *the hell out of* NP
- “Intensifying” function
- Recent construction: first instances in the COHA from the 1930s

You scared the hell out of me!

Then I [...] avoided the hell out of his presence

But you drove the hell out of it!

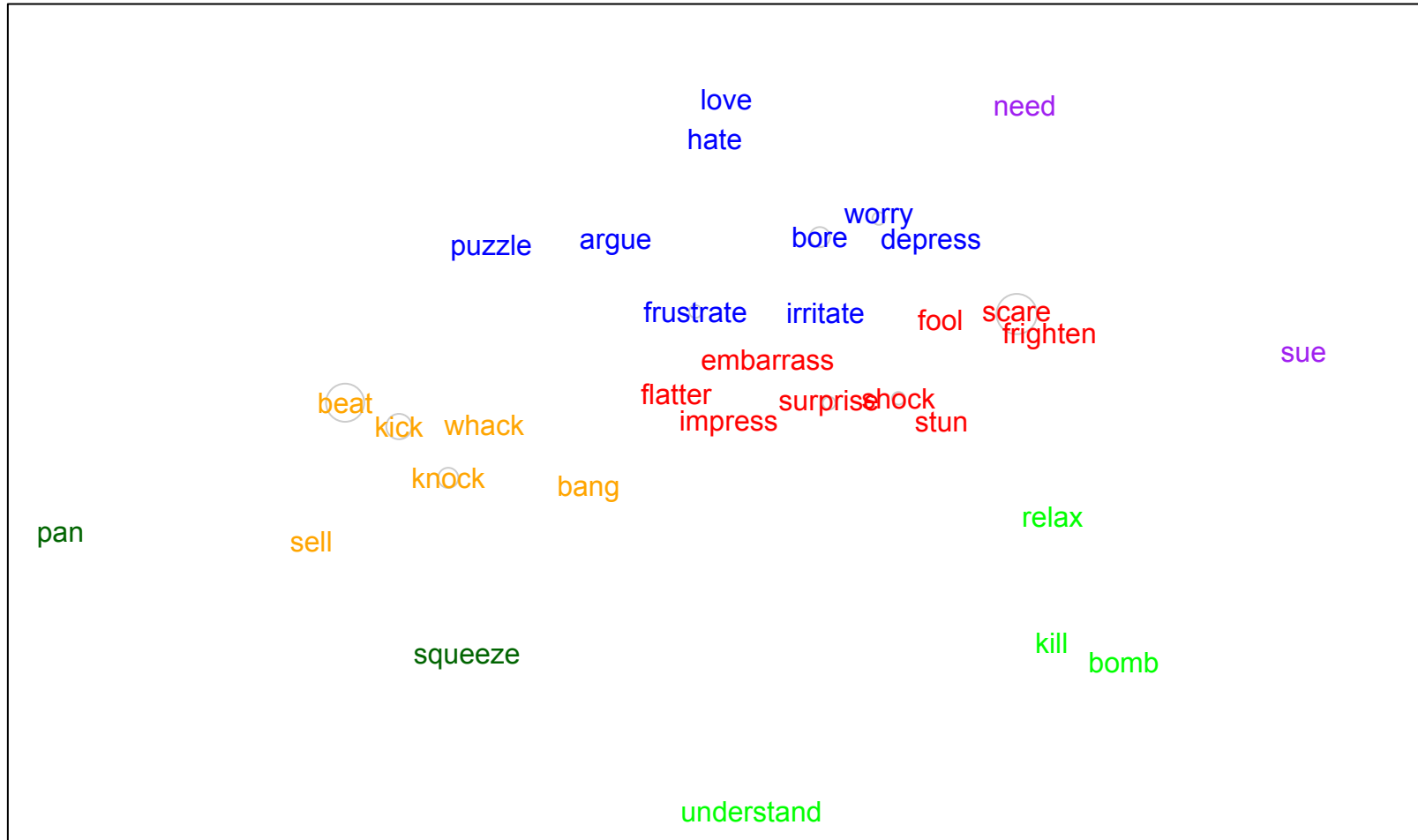
I've been listening the hell out of your tape.

I voice the hell out of 'b' (Phillip Hamrick at GURT 2014, Georgetown)

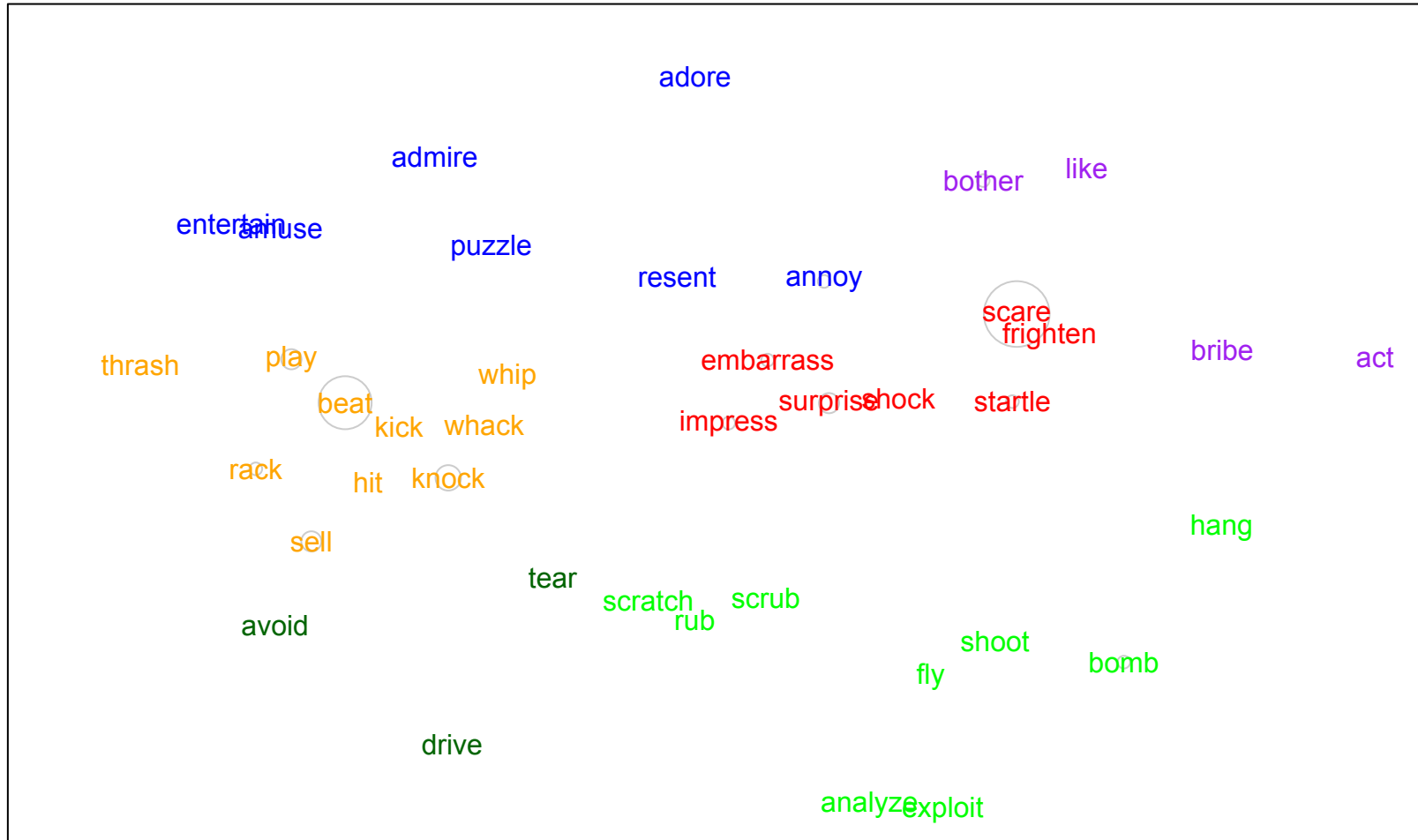
1930s-1940s



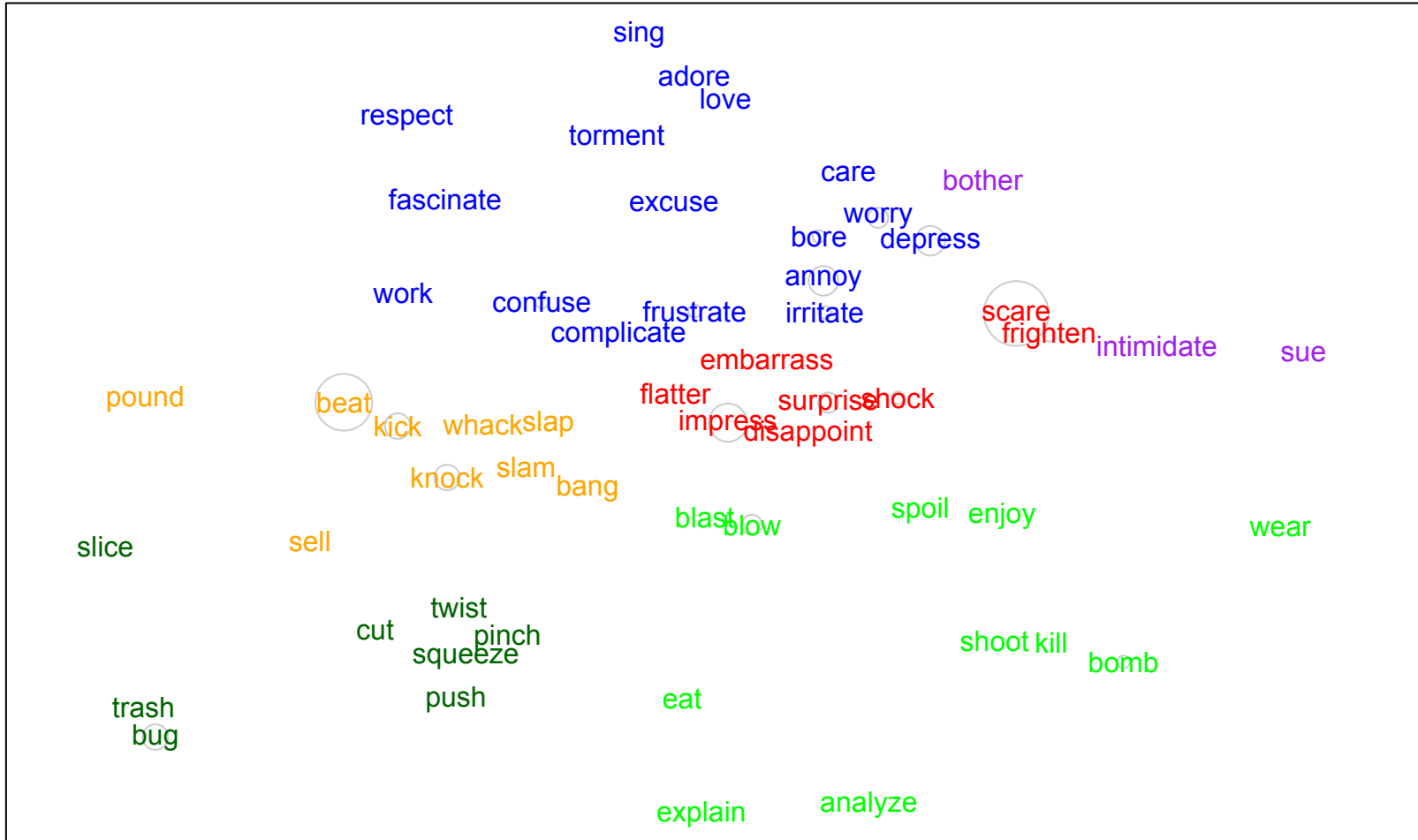
1950s-1960s



1970s-1980s



1990s-2000s



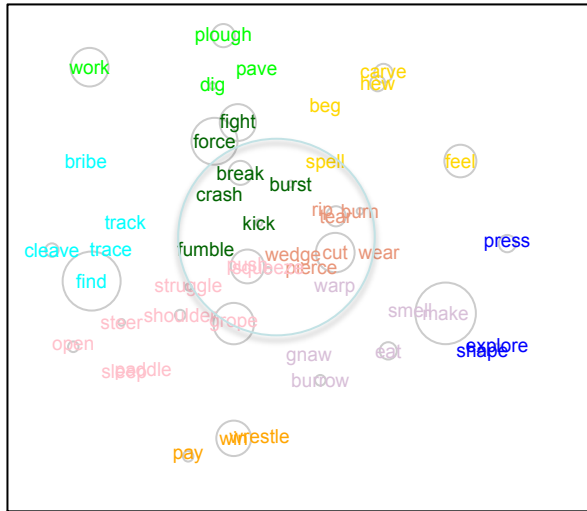
Observations

- Two domains of predilections: psych-verbs and verbs of hitting
- Other regions of the semantic space are more sparsely populated
- In line with previous findings on syntactic productivity
 - E.g., Suttle and Goldberg (2011)
 - Densely populated regions are more likely to attract new members
 - New verbs appear either close to or inside a cluster

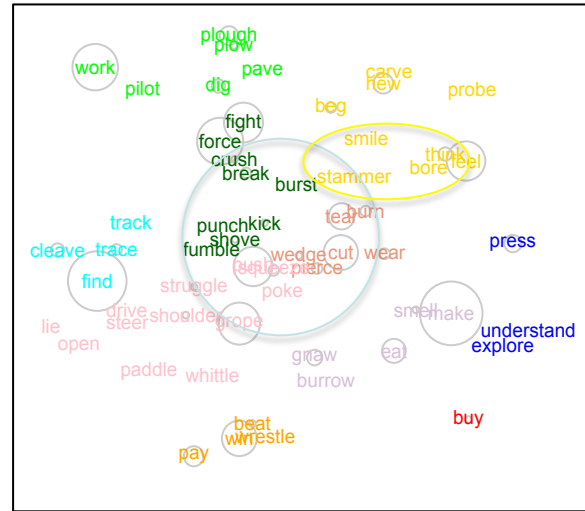
Example 2: the way-construction

- Verb *one's way* PP
- Describes motion of the subject referent
- Focus on the 'means' interpretation
 - The action causes or enables motion
They hacked their way through the jungle
 - As opposed to manner interpretation
e.g., *They limped their way to the door*
- In diachrony: increasingly abstract causation
(Israel 1996, Mondorf 2011)
e.g., *The chef chopped and diced his way to fame*

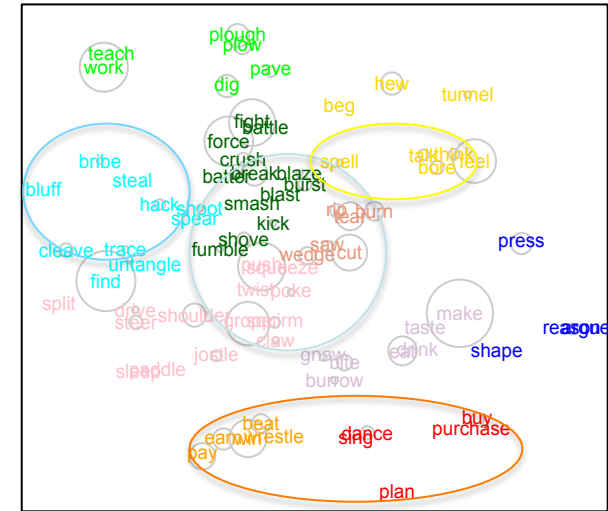
1830-1850



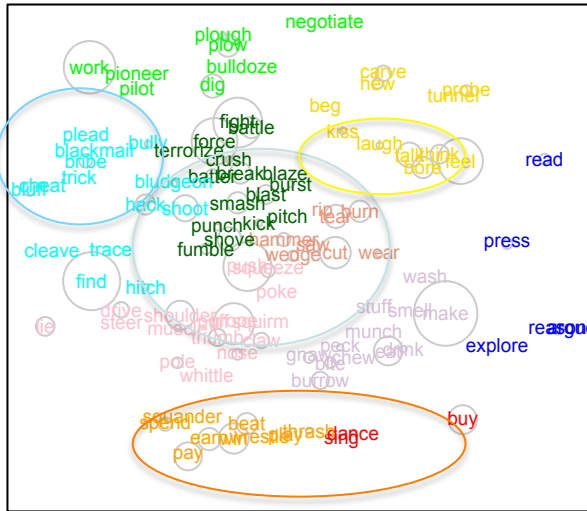
1860-1880



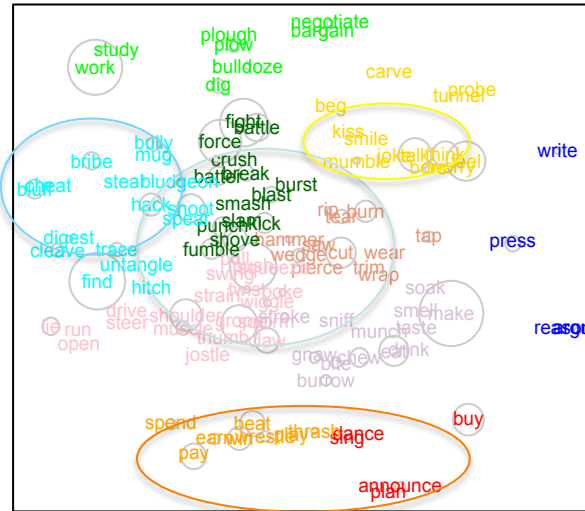
1890-1910



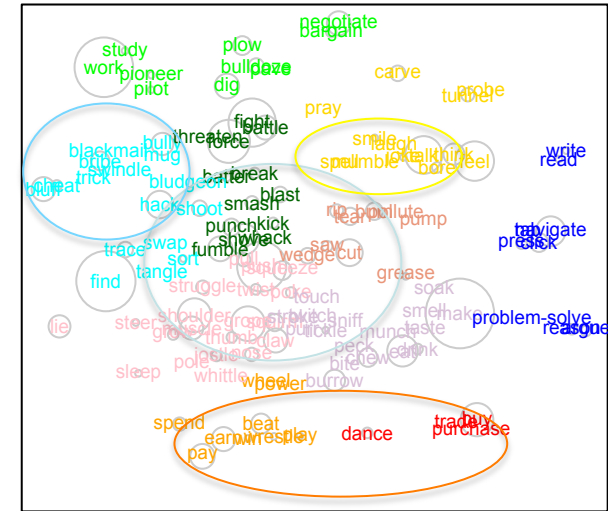
1920-1940



1950-1970



1980-2000



Conclusions

- Distributional semantics is appropriate for the study of syntactic productivity in diachrony
- Benefits:
 - Turns the informal notion of meaning into a quantified representation
 - Fully automatic and data-driven
 - Virtually no limit on the number of items to be considered
 - Enables the use of visualization techniques and statistical analysis
- Distribution-based account consistent with current views
- Promising approach to the study of syntactic productivity

I thank the hell out of you!

florent.perek@unibas.ch
<http://www.fperek.net>

- Bybee, J. & Eddington, D. (2006). A usage-based approach to Spanish verbs of 'becoming'. *Language*, 82(2), 323–355.
- Davies, M. (2010). *The Corpus of Historical American English: 400 million words, 1810-2009*. Available online at <http://corpus.byu.edu/coha/>
- Firth, J.R. (1957). A synopsis of linguistic theory 1930-1955. In *Studies in Linguistic Analysis*, pp. 1-32. Oxford: Philological Society.
- Goldberg, A. (1995). *Constructions: A construction grammar approach to argument structure*. Chicago: University of Chicago Press.
- Israel, M. (1996). The way constructions grow. In A. Goldberg (ed.), *Conceptual structure, discourse and language*. Stanford, CA: CSLI Publications, 217-230.
- Miller, G. & W. Charles (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1), 1-28.
- Mondorf, B. (2011). Variation and change in English resultative constructions. *Language Variation and Change*, 22, 397–421.
- Perek, F. (2014). Vector spaces for historical linguistics: Using distributional semantics to study syntactic productivity in diachrony. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, Maryland USA, June 23-25 2014*.
- Perek, F. (to appear). Using distributional semantics to study syntactic productivity in diachrony: A case study. *Linguistics*.
- Ross, J. (1973). Nouniness. In O. Fujimura (ed.), *Three Dimensions of Linguistic Research*. TEC Company Ltd.
- Suttle, L. & Goldberg, A. (2011). The partial productivity of constructions as induction. *Linguistics*, 49(6), 1237–1269.
- Van der Maaten, L. & Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9, 2579-2605.