UNIVERSITY OF BIRMINGHAM | COLLEGE OF ARTS AND LAW

A distributional semantic approach to identifying stages in constructional productivity change

Florent Perek

University of Birmingham

Overview

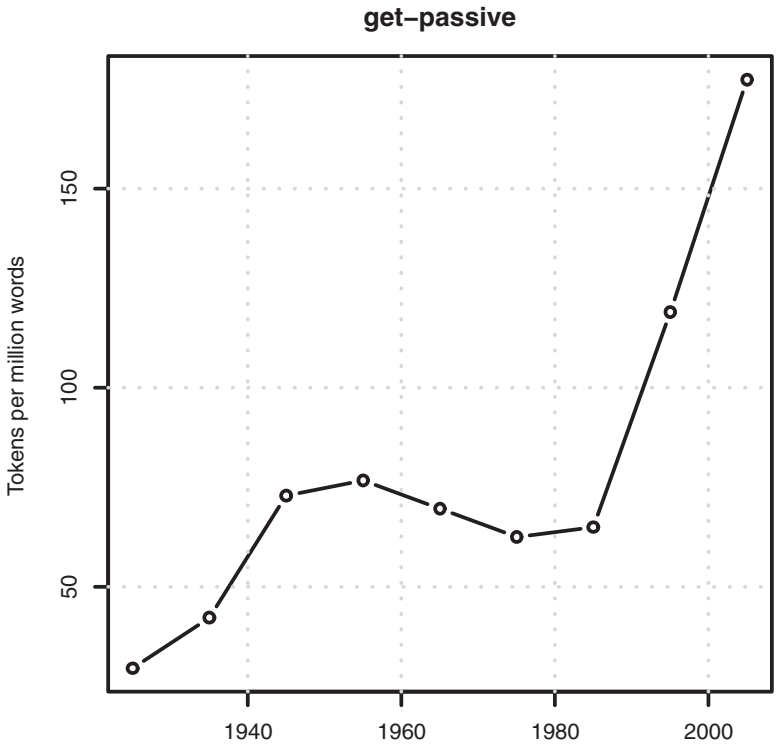
- New method for diachronic studies
- Aim: identify stages of language change in the productivity of constructions
- Combines variability-based neighbour clustering and distributional semantics
- Case study on the recent history of the *way*-construction

Usage-based approaches to the study 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

- Frequency of passive constructions from the 1920s onwards (TIMES corpus; source: Hilpert 2013: 30)



Hilpert, M. (2013). *Constructional Change in English. Developments in Allomorphy, Word Formation, and Syntax*. Cambridge: Cambridge University Press

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**

Gries, S., & Hilpert, M. (2008). The Identification of Stages in Diachronic Data: Variability-based Neighbor Clustering. *Corpora*, 3, 59–81.

The VNC algorithm

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

1. Look at all pairs of adjacent periods (e.g, 1830s-1840s, 1840s-1850s, 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 (Hilpert 2013: 36)

Hilpert, M. (2013). *Constructional Change in English. Developments in Allomorphy, Word Formation, and Syntax*. Cambridge: Cambridge University Press

VNC

- Most applications of VNC so far are based on quantitative variables:
  - Frequencies: tokens, types, hapax legomena etc.
  - Frequency distributions of lexical items
  - Distinctive collexeme analysis
- Main novelty of this work: include semantic information
- Especially appropriate for the study of productivity

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. (from 16th century)

  She talked her way into the club. (from 19th century)
- Type frequency often taken as an indicator of productivity
  - Number of different items, but not a measure of how different these items are
  - Need to consider the semantic diversity of the distribution

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

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)
- Distributional Semantic Models capture the meaning of words through their distribution in large corpora

Firth, J.R. (1957). A synopsis of linguistic theory 1930-1955. In *Studies in Linguistic Analysis*, pp. 1-32. Oxford: Philological Society.

Lenci, A. (2008). Distributional semantics in linguistic and cognitive research. *Rivista di Linguistica*, 20(1), 1–31.

Miller, G. & W. Charles (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, *6*(1), 1-28.

“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

the **upper crust;**	*cut*	a **lip** in it ; and ornament
growing **season.** “I	*spend*	a **lot** of my garden time
and disdainful **port;**	*looked*	intrepidly and indignantly
mocking me? What! I	*marry*	a **woman** sixty-four years old
that they no **longer**	*fight*	against it ; it is embalmed

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

Distributional semantic model

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

	(column1)	*(column2)*	*(column3)*		*(column300)*
find	15.59443	-2.022215	0.561186	...	-0.5778517
carry	21.82777	4.714768	-11.974389	...	-0.5226300
answer	11.66246	2.008967	8.810539	...	-0.2389049
push	22.09577	13.130336	-6.027978	...	0.8539545
...	...	...	...	...	...

- Each column is a distributional-semantic feature
- Semantically similar words tend to have similar values in the same features

Distributional period clustering

- Proposal: use distributional semantic to build representations of the semantic range of a construction
- Case study: the *way*-construction
  - E.g., *They pushed their way through the crowd*
  - Data: all instances in the COHA between 1830 and 2009
  - Manually filtered and annotated for constructional meaning:

    Path-creation: the verb describes what enables motion

    They hacked their way through the jungle.

    Manner: the verb describes the manner of motion

    They trudged their way through the snow.

Period vectors

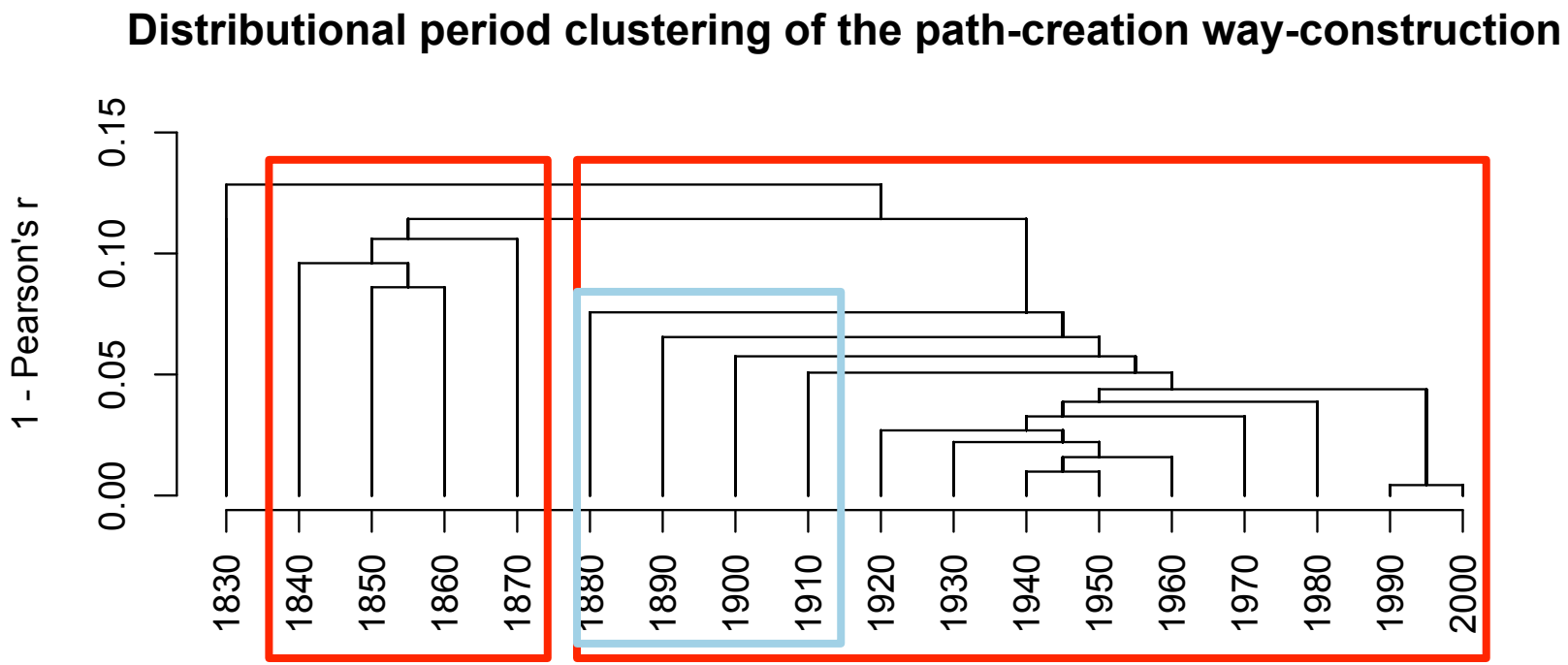
- For each period, extract the semantic vector of each verb in the distribution of the construction
- Add all vectors and divide by the number of verbs: this is the period vector.

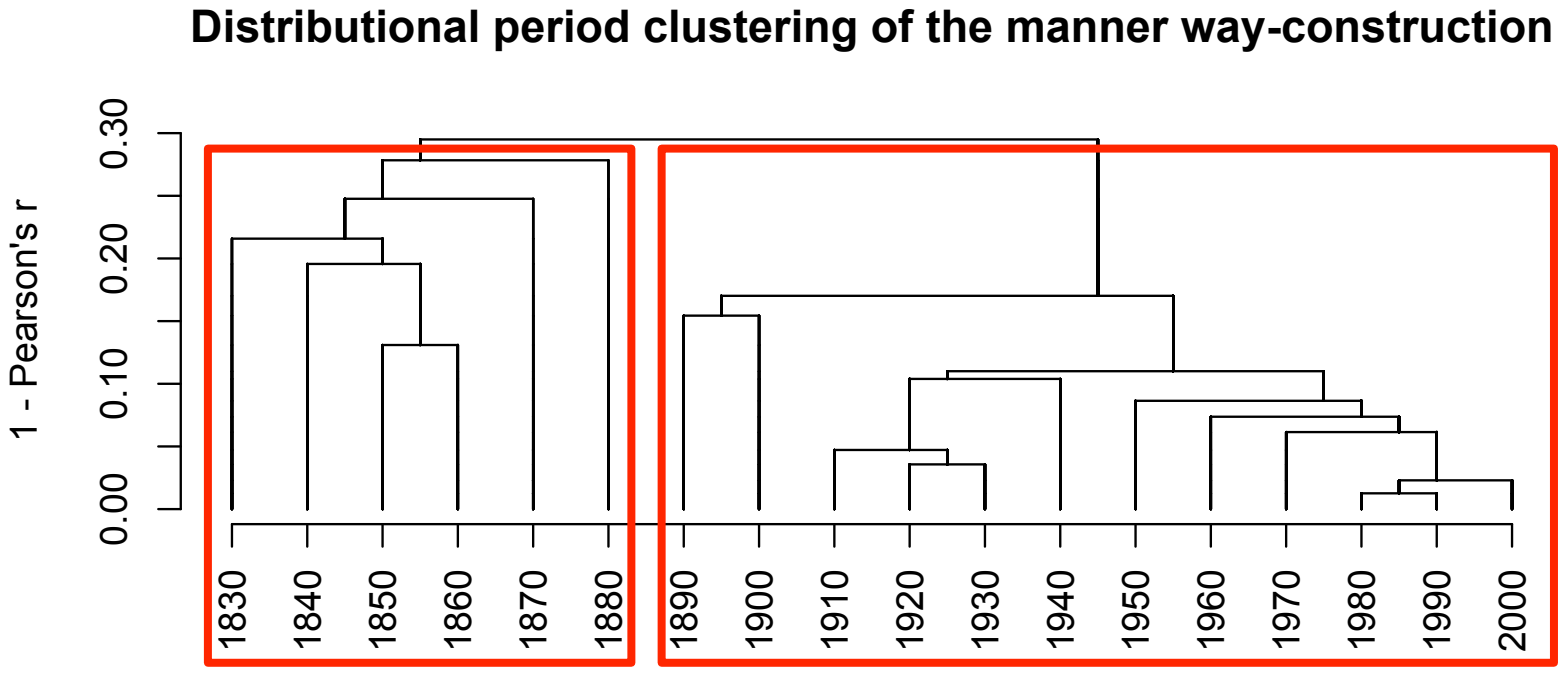
		(column1)	*(column2)*	*(column3)*		*(column300)*
	make	14.09814	-4.231832	-1.844898	...	0.06963598
	find	15.59443	-2.022215	0.561186	...	-0.5778517
	push	22.09577	13.130336	-6.027978	...	0.8539545
Sum		51.78834	6.876289	-7.311691	...	0.3457388
/3		**17.26278**	**2.292096**	**-2.43723**	...	**0.1152463** ← period vector

- "Semantic average" of the distribution.
- Features of the period vector reflect semantic properties of the verbs attested in the period

Distributional period clustering

- The VNC algorithm is run on the period vectors
- Similarity between periods is measured by Pearson's *r*
- 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.


Distributional period clustering of the path-creation way-construction
1 - Pearson's r
0.00
0.05
0.10
0.15
1830
1840
1850
1860
1870
1880
1890
1900
1910
1920
1930
1940
1950
1960
1970
1980
1990
2000


Distributional period clustering of the manner way-construction
1 - Pearson's r
0.00
0.10
0.20
0.30
1830
1840
1850
1860
1870
1880
1890
1900
1910
1920
1930
1940
1950
1960
1970
1980
1990
2000

Interpreting period clustering

- How to characterize each period?
  - The distributional-semantic features are highly abstract and not directly interpretable
  - The only way to interpret semantic changes is to look at the verb themselves
- How do verbs in each period relate to the semantic range of their period vs. the surrounding periods?

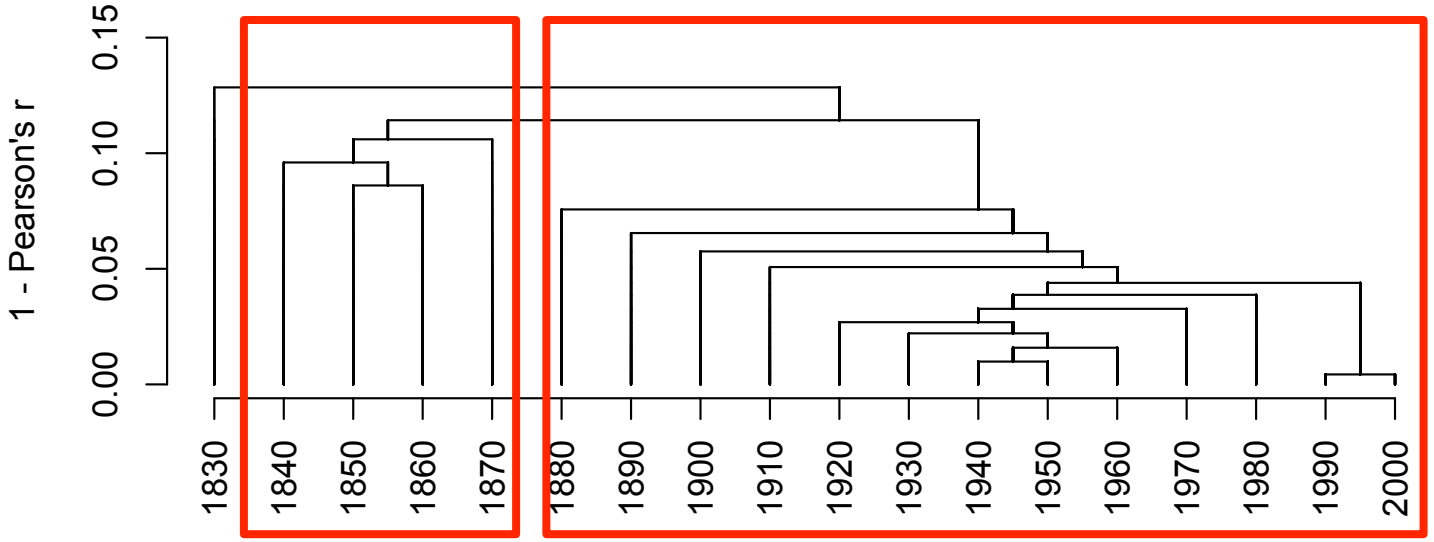
Interpreting period clustering

- For all verbs in a period, calculate the difference between:
  - The similarity of the verb vector to the period vector
  - And the similarity of the verb vector to a surrounding period

    i.e., $\textit{similarity}(V_{period}, V_{verb}) - \textit{similarity}(V_{period+1}, V_{verb})$

    or $\textit{similarity}(V_{period}, V_{verb}) - \textit{similarity}(V_{period-1}, V_{verb})$
  - Similarity measured by Pearson's *r*
- Positive differences indicate that the verb is more typical of that period than of the neighbouring period
- The verbs with the highest differences should provide an indication of semantic change in either direction

Distributional period clustering of the path-creation way-construction



Many concrete, physical actions: exertion of a force, change of state, etc.

Literal creation of a physical path

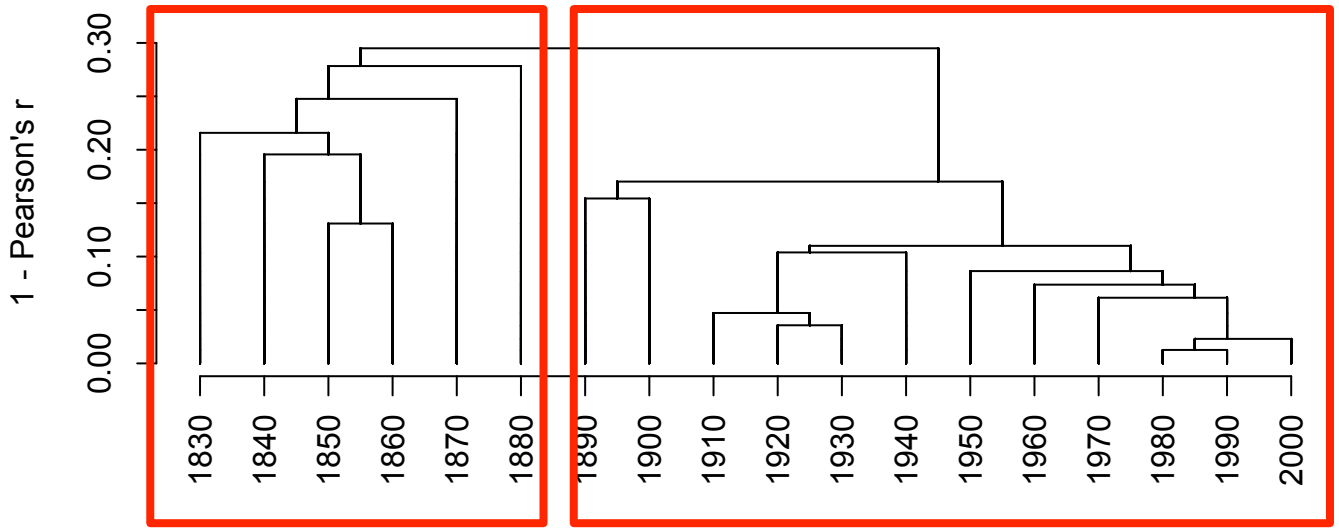
pierce	0.0626
rend	0.0593
tear	0.0512
trace	0.0466
break	0.0457
probe	0.0440
strike	0.0425
conquer	0.0402
rip	0.0400
explore	0.0397
shape	0.0394
crush	0.0367

talk	0.0958
laugh	0.0937
joke	0.0833
chat	0.0792
kid	0.0787
smile	0.0722
chatter	0.0716
bawl	0.0683
shrug	0.0683
nod	0.0679
grin	0.0660
mumble	0.0660

More abstract actions: communication, social interaction, etc.

Creation of a metaphorical path

Distributional period clustering of the manner way-construction



Motion involving difficulty: obstacles, difficult terrain, clumsiness

toil	0.0596
plow	0.0506
plough	0.0498
tread	0.0456
plod	0.0454
wind	0.0380
ply	0.0339
climb	0.0335
thread	0.0323
urge	0.0295
clamber	0.0271
trudge	0.0216

bob	0.0667
swirl	0.0527
twirl	0.0495
blink	0.0480
filter	0.0462
stomp	0.0436
spin	0.0426
skim	0.0424
strut	0.0410
rock	0.0409
curl	0.0400
bounce	0.0397

Difficulty component less prominent, more “neutral” manners of motion

Verbs describing movements of the agent

Summary

- Period clustering identifies two broad semantic changes
- 1) in the path-creation *way*-construction
  - Shift from physical path creation to more abstract means
  - Started in the 1880s, gradual expansion
- 2) in the manner *way*-construction
  - Shift from difficult motion from general manner of motion
  - Started in the 1890s
- In line with the findings of Perek (to appear)

Perek, F. (to appear). Recent change in the productivity and schematicity of the *way*-construction: a distributional semantic analysis. To appear in *Corpus Linguistics and Linguistic Theory*.

Conclusion

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

UNIVERSITY OF BIRMINGHAM | COLLEGE OF ARTS AND LAW

Thanks for your attention!

f.b.perek@bham.ac.uk

www.fperek.net