Periodization of constructional productivity in diachronic corpora

Florent Perek
University of Birmingham
Overview

- New method for diachronic studies
- Aim: identify stages of language change in the productivity of grammatical constructions
- Two case studies
Corpus-based studies of language change

- Typical corpus-based studies of language change
  - Extract tokens from a diachronic corpus
  - Classify these tokens according to some criterion
  - Compare the state of the language at different points in time

- Assess stages of language change
  - When was it relatively stable, and for how long?
  - When did it change (and how)?
Manual periodization

- Normalised frequency of the *hell*-construction in the COHA

  "Verb the hell out of", e.g., *You scared the hell out of me!*

![Diagram showing the normalised frequency of the 'hell' construction from 1930 to 2000. The frequency increases over time, with a peak around 1980 and a slight decrease around 1990.](image-url)
Problems with manual periodization

- Stages are not always clear to discern
- Potentially subjective: what are the criteria for splitting periods?
  - Different possible groupings for the same data
  - Comparison between studies
- More complex when multiple variables are considered
  e.g., token frequency + type frequency
Periodization

- This problem was first exposed by Gries & Hilpert (2008).
- They introduce “variability-based neighbour clustering” (VNC) as a method for automatic periodization.
- Variant of agglomerative clustering algorithm:
  - Periods are grouped according to their similarity, following some pre-defined criteria.
  - *Only time-adjacent periods can be merged.*

The VNC algorithm

- Starting point: data partitioned into “natural” time periods (years, decades, etc.)

1. Look at all pairs of adjacent periods (e.g., 1930s-1940s, 1940s-1950s, etc.). Measure their similarity according to some quantifiable property/ies.

2. Merge the two periods that are the most similar.

3. Calculate the properties of the merger as the mean values of its constituent periods.

- Repeat until all periods have been merged.
VNC: an example

- VNC with one variable: frequency of the *hell*-construction
VNC

- Two kinds of uses of VNC in the literature
  - To partition data in a principled way for further analysis
  - To uncover patterns of change and/or compare changes

- So far mostly based on quantitative variables
  - Frequencies: tokens, types, hapax legomena, etc.
  - Frequency distributions of lexical items, collexeme analysis

- Lines up with usage-based linguistics: grammatical representations are shaped by frequency

- Frequency = good starting point for looking at the history of constructions, but do not tell the whole story
Productivity

- Especially true for the study of productivity
  - The property of a construction to attract new lexical fillers
  - E.g., verbs in the way-construction (Israel 1996)
    - *They hacked their way through the jungle.* (16th century)
    - *She talked her way into the club.* (19th century)

- Type frequency often taken as an indicator of productivity
  - Number of different items, but not how different they are
  - Need to consider the semantic diversity of the distribution

Operationalizing word meaning

- Distributional semantics (Lenci 2008)
  - “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)

- Captures the meaning of words through their distribution in a large corpus

- Proposal: use distributional semantics to build representations of the semantic range of a construction

“Bag of words” approach

- Distributional data extracted from COHA (Davies 2010); 400 MW from 1810 to 2009
- Collocates of all verbs in a 2-word window
- Restricted to the 10,000 most frequent nouns, verbs, adjectives and adverbs

## Distributional semantic model

- Co-occurrence frequencies turned into PPMI scores
- 10,000 columns of the co-occurrence matrix reduced to 300 distributional-semantic features with SVD
- In the distributional semantic model, each verb corresponds to an array of 300 values, i.e., a vector

<table>
<thead>
<tr>
<th></th>
<th>(column1)</th>
<th>(column2)</th>
<th>(column3)</th>
<th>(column300)</th>
</tr>
</thead>
<tbody>
<tr>
<td>find</td>
<td>15.59443</td>
<td>-2.022215</td>
<td>0.561186</td>
<td>...</td>
</tr>
<tr>
<td>carry</td>
<td>21.82777</td>
<td>4.714768</td>
<td>-11.974389</td>
<td>...</td>
</tr>
<tr>
<td>answer</td>
<td>11.66246</td>
<td>2.008967</td>
<td>8.810539</td>
<td>-0.5226300</td>
</tr>
<tr>
<td>push</td>
<td>22.09577</td>
<td>13.130336</td>
<td>-6.027978</td>
<td>0.8539545</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

- Semantically similar words tend to have similar values in the same features
Period vectors

- For each period, extract the semantic vector of each verb in the distribution of the construction
- Add all vectors and divide by the number of verbs: this is the period vector

<table>
<thead>
<tr>
<th>(column1)</th>
<th>(column2)</th>
<th>(column3)</th>
<th>(column300)</th>
</tr>
</thead>
<tbody>
<tr>
<td>make</td>
<td>14.09814</td>
<td>-4.23132</td>
<td>-1.84498</td>
</tr>
<tr>
<td>find</td>
<td>15.59443</td>
<td>-2.02215</td>
<td>0.51186</td>
</tr>
<tr>
<td>push</td>
<td>22.09577</td>
<td>13.13033</td>
<td>-6.027978</td>
</tr>
<tr>
<td>Sum</td>
<td>51.7834</td>
<td>6.876289</td>
<td>-7.31191</td>
</tr>
<tr>
<td>/3</td>
<td>17.26278</td>
<td>2.292096</td>
<td>-2.43723</td>
</tr>
</tbody>
</table>

- “Semantic average” of the distribution; reflects semantic properties of the verbs attested in the period
Distributional period clustering

- The VNC algorithm is run on the period vectors.
- Similarity is measured by cosines between vectors.
- The output dendrogram shows the semantic history of the construction:
  - Early mergers correspond to periods of semantic stability.
  - Late mergers of large clusters indicate semantic shifts.
Two case studies

- Both using COHA, focusing on verbs in two constructions
- The **hell**-construction \( V \) **the hell out of** NP
  
  \[
  \text{You scared the hell out of me!}
  \]
  
  \[
  \text{I enjoyed the hell out of that show.}
  \]
  
  \[
  \text{They beat the hell out of him.}
  \]

- The **way**-construction \( V \) one’s way PP
  
  \[
  \text{They hacked their way through the jungle.}
  \]
  
  \[
  \text{She talked her way into the club.}
  \]

  Restricted to the “path-creation” interpretation: the verb describes an action that enables motion

  (vs. manner: \text{They trudged their way through the snow})
The *hell*-construction

**VNC dendrogram**

**Token frequency (per million words)**

**Type frequency**

**Hapax legomena**
The *hell*-construction

- The shape of the dendrogram reflects gradual expansion rather than brutal shifts (cf. Perek 2014, 2016)
- Construction centered on the same semantic classes, with new members joining the periphery
- Vs. two-way split obtained with quantitative measures
- Questions the practice of using quantitative data for the initial partitioning


The *way*-construction

1830s – 1870s
Concrete, physical actions, literal creation of a path: *hew*, *shape*, *explore*, *carve*, *track*, *enforce*, *shoulder*, etc.

1880s: transition period
More abstract verbs than the previous period: *buy*, *smell*, *stammer*, *think*, *pay*, etc.

1890s – 2000s
More abstract: communication, social interaction, etc.: *joke*, *bellow*, *chatter*, *snarl*, *spit*, *laugh*, *talk*, *bully*, etc.

1880s
More concrete verbs than the next period: *bore*, *pierce*, *feel*, *wear*, *melt*, *trace*, *burn*, etc.
The *way*-construction

- Change from mostly concrete to more abstract verbs (in line with Israel 1996, Perek aop)
- How does distributional semantics compare to collostructional analysis for periodization?
  - Which verbs occur more distinctively frequently in each decade than in the others? (Hilpert 2006)
  - Each verb receives an association score in each decade
  - The distribution of collexemes can be used as input for VNC (Hilpert 2012): change in lexico-grammatical associations

Perek, F. (ahead-of-print). Recent change in the productivity and schematicity of the *way*-construction: a distributional semantic analysis. *Corpus Linguistics and Linguistic Theory*. 
VNC with collostructional analysis

Physical change of state: cut, hew, tear, cleave, break, pierce, burst, etc.
Semantically neutral verbs: take, find, win, make

Haphazard list of more abstract verbs:
earn, sing, advertise, brew, declaim, experiment
work, pick (1930s-2000s)
talk, buy, negotiate, lie (1910s-1920s)
VNC with collostructional analysis

- Some evidence of a shift from concrete to abstract verbs
- But it is attested later than in the distributional VNC
- Semantic classes are less clearly identifiable
- With collostructional analysis, the detection of changes is highly dependent on token frequency
  - Frequency associations are not always semantically relevant
  - “Real” change is only exemplified by high-frequency types
  - The timing of these changes is delayed, until sufficient frequency is reached
Conclusion

- Distributional period clustering captures semantic changes in the productivity of constructions
- Represents a step forward from regular VNC
- Results confirm previous studies
- Two advantages
  - Semantic changes are inferred mathematically rather than assessed impressionistically
  - Changes can be dated more precisely

... paper (with Martin Hilpert) under review, downloadable at www.fperek.net
Thanks for your attention!

f.b.perek@bham.ac.uk
www.fperek.net