UNIVERSITY OF BIRMINGHAM | COLLEGE OF ARTS AND LAW

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

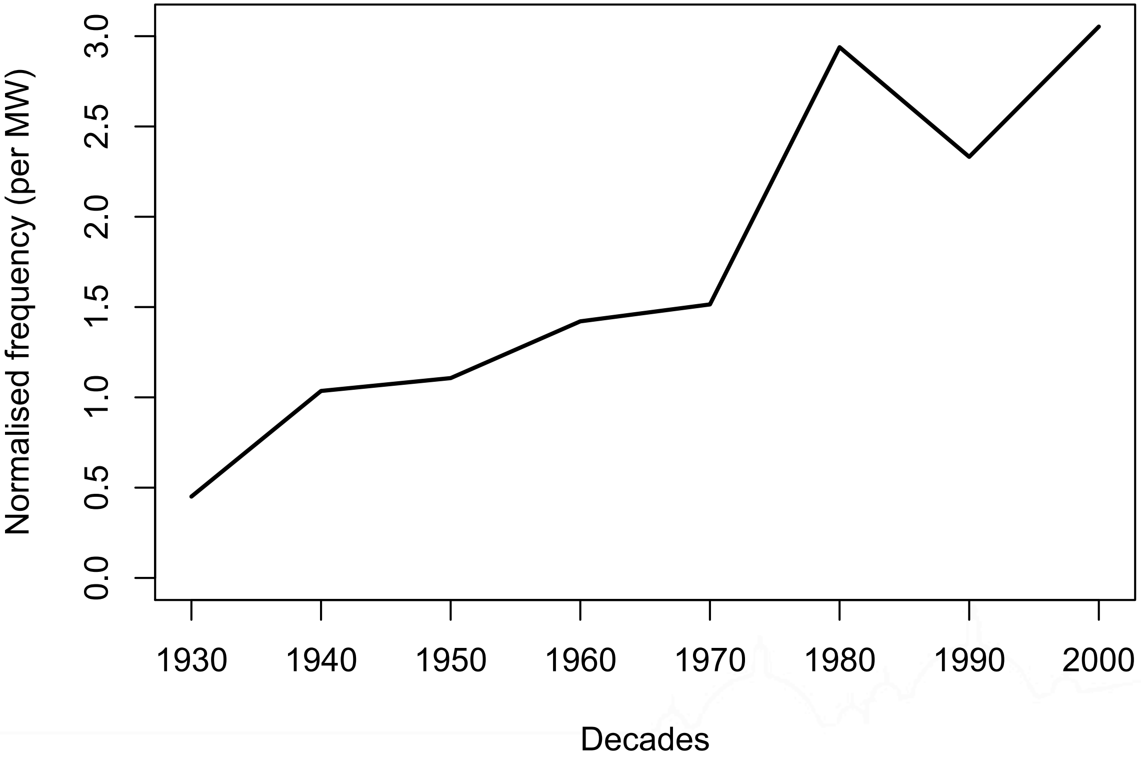
Corpus-based studies of language change

- □ 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!*



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

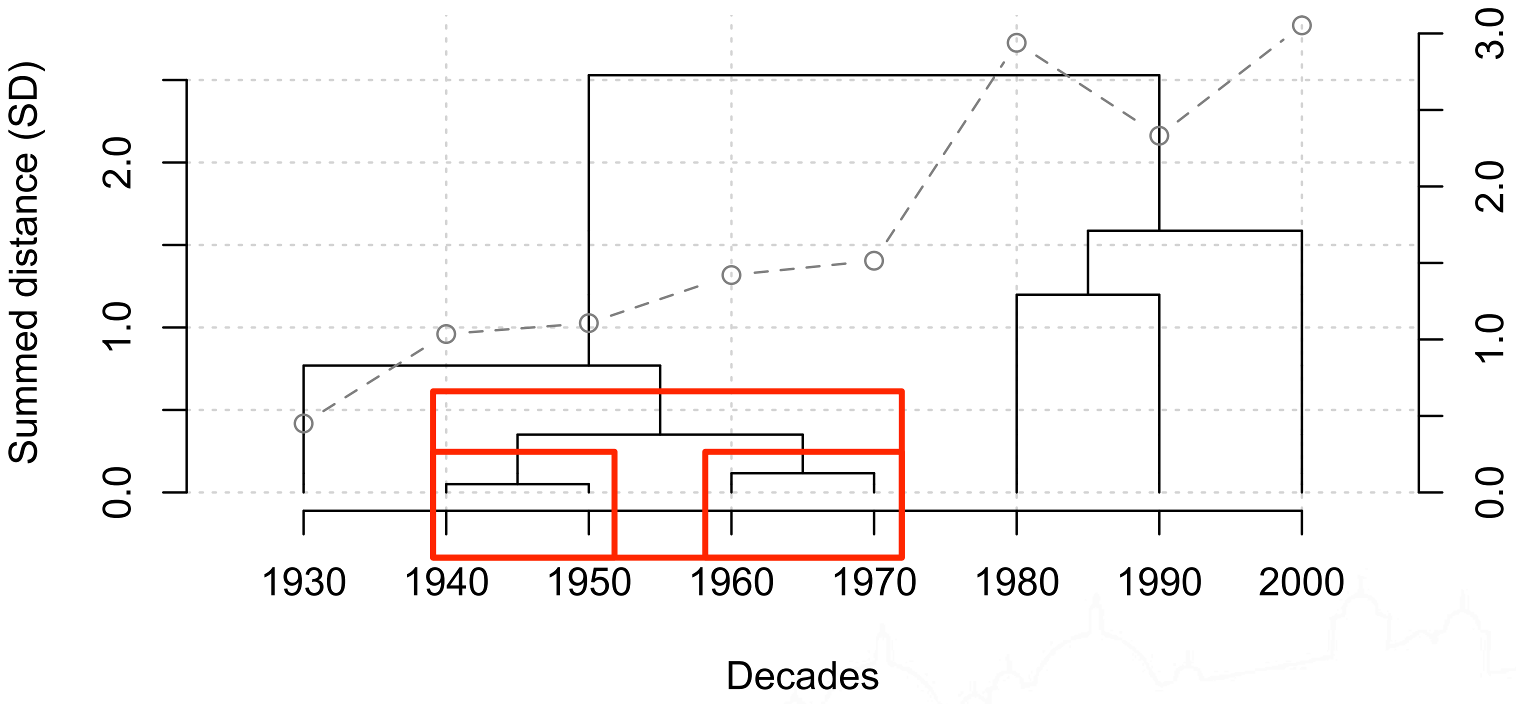
- Starting point: data partitioned into “natural” time periods (years, decades, etc.)

1. Look at all pairs of adjacent periods (e.g., 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

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)
- □ 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

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 distributional-semantic features 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
...       ...      ...         ...        ...  ...
```

- □ 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

		(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; 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

  You scared the hell out of me!

  I enjoyed the hell out of that show.

  They beat the hell out of him.

- The *way*-construction V *one's way* PP

  They hacked their way through the jungle.

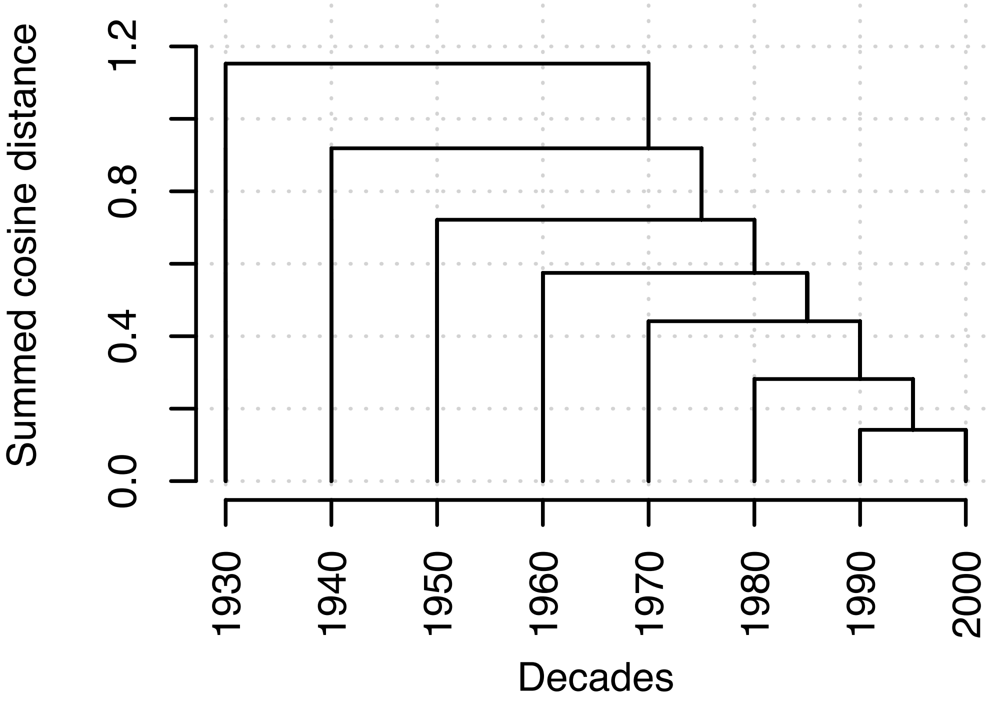
  She talked her way into the club.

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

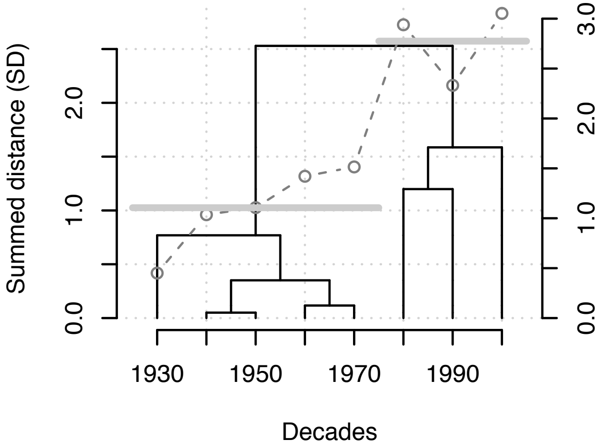
  (vs. manner: *They trudged their way through the snow*)

The *hell*-construction

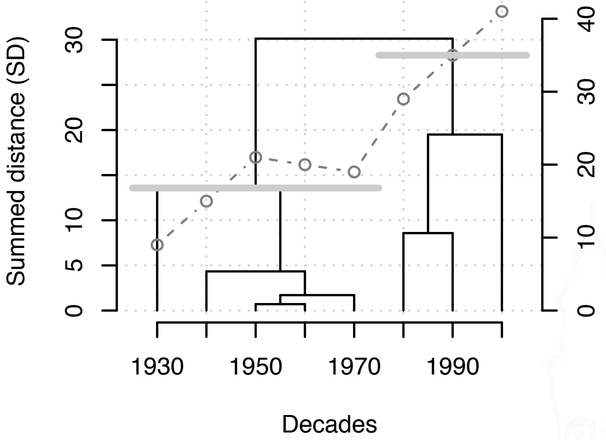
VNC dendrogram



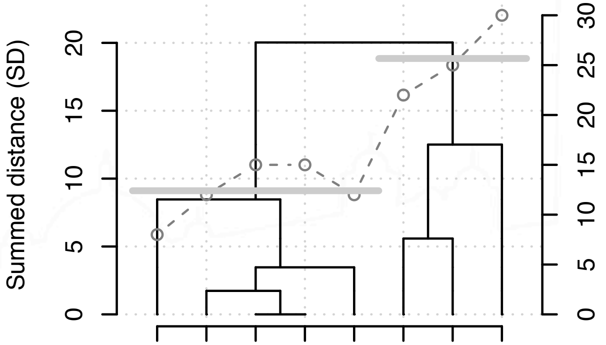
Token frequency (per million words)



Type frequency



Hapax legomena



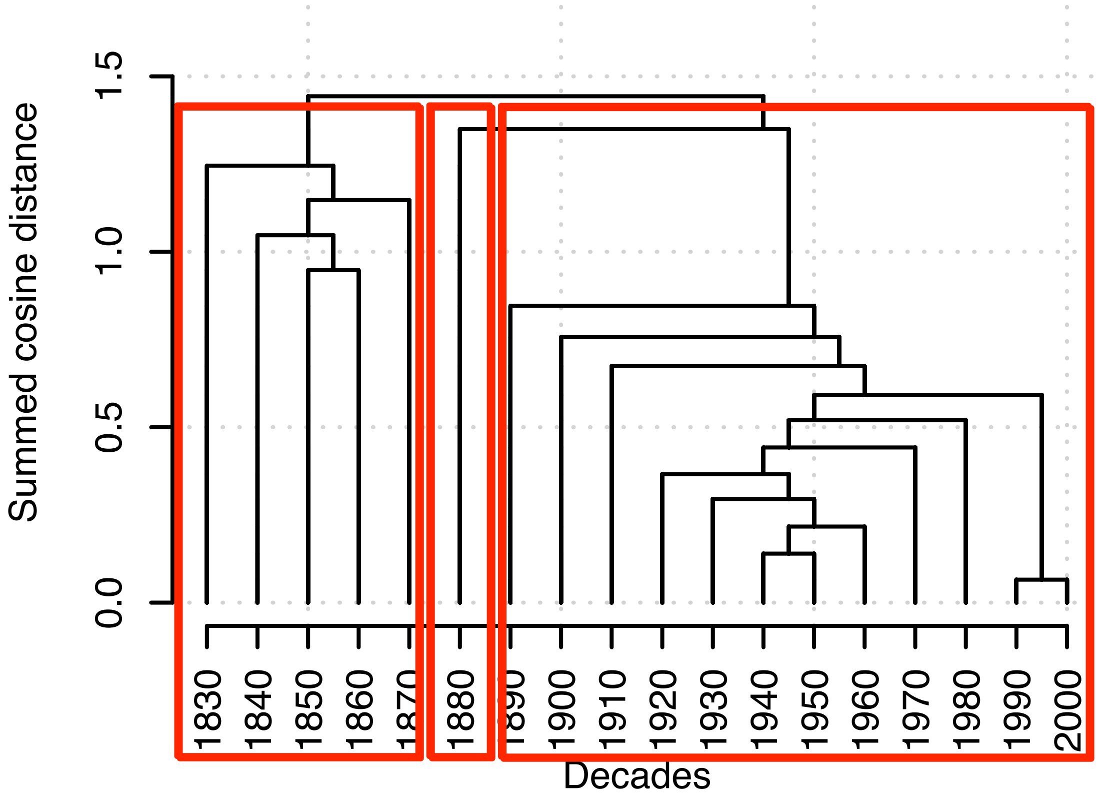
The *hell*-construction

- 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

Perek, F. (2014). Vector spaces for historical linguistics: Using distributional semantics to study syntactic productivity in diachrony. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, Maryland USA, June 23-25 2014* (pp. 309-314).

Perek, F. (2016). Using distributional semantics to study syntactic productivity in diachrony: A case study. *Linguistics*, 54(1), 149–188.

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*, *begin*, *think*, *pay*, etc.

More concrete verbs than the next period: *pierce*, *feel*, *wear*, *melt*, *trace*, *burn*, etc.

1890s – 2000s

Most abstract: communication, social interaction, etc.

joke, *bellow*, *chatter*, *snarl*, *spit*, *laugh*, *talk*, *bully*, etc.

The *way*-construction

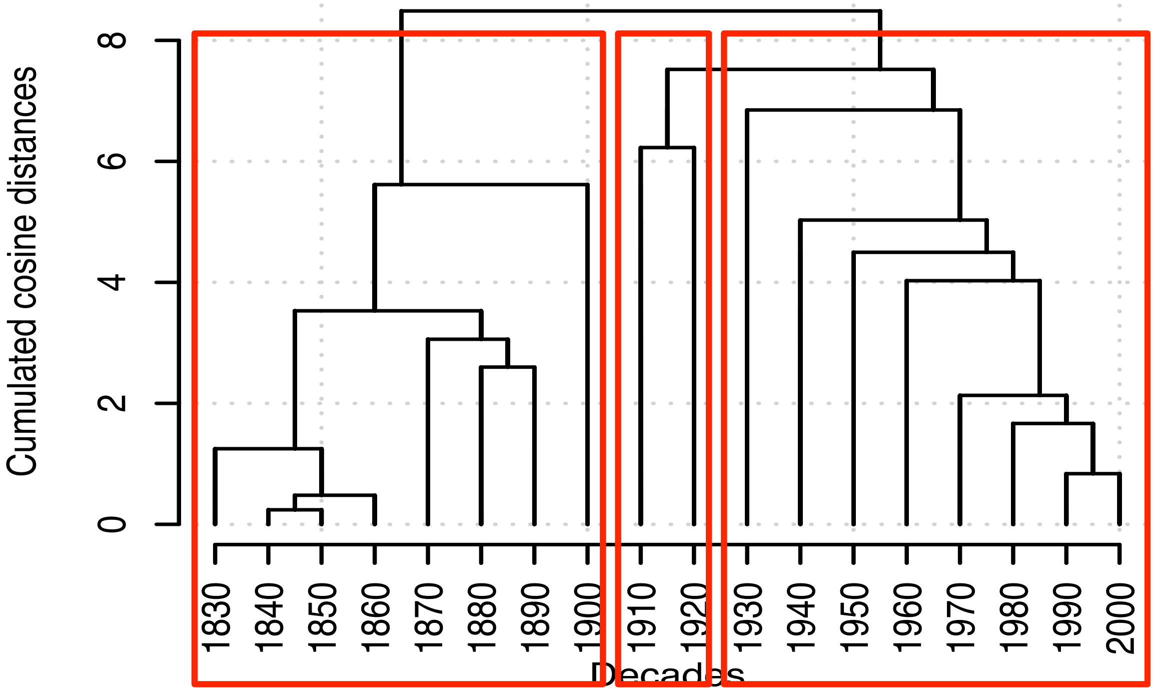
- Change from mostly concrete to more abstract verbs (in line with Israel 1996, Perek aop)
- How does distributional semantics compare to collostructional analysis for periodization?
  - Which verbs occur more distinctively frequently in each decade than in the others? (Hilpert 2006)
  - Each verb receives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

Hilpert, M. 2006. Distinctive collexeme analysis and diachrony. *Corpus Linguistics and Linguistic Theory* 2(2). 243–57.

Hilpert, M. 2012. Diachronic collostructional analysis. How to use it, and how to deal with confounding factors. In K. Allan & J. Robynson (eds.), *Current Methods in Historical Semantics*, 133–160. Berlin: Mouton de Gruyter.

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*
(1910s-1920s)

talk, *buy*, *negotiate*, *lie*
(1930s-2000s)

work, *pick* (1930s-2000s)

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

UNIVERSITY OF BIRMINGHAM | COLLEGE OF ARTS AND LAW

Thanks for your attention!

f.b.perek@bham.ac.uk

www.fperek.net