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.*

Source: Corpus of Historical American English (COHA, Davies 2010)
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-consumming 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)
1970s-1980s

- admire
- analyze
- annoy
- avoid
- beat
- bomb
- bother
- bribe
- drive
- entertain
- exhaust
- fly
- fracture
- fry
- lament
- love
- mask
- offset
- poison
- scandal
- shackle
- shock
- shoot
- startle
- surprise
- surprise
- transform
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
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


