Vector spaces for historical linguistics

Using distributional semantics to study syntactic productivity in diachrony

Florent Perek

Princeton University
Department of Psychology

fperek@princeton.edu
http://www.fperek.net

Syntactic productivity

- Property of a construction to attract new lexical fillers
- The distribution of constructions may vary over time
  - e.g., verb slot in 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.

Previous research

- Points to a strong semantic component in syntactic productivity
  - Productivity depends on the structure of the semantic space

    cf. Barðdal (2008), Bybee (2010), Bybee & Eddington (2006), Bybee & Thompson (1997), Suttle & Goldberg (2011), Wonnacott et al. (2012)
  - The likelihood of a novel use increases with the number and semantic diversity of attested types and the similarity with semantic neighbors
- How to operationalize semantics?
  - In previous studies: introspection, semantic norming
  - Proposal: use distributional semantics (Lenci 2008; Turney and Pantel 2010)

Case study: The "*hell*-construction"

- V *the hell out of* NP, e.g., *You scared the hell out of me!*
- Intensifying function (broadly defined)
- *Scare* and *beat* most typical, but also a wide range of other verbs:

  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 know the hell out of women!

The *hell*-construction in diachrony

- Data from the COHA (Davies 2010)
- 362 tokens, 105 verbs from 1930 to 2009
- Goal: track the semantic development of the construction by using distributional semantics

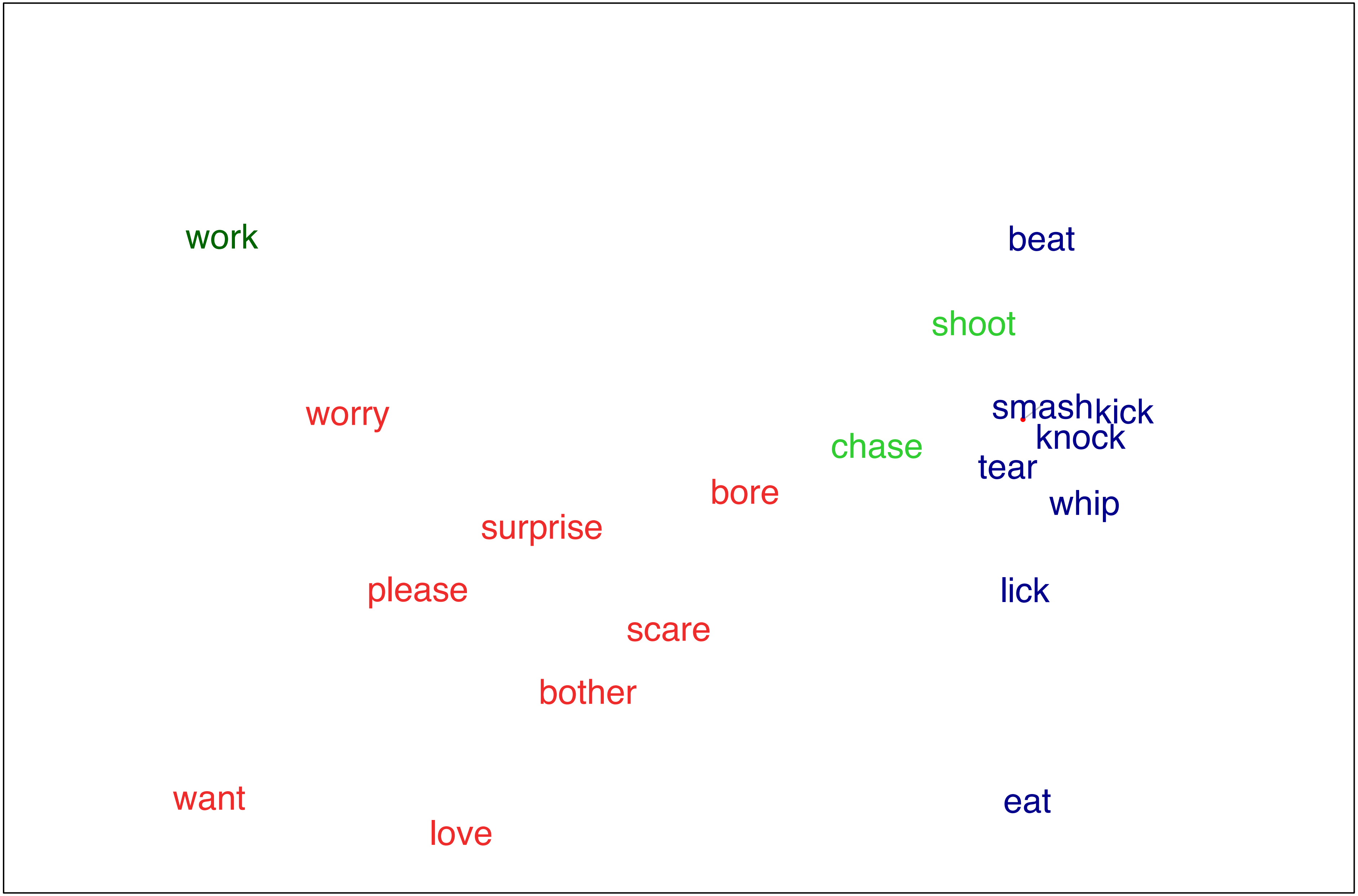
Vector-space model

- Captures how the verbs in the *hell*-construction are semantically related
- Built with DISSECT toolkit (Dinu et al. 2013)
- Based on lexical co-occurrences
  - Data from COCA (~450MW; Davies 2008)
  - Only the 92 verbs with F>2000
  - Collocates in 5-word window, lemmatized and PoS-tagged (Schmid 1994)
  - Nouns, verbs, adjectives, and adverbs from the 5,000 most frequent words
- Weighing scheme: Point-wise Mutual Information
- Cosine distance to compute distance matrix between the 92 verbs

Visualization

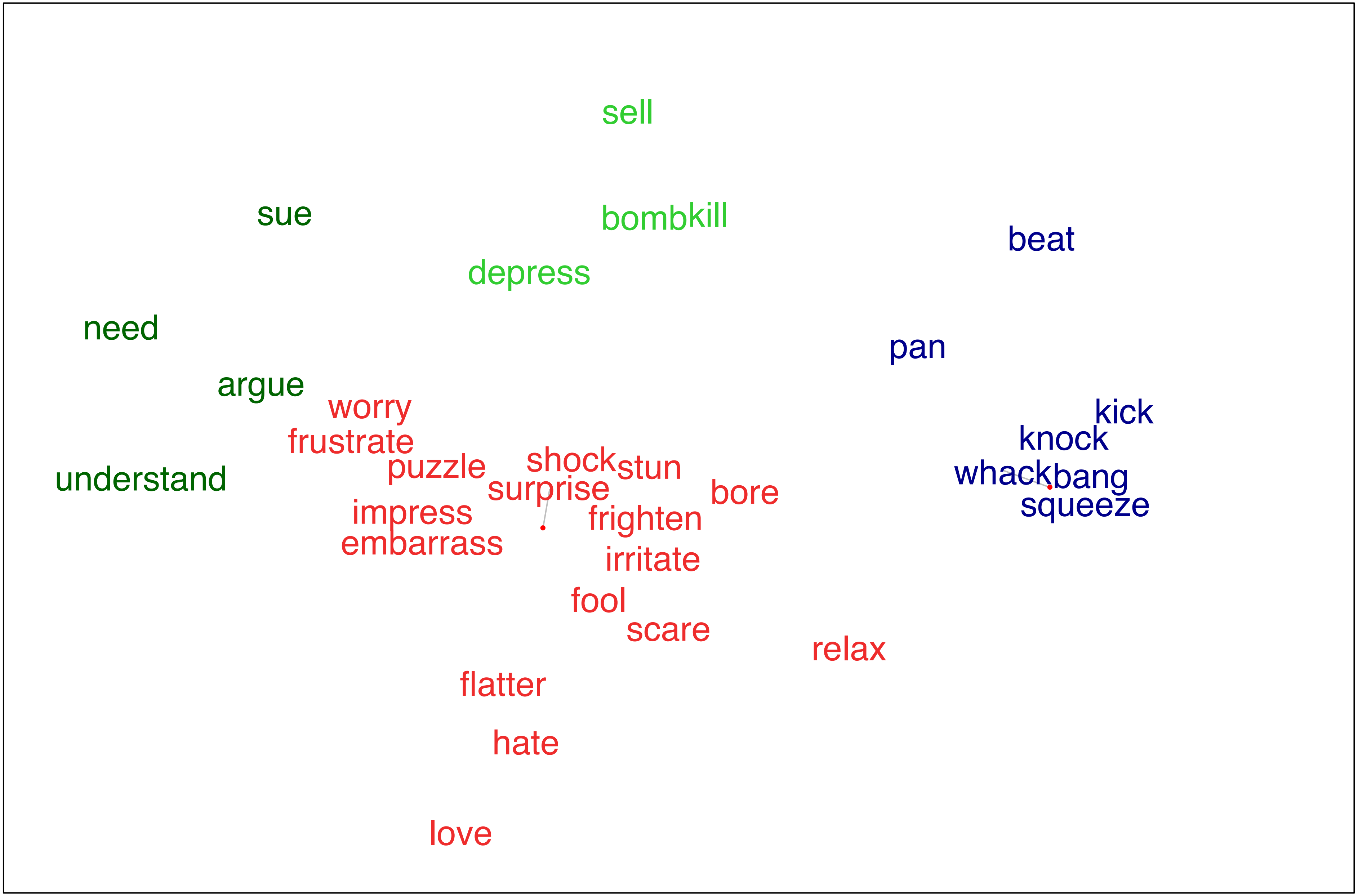
- Multidimensional scaling (MDS) to plot the semantic space
  - Places objects in a 2-dimensional space such that the between-object distances are preserved as well as possible
  - Converts distance matrix to set of coordinates
- Four plots for each 20-year period
  - 1930-1949
  - 1950-1969
  - 1970-1989
  - 1990-2009

1930s – 1940s

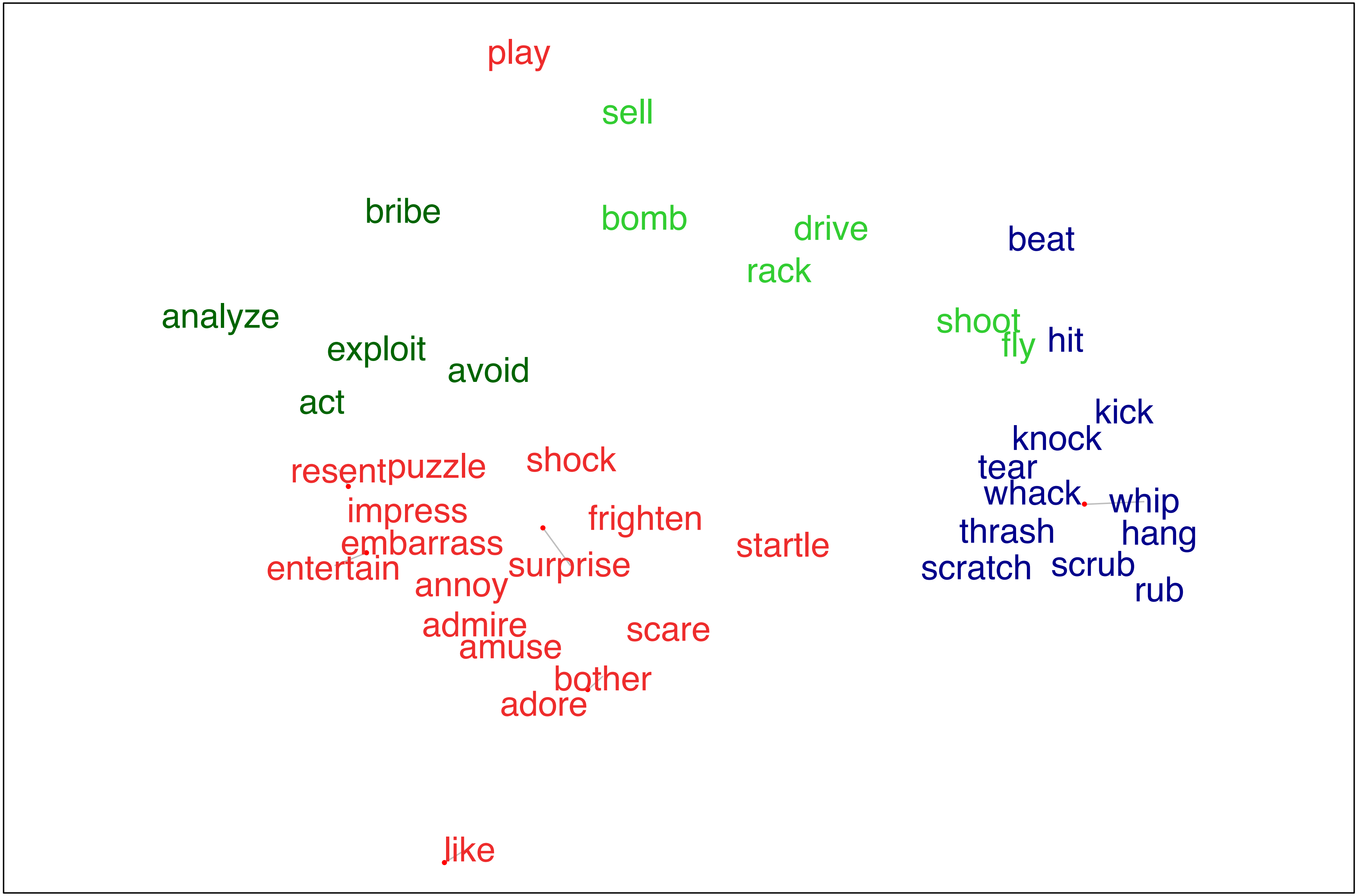


work
beat
shoot
worry
smash
kick
knock
chase
tear
bore
whip
surprise
please
lick
scare
bother
want
eat
love

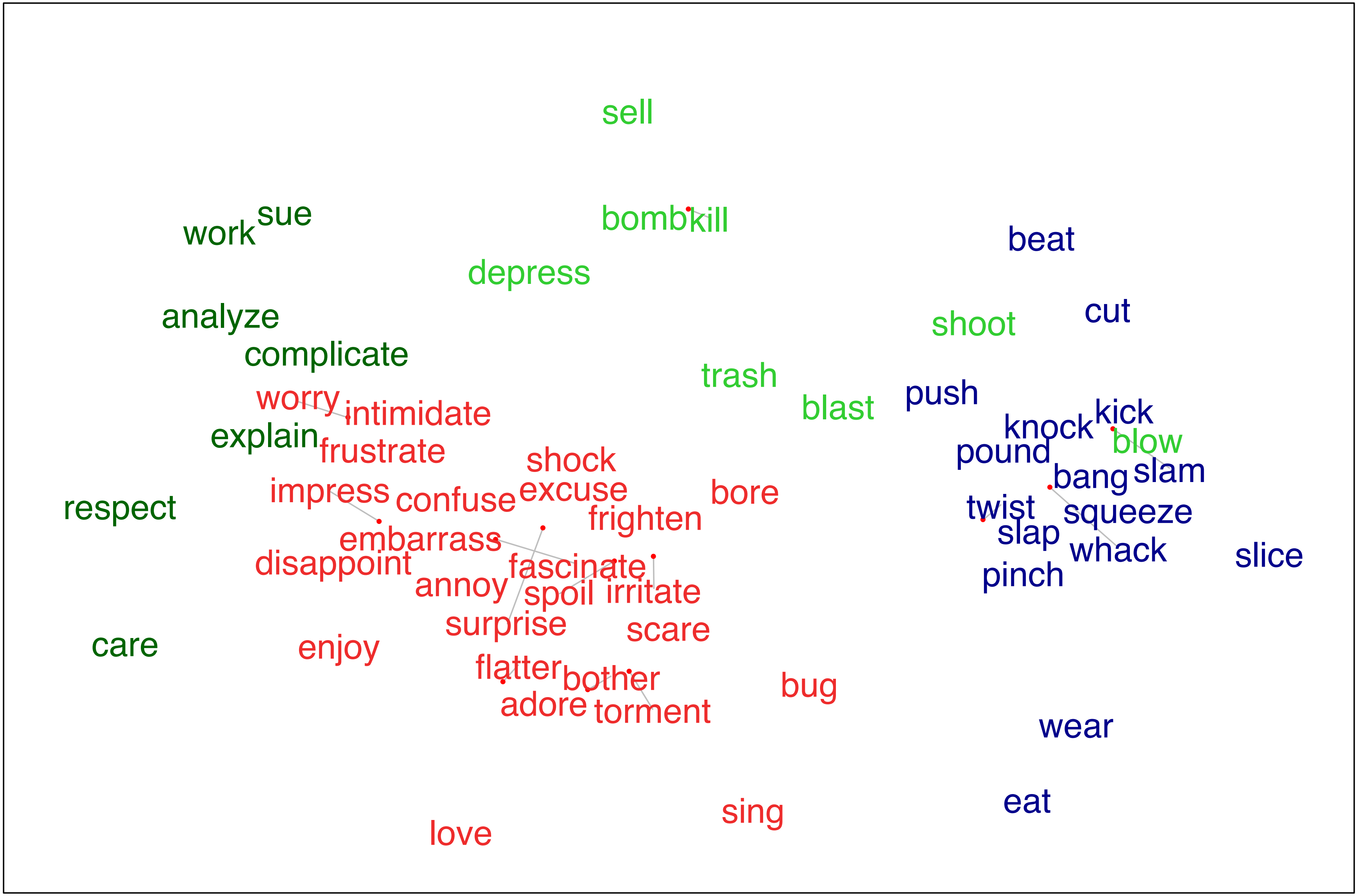
1950s – 1960s

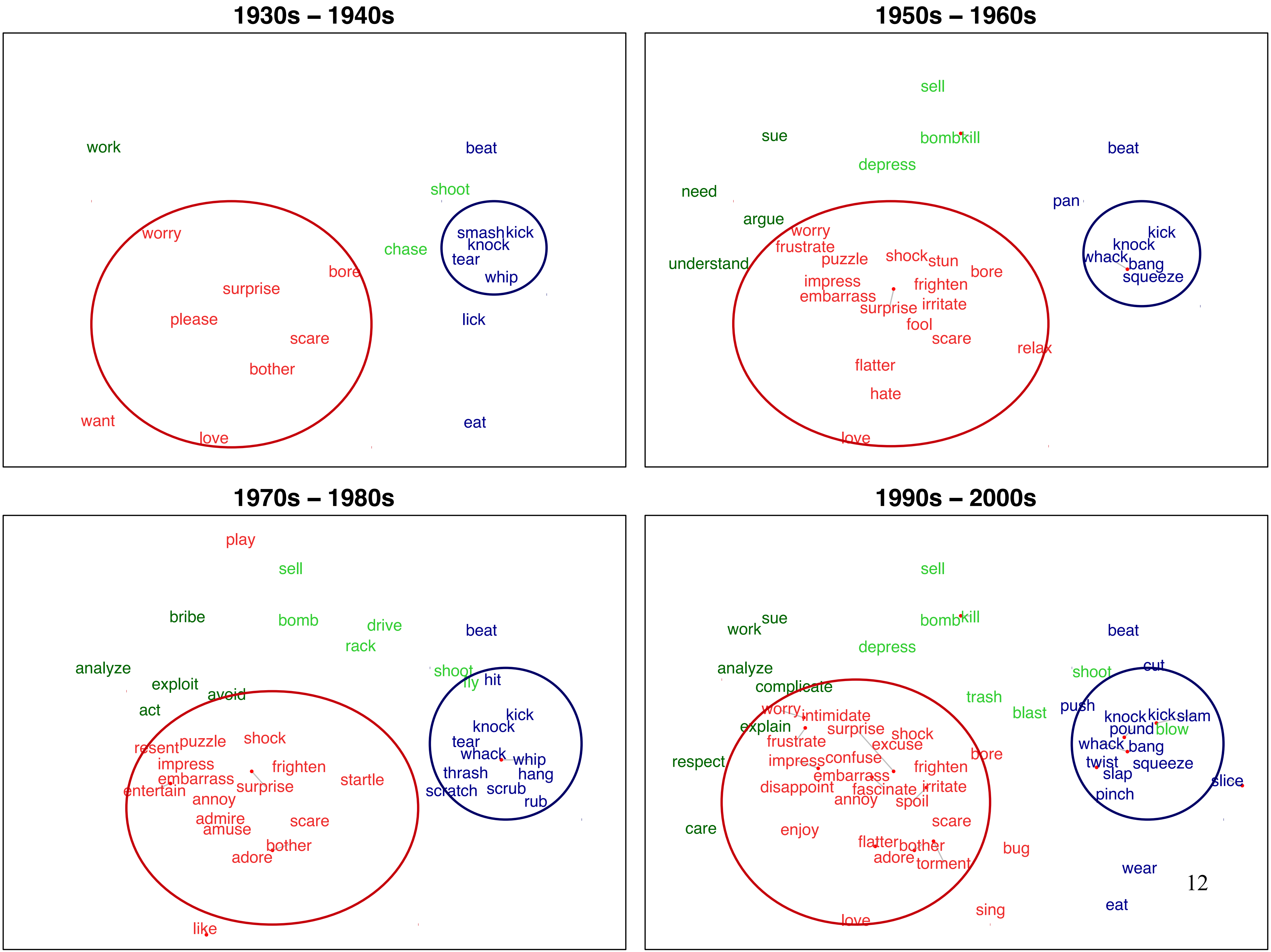

sell
sue
bomb
kill
beat
depress
need
pan
argue
worry
kick
frustrate
knock
shock
stun
puzzle
understand
whack
bang
surprise
bore
squeeze
impress
frighten
embarrass
irritate
fool
scare
relax
flatter
hate
love

1970s – 1980s


play
sell
bribe
bomb
drive
beat
rack
analyze
shoot
exploit
fly
hit
avoid
act
kick
knock
resent
puzzle
shock
tear
whack
whip
impress
frighten
thrash
hang
embarrass
startle
entertain
surprise
scratch
scrub
annoy
rub
admire
scare
amuse
bother
adore
like

1990s – 2000s


sell
work
sue
bomb
kill
beat
depress
analyze
cut
shoot
complicate
trash
worry
push
intimidate
blast
explain
knock
kick
blow
frustrate
pound
shock
slam
bang
impress
excuse
confuse
bore
respect
twist
frighten
squeeze
embarrass
slap
disappoint
whack
slice
fascinate
annoy
spoil
irritate
pinch
surprise
scare
care
enjoy
flatter
bother
bug
adore
torment
wear
eat
sing
love



1930s – 1940s

1950s – 1960s

1970s – 1980s

1990s – 2000s

Summary

- Distribution-based account in line with previous research
  - Densely populated regions are more likely to attract new members
  - New verbs tend to appear either close to or inside a cluster

- Another benefit of the distributional approach:
  - Vector representations allow quantification of properties of the sem. space
  - This enables the use of statistical analysis (e.g., logistic regression)
  - e.g., effect of space density on the probability of occurrence of a new item

Conclusion

- Distributional semantics is appropriate for the study of syntactic productivity in diachrony; benefits:
  - Fully automatic and data-driven
  - Virtually no limit on the number of items to be considered
  - Enables exploratory analysis and inferential statistics
- Promising application of a computational linguistic technique for diachronic studies

I thank the hell out of you!

Barðdal, J. (2008). *Productivity: Evidence from Case and Argument Structure in Icelandic*. Amsterdam: John Benjamins.

Bybee, J. (2010). *Language, Usage and Cognition*. Cambridge: Cambridge University Press.

Bybee, J. & D. Eddington (2006). A usage-based approach to Spanish verbs of 'becoming'. *Language* 82 (2), 323–355.

Bybee, J. & S. Thompson (1997). Three frequency effects in syntax. *Berkeley Linguistics Society* 23, 65–85.

Davies, M. (2008). *The Corpus of Contemporary American English: 450 million words, 1990-present*. Available online at http://corpus.byu.edu/coca/

Davies, M. (2010). *The Corpus of Historical American English: 400 million words, 1810-2009*. Available online at http://corpus.byu.edu/coha/

Dinu, G., N. Pham and M. Baroni (2013). DISSECT: DIStributional SEmantics Composition Toolkit. In Proceedings of the System Demonstrations of ACL 2013 (51st Annual Meeting of the Association for Computational Linguistics). East Stroudsburg PA: ACL, 31-36.

Israel, M. (1996). The way constructions grow. In A. Goldberg (ed.), *Conceptual structure, discourse and language*. Stanford, CA: CSLI Publications, 217-230.

Lenci, A. (2008). Distributional semantics in linguistic and cognitive research. *Rivista di Linguistica* 20.1, 1-31.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of International Conference on New Methods in Language Processing, Manchester, UK.*

Suttle, L. & A. Goldberg (2011). The partial productivity of constructions as induction. *Linguistics* 49 (6): 1237–1269.

Turney, P. and P. Pantel (2010). From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research* 37, 141-188.

Wonnacott, E., J. Boyd, J. Thompson & A. Goldberg (2012). Input effects on the acquisition of a novel phrasal construction in 5 year olds. *Journal of Memory and Language* 66: 458–478.