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 16\(^{th}\) century
      
      *They hacked their way through the jungle.*
    - Abstract means of reaching a goal only appear in the 19\(^{th}\) 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
  – 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
1990s – 2000s

love

respect
care

analyze
complicate
explain
worry

work
sue

intimidate
frustrate
impress

bore
frighten

enjoy
flatter

adore

bother

fascinate

disappoint

imper

anxiety

annoy

shock

spoil

irritate

surprise

scare

flatter

bother

adore

torment

bug

love

sing

respect

care

analyze
complicate
explain
worry

work
sue

intimidate
frustrate
impress

bore
frighten

enjoy
flatter

adore

bother

fascinate

disappoint

imper

anxiety

annoy

shock

spoil

irritate

surprise

scare

flatter

bother

adore

torment

bug

love

sing
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!


