Using vector-space models to visualize the semantic distribution of argument structure constructions

Florent Perek
Universität Basel & Princeton University
florent.perek@unibas.ch

Overview

- Problem: using semantics in corpus studies of argument structure
- Description of the present approach
  - Based on vector-space semantics
  - Applied to the study of syntactic productivity in diachrony
- Case study

Introduction

- Most theories of argument structure rely on semantics
  - Levin (1993): “grammatically relevant features of meaning”
  - Pinker (1989), Jackendoff (1990): event structure + linking rules
  - Construction grammar: principle of semantic coherence (Goldberg 1995)

    A verb can be used in an argument structure construction if its meaning is compatible with the meaning of the construction

Introduction

- How do we assess the semantics of verbs?
  - Meaning is not directly observable
  - No real consensus on what verb meaning consists of
- Especially problematic for quantitative corpus studies
  - A corpus only gives access to word forms
  - How do we factor in semantics?

Introduction

- First (and most common) solution: manual semantic annotation
  - Based on the semantic intuitions of the analyst
  - Meaning is harder to annotate than grammatical features
    - Time-consuming
    - Unclear criteria, can be highly subjective
    - Categorical data: how do we quantify semantic similarity?

Introduction

- Second solution: norming study
  - Semantic information collected from a group of native speakers

    e.g., how similar are words in a pair? (Bybee and Eddington 2006)
  - Also hard and time-consuming
  - For semantic similarity, restricted to a limited set of words
    - Judgments must be collected for every pair
    - The number of judgments to be collected increases exponentially with the number of words

Introduction

- A third alternative?
  - Computational linguistics provides techniques for automatic annotation
  - Many of them are commonly employed to annotate corpora: part-of-speech tagging, syntactic parsing, dependency annotations
  - Ways to handle semantics have also been devised
- Vector-space models: based on the distributional hypothesis

  "You shall know a word by the company it keeps." (Firth 1957: 11)
- Words that occur in similar contexts tend to have related meanings
- Therefore, a way to access the meaning of words is through their distributions

Vector-space models

- First step: building a word co-occurrence matrix from a corpus
  - The matrix is filled by counting, for each occurrence of each word, the frequency of co-occurrence of other words within a set context window
  - Function words are usually ignored
  - e.g., *kiss* (+/-5-word window)

```
   frizzy hair, felt him kiss me on the cheek with
  give her my flowers, to kiss her hand, but did not
  Diana, and bent down to kiss him on the cheek. She
towards him and tried to kiss her on the mouth. Over
     and I lean over and kiss her on the mouth. I
```

frizzy	1
hair	1
feel	1
cheek	2
give	1
flower	1
hand	1
Diana	1
bend	1
try	1
mouth	2
lean	1

Vector-space models

- Results in a Nwords x Nwords matrix
- Dimensionality reduction is usually employed on the co-occurrence matrix
  - i.e., the matrix is transformed so as to contain fewer columns, singling out the most salient contextual features of word distributions
  - Computationally more tractable
  - Preserves only the most informative aspects of word distributions
- Each word is associated with a row of the matrix
- Distributional hypothesis:
  - Semantic distance between words is a function of their distributional similarity
  - Similarity between rows approximate semantic similarity
  - It can be quantified by mathematical measures

Vector-space models

- Benefits of vector-space models
  - The informal notion of semantic representation is turned into an empirically testable semantic model
  - In that model, semantic similarity can be quantified
- They have some psychological reality
  - Shown to correlate positively with human performance on various tasks: synonymy judgments, word association, semantic priming, … (Lund et al. 1995, Landauer et al. 1998)
  - Models based on both experiential and distributional information perform even better than models based on either kind (Andrews et al. 2008)

Vector-space models

- Only recently applied to address linguistic research questions
- To inductively determine semantic classes
  - Gries and Stefanowitsch (2010): clustering of the verbs in the distribution of constructions according to their frequent collocates
  - Levshina and Heylen (in press): identification of contrasting sets of semantic classes for the causees involved in Dutch causative constructions
- Can vector-space models also be used to investigate syntactic productivity?
  - i.e., the property of syntactic constructions to be combined with new words
  - In diachrony, productivity corresponds to the expansion of the lexical distribution of constructions

Case study

- My case study: the transitive *hell*-construction
  - V *the hell out of* NP

    You scared the hell out of me!
  - Conveys an intensifying function

    scare the hell out of someone = ‘scare someone very much’
  - Some frequent “vulgar” variants: *crap*, *fuck*, *shit* instead of *hell*
  - Lends itself nicely to a construction grammar analysis
    - The intensifying function is conveyed by the whole syntactic pattern
    - This pattern contains a verb slot: how did it evolve over time?

The *hell*-construction in diachrony

- Data from the Corpus of Historical American English (COHA; Davies 2010)
  - ~20 MW of AmE from each decade between 1810 and 2009
  - Written data balanced for genre: fiction, magazines, newspapers, non-fiction
- Query for the string "V the hell|crap|fuck|shit out of"
- Manual filtering (ruled out sentences like *get the hell out of here*)
- Clearly centered on two verbs: *scare* and *beat* (30% and 25% in the 2000s)
- But it can be used with a wide and diverse range of verbs:

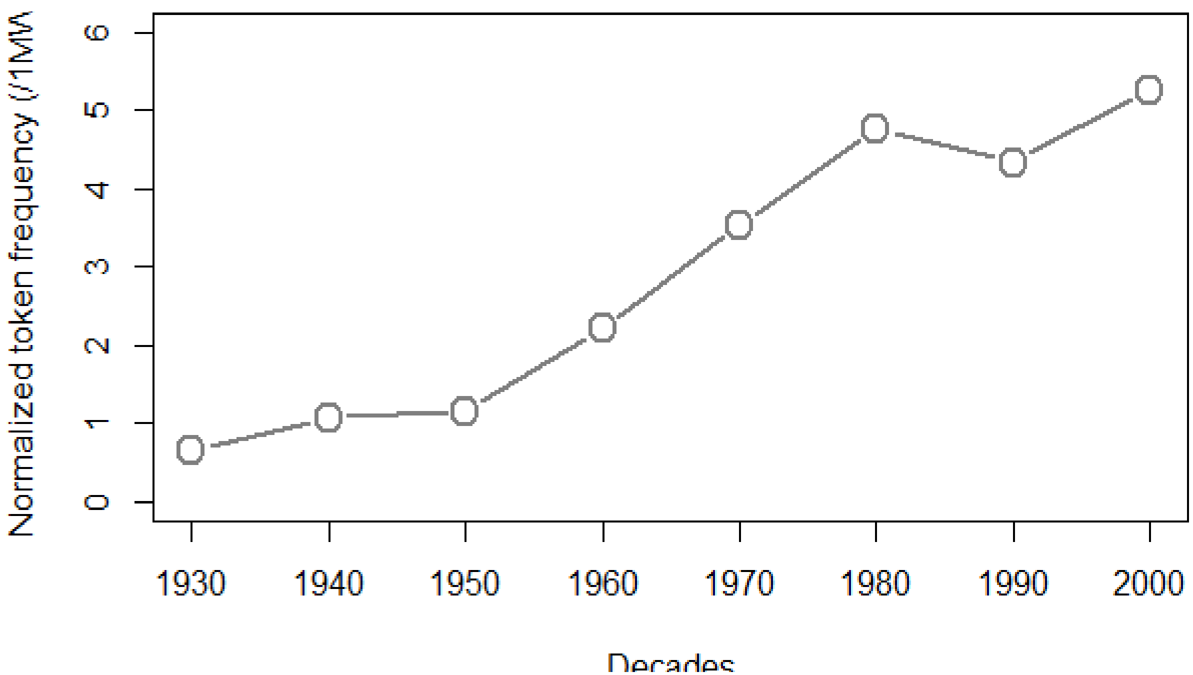
  Then I [...] avoided the hell out of his presence

  But you drove the hell out of it!

  The Russians understood the hell out of that.

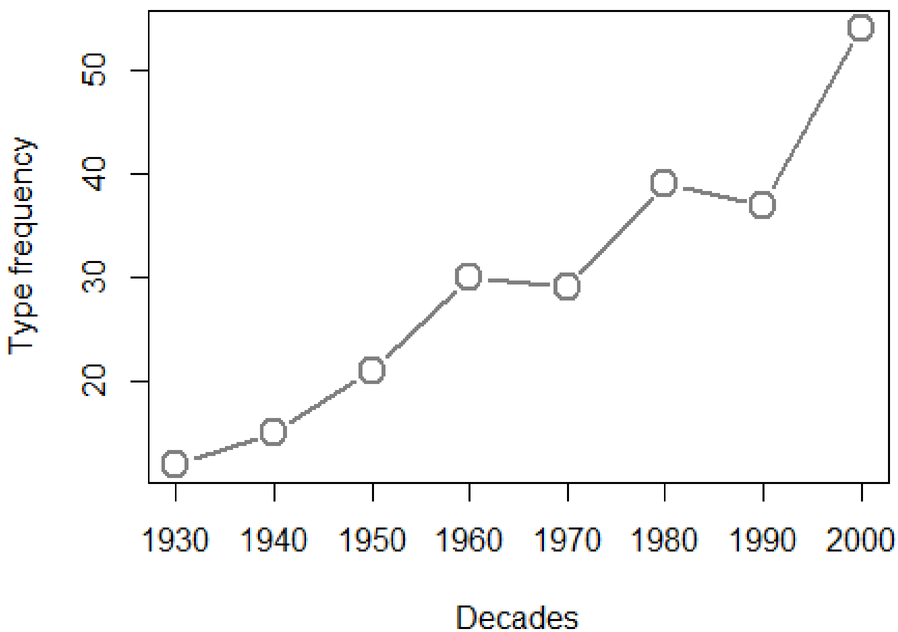
The *hell*-construction in diachrony

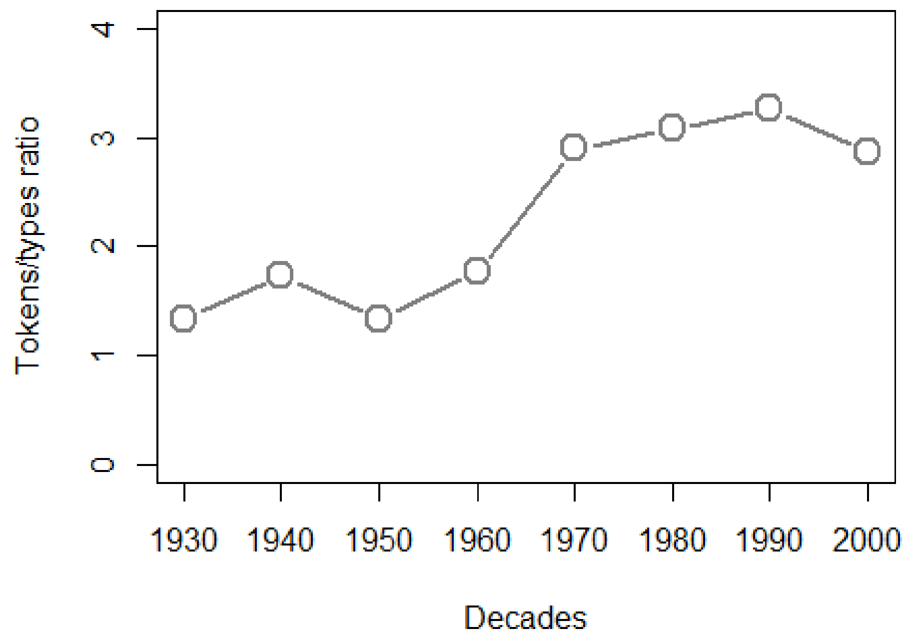
- A recent construction: first attestations from the 1930s
- The construction has been steadily increasing in frequency ever since



The *hell*-construction in diachrony

- More and more verb types are used in the construction





The *hell*-construction in diachrony

- Type frequencies reflect an increase in productivity
- But they do not reflect the “structure” of that productivity
  - What kinds of verbs joined the distribution, and when?
  - Are there particular semantic domains preferred by the construction?
  - Did that change over time, and how?
  - In how far do these data line up with current hypotheses on productivity?

    i.e., productivity is promoted by:

    - High type frequency (Wonnacott et al. 2012; Zeschel 2012)

      i.e., a high number of different lexical items occurring in the slots of the construction
    - High semantic variability (Barðdal 2008; Suttle and Goldberg 2011)

      i.e., how different these items are

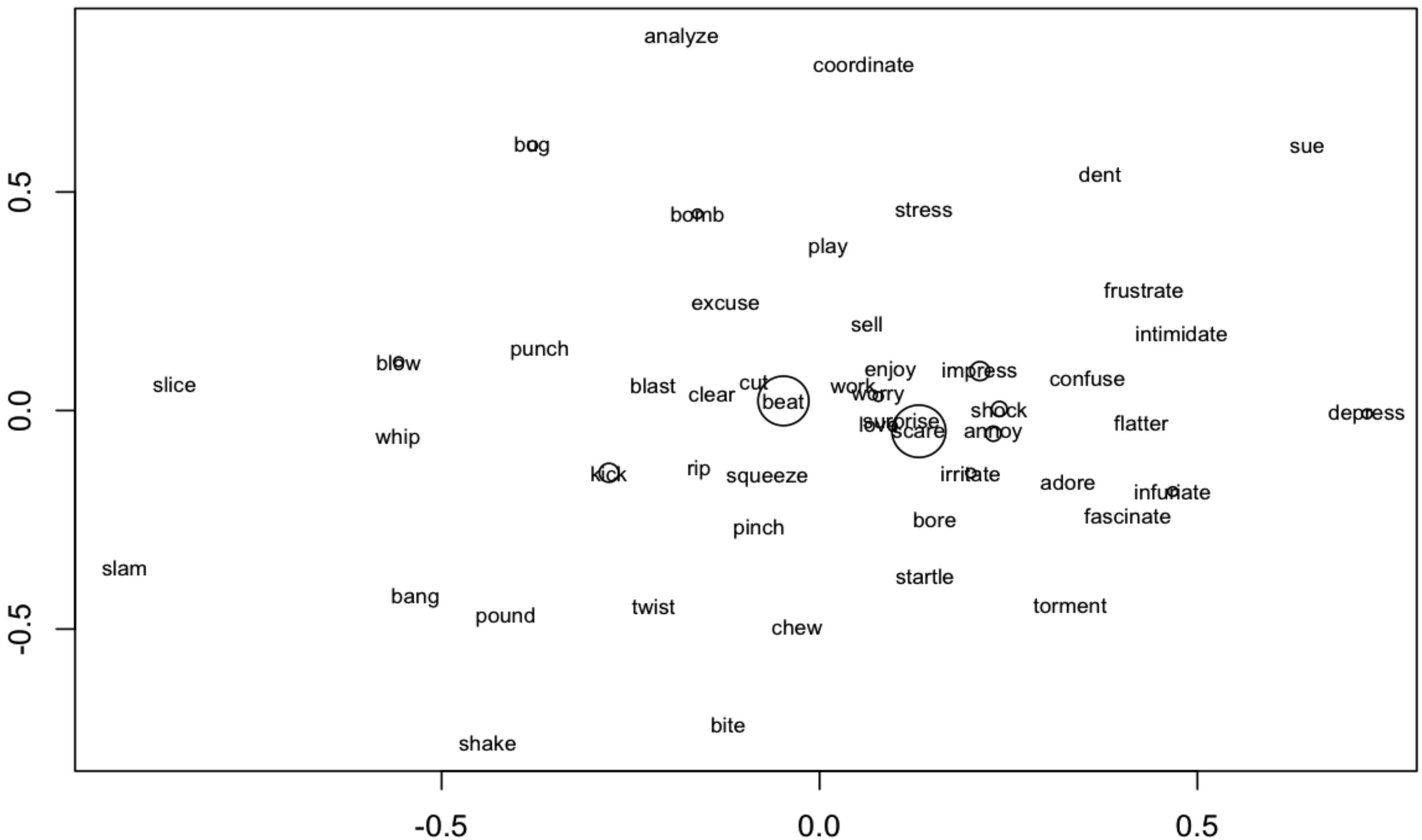
The *hell*-construction in diachrony

- A vector-space model to assess the meaning of the verbs in the *hell*-construction
  - Trained on the written part of the BNC (~90 MW)
  - Co-occurrence window of +/- 2 words
  - Stop-words: function words, highly frequent modifiers (*really*, *again*, …), low-frequency words (F < 100)
  - S-Space package (Java): https://github.com/fozziethebeat/S-Space
  - Random Indexing algorithm (Sahlgren 2001)

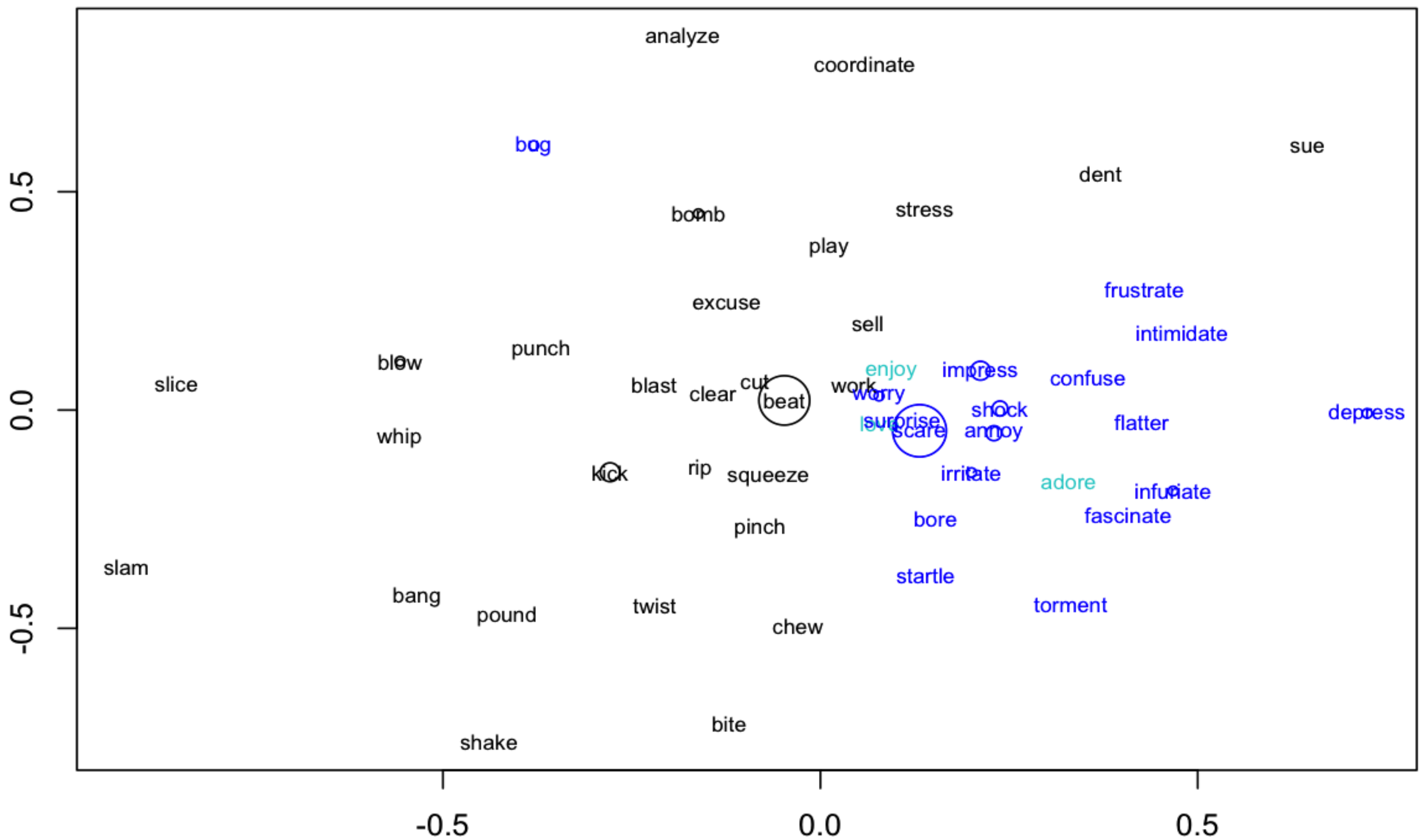
The *hell*-construction in diachrony

- How to visualize the semantic space?
  - By means of multidimensional scaling (MDS)
    - Aims to place objects in a space with 2 (or more) dimensions such that the between-object distances are preserved as well as possible
    - Each object is assigned coordinates in 2 dimensions
  - Semantic similarity matrix: pairwise similarities between verbs are computed (cosine distance)
  - The matrix is submitted to MDS and the coordinates of the verbs are plotted

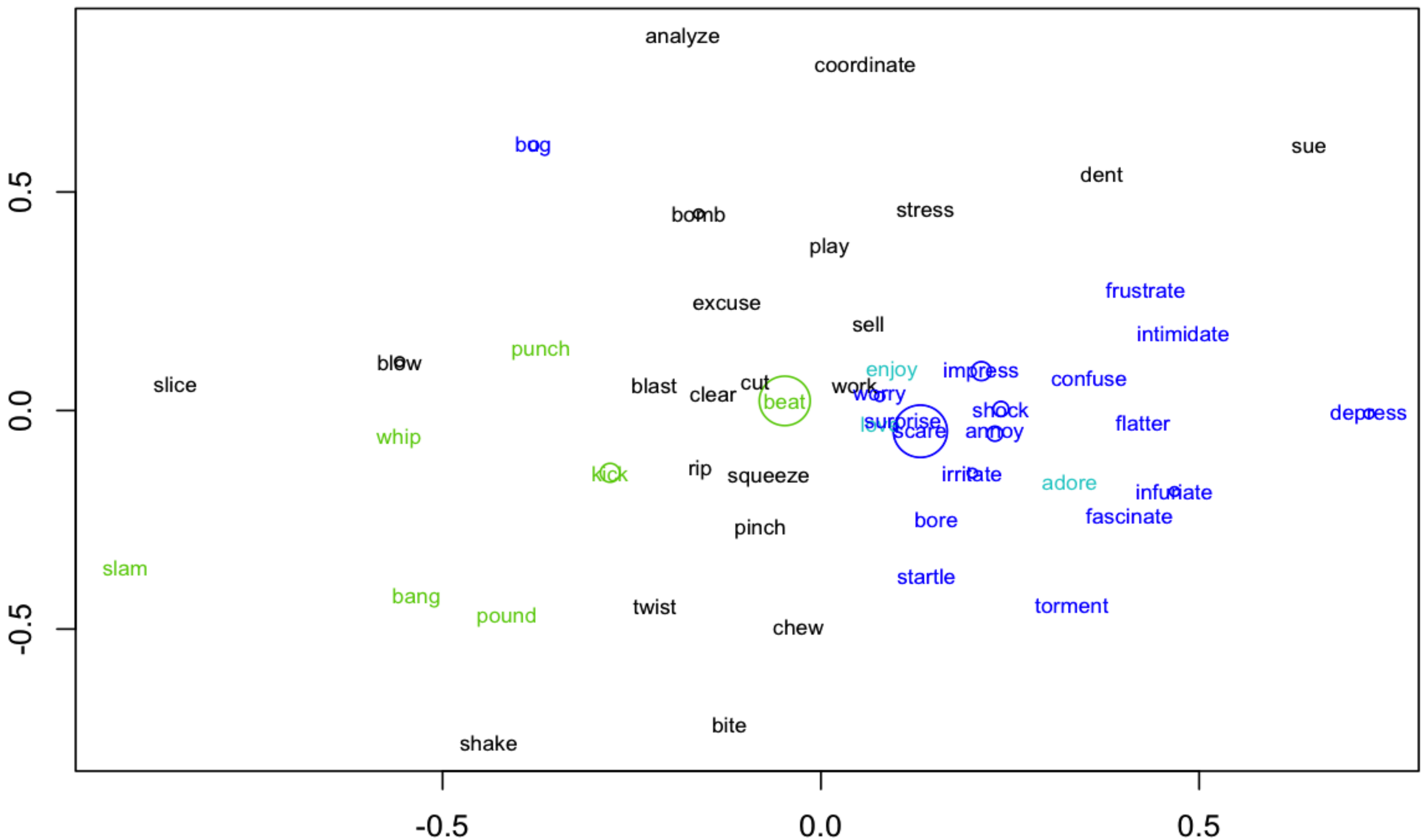
Example: The *hell*-construction in the 2000s



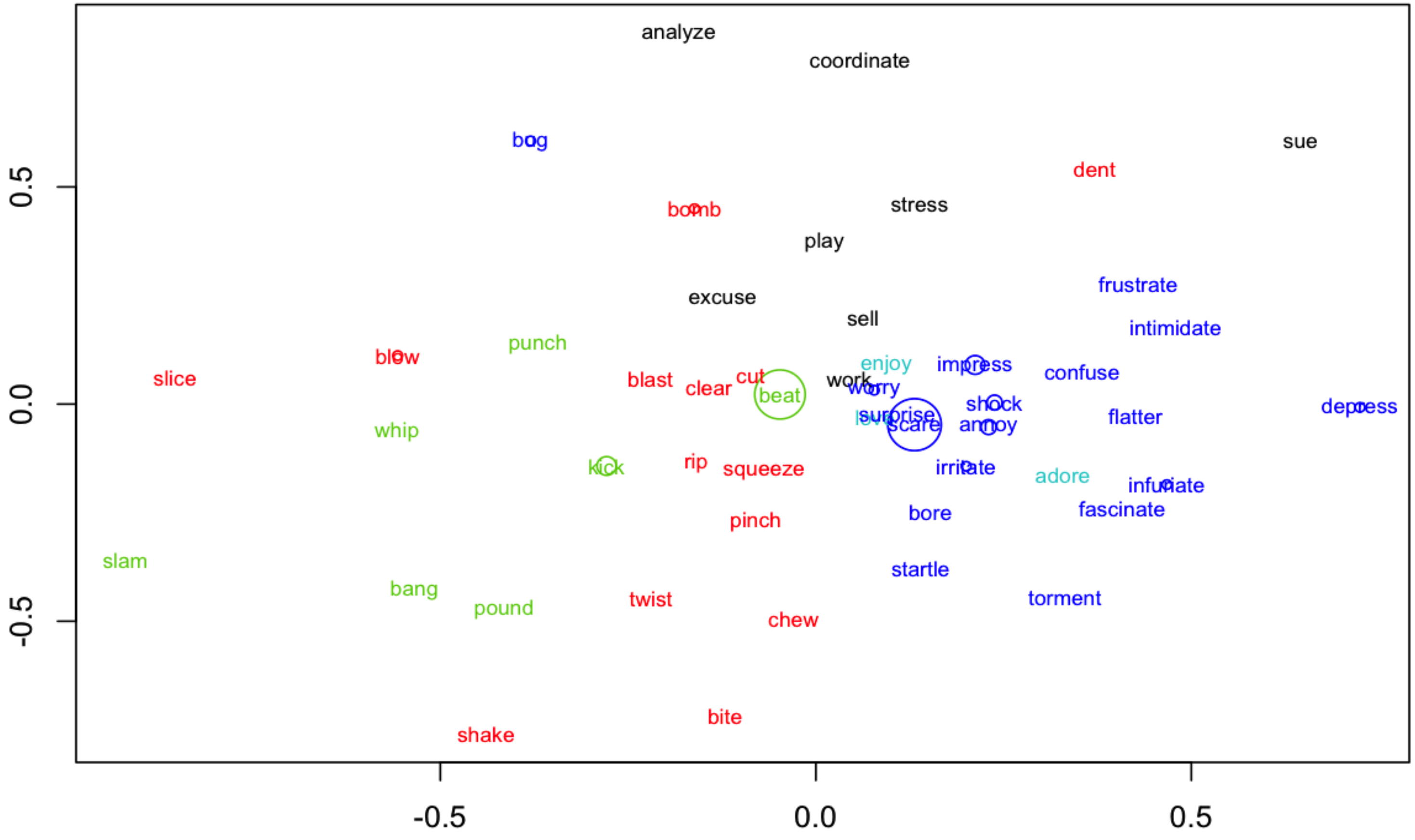
Example: The *hell*-construction in the 2000s



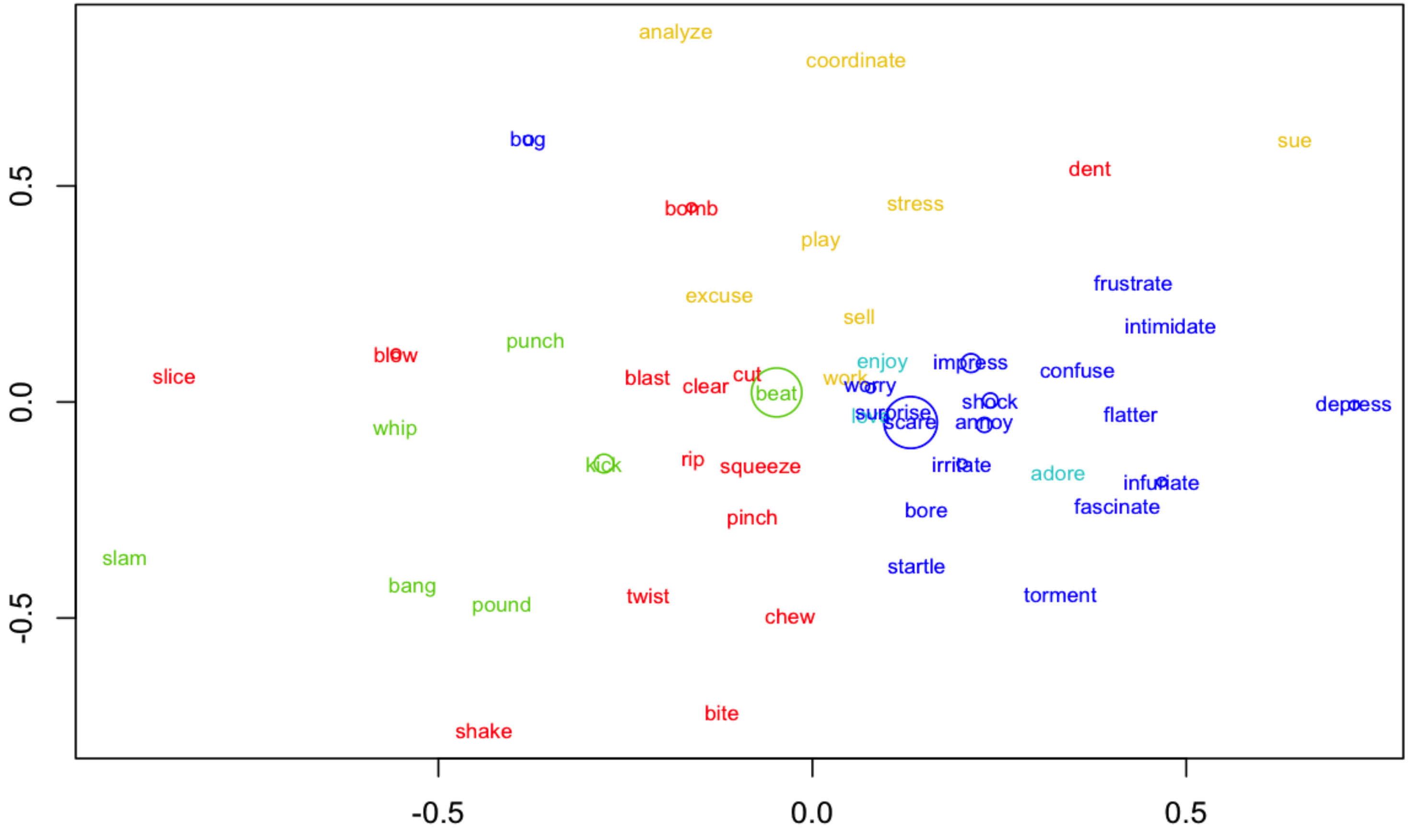
Example: The *hell*-construction in the 2000s



Example: The *hell*-construction in the 2000s



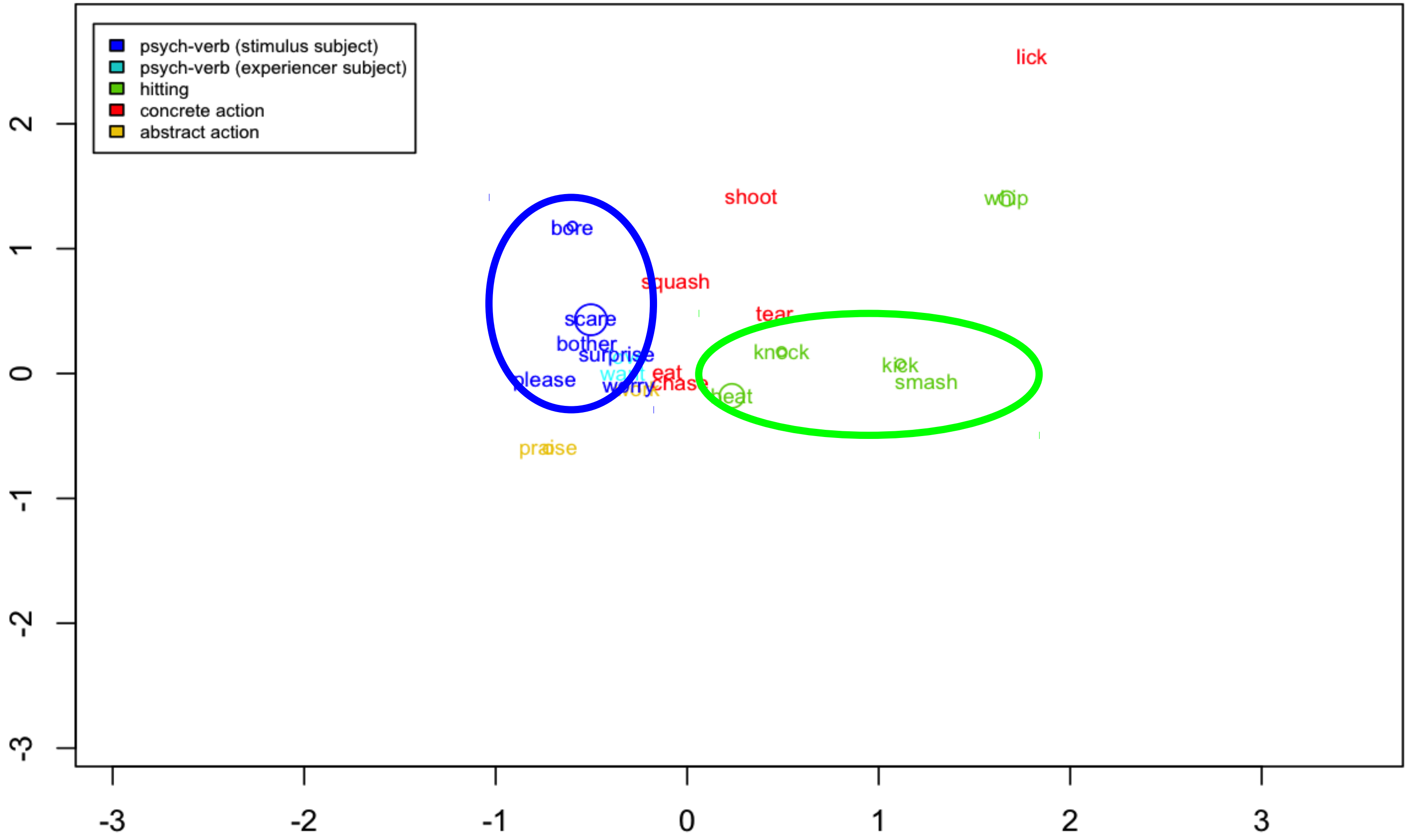
Example: The *hell*-construction in the 2000s



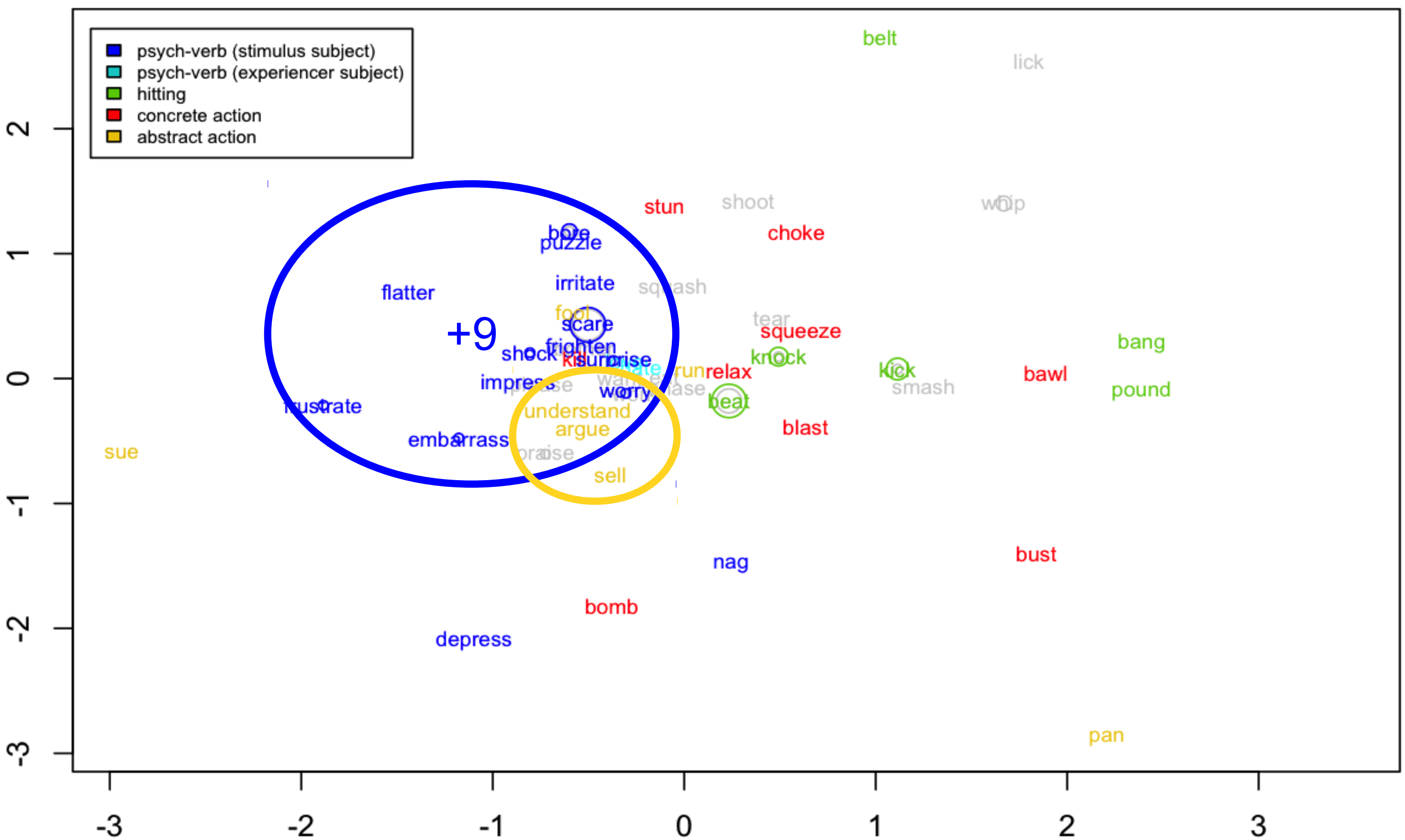
The *hell*-construction in diachrony

- Diachronic analysis: periods of 20 years
  - 1930s-1940s
  - 1950s-1960s
  - 1970s-1980s
  - 1990s-2000s
- By comparing these periods, we can follow the semantic development of the construction
- Similar to Hilpert's (2011) motion charts, but with semantics

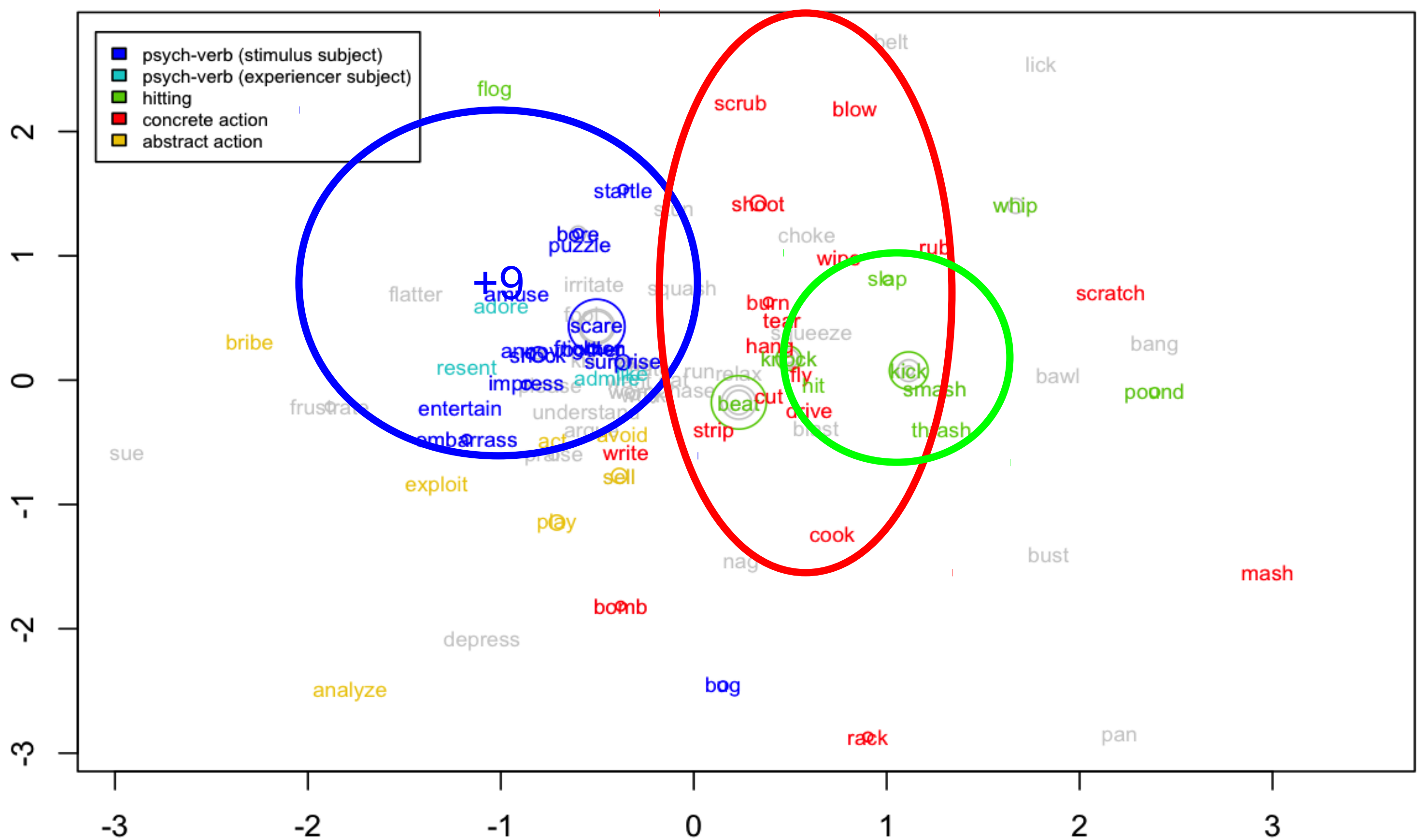
The *hell*-construction in diachrony: 1930s-1940s



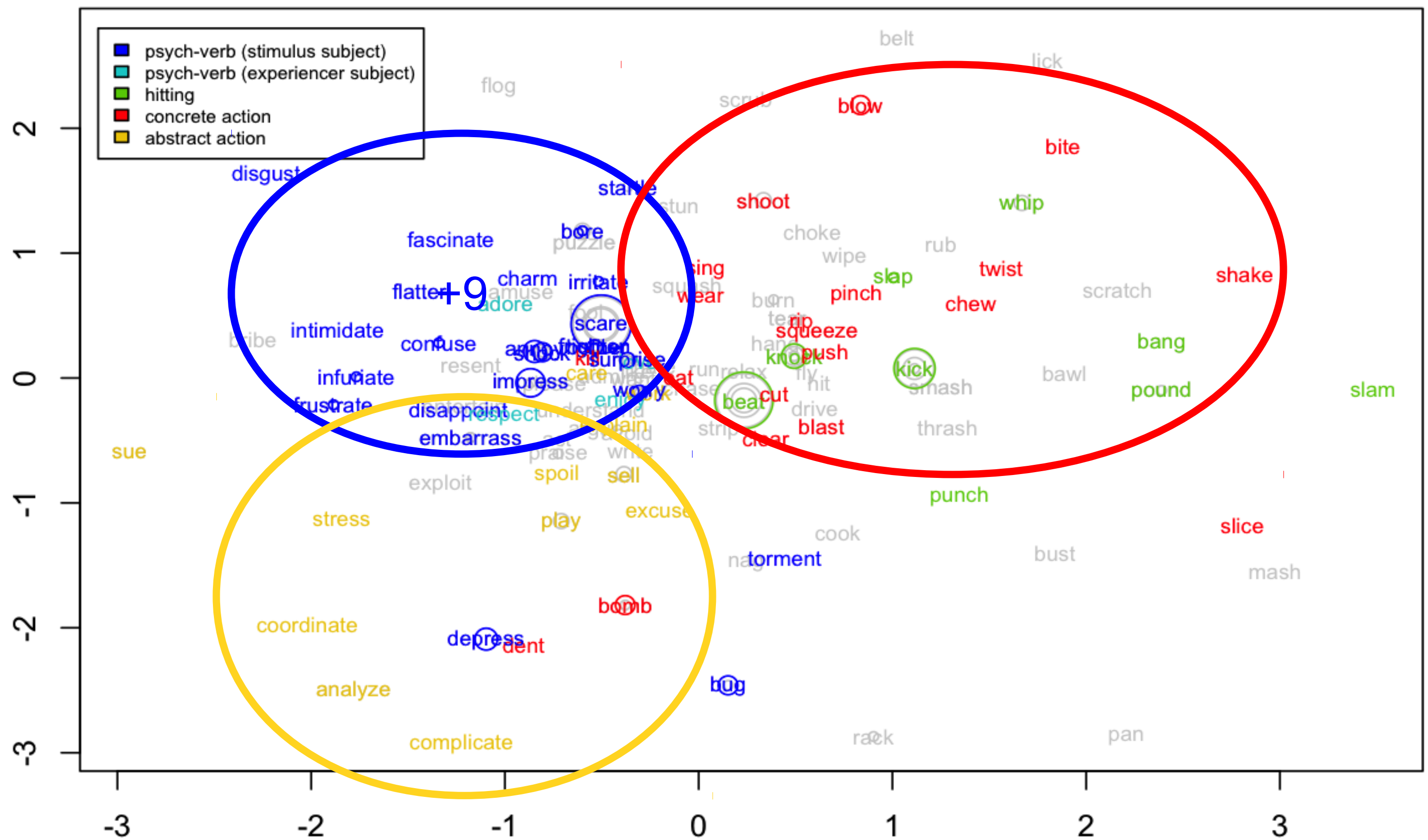
The *hell*-construction in diachrony: 1950s-1960s



The *hell*-construction in diachrony: 1970s-1980s



The *hell*-construction in diachrony: 1990s-2000s



Summary

- The domain of predilection is clearly psych-verbs (esp. with stimulus subject)
  - They form a tight cluster from the start
  - One of the most populated area at all times
  - They regularly attract new members
- What about the other semantic domains?
  - More sparsely populated, i.e., low type frequency, high semantic variability
  - Markedly fewer in the 1930s-1940s (psych-verbs dominate)
  - They first attract few new members
  - Their productivity increases later

e.g., in the 1970s-1980s for concrete actions
in the 1990s-2000s for abstract actions

Summary

- The results are in line with current usage-based models of productivity
  - i.e., productivity is a function of type frequency and semantic variability
  - Densely populated regions are the most likely to attract new members
  - In sparsely populated regions, a "critical mass" of verbs is needed for productivity to take off
  - Token frequency is not a particularly strong source of productivity
    - e.g., the frequent verb of hitting *kick* remains largely isolated throughout
    - Only a few new verbs appear sporadically in its neighborhood: presumably analogical extensions from a salient model (Barðdal 2008; Bybee and Eddington 2006)

Limits of the present approach

- BNC: British English; AmE data should be used
- Broad semantic domains are visible, but the inner structure of these domains is not always properly reflected

  e.g., *startle* and *surprise* are far apart
- Does not handle polysemy
- How to improve the model?
  - Include grammatical information: dependency-based vector-space models (Padó and Lapata 2007)
  - Use automatic sense induction (Purandare and Pedersen 2004) to handle polysemy in a data-driven way

Conclusion

- A promising approach for the study of syntactic productivity
- The range of other questions it could address is yet to be explored
- Not by any way perfect, but even this coarse model gives a pretty clear picture
  - Verbs from the same broad classes occupy the same regions
  - Follows the predictions of current models of productivity
- Benefits
  - Instantaneously reveals patterns of productivity
  - The visualization allows to easily spot semantic changes in diachrony

Thanks for your attention!

Andrews, M., Vigliocco, G. & D. Vinson (2009). Integrating Experiential and Distributional Data to Learn Semantic Representations. *Psychological Review* 116(3), 463-498.

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

Bybee, J. and D. Eddington (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.

Gries, S. and A. Stefanowitsch (2010). Cluster analysis and the identification of collexeme classes. In S. Rice & J. Newman (eds.), *Empirical and experimental methods in cognitive/functional research*, 73-90. Stanford, CA: CSLI.

Hilpert, M. (2011). Dynamic visualizations of language change: Motion charts on the basis of bivariate and multivariate data from diachronic corpora. *International Journal of Corpus Linguistics* 16(4), 435–461.

Landauer, T. K., Foltz, P. W., and Laham, D. (1998). Introduction to Latent Semantic Analysis. *Discourse Processes* 25, 259-284.

Levshina, N. and K. Heylen. (in press). A Radically Data-driven Construction Grammar: Experiments with Dutch causative constructions. In R. Boogaart, T. Colleman & G. Rutten (Eds.), *Constructions in Germanic – Extending the scope*.

Lund, K., Burgess, C. and R. Atchley (1995). Semantic and associative priming in a high-dimensional semantic space. *Cognitive Science Proceedings* (LEA), 660-665.

Padó, S. and M. Lapata (2007). Dependency-based Construction of Semantic Space Models. *Computational Linguistics* 33(2), 161-199.

Purandare, A. and T. Pedersen (2004). Word Sense Discrimination by Clustering Contexts in Vector and Similarity Spaces. In *Proceedings of the Conference on Computational Natural Language Learning (CoNLL), May 6-7, 2004, Boston, MA*, pp. 41-48.

Sahlgren, M. (2001). Vector-based semantic analysis: Representing word meanings based on random labels. In *Proceedings of the ESSLLI 2001 Workshop on Semantic Knowledge Acquisition and Categorisation, Helsinki, Finland*.

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

Wonnacott, E., J. Boyd, J. Thompson and 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.

Zeschel, A. (2012). *Incipient productivity. A construction-based approach to linguistic creativity*. Berlin/New York: de Gruyter.