UNIVERSITY OF BIRMINGHAM | COLLEGE OF ARTS AND LAW

Big Data, Big Meaning

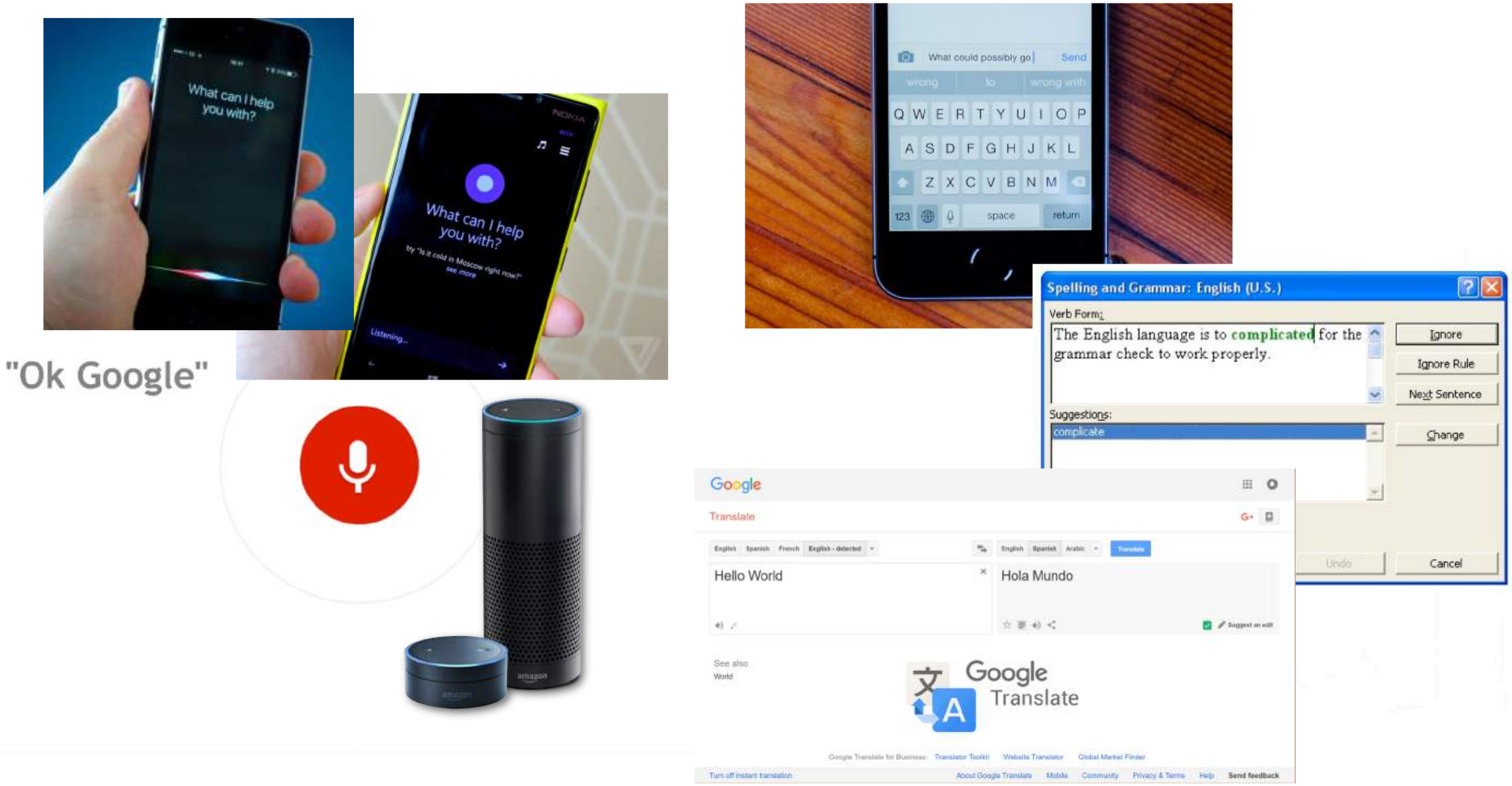
Using distributional semantics in linguistic research

Florent Perek

University of Birmingham

Natural language processing (NLP)

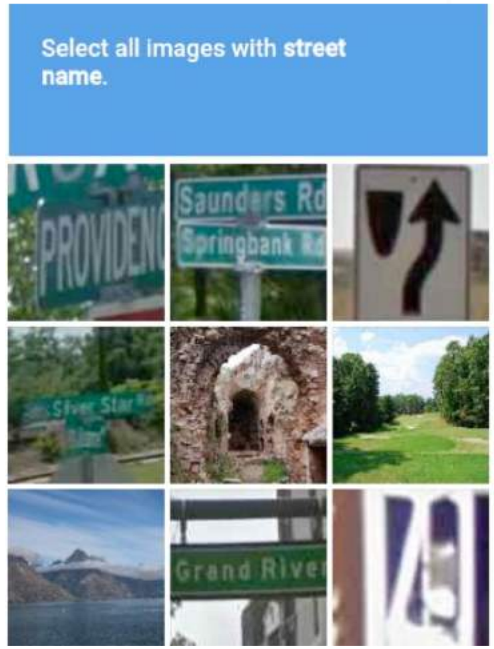
Programming computers to use human language



Natural language processing (NLP)

- NLP is everywhere
- Fast change that happened over the last 10-15 years
  - Increasingly advanced statistical processing
  - Big Data

$$\begin{array}{r} 12765 \\ \times \quad 964 \\ \hline ? \end{array}$$



NLP and linguistics

- NLP has produced many techniques to process large amount of data and extract linguistic information from it
- Linguistic research can benefit a lot from these techniques
- Case in point: distributional semantics

I am fluent in over six million forms of communication

Distributional semantics

"You shall know a word by the company it keeps"

Firth (1957: 11)

Semantic knowledge → knowing *when* to use words

Contexts of use are a source of semantic information

Firth, J.R. (1957). A synopsis of linguistic theory 1930-1955. In *Studies in linguistic analysis (Special volume of the Philological Society)*, 1–32. Oxford: Blackwell.

Guess the missing word...

that. He was stood in front of me in the	**xxxxxx**	queue the other day and [unclear] .
On the station he bought a	**xxxxxx**	and a cup of tea. He was surprised
be located, how to prepare a salami	**xxxxxx**	, and what to do if you should come
was quite expensive so I've bought a	**xxxxxx**	in the shop instead. That's a normal
probably use to describe an indifferent	**xxxxxx**	. 'A bit too smooth, though.' 'He
nowhere till I've had a hot pastrami	**xxxxxx**	.' We crowded into a mêlée like the
that knowing how to make a Marmite	**xxxxxx**	would be enough. I pressed on. The
for a stroll to the pub for a drink and a	**xxxxxx**	, they had spent nearly seventeen
but I weren't sure if it was my fish paste	**xxxxxx**	or not! Shit! Just got a whiff as soon
fat-free yoghurt. Supper Wholemeal	**xxxxxx**	with low-fat cream cheese and bana
and if not, whether he should get a	**xxxxxx**	in a pub instead, and if so, whether
of there. Well I like to have a toasted	**xxxxxx**	for dinner. I forget about it. Yeah, bu
up a [pause] plate Mhm. and I took the	**xxxxxx**	over Mhm. and I eat it and I went,

Sandwich

that. He was stood in front of me in the	**sandwich**	queue the other day and [unclear] .
On the station he bought a	**sandwich**	and a cup of tea. He was surprised
be located, how to prepare a salami	**sandwich**	, and what to do if you should come
was quite expensive so I've bought a	**sandwich**	in the shop instead. That's a normal
probably use to describe an indifferent	**sandwich**	. 'A bit too smooth, though.' 'He
nowhere till I've had a hot pastrami	**sandwich**	.' We crowded into a mêlée like the
that knowing how to make a Marmite	**sandwich**	would be enough. I pressed on. The
for a stroll to the pub for a drink and a	**sandwich**	, they had spent nearly seventeen
but I weren't sure if it was my fish paste	**sandwich**	or not! Shit! Just got a whiff as soon
fat-free yoghurt. Supper Wholemeal	**sandwich**	with low-fat cream cheese and bana
and if not, whether he should get a	**sandwich**	in a pub instead, and if so, whether
of there. Well I like to have a toasted	**sandwich**	for dinner. I forget about it. Yeah, bu
up a [pause] plate Mhm. and I took the	**sandwich**	over Mhm. and I eat it and I went,

Guess the missing word...

then [unclear] It was really part of the	**xxxxxx**	. Mhm. Mhm. So did he [unclear] Do
together.' Hastings knew he'd got the	**xxxxxx**	on Sunday night, and while he
to be generally accepted that their	**xxxxxx**	is by no means sinecure. Accordingly,
the Hawick-based knitters showed	**xxxxxx**	opportunities for at least 50 skilled
poet John Wain was clearly doing his	**xxxxxx**	. One aspect of the Lewis regime
Write-in: I could do a better	**xxxxxx**	if I knew more about Line
he was sacked from the manager's	**xxxxxx**	at Preston in 1981 he immediately told
Arena today: 'I go out and do a	**xxxxxx**	on anyone who is giving our top
to be 'professionalized', experts at our	**xxxxxx**	. But sadly our world suffers because
in the structure clearly identified by	**xxxxxx**	descriptions and departmental
As is the norm in such projects, every	**xxxxxx**	turned out twice as extensive and
grinning from ear to ear with his latest	**xxxxxx**	. He has landed a plum role as the
courses (part teaching, part practical	**xxxxxx**	experience), while universities tend to

Job

then [unclear] It was really part of the	**job**	. Mhm. Mhm. So did he [unclear] Do
together.' Hastings knew he'd got the	**job**	on Sunday night, and while he
to be generally accepted that their	**job**	is by no means sinecure. Accordingly,
the Hawick-based knitters showed	**job**	opportunities for at least 50 skilled
poet John Wain was clearly doing his	**job**	. One aspect of the Lewis regime
Write-in: I could do a better	**job**	if I knew more about Line
he was sacked from the manager's	**job**	at Preston in 1981 he immediately told
Arena today: 'I go out and do a	**job**	on anyone who is giving our top
to be 'professionalized', experts at our	**job**	. But sadly our world suffers because
in the structure clearly identified by	**job**	descriptions and departmental
As is the norm in such projects, every	**job**	turned out twice as extensive and
grinning from ear to ear with his latest	**job**	. He has landed a plum role as the
courses (part teaching, part practical	**job**	experience), while universities tend to

Guess the missing word...

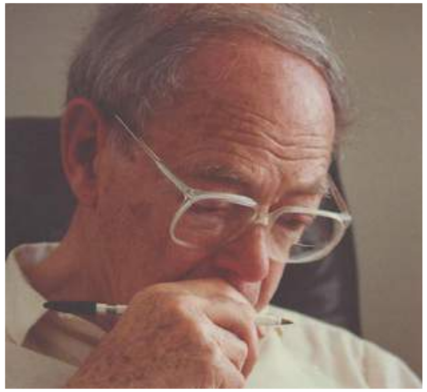
of government. We will give a Cabinet	**xxxxxx**	responsibility for the Citizen's
the notes on clauses. I hope that the	**xxxxxx**	will clear that up when he replies.
rapidly-declining stocks. Fisheries	**xxxxxx**	John Crosbie denied, however, that
was elected as LDP leader and Prime	**xxxxxx**	in August 1989 [see pp. 36849-50].
party. My right hon. Friend the Prime	**xxxxxx**	was absolutely right to describe it as
in May or June [see ED 67]. Fisheries	**xxxxxx**	Jan Henry Olsen said a quota would
Dame Cath Tizard. Prime	**xxxxxx**	: Jim Bolger (since October 1990;
initiative on Aug. 15 the Iranian Foreign	**xxxxxx**	, Ali Akbar Vellayati, described it as
The new Science and Technology	**xxxxxx**	sees information as an instrument of
Majorism isn't working? The Prime	**xxxxxx**	as the right hon. Gentleman is now
it's a moral problem, problem. The	**xxxxxx**	said, no it isn't, it's an economic
civil servant Sir Humphrey would tell his	**xxxxxx**	whenever the hapless Hacker
trade union paper Hodolmor, the new	**xxxxxx**	of Labour, Choyjamtsyn

Minister

of government. We will give a Cabinet	**minister**	responsibility for the Citizen's
the notes on clauses. I hope that the	**minister**	will clear that up when he replies.
rapidly-declining stocks. Fisheries	**minister**	John Crosbie denied, however, that
was elected as LDP leader and Prime	**Minister**	in August 1989 [see pp. 36849-50].
party. My right hon. Friend the Prime	**Minister**	was absolutely right to describe it as
in May or June [see ED 67]. Fisheries	**Minister**	Jan Henry Olsen said a quota would
Dame Cath Tizard. Prime	**Minister**	: Jim Bolger (since October 1990;
initiative on Aug. 15 the Iranian Foreign	**minister**	, Ali Akbar Vellayati, described it as
The new Science and Technology	**minister**	sees information as an instrument of
Majorism isn't working? The Prime	**Minister**	as the right hon. Gentleman is now
it's a moral problem, problem. The	**minister**	said, no it isn't, it's an economic
civil servant Sir Humphrey would tell his	**minister**	whenever the hapless Hacker
trade union paper Hodolmor, the new	**minister**	of Labour, Choyjamtsyn

Distributional semantics

“[I]f we consider words or morphemes A and B to be more different in meaning than A and C, then we will often find that the distributions of A and B are more different than the distributions of A and C. In other words, difference of meaning correlates with difference of distribution.”



Harris (1954: 156)

Harris, Z. (1954). Distributional structure. *Word* 10(23). 146–162.

Example: *drink* and *sip*

Sentences from the COCA corpus:

the pizzeria for a while,	**drinking**	a beer at a table
hell, I'd meet you,	**drink**	a glass of beer or
books. She changed her dress,	**drank**	a glass of cold water
Willie picks up his cup,	**drinks**	some coffee, and leaves with
men picked up their beers,	**sipped**	them, and put them back
to trust his intuition. She	**sipped**	from the champagne glass and
food itself. Even when he	**sipped**	his cold beer, it was
Emily was no different. Kate	**sipped**	from her water bottle, then

Example: *drink* and *sip*

the pizzeria for a while,	**drinking**	a **beer** at a table
hell, I'd meet you,	**drink**	a glass of **beer** or
books. She changed her dress,	**drank**	a glass of cold **water**
Willie picks up his cup,	**drinks**	some **coffee**, and leaves with
men picked up their **beers**,	**sipped**	them, and put them back
to trust his intuition. She	**sipped**	from the **champagne** glass and
food itself. Even when he	**sipped**	his cold **beer**, it was
Emily was no different. Kate	**sipped**	from her **water** bottle, then

Beverages

Example: *drink* and *sip*

the pizzeria for a while,	**drinking**	a **beer** at a table
hell, I'd meet you,	**drink**	a **glass** of **beer** or
books. She changed her dress,	**drank**	a **glass** of cold **water**
Willie picks up his **cup**,	**drinks**	some **coffee**, and leaves with
men picked up their **beers**,	**sipped**	them, and put them back
to trust his intuition. She	**sipped**	from the **champagne glass** and
food itself. Even when he	**sipped**	his cold **beer**, it was
Emily was no different. Kate	**sipped**	from her **water bottle**, then

Beverages

Containers for beverages

Example: *drink* and *sip*

the pizzeria for a while,	**drinking**	a beer at a table
hell, I'd meet you,	**drink**	a glass of beer or
books. She changed her dress,	**drank**	a glass of cold water
Willie picks up his cup,	**drinks**	some coffee, and leaves with
men picked up their beers,	**sipped**	them, and put them back
to trust his intuition. She	**sipped**	from the champagne glass and
food itself. Even when he	**sipped**	his cold beer, it was
Emily was no different. Kate	**sipped**	from her water bottle, then

Beverages

Containers for beverages

Drinking and dining

‘Bag-of-words’ approach

Based on the frequency of co-occurrence between words in a large corpus

Count how many times each word occurs with each other word within a set context window

E.g., collocates of the verbs *answer*, *carry*, *push*, *reply*, and *tell* within a +/- 2 word window in the COHA corpus (400 MW)

	question	lift	heavy	softly	…
answer	5854	44	13	119	…
carry	56	66	512	27	…
push	41	28	58	27	…
reply	201	40	3	66	…
tell	229	16	36	81	…

‘Bag-of-words’ approach

Co-occurrence counts often replaced by association scores

I.e., how strong is the association between two words, given the individual frequency of these words?

Typical association measure: Positive Pointwise Mutual Information (PPMI)

	question	lift	heavy	softly	...
answer	3.8523	1.0399	0	1.1807	...
carry	0	1.1074	2.21	0	...
push	0	1.3181	1.1003	0.4276	...
reply	0.7709	1.2347	0	0.8814	...
tell	0	0	0	0	

‘Bag-of-words’ approach

The rows of the matrix are called vectors

→ **vector space models**

	question	lift	heavy	softly	...
answer	3.8523	1.0399	0	1.1807	...
carry	0	1.1074	2.21	0	...
push	0	1.3181	1.1003	0.4276	...
reply	0.7709	1.2347	0	0.8814	...
tell	0	0	0	0	

vector

The matrix is often reduced to a lower number of dimensions (e.g., by means of Singular Value Decomposition)

‘Bag-of-words’ approach

	(column 1)	(column 2)	(column 3)		(column 300)
answer	11.662463	2.00896724	8.810539	…	-0.2389049
carry	21.827765	4.71476816	-11.974389	…	-0.52263
push	22.095771	13.130336	-6.027978	…	0.8539545
reply	15.407709	1.90698674	13.22548	…	-0.246191
tell	7.926409	0.06556502	4.79983	…	-0.3177306

Abstract distributional-semantic features corresponding to a large set of collocates

Vectors with similar values are expected to correspond to words with similar meaning

Similarity

Semantic similarity is measured by mathematical similarity between word vectors

Most common measure: cosine

1: the vectors are identical

0: maximally dissimilar

	answer	carry	push	reply	tell
answer	1	0.1871	0.2960	0.9241	0.6461
carry	0.1871	1	0.5787	0.1622	0.1514
push	0.2960	0.5787	1	0.2581	0.2314
reply	0.9241	0.1622	0.2581	1	0.6774
tell	0.6461	0.1514	0.2314	0.6774	1

Benefits

- Data-driven: more objective than ‘intuitive’ approach
- No manual intervention needed
- No limits on the number of lexical items
- Precise quantification
- Robust, adequately reflects semantic intuitions
  - Correlates with human performance in various tasks (e.g., Landauer et al. 1998, Lund et al. 1995)
  - Evidence for psychological adequacy (Andrews & Vigliocco 2008)

Andrews, Mark, Gabriella Vigliocco & David P. Vinson. 2009. Integrating Experiential and Distributional Data to Learn Semantic Representations. *Psychological Review* 116(3). 463–498.

Landauer, Thomas K., Peter W. Foltz & Darrell Laham. 1998. Introduction to Latent Semantic Analysis. *Discourse Processes* 25. 259–284.

Lund, Kevin, Curt Burgess & Ruth A. Atchley. 1995. Semantic and associative priming in a high-dimensional semantic space. In *Cognitive Science Proceedings (LEA)*, 660–665.

Using distributional semantics

- Distributional semantics is a robust way to capture semantic similarity, widely used in NLP
- How can it be used in linguistic research? Two methods:
  - Distributional semantic plots

    To visualize the semantic spread of a set of words
  - Distributional clustering

    To partition semantic development into stages
- Case studies in historical linguistics

Productivity

- The range of lexical items that can be used in the slots of a construction
- E.g., verbs in the "*hell*-construction": V *the hell out of* NP (Perek 2014, 2016)

  *You **scared** the hell out of me!*

  *I **enjoyed** the hell out of that show!*

  *But you **drove** the hell out of it!*

  *I've been **listening** the hell out of your tape.*

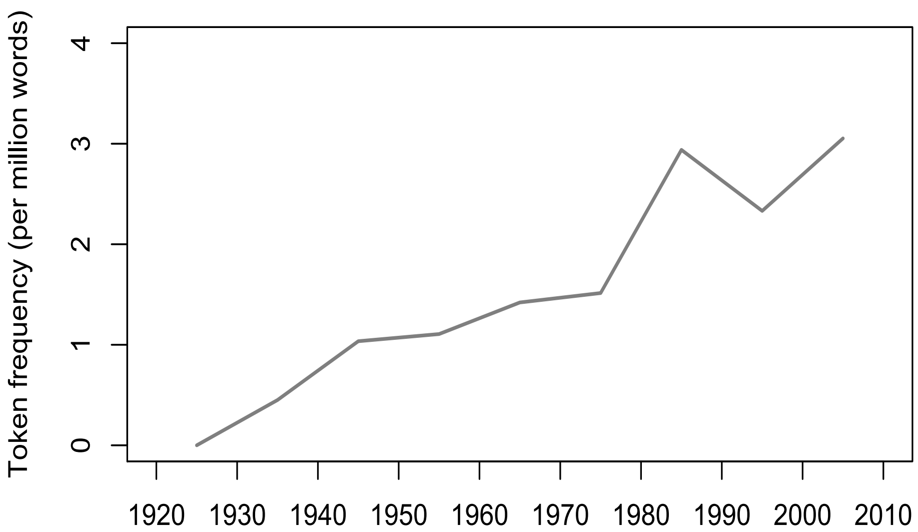
  *I **voiced** the hell out of 'b'* (heard at GURT 2014, Georgetown)

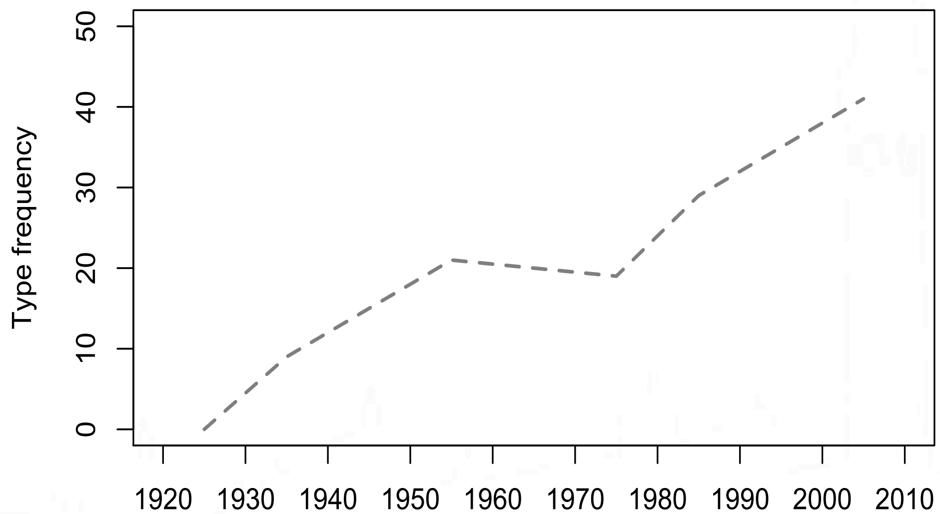
Perek, F. (2014). Vector spaces for historical linguistics: Using distributional semantics to study syntactic productivity in diachrony. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, Maryland USA, June 23-25 2014* (pp. 309-314).

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

The *hell*-construction in the COHA

- Recent construction: first instances in the 1930s
- Increasingly popular
- More and more verbs in the construction
- But how different are these verbs?

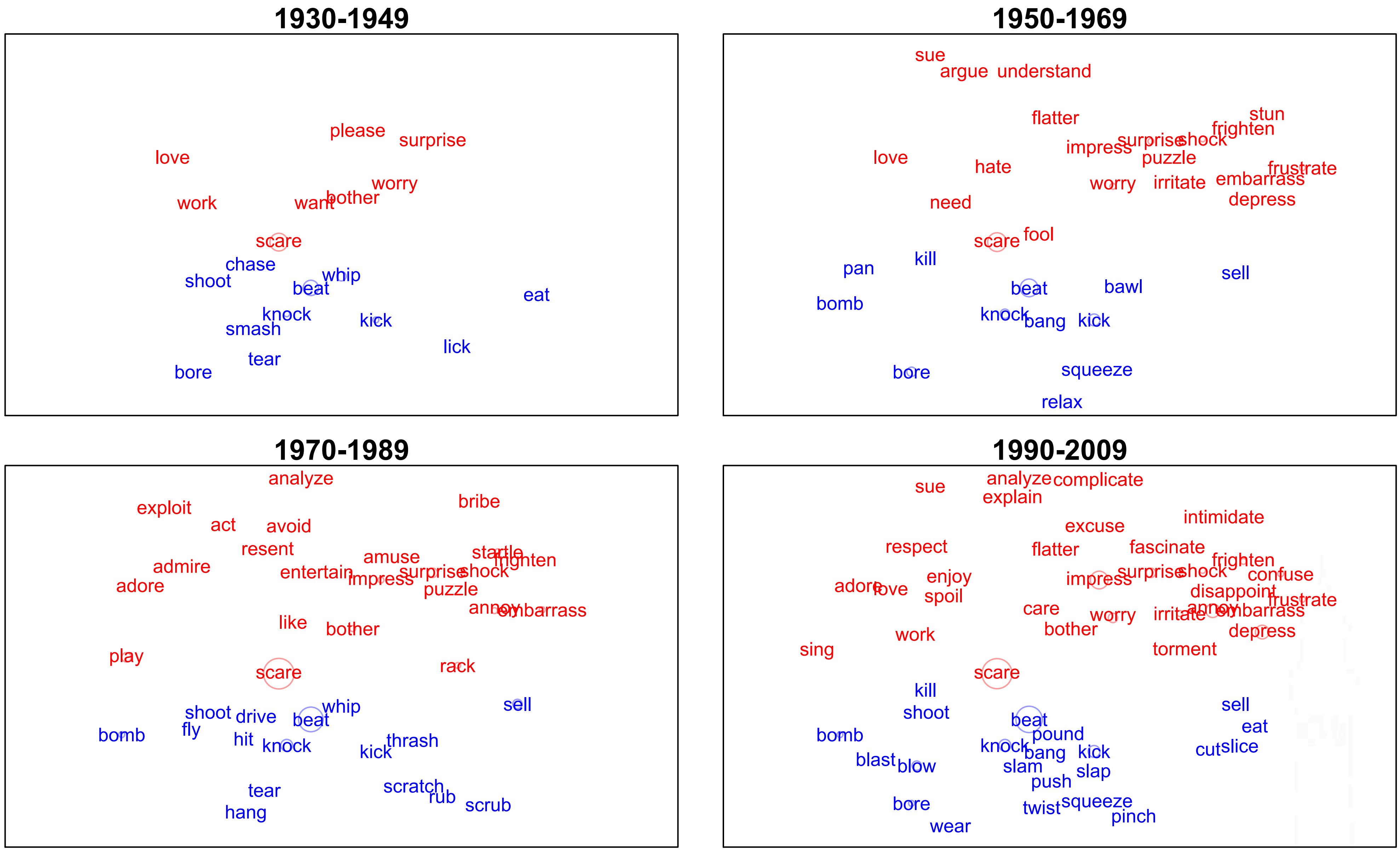




Distributional semantic plots

- Method to visualise the semantic space filled by a certain set of words
- Pairwise semantic distances are derived from a distributional semantic model
- Converted to a set of coordinates and plotted
  - E.g., with multidimensional scaling (MDS) or *t*-SNE (Van der Maaten & Hinton 2008)
  - Place objects in a 2-dimensional space such that the between-object distances are preserved as well as possible

Van der Maaten, L. & Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, *9*, 2579-2605.



Red: emotions, feelings, thoughts, mental activities

Blue: violent contact, exertion of force

Two domains of predilection

- Cognition verbs

  bother, *disappoint*, *shock*, *startle*, *worry*

  adore, *enjoy*, *impress*, *love*, *want*

  analyze, *explain*, *understand*

- Verbs of hitting and other forceful actions

  beat, *knock*, *hit*, *kick*, *slap*

  push, *squeeze*, *twist*

  blast, *kill*, *shoot*

The *way*-construction

- Verb *one's way* PP (Perek 2016)

  We pushed our way into the pub.

- Focus on the "path-creation" use: the verb refers to the means what enables motion of the subject

  They hacked their way through the jungle.

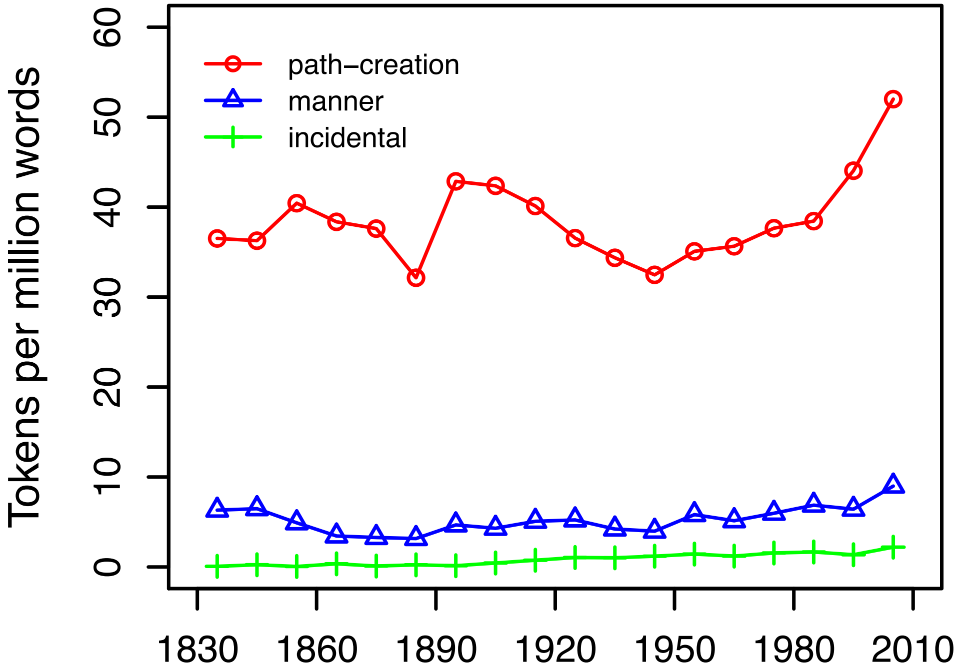
- Vs. "manner" or "incidental-action"

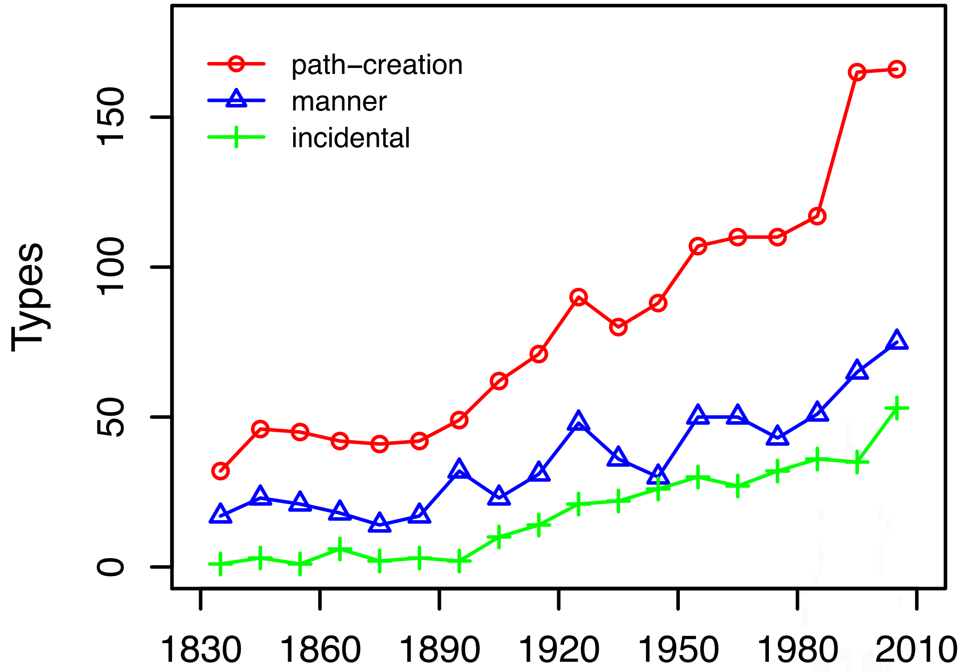
  They trudged their way through the snow.

  He whistled his way across the room.

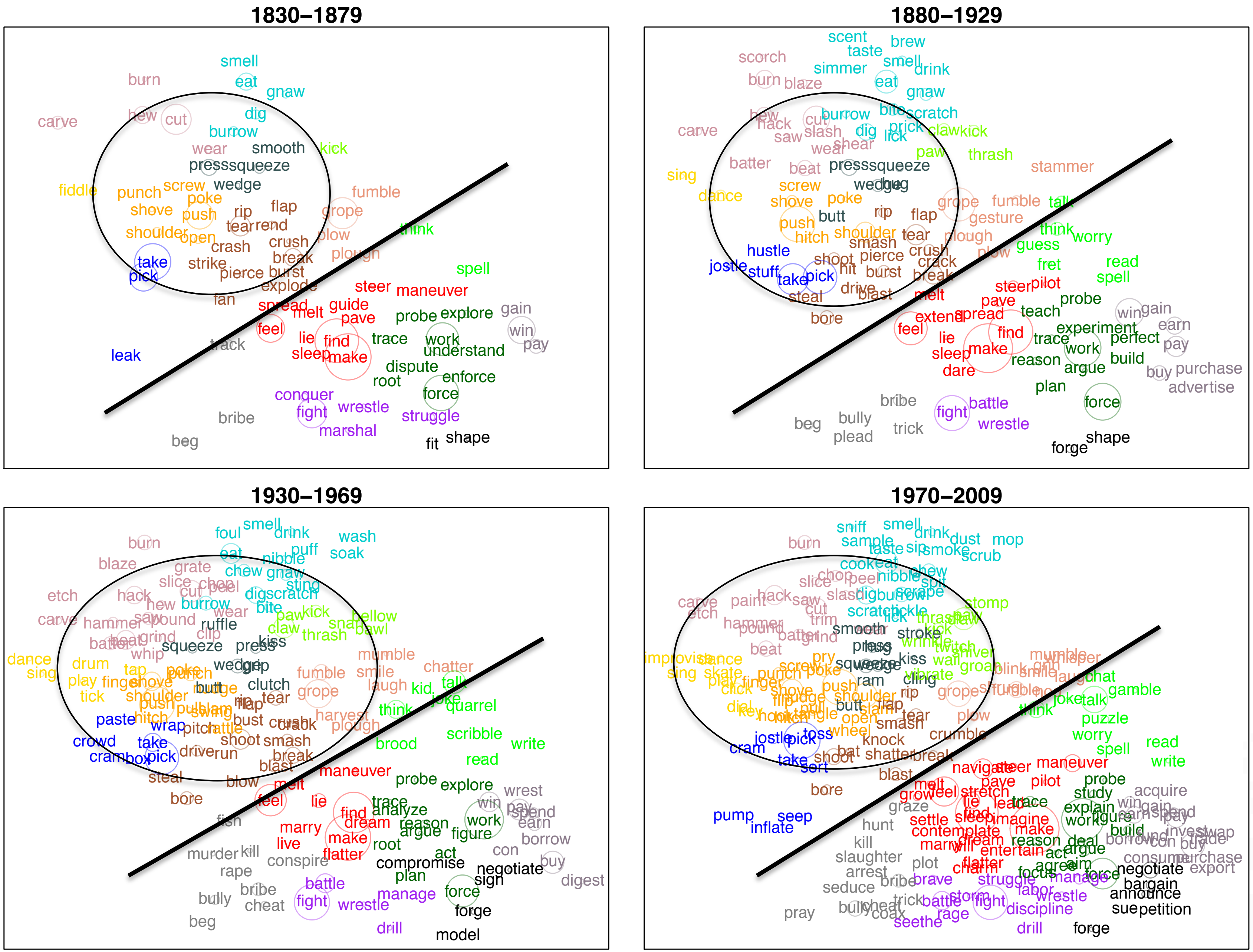
Perek, F. (2016). Recent change in the productivity and schematicity of the way-construction: a distributional semantic analysis. *Corpus Linguistics and Linguistic Theory* (ahead-of-print).

Data



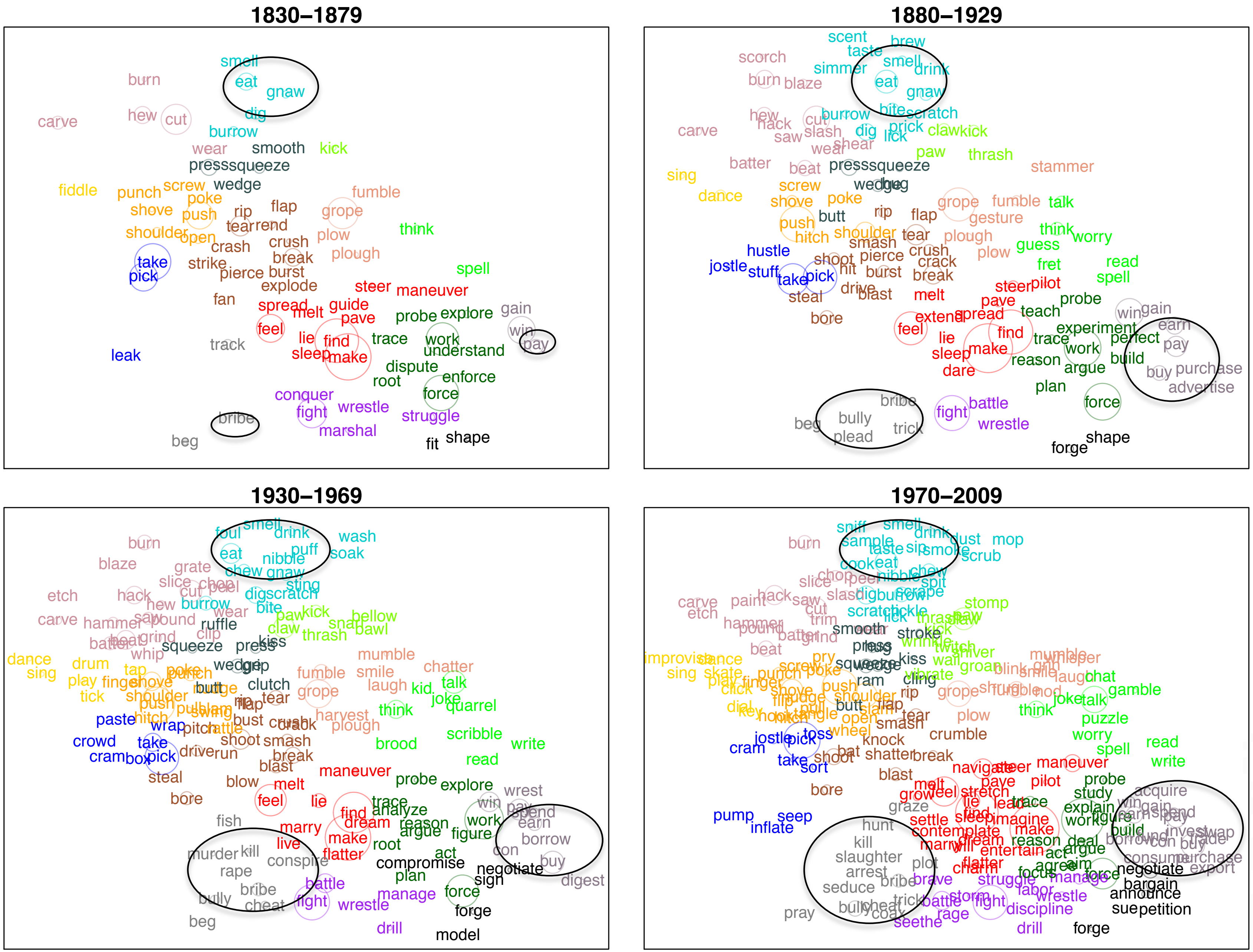


- Relatively stable in frequency
- More and more verbs are used in the construction

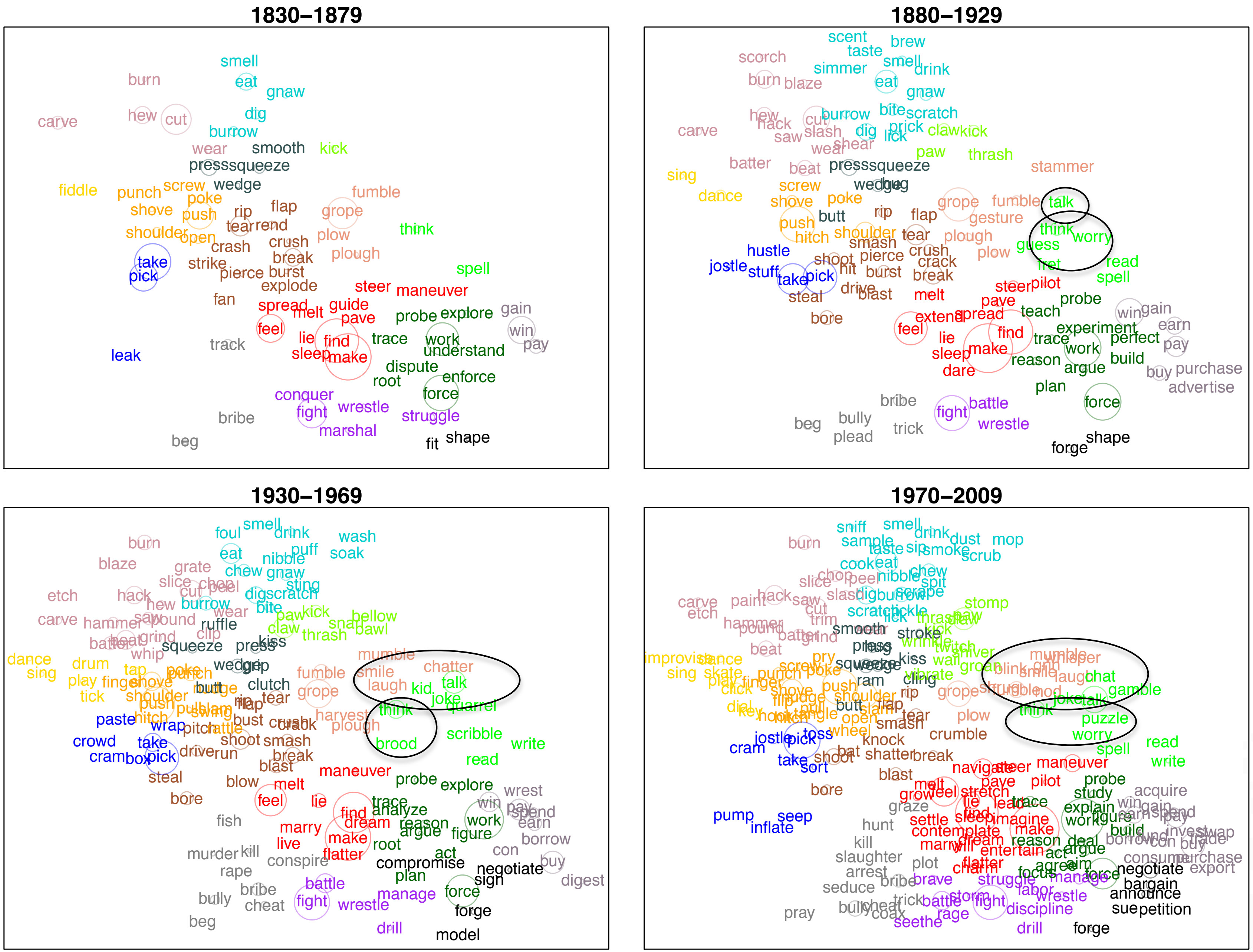


Clear concrete/abstract divide in the distributional semantic plot

Higher density of verbs describing forceful actions (*cut*, *push*, *kick*, ..), especially in earlier periods



From period 2 onwards: ingestion (*eat*, *drink*, *nibble*, *puff*, *sip*, *smoke, ..*), commerce & finance (*buy*, *export*, *fund*, *invest*, *pay*, *spend, ..*), misconduct (*bribe*, *bully*, *cheat*, *conspire*, *kill*, *murder*, *plot*, *rape*, *trick, ..*)



From period 3 onwards: social interaction (*chat*, *chatter*, *joke*, *kid*, *nod*, *quarrel*, *talk*), emotion (*grin*, *laugh*, *smile*, *shrug*, *laugh*), cognition (*brood*, *fret*, *puzzle*, *think*, *worry*)

The path-creation sense

- Many new verb classes refer to unusual ways to cause motion: interaction, commerce, cognition, etc.
- These new uses involve abstract, metaphorical motion:

  [T]hey talk about Uncle Paul having ***bought his way into the Senate****!*

  I sit and watch […], ***grazing my way through a muffuletta****.*

- Main semantic development: the construction becomes more and more open to encoding abstract motion

Periodization

- Distributional semantic plots are a useful tool to observe the development of constructions
- However, it is limited by the arbitrary division of the data
  - Periods of same length
  - Might not be consistent with regards to semantics
- Changes are assessed impressionistically rather than inferred quantitatively
- This relates to the problem of periodization: how to reliably identify stages of change in the data?

Periodization

- Gries & Hilpert (2008) “variability-based neighbour clustering” (VNC): 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.

Distributional clustering

- VNC on the basis of the meaning of words attested in a construction at different points in time (Perek & Hilpert 2017)
- Proposal:
  - Use distributional semantics to build representations of the semantic range of a construction
  - Submit these representations to VNC

Perek, F. & Hilpert, M. (2017). A distributional semantic approach to the periodization of change in the productivity of constructions. *International Journal of Corpus Linguistics* 22(4), 490–520.

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

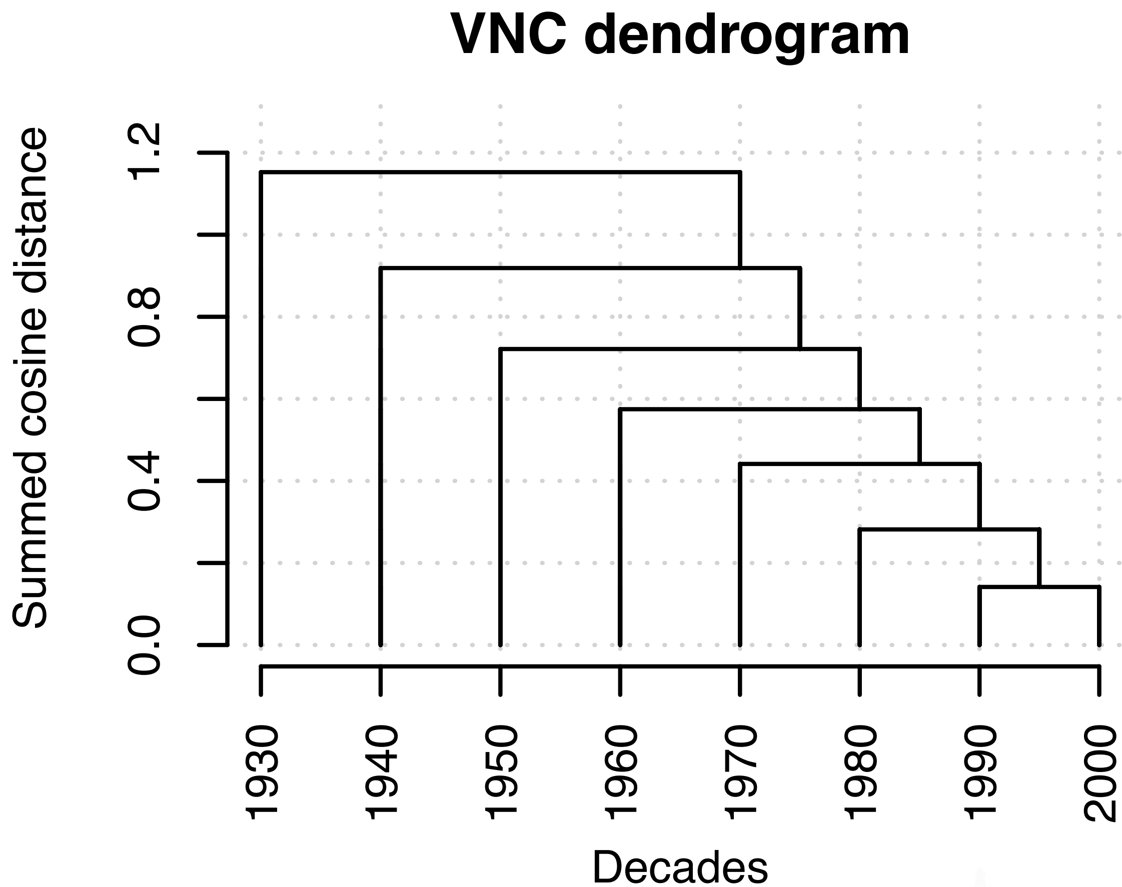
The distributional clustering algorithm

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

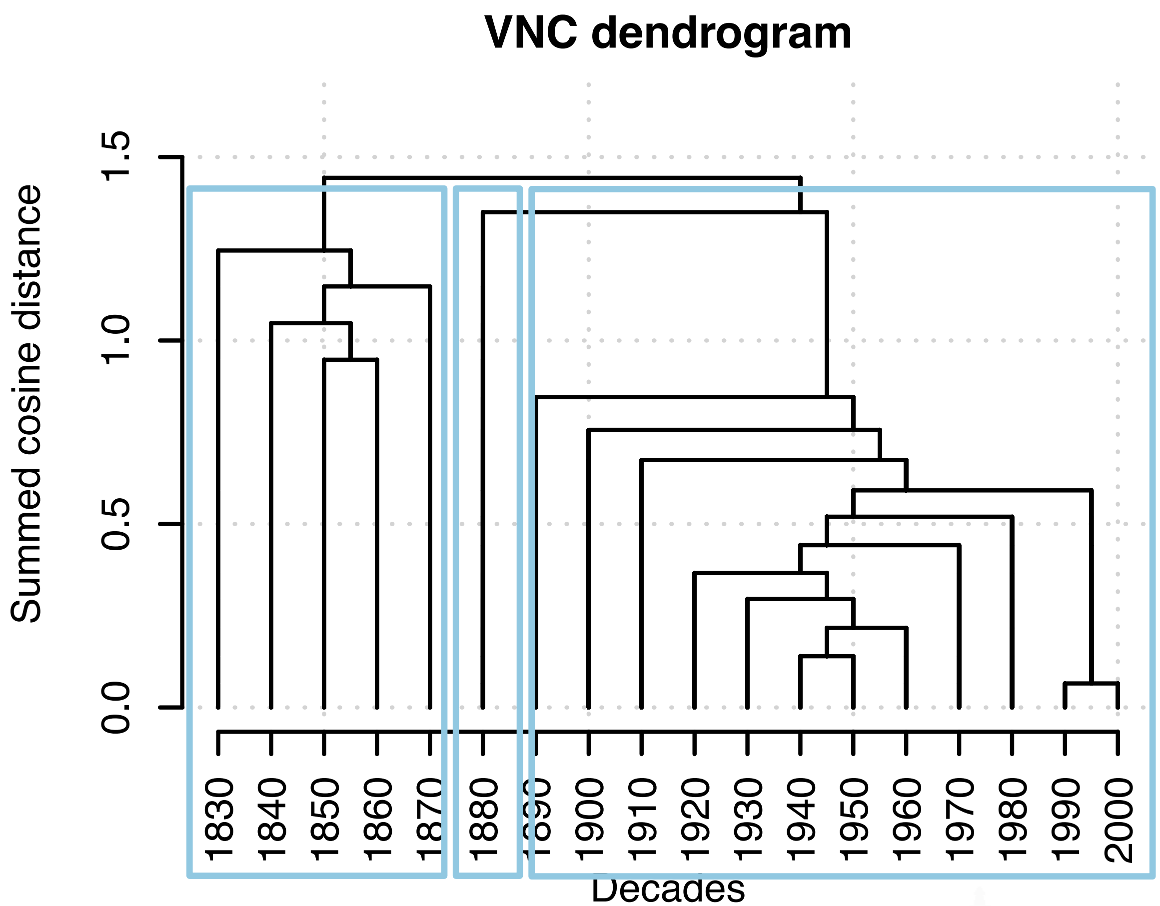
1. Measure the similarity between the period vectors of all pairs of adjacent periods (e.g, 1830s-1840s, 1840s-1850s, etc.).
2. Merge the two periods that are the most similar.
3. Calculate the period vector of the merger as the mean between the vectors of its constituent periods.

- Repeat until all periods have been merged.

The *hell*-construction



The path-creation *way*-construction



Interim summary

- The shapes of the dendrograms indicate different historical scenarios:
  - *Hell*-construction: gradually expanding construction
  - *Way*-construction: variations in distribution
- 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 verbs themselves

Interpreting the dendrograms

- 1830s – 1870s

  hew, *shape*, *explore*, *carve*, *track*, *enforce*, *shoulder*, etc.

  Concrete, physical actions, literal creation of a path

- 1890s – 2000s

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

  More abstract: communication, social interaction, etc.

- 1880s: transition period

  guess, *buy*, *smell*, *stammer*, *beg*, *think*, *pay*, etc.

  bore, *pierce*, *feel*, *wear*, *melt*, *trace*, *burn*, etc.

- Gradual change from mostly concrete to more abstract verbs, in line with previous findings

Summary

- Distributional period clustering provides precise quantitative measurement to impressionistic observations
- Models different kinds of change with dendrograms
- Results are in line with semantic plots, but the timing of changes is measured more objectively

Conclusion

- Distributional semantics is a promising tool for studies on productivity (and more)
- Turns the informal notion of meaning into a quantified representation
- Gives a semantic interpretation to changes in productivity

Theory?

- Such methods can inform theories of language change
- For instance, in diachronic construction grammar (Traugott & Trousdale 2013)
  - Grammar seen as inventory of form-meaning pairs, related in a taxonomic hierarchy (Goldberg 1995)
  - In diachrony: creation of new constructions, changes in existing ones, change in relations between constructions
- The *hell*-construction becomes more productive
- The *way*-construction becomes more productive *and* more schematic

Goldberg, A. (1995). *Constructions: A construction grammar approach to argument structure*. Chicago: University of Chicago Press.

Traugott, E. & G. Trousdale (2013). *Constructionalization and Constructional Changes*. Oxford: Oxford University Press.

UNIVERSITY OF BIRMINGHAM | COLLEGE OF ARTS AND LAW

Thanks for your attention!

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