Variability-based neighbor clustering with historical corpus data:

Results, new applications, and future directions

Martin Hilpert

UNIVERSITÉ DE NEUCHÂTEL

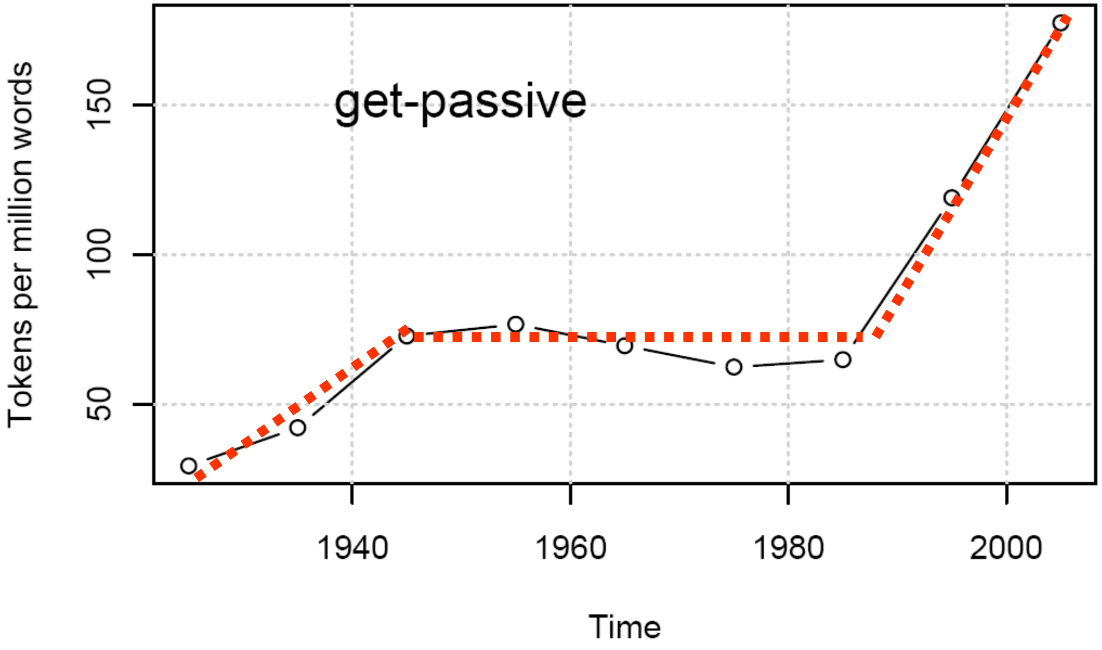
It would be nice if we had a method allowing us to divide a development in language change into a sequence of stages.

Gries & Hilpert (2008) *Corpora* 3/1

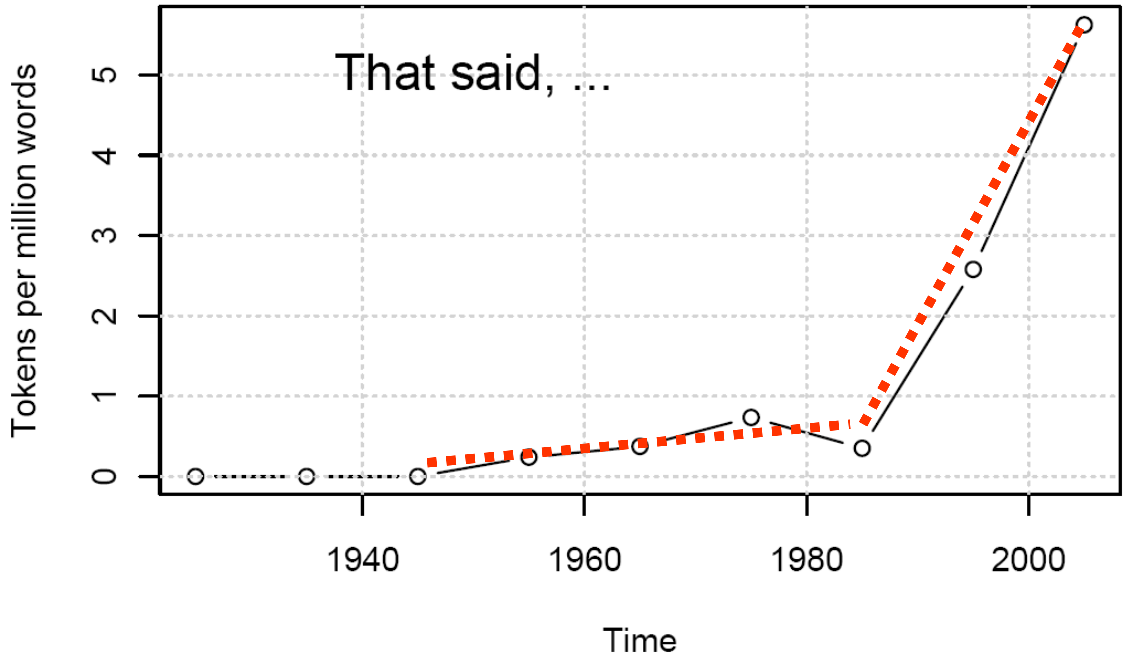
Variability-based neighbor clustering:

A technique that allows us to partition a historical development into a sequence of stages

Three stages?



Two stages?

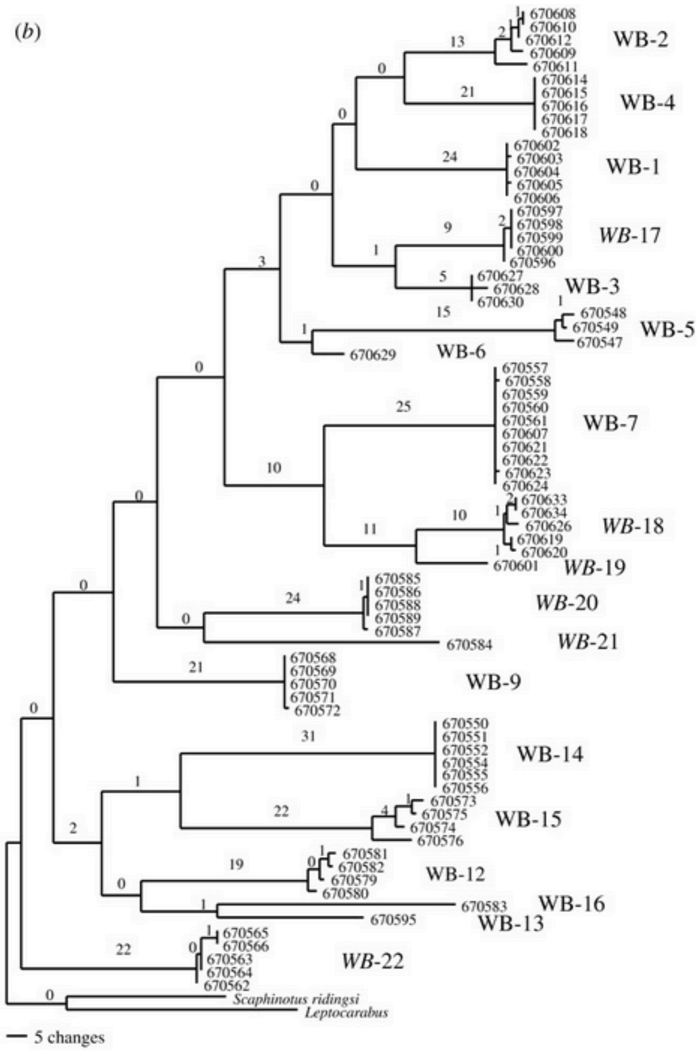

That said, ...
Tokens per million words
0
1
2
3
4
5
1940
1960
1980
2000
Time

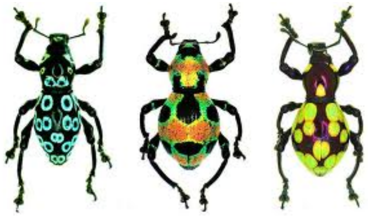
Partitioning historical data

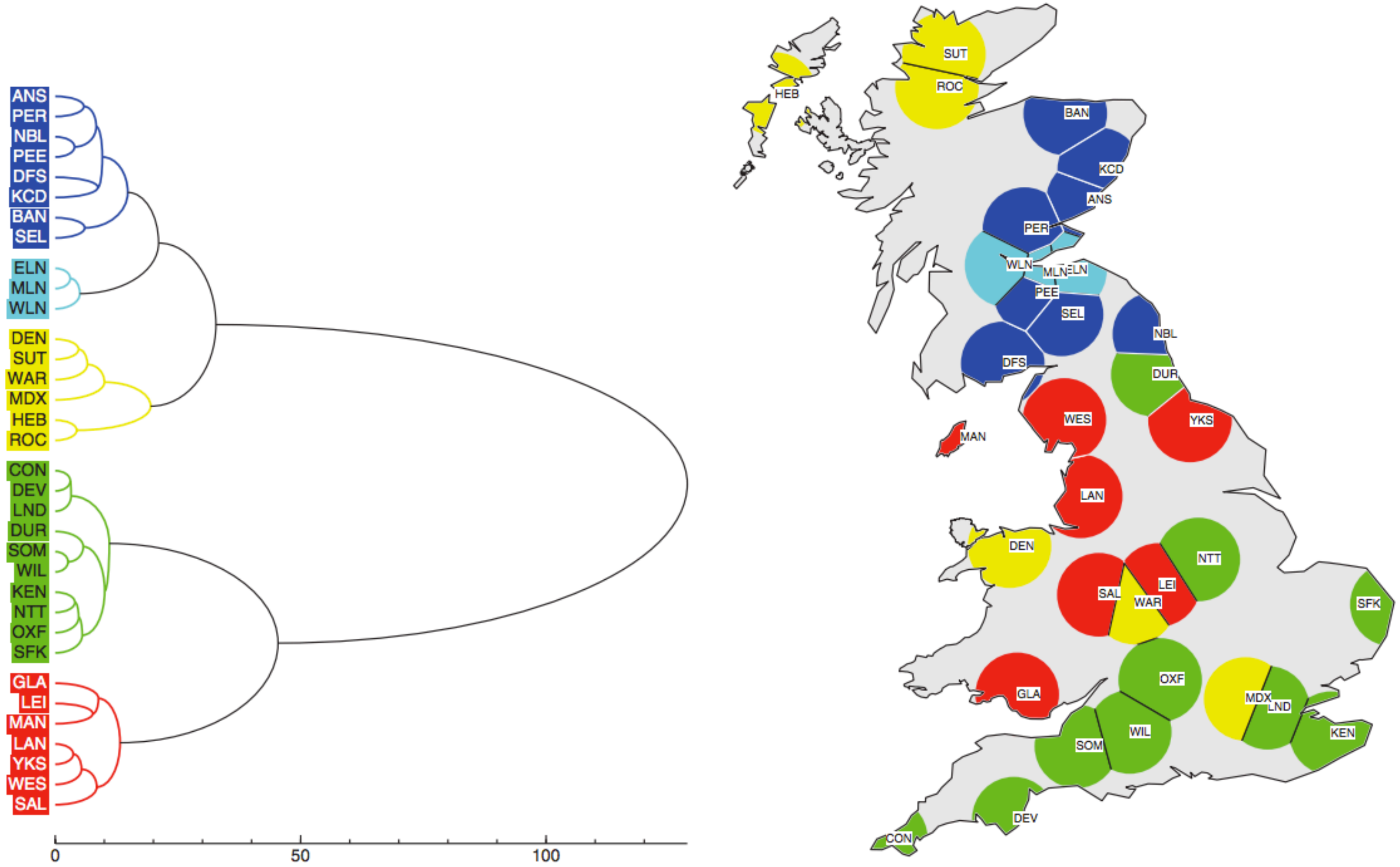
- Diachronic corpus work divides data into sequential periods (centuries, half-centuries, decades, ...)
- Linguistic change can move in fits, bumps, and U-shaped curves
- Averaging over a given period may be misleading
- Different time slices >> different results
- Ideal: dividing the corpus into time slices on the basis of the phenomenon that is studied (data-driven)
- One way to find structures in large bodies of data: hierarchical clustering

Hierarchical Clustering

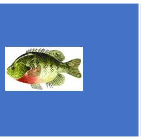
- A technique to find categories in sets of items that are similar to varying degrees.

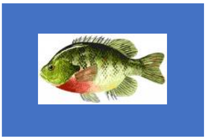
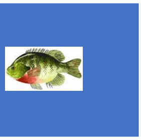
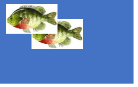

(b)
1
2
670608
670610
670612
670609
670611
13
WB-2
0
21
670614
670615
670616
670617
670618
WB-4
0
24
670602
670603
670604
670605
670606
WB-1
0
9
2
670597
670598
670599
670600
670596
WB-17
1
5
670627
670628
670630
WB-3
3
15
1
670548
670549
670547
WB-5
1
670629
WB-6
0
25
670557
670558
670559
670560
670561
670607
670621
670622
670623
670624
WB-7
10
0
11
10
2
1
670633
670634
670626
670619
670620
1
WB-18
670601
WB-19
0
24
1
670585
670586
670588
670589
670587
WB-20
0
670584
WB-21
21
670568
670569
670570
670571
670572
WB-9
0
31
670550
670551
670552
670554
670555
670556
WB-14
1
22
1
4
670573
670575
670574
670576
WB-15
2
19
0
1
670581
670582
670579
670580
WB-12
0
1
670583
WB-16
670595
WB-13
22
0
1
670565
670566
670563
670564
670562
WB-22
0
Scaphinotus ridingsi
Leptocarabus
5 changes

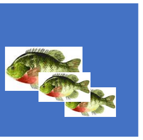
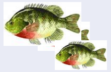


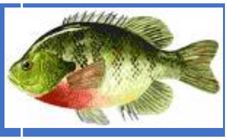
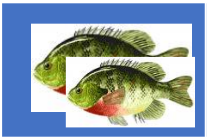
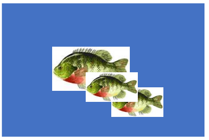
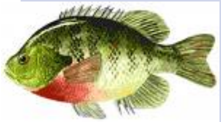
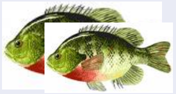
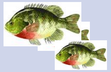


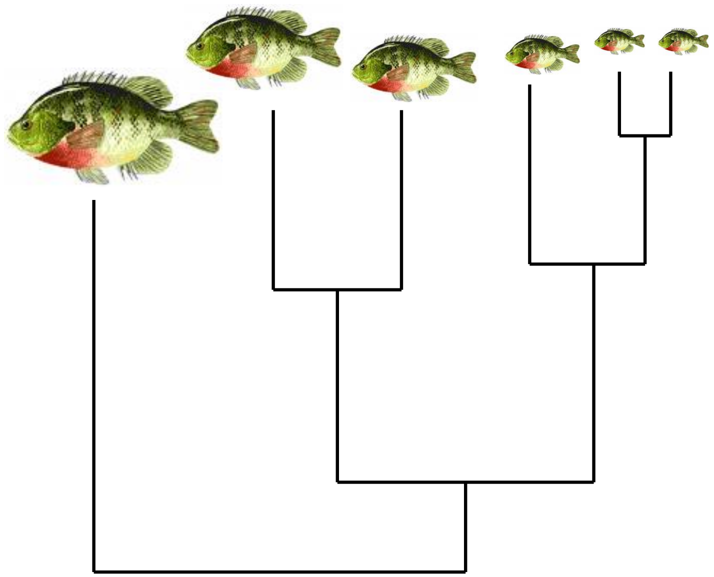
ANS
PER
NBL
PEE
DFS
KCD
BAN
SEL
ELN
MLN
WLN
DEN
SUT
WAR
MDX
HEB
ROC
CON
DEV
LND
DUR
SOM
WIL
KEN
NTT
OXF
SFK
GLA
LEI
MAN
LAN
YKS
WES
SAL
0
50
100
SUT
ROC
HEB
BAN
KCD
ANS
PER
WLN
MLN
ELN
PEE
SEL
NBL
DFS
DUR
WES
YKS
MAN
LAN
DEN
NTT
LEI
SAL
WAR
SFK
OXF
GLA
MDX
LND
KEN
WIL
SOM
DEV
CON

						
	0	5	7	9	10	10.5
		0	2	6	7	7.5
			0	2	3	3.5
				0	1	1.5
					0	0.5
						0

					
	0	5	7	9	10.25
		0	2	6	7.25
			0	2	3.25
				0	1.25
					0

				
	0	5	7	9.625
		0	2	6.625
			0	2.625
				0

			
	0	6.5	9.625
		0	6.125
			0

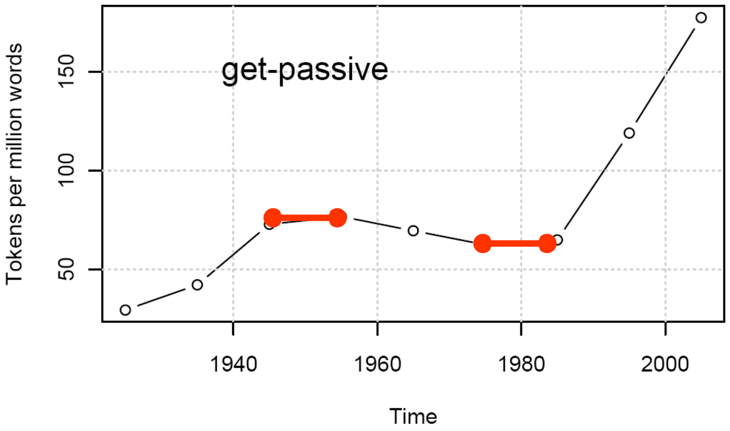


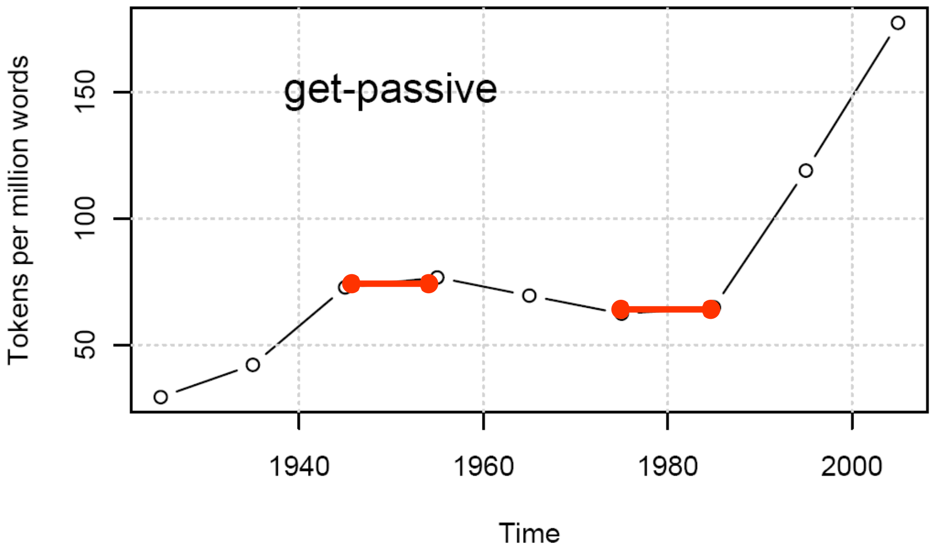
Hierarchical Clustering

- Idea: Use clustering to find out how the development of a given linguistic unit can be divided into stages.
  - Data from different historical periods are coded for a parameter (frequency, range of collocates,...) and grouped according to their similarity.
- Problem: Clustering algorithms are blind to temporal sequence.
  - If, for instance, 1993 is more similar to 2000 than to 1994, we end up with nonsensical clusters.
- Proposal: Variability-based Neighbor Clustering (VNC)
  - Only temporally adjacent nodes are allowed to merge.

Variability-based Neighbor Clustering

- find the two closest neighbors
- merge them and take the mean value
- now find again the two closest neighbors
- merge them and take the mean value
- ...
- until all periods are merged

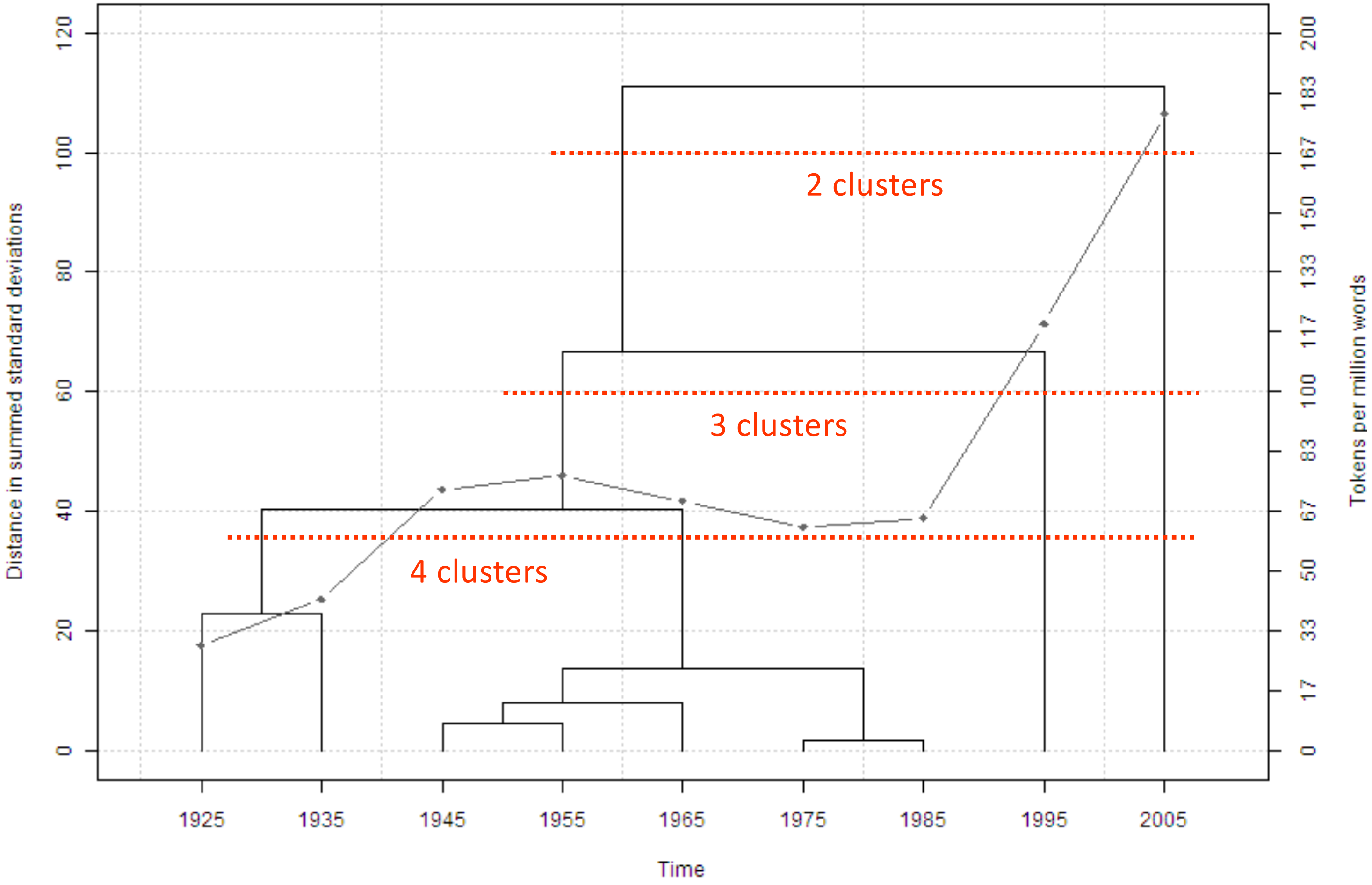




Decade	1920s	1930s	1940s	1950s	1960s	1970s	1980s	1990s	2000s
Tokens per MW	29.5	42.2	72.8	76.7	69.6	62.4	64.9	119.0	177.3

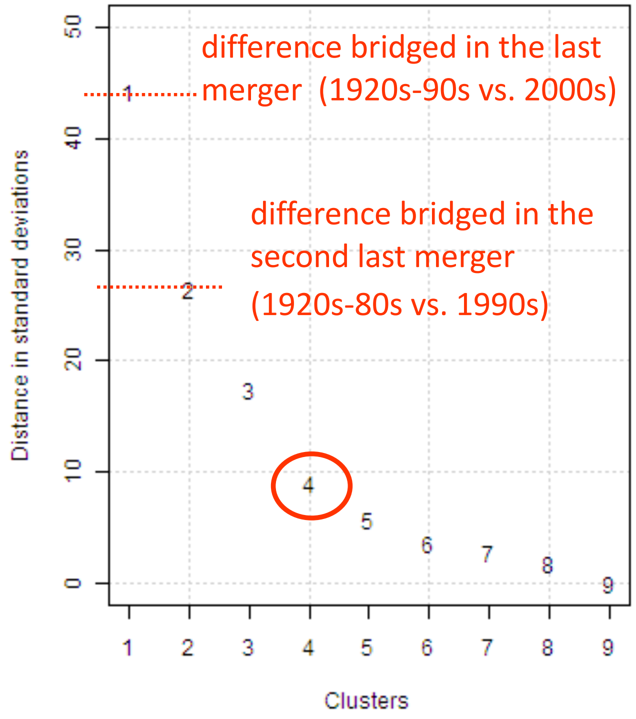
Decade	1920s	1930s	1940s	1950s	1960s	1970s and 1980s	1990s	2000s
Tokens per MW	29.5	42.2	72.8	76.7	69.6	63.65	119.0	177.3

Decade	1920s	1930s	1940s and 1950s	1960s	1970s and 1980s	1990s	2000s
Tokens per MW	29.5	42.2	74.75	69.6	63.65	119.0	177.3



120
100
80
60
40
20
0
Distance in summed standard deviations
200
183
167
150
133
117
100
83
67
50
33
17
0
Tokens per million words
1925
1935
1945
1955
1965
1975
1985
1995
2005
Time
2 clusters
3 clusters
4 clusters

How many clusters?



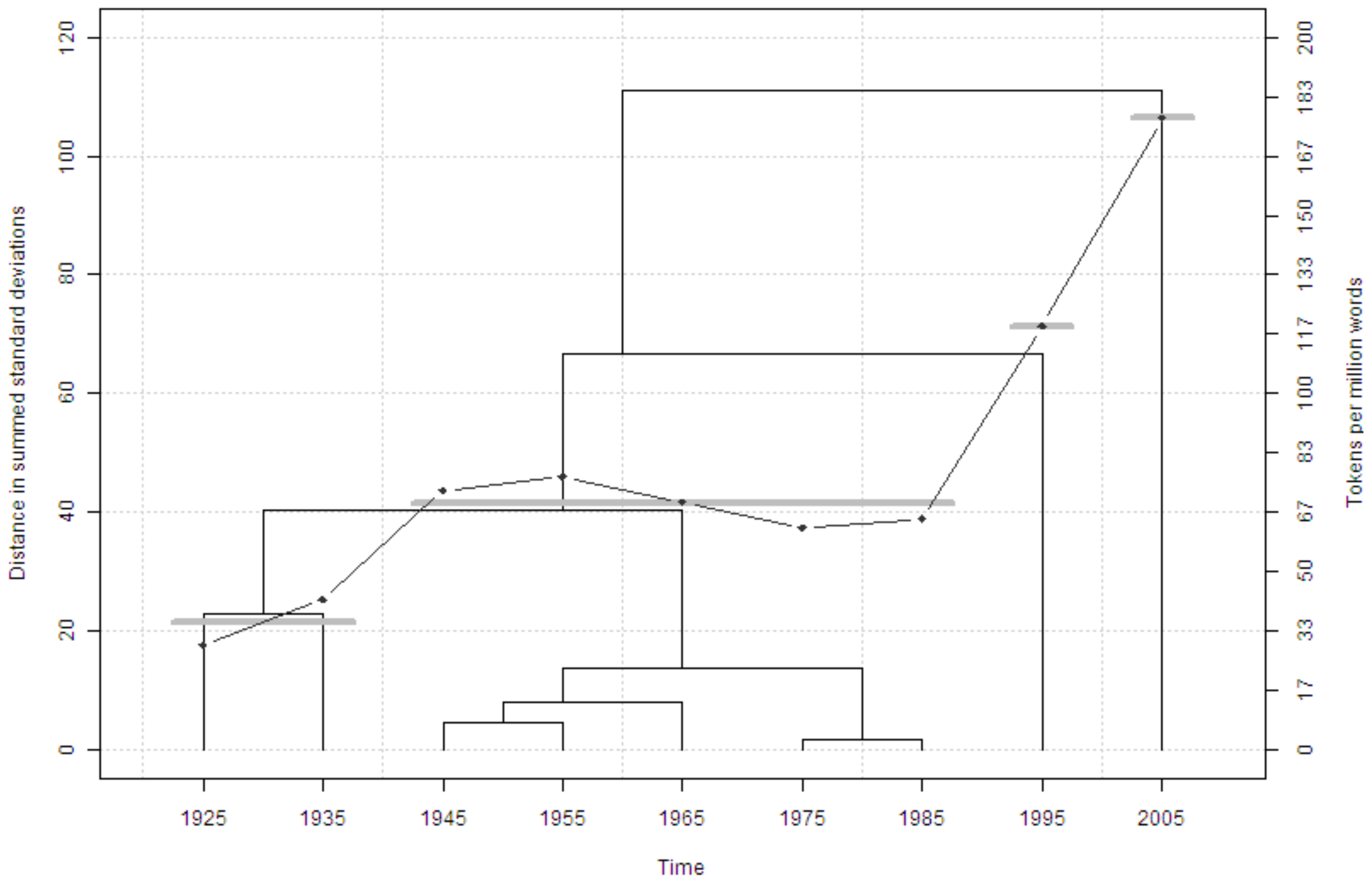
difference bridged in the last merger (1920s-90s vs. 2000s)

difference bridged in the second last merger (1920s-80s vs. 1990s)

as low as possible

as few as possible

4-cluster solution for the *get*-passive



VNC: interim conclusions

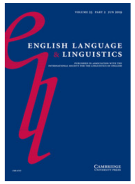
- VNC shows that
  - the trend has four different temporal stages
  - provides their lengths
  - provides their average frequencies
- VNC can detect structure that may otherwise go unnoticed / be hard to characterize objectively

This seems complicated.
Is it really worth the effort?

It is less complicated than it seems. In addition to giving you a sequence of stages, it has further benefits!

Dectecting outliers with VNC

- Frequencies from diachronic corpus data are often messy.
- There is no principled way to identify certain data points as outliers.
- Picking outliers manually is dangerous because you might fit the data to your expectations.
- VNC provides a solution.

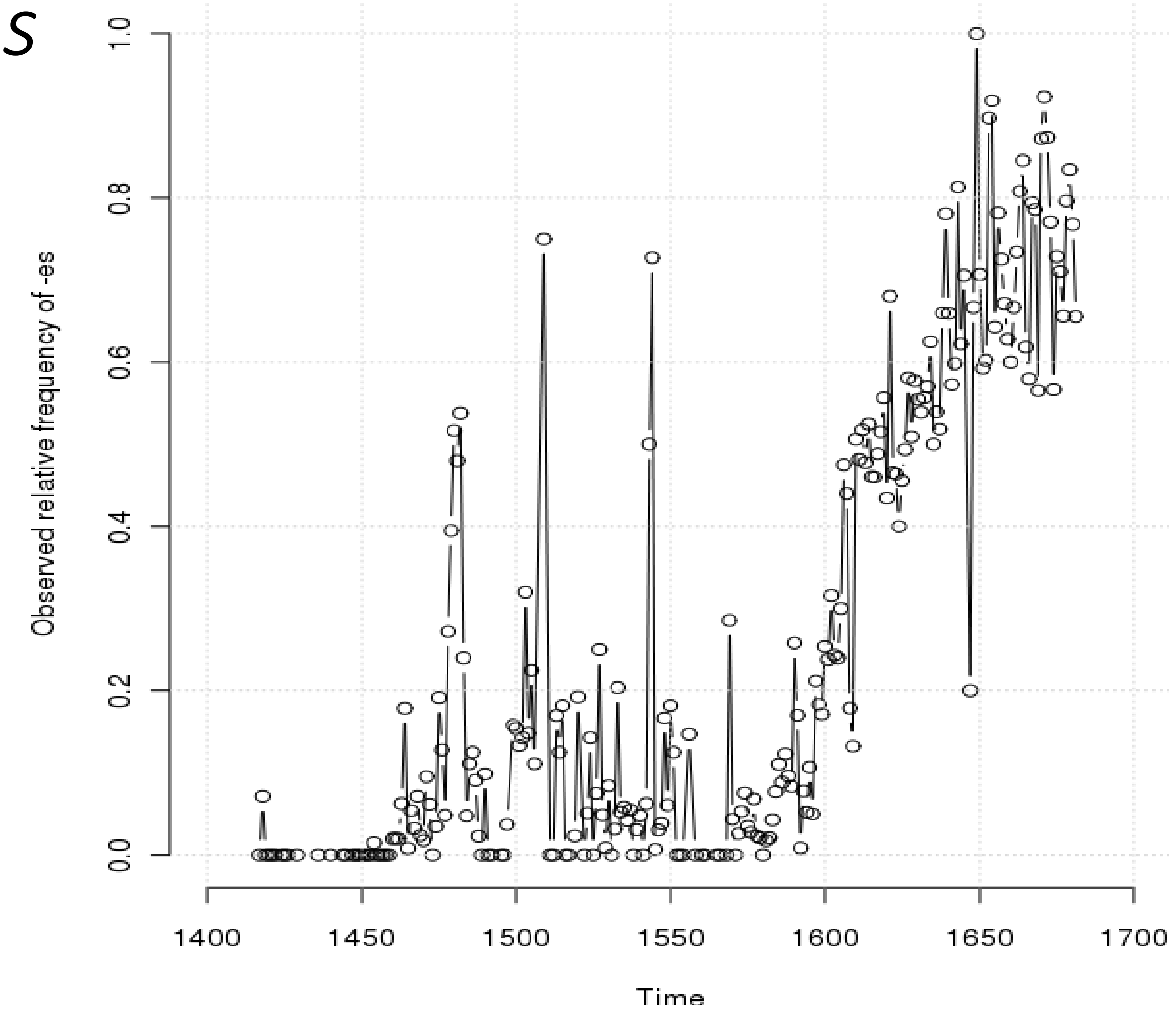


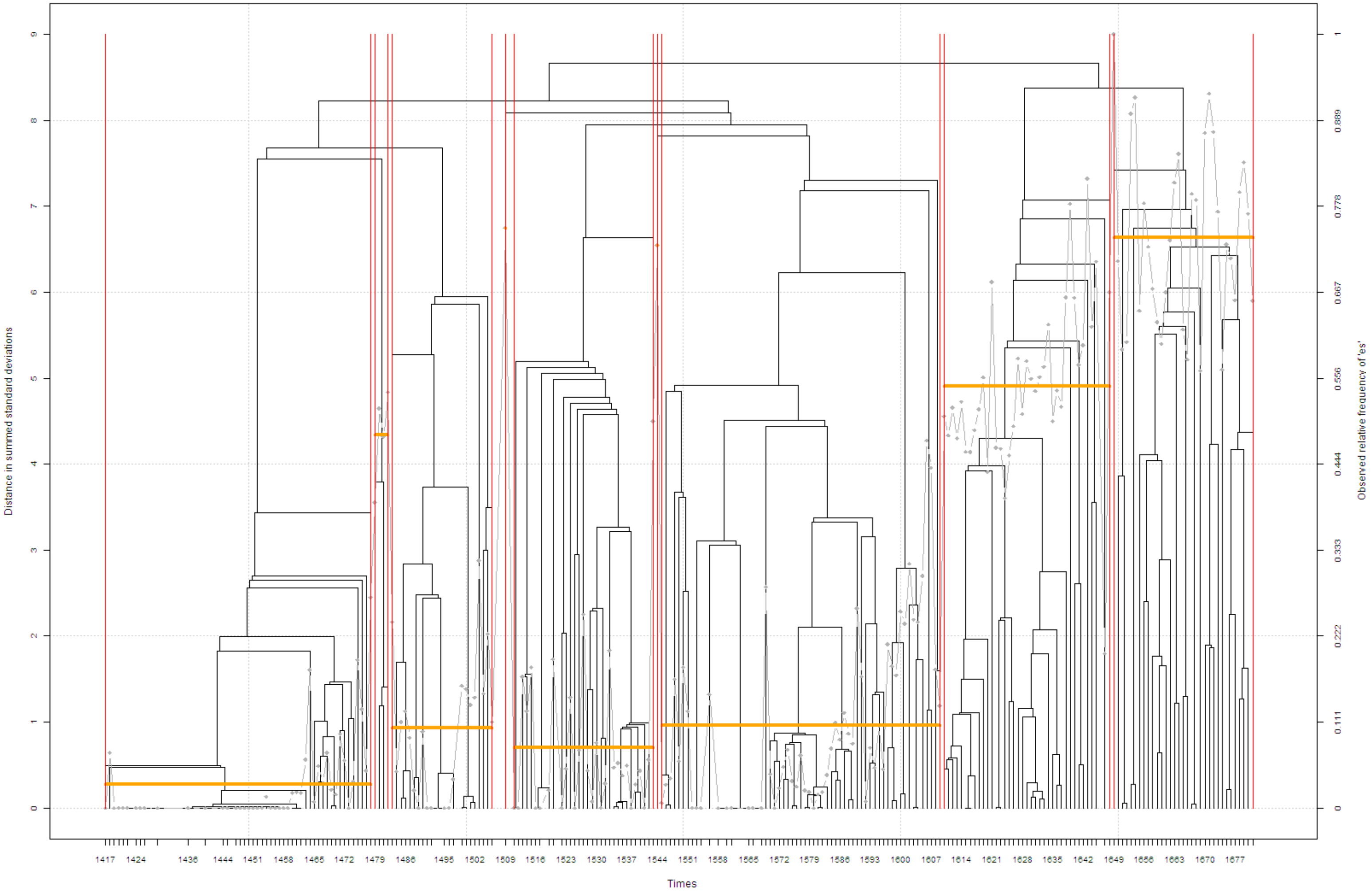
Gries & Hilpert (2010) *English Language and Linguistics* 14/3

From *giveth* to *gives*

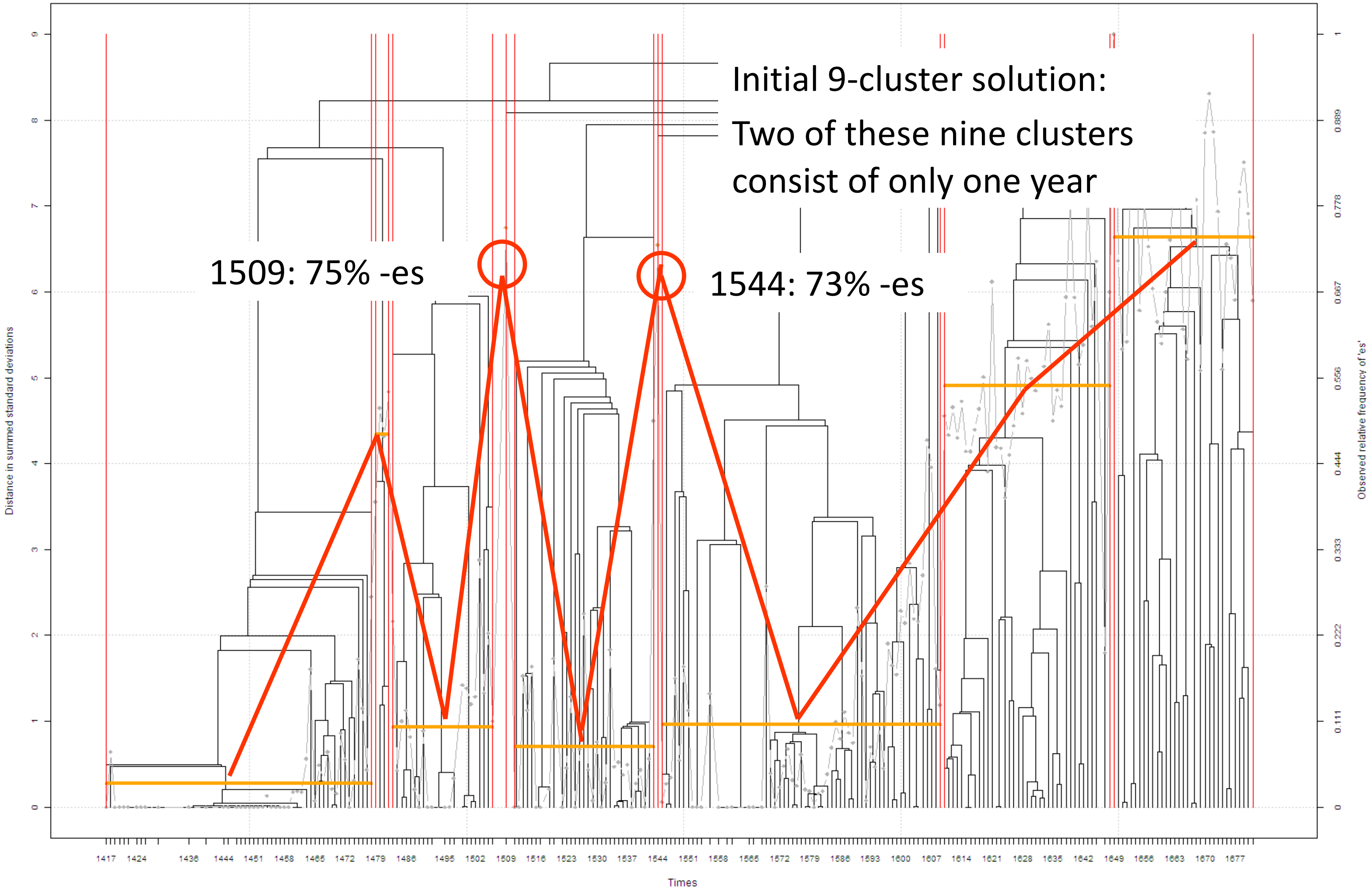
~270 data points

relative frequencies of –*es* range from 0 to 1





Distance in summed standard deviations
0
1
2
3
4
5
6
7
8
9
Observed relative frequency of 'es'
0
0.111
0.222
0.333
0.444
0.556
0.667
0.778
0.889
1
1417
1424
1436
1444
1451
1458
1465
1472
1479
1486
1495
1502
1509
1516
1523
1530
1537
1544
1551
1558
1565
1572
1579
1586
1593
1600
1607
1614
1621
1628
1635
1642
1649
1656
1663
1670
1677
Times

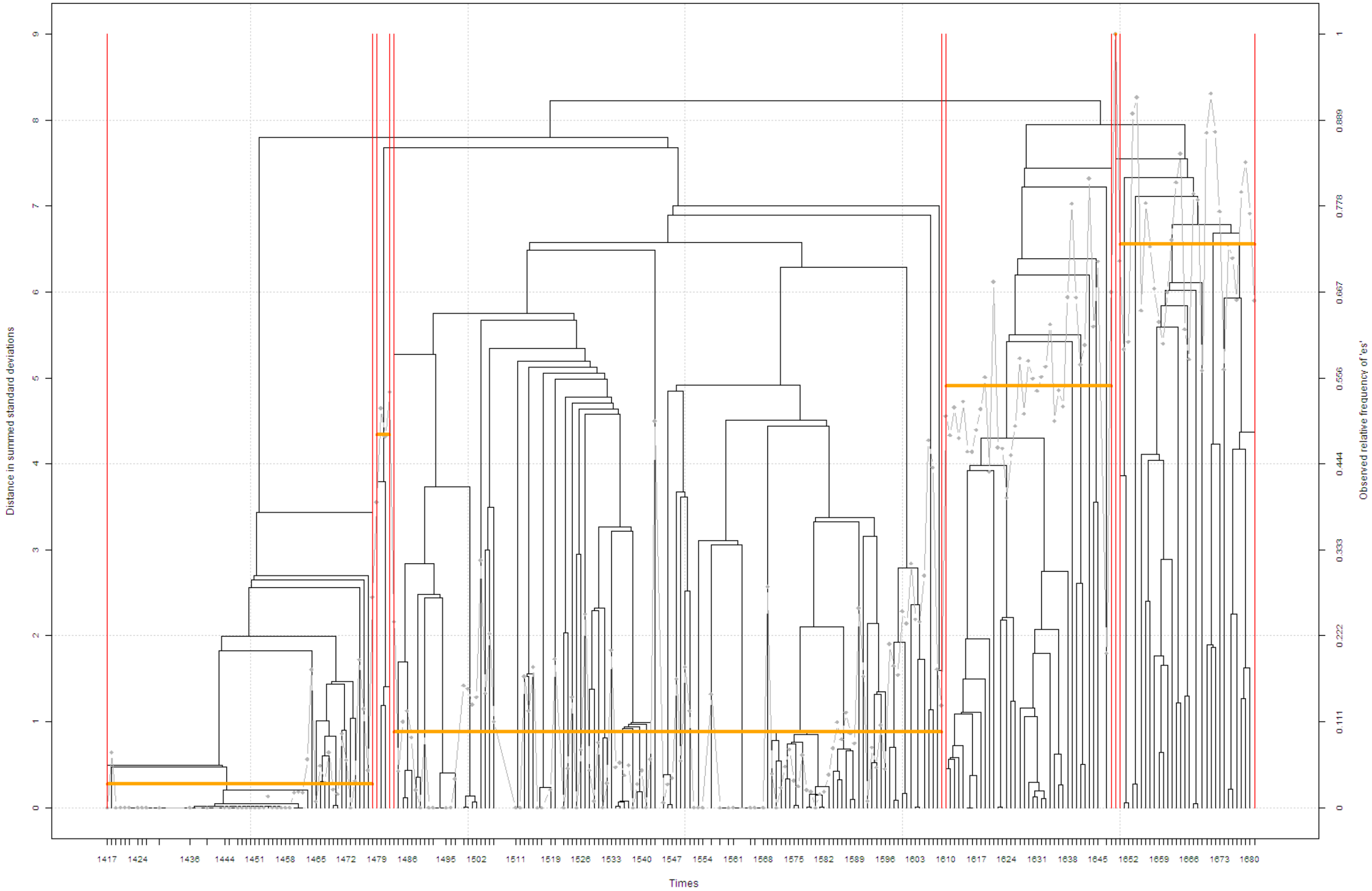


Initial 9-cluster solution:

Two of these nine clusters consist of only one year

1509: 75% -es

1544: 73% -es

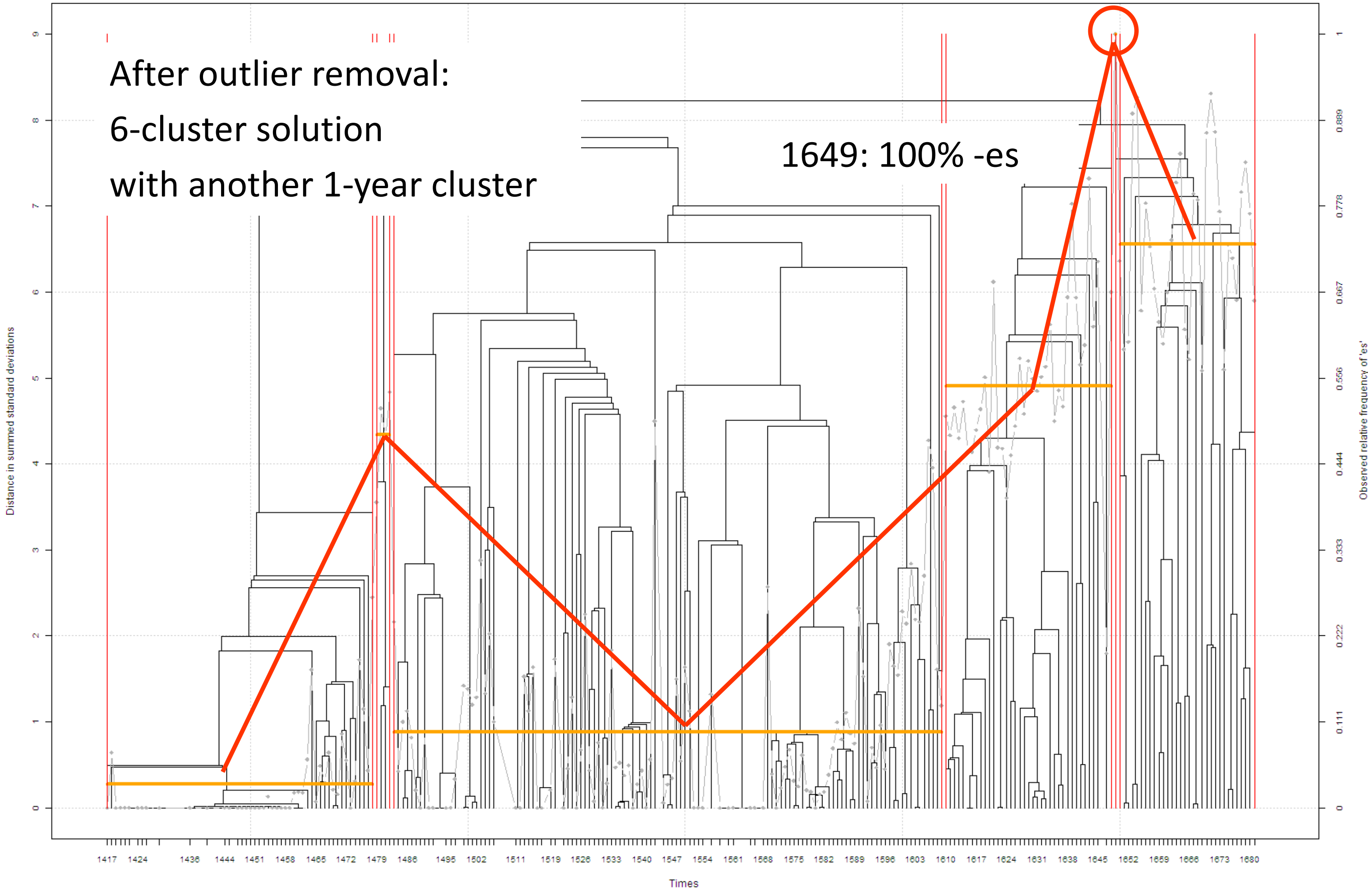

Distance in summed standard deviations
0
1
2
3
4
5
6
7
8
9
1417
1424
1436
1444
1451
1458
1465
1472
1479
1486
1495
1502
1511
1519
1526
1533
1540
1547
1554
1561
1568
1575
1582
1589
1596
1603
1610
1617
1624
1631
1638
1645
1652
1659
1666
1673
1680
Times
0
0.111
0.222
0.333
0.444
0.556
0.667
0.778
0.889
1
Observed relative frequency of 'es'

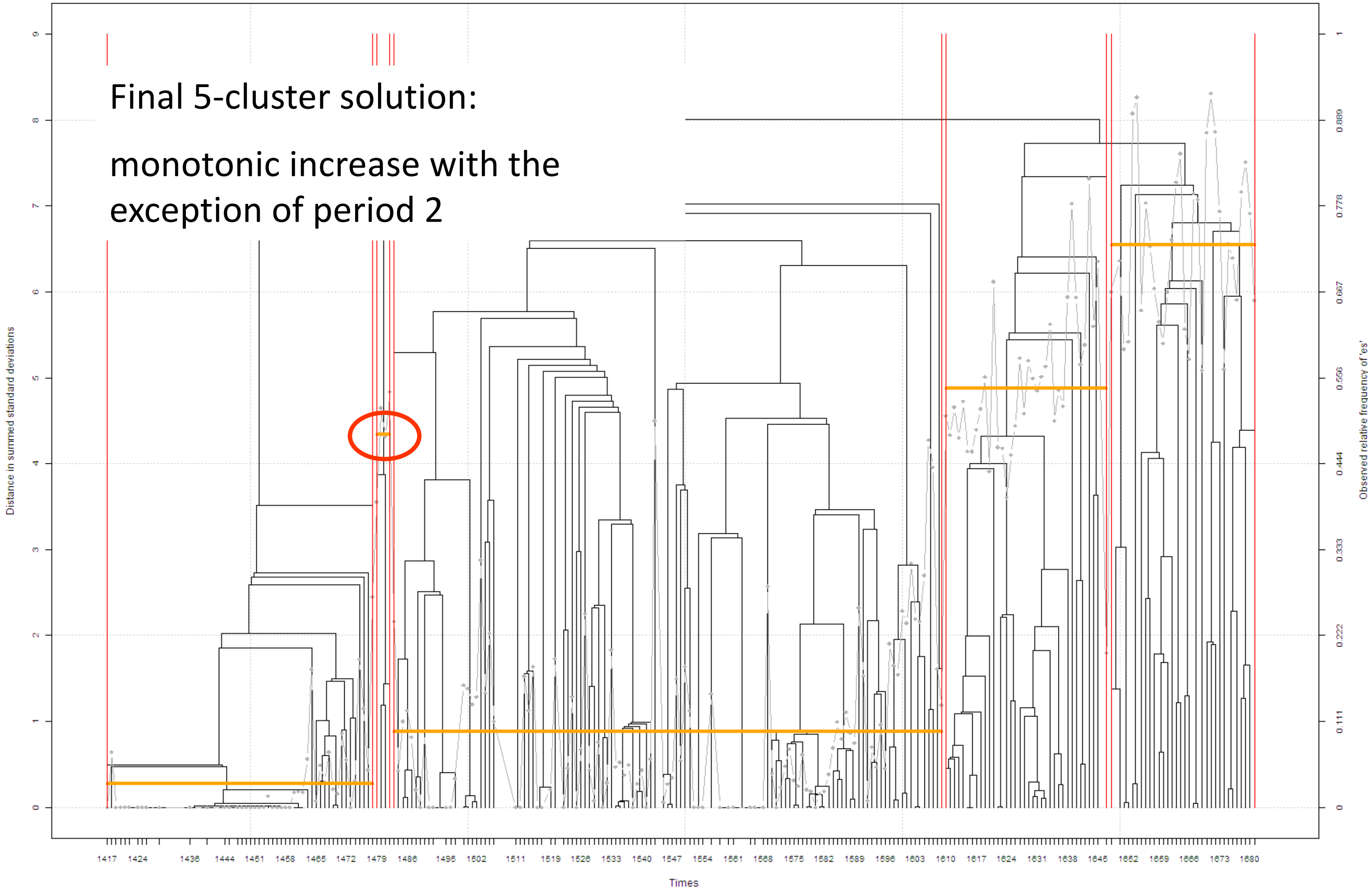
After outlier removal:

6-cluster solution

with another 1-year cluster

1649: 100% -es





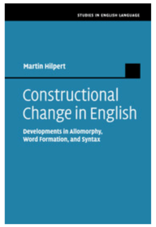
Final 5-cluster solution:

monotonic increase with the exception of period 2

Dectecting outliers with VNC

- Especially when year-by-year data from historical sources is used (OED, TIME, PCEEC, PPCEME, etc.), searches often yield extreme and odd data points.
- If data points are really bad neighbors, VNC will find them.
- They can then be evicted.

Sometimes the most important change in a development is not token frequency. What about type frequency and different measures of productivity?



STUDIES IN ENGLISH LANGUAGE

Martin Hilpert

Constructional Change in English

Developments in Allomorphy, Word Formation, and Syntax

Hilpert (2013) *Cambridge University Press*

VNC with similarity measures other than token frequency

The V-ment construction

- Combination of a lexical stem and a suffix with the phonemic structure [mənt].
- The stem strongly tends to be verbal (judgment, punishment, but of course: merriment, scholarment).
- Typically conveys the meaning of an action (adjustment), the result of an action (assortment), or the means to accomplish an action (refreshment).

A very short history of -ment

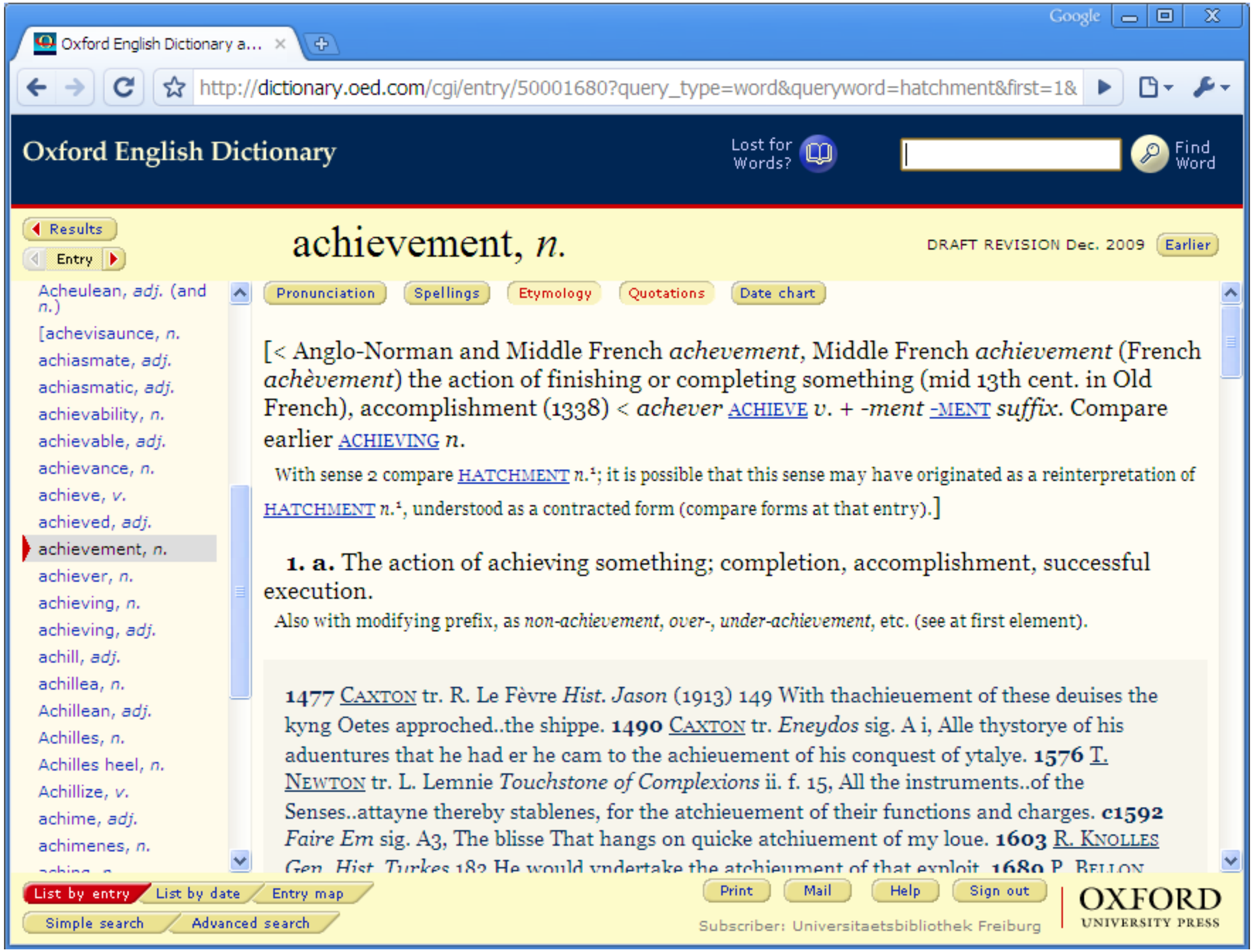
- Isolated Latin loans during OE
- Wave of French loans after 1066
- Nativization between 1250 and 1350
- Rate of new loans recedes after 1600
- Overall productivity recedes
- In PDE, a residue of ~1000 types remains, but the construction is non-productive

  (jogment?, kissment?)

Are there stages in this development that VNC can identify?

Data

- retrieve all types from the Oxford English Dictionary
- retrieve all quotations with these types from the OED



achievement, *n.*

[< Anglo-Norman and Middle French *achevement*, Middle French *achievement* (French *achèvement*) the action of finishing or completing something (mid 13th cent. in Old French), accomplishment (1338) < *achever* ACHIEVE *v.* + *-ment* -MENT *suffix*. Compare earlier ACHIEVING *n.*

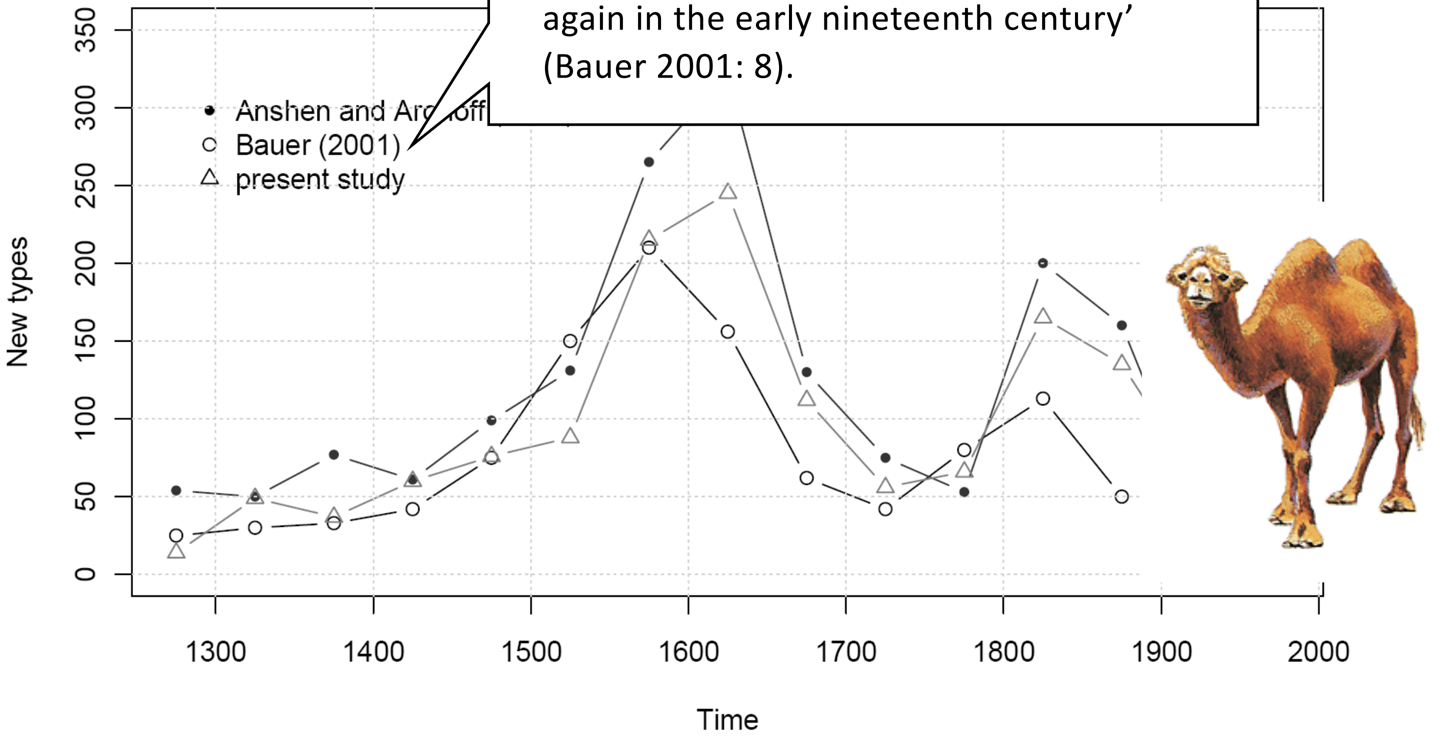
With sense 2 compare HATCHMENT *n.*¹; it is possible that this sense may have originated as a reinterpretation of HATCHMENT *n.*¹, understood as a contracted form (compare forms at that entry).]

1. a. The action of achieving something; completion, accomplishment, successful execution.

Also with modifying prefix, as *non-achievement*, *over-*, *under-achievement*, etc. (see at first element).

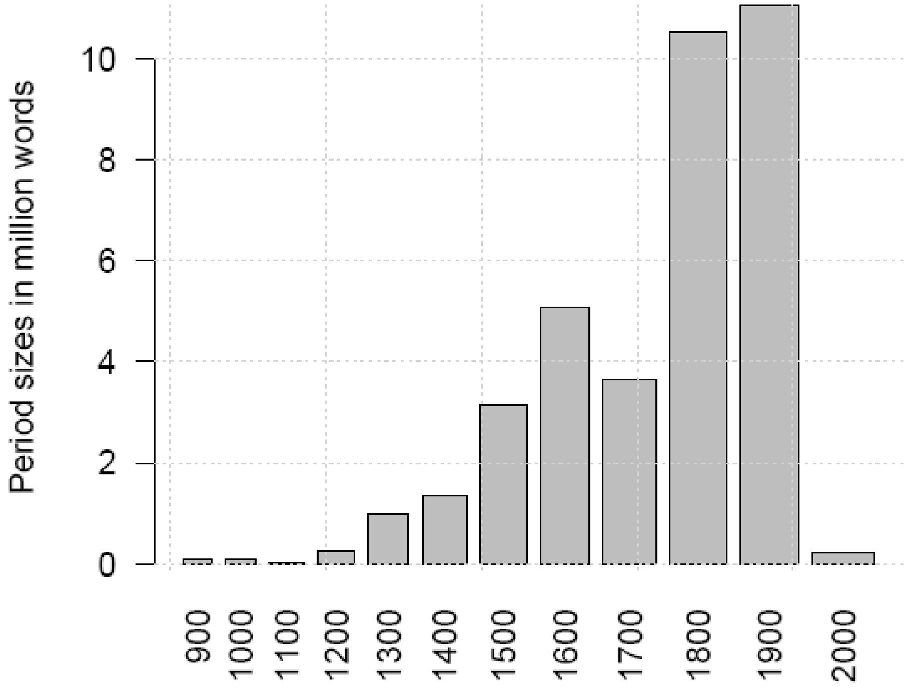
1477 CAXTON tr. R. Le Fèvre *Hist. Jason* (1913) 149 With thachieuement of these deuises the kyng Oetes approched..the shippe. **1490** CAXTON tr. *Eneydos* sig. A i, Alle thystorye of his aduentures that he had er he cam to the achieuement of his conquest of ytalye. **1576** T. NEWTON tr. L. Lemnie *Touchstone of Complexions* ii. f. 15, All the instruments..of the Senses..attayne thereby stablenes, for the atchieuement of their functions and charges. **c1592** *Faire Em* sig. A3, The blisse That hangs on quicke atchiuement of my loue. **1603** R. KNOLLES *Gen. Hist. Turkes* 182 He would vndertake the atchieument of that exploit. **1680** P. BELLON

Data from the OED (~1400 types)

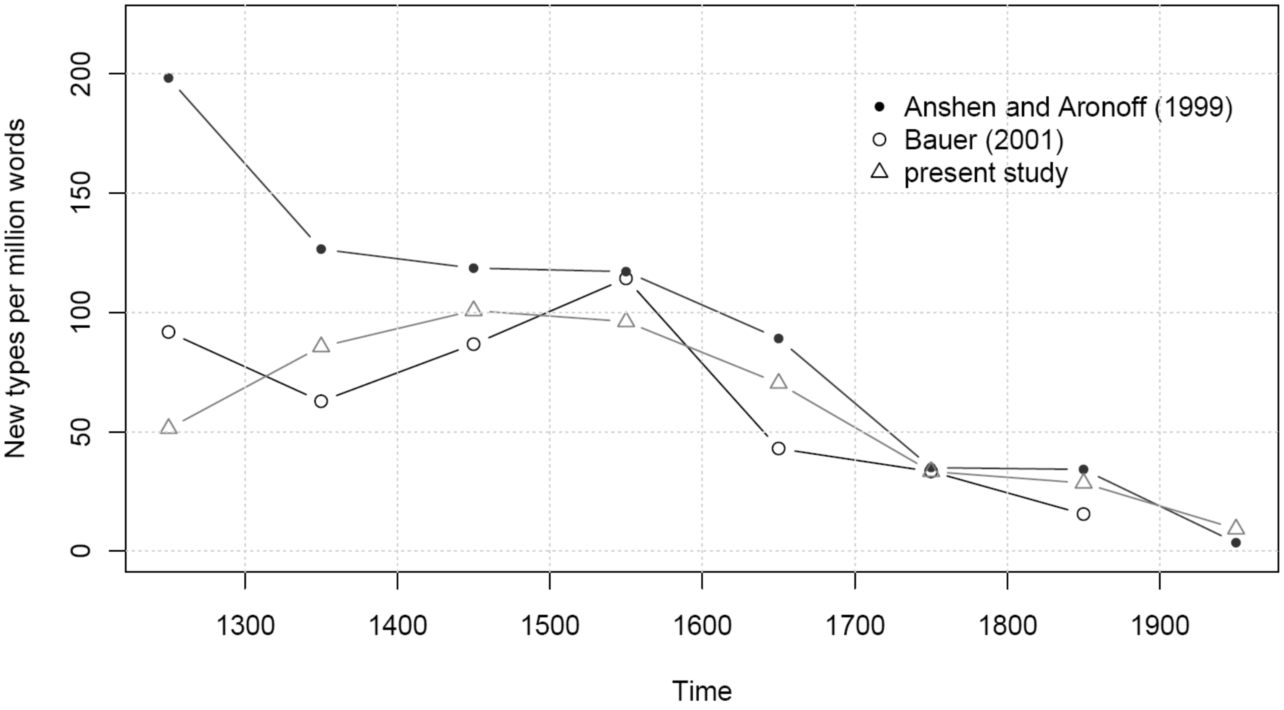


‘The productivity of –ment peaks twice: first in the early seventeenth century and again in the early nineteenth century’ (Bauer 2001: 8).

Words in OED quotations

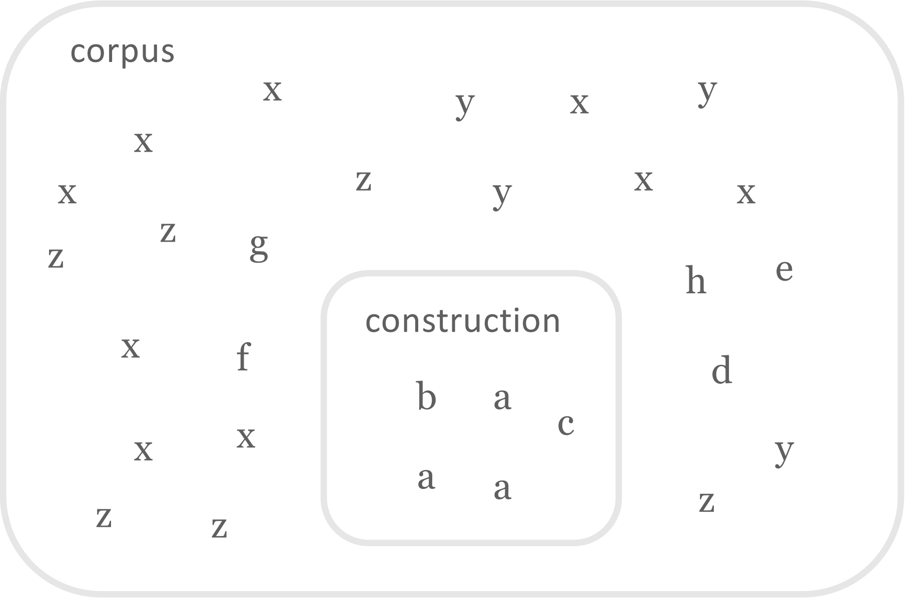


Normalized type frequencies



Expanding productivity

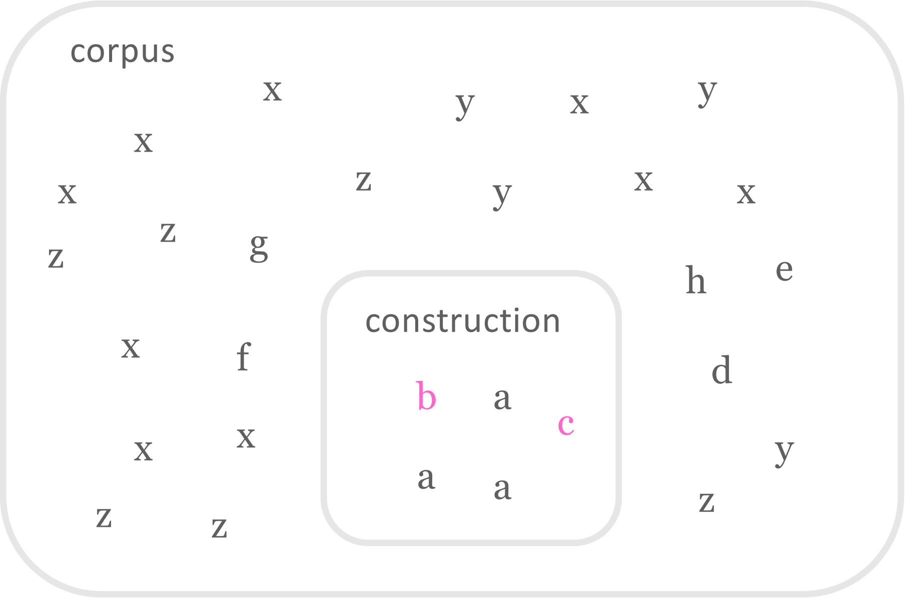
- computed as a ratio:



Expanding productivity

- computed as a ratio:

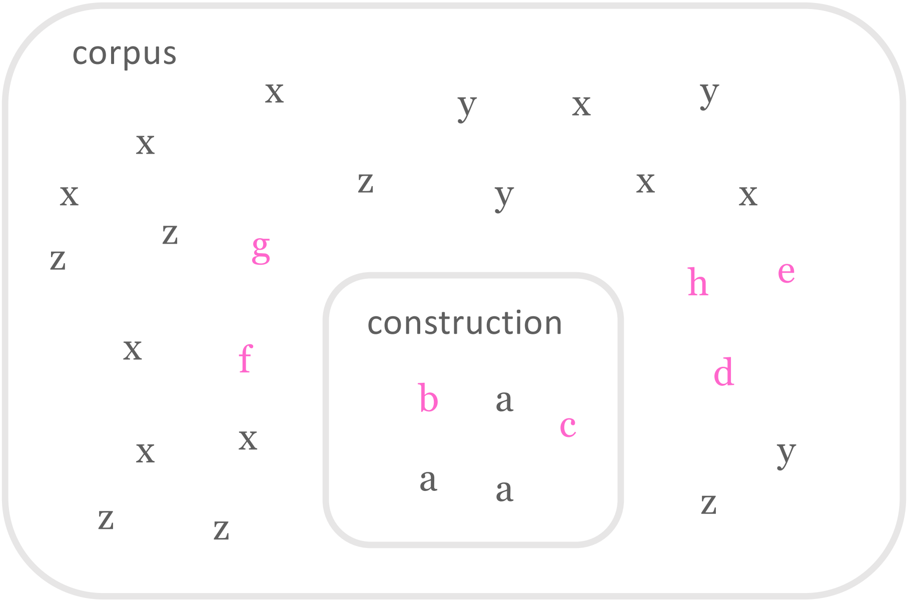
  all hapax legomena of a construction



Expanding productivity

- computed as a ratio:

$$\frac{\text{all hapax legomena of a construction}}{\text{all hapax legomena of the corpus}}$$

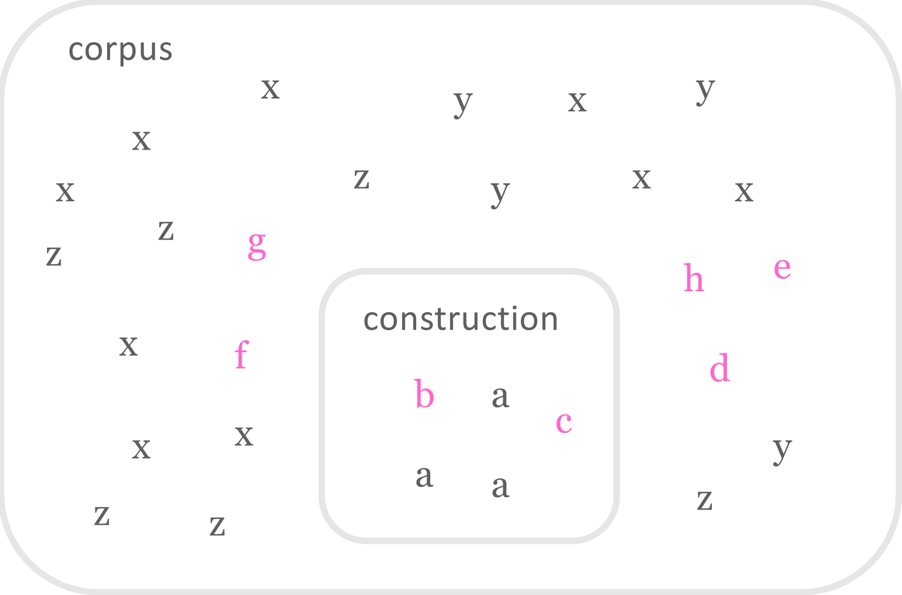


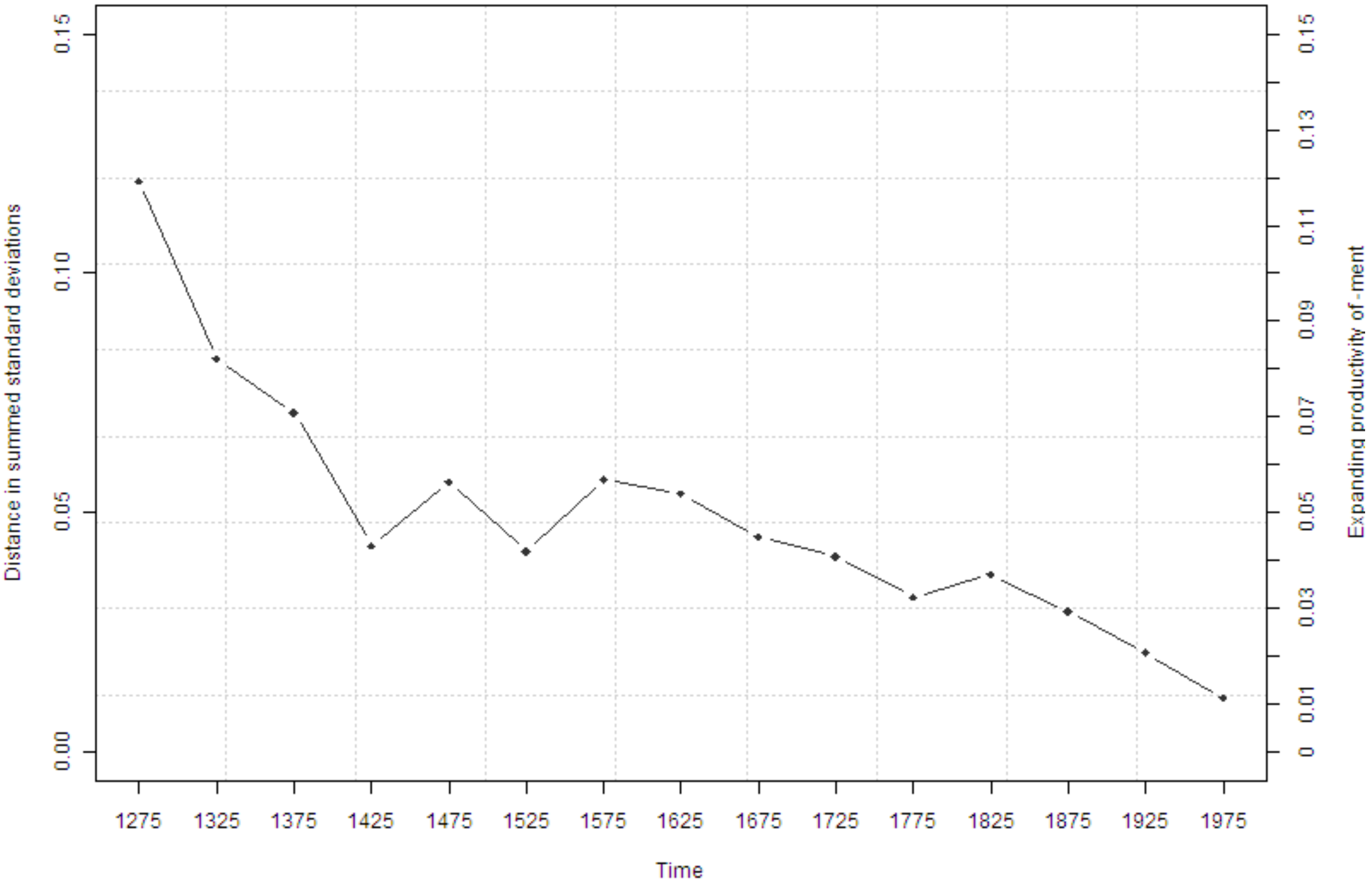
Expanding productivity

- computed as a ratio:

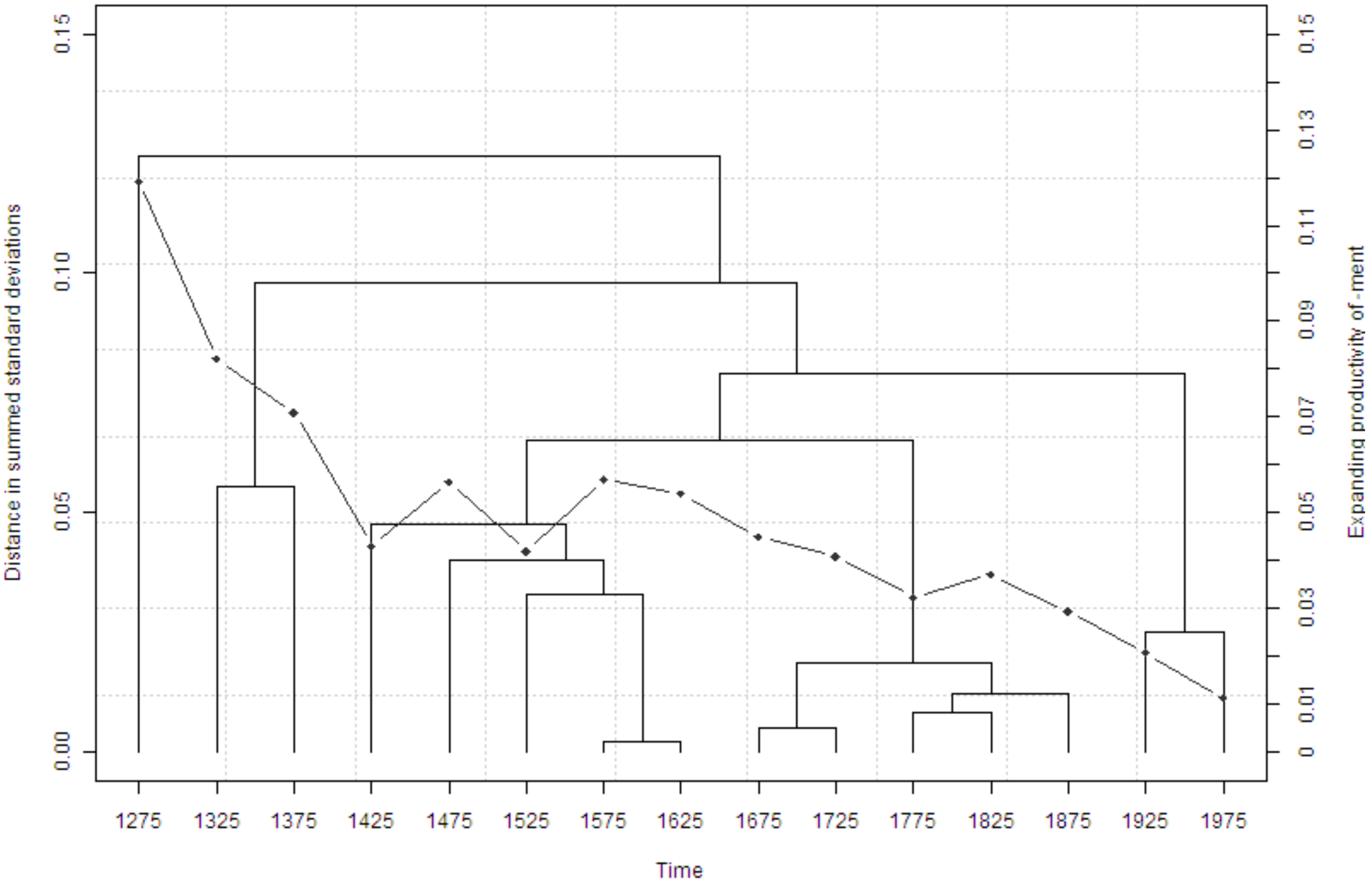
$$\frac{\text{all hapax legomena of a construction}}{\text{all hapax legomena of the corpus}}$$

- in this example 2/7 = 0.29
- the construction holds 29% of the creative business in the corpus

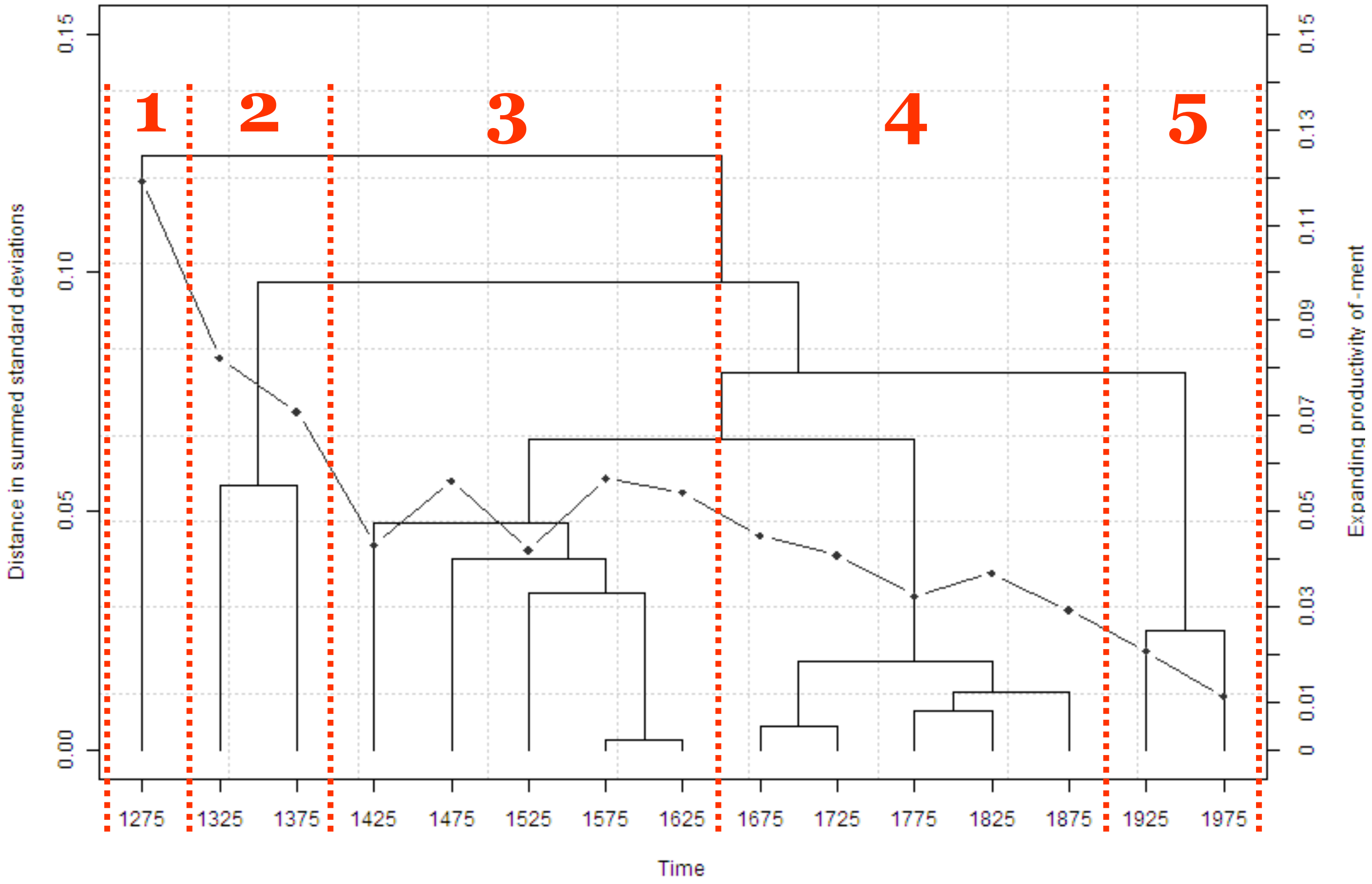




Distance in summed standard deviations
0.00
0.05
0.10
0.15
Expanding productivity of -ment
0
0.01
0.03
0.05
0.07
0.09
0.11
0.13
0.15
1275
1325
1375
1425
1475
1525
1575
1625
1675
1725
1775
1825
1875
1925
1975
Time



0.15
0.10
0.05
0.00
Distance in summed standard deviations
0.15
0.13
0.11
0.09
0.07
0.05
0.03
0.01
0
Expanding productivity of -ment
1275
1325
1375
1425
1475
1525
1575
1625
1675
1725
1775
1825
1875
1925
1975
Time


1
2
3
4
5
Distance in summed standard deviations
0.00
0.05
0.10
0.15
Expanding productivity of -ment
0
0.01
0.03
0.05
0.07
0.09
0.11
0.13
0.15
1275
1325
1375
1425
1475
1525
1575
1625
1675
1725
1775
1825
1875
1925
1975
Time

Analytical steps

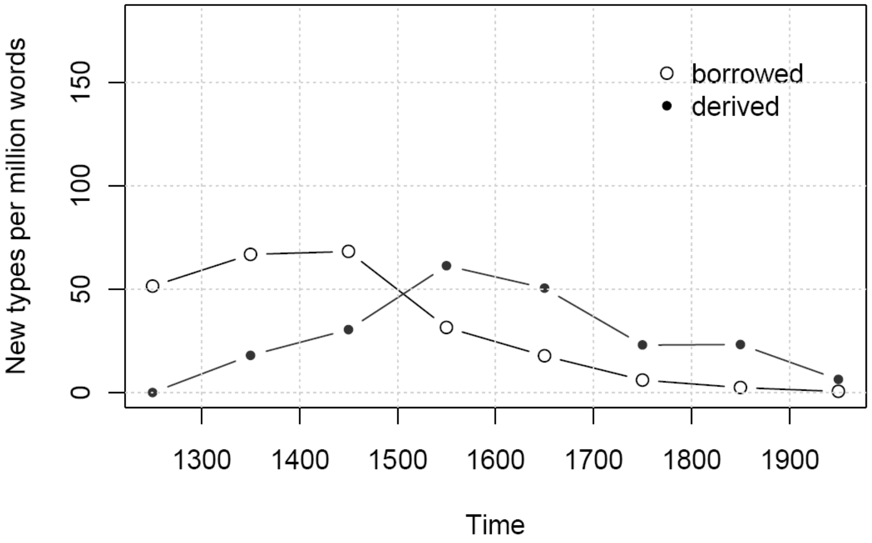
- determine relevant variables
- annotate all 1400 types in the database for these variables
- explore whether patterns of variation change over time, using a multivariate analysis

Variable 1: Etymological source

Is a form borrowed or derived?

B: achievement, detachment, enforcement

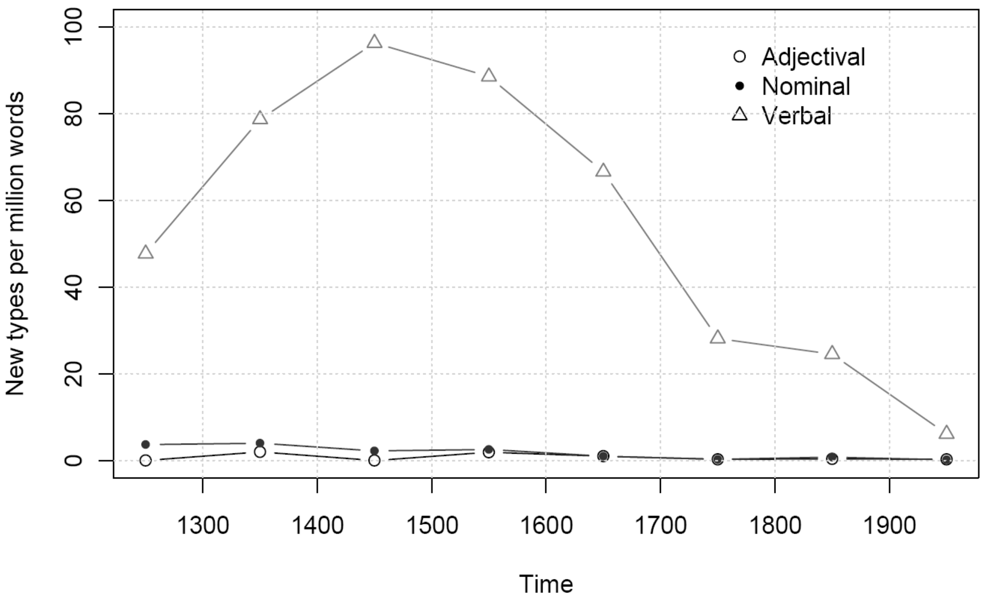
D: bickerment, erasement, shipment



Variable 2: Stem type

What is the lexical category of the stem?

V:	achievement, enforcement
A:	merriment, unruliment
N:	scholarment, utensilment



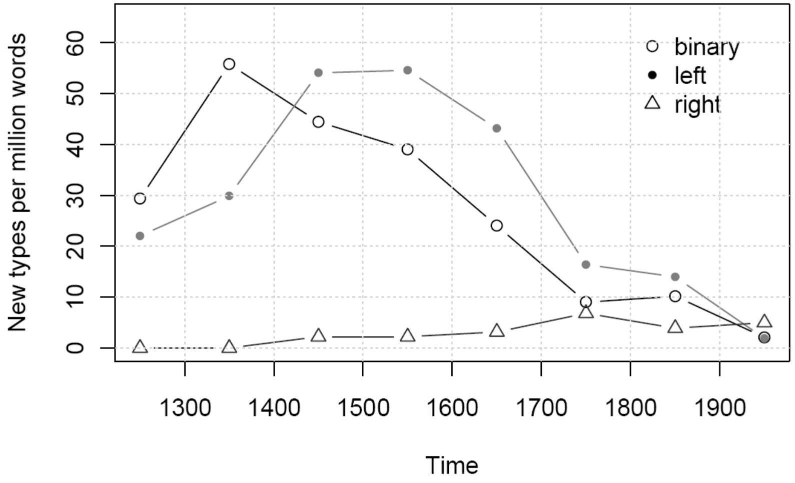
Variable 3: Branching structure

What is the internal hierarchical structure?

Binary: judgment, treatment

Left-branching: [en+rich]ment, [be+little]ment

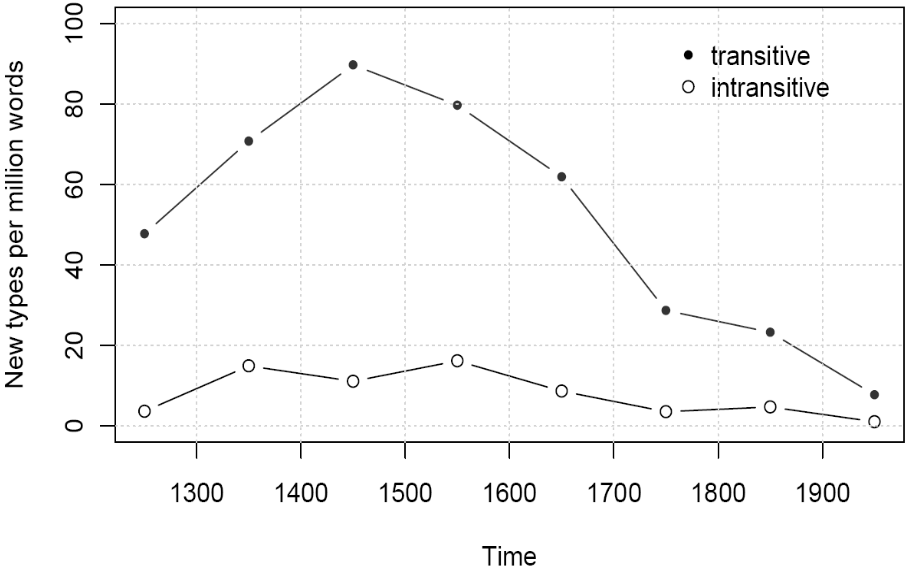
Right-branching: eco[manage+ment], non[agree+ment]



Variable 4: Transitivity

Does the form evoke an entity that is acted upon?

Transitive:	arousement, punishment
Intransitive:	flourishment, merriment



Variable 5: Semantic types

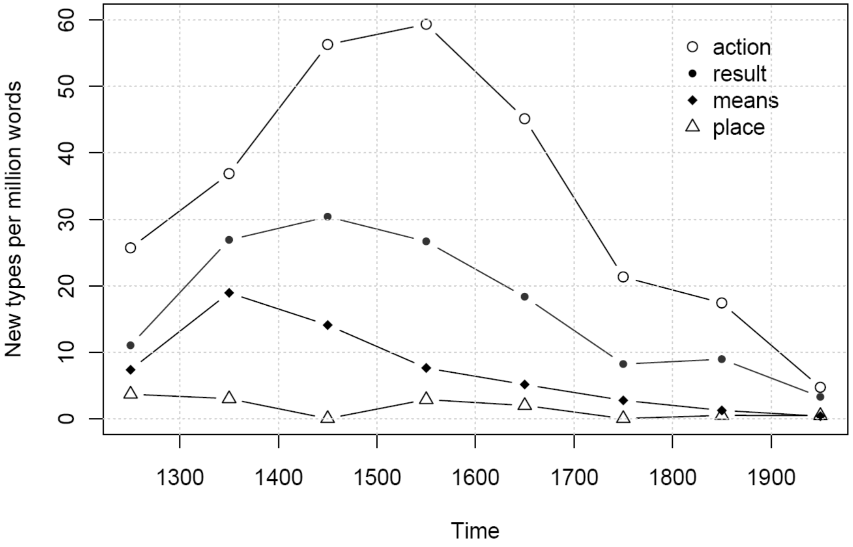
Which overall meaning is conveyed by the form?

Action: confrontment, dismantlement

Result: settlement, scholarment

Means: ornament, refreshment

Place: parliament, environment



Analysis

transitivity	binary	left	right
intransitive	130	56	15
transitive	404	670	132

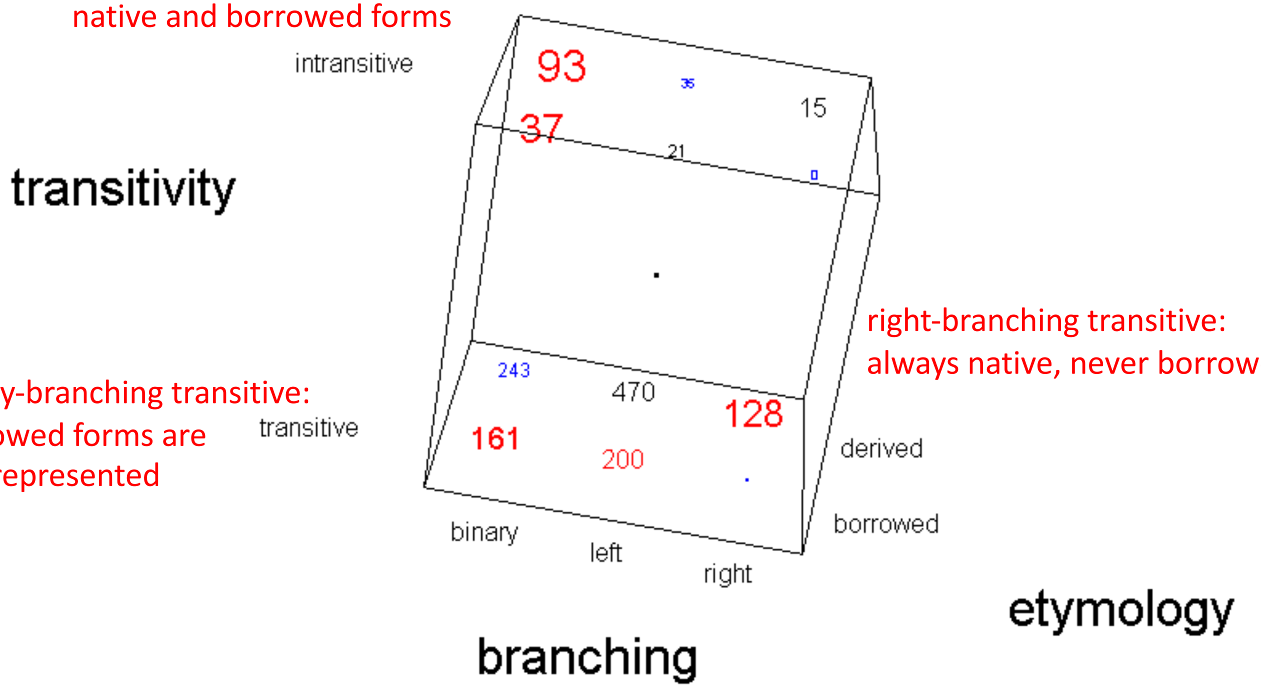
branching

transitivity

	binary	left	right
intransitive	settlement 130	56	15
transitive	treatment, punishment 404	enlargement 670	132

branching

binary-branching intransitive:
native and borrowed forms



right-branching transitive:
always native, never borrowed

binary-branching transitive:
borrowed forms are
overrepresented

Configural Frequency Analysis

- cross-tabulate all 1400 types for the following variables:

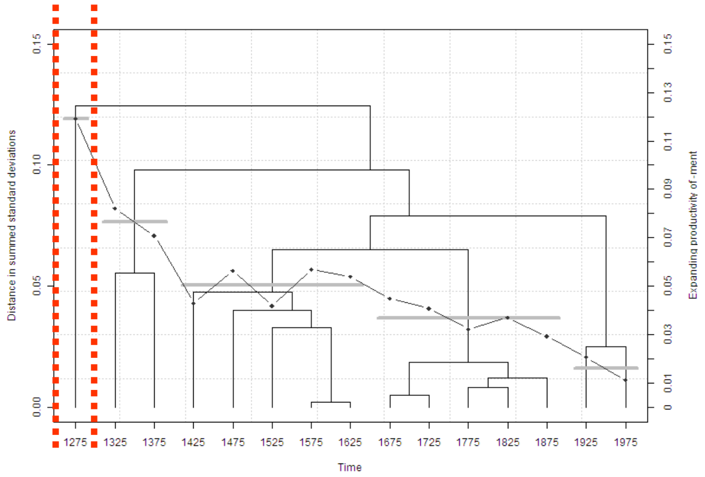
VARIABLE	VALUES
Period:	1,2,3,4,5
Source:	borrowed, derived
Stem:	verb, noun, adjective
Branching:	binary, left, right
Transitivity:	transitive, intransitive
Semantics:	action, result, means, place

- determine configurations of values that occur with greater than chance frequency
- see if early types differ from later types

Results

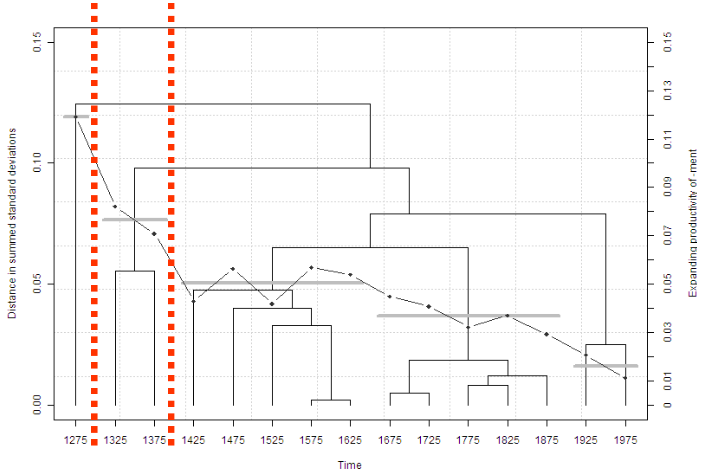
1250-1299

- Type1: commencement
  - borrowed, transitive verbal stem
  - imprisonment, confirmment, enchantment, judgment, ...
  - consonant with previous claims that early ment-types typically had transitive verbs as hosts (Gadde 1910, Dalton-Puffer 1996)



1300-1399

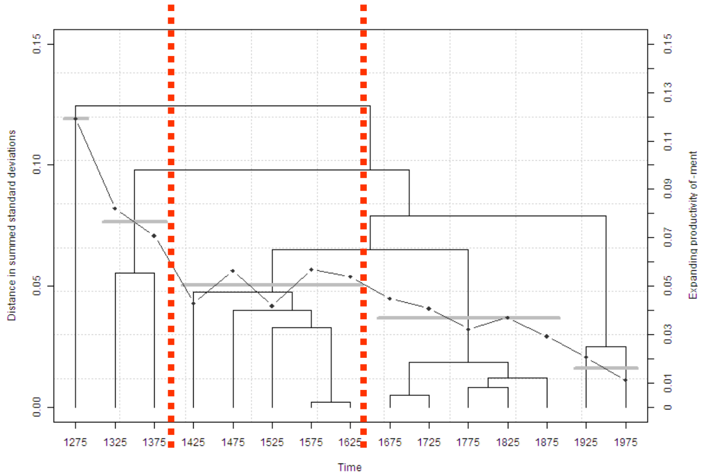
- Type 2: ointment
  - borrowed, verbal, transitive, binary, means
  - vestment, supplement, ornament, ...
  - semantic type of means is not very frequent but rises to a moderate level during the 14th century
  - the forms are classified as transitive because in each case, a ‘patient’ can be identified



- Type 3: vesselment
  - borrowed, nominal, transitive, binary, means
  - monument, odorament, and vesselment
  - highly infrequent, but still more frequent than expected
  - both nominal and means are rare, their combination rarer still

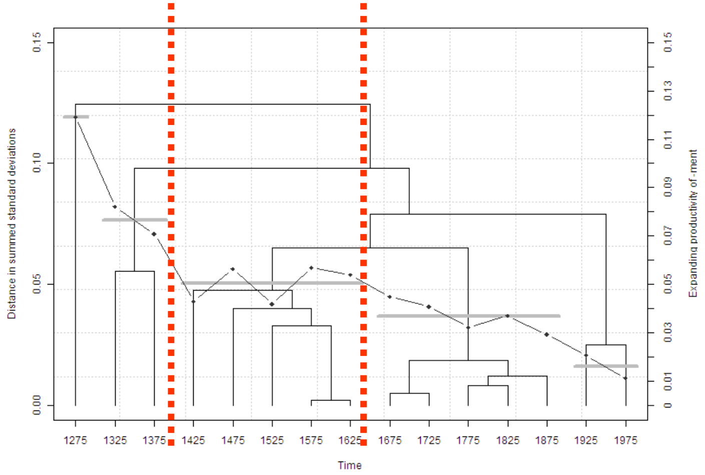
1400-1650

- Type 4: enlargement
  - derived, verbal, transitive, left, action
  - disbursement, misusement, renewment, ...
  - the most frequent configuration in the database (174 instances in period 3 alone, 339 in total)
  - Plag (1999: 16): unattested forms such as encodement or envisionment sound fully acceptable to modern speakers
  - Type 4 explains this: neologisms are OK if the host is a prefixed transitive verb



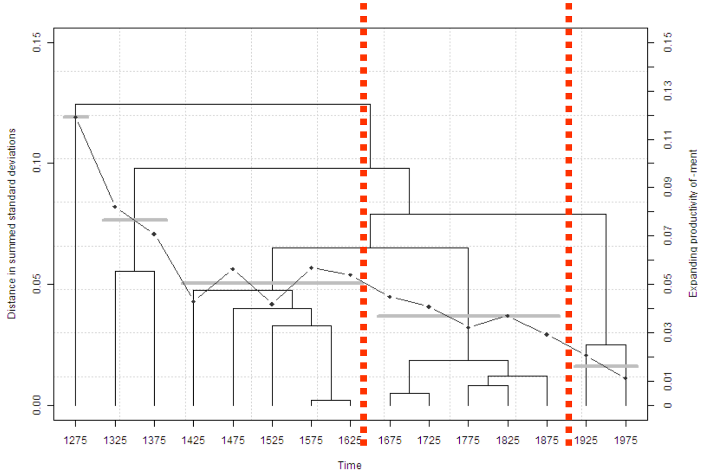
1400-1650

- Type 5: merriment
  - adjectival, derived, action, intransitive, binary
  - coldment, dreariment, jolliment, justment, and wariment
  - genuinely English pattern that is not based on borrowed coinages
  - an innovative but short-lived fad; all types coined between 1548 and 1611



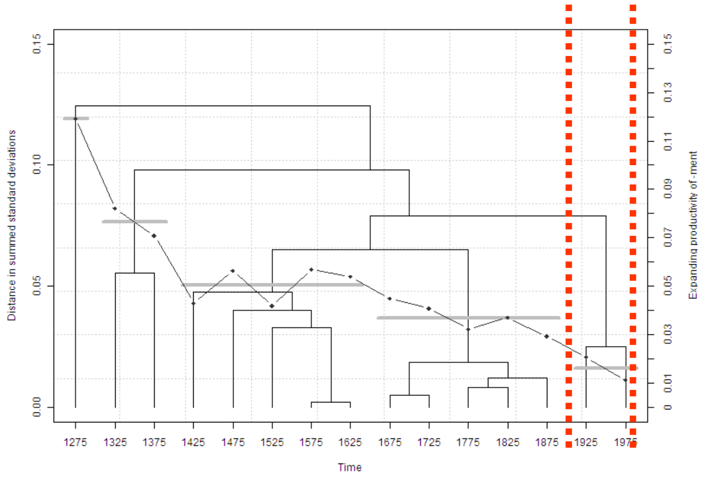
1650-1899

- Type 6: disembodiment
  - right-branching, verbal, action, transitive, derived,
  - maltreatment, overenrichment, reemplacement, selfchastisement
  - typically coined on the basis of Type 4 (enlargement) forms
  - outgrowth of the V-ment prototype
  - independent of the productivity of the suffix –ment: host element is an already existing form of the V-ment construction
  - hence, this type can continue to thrive while other types of the V-ment construction are in demise



1900-2000

- Type 7: semiretirement
  - right-branching, verbal, transitive, derived, result
  - malnourishment, misalignment, noninvolvement, semiretirement, …
  - this type dodges the strong bias towards the meaning of action
  - metonymic shift from actions to results is a recent semantic trend that is confined to right-branching structures



VNC on the basis of distributional semantic information

cabbage

jacket

sheep

mouse

beans

trousers

shirt

pig

potatoes

wheat

hat

boots

cow

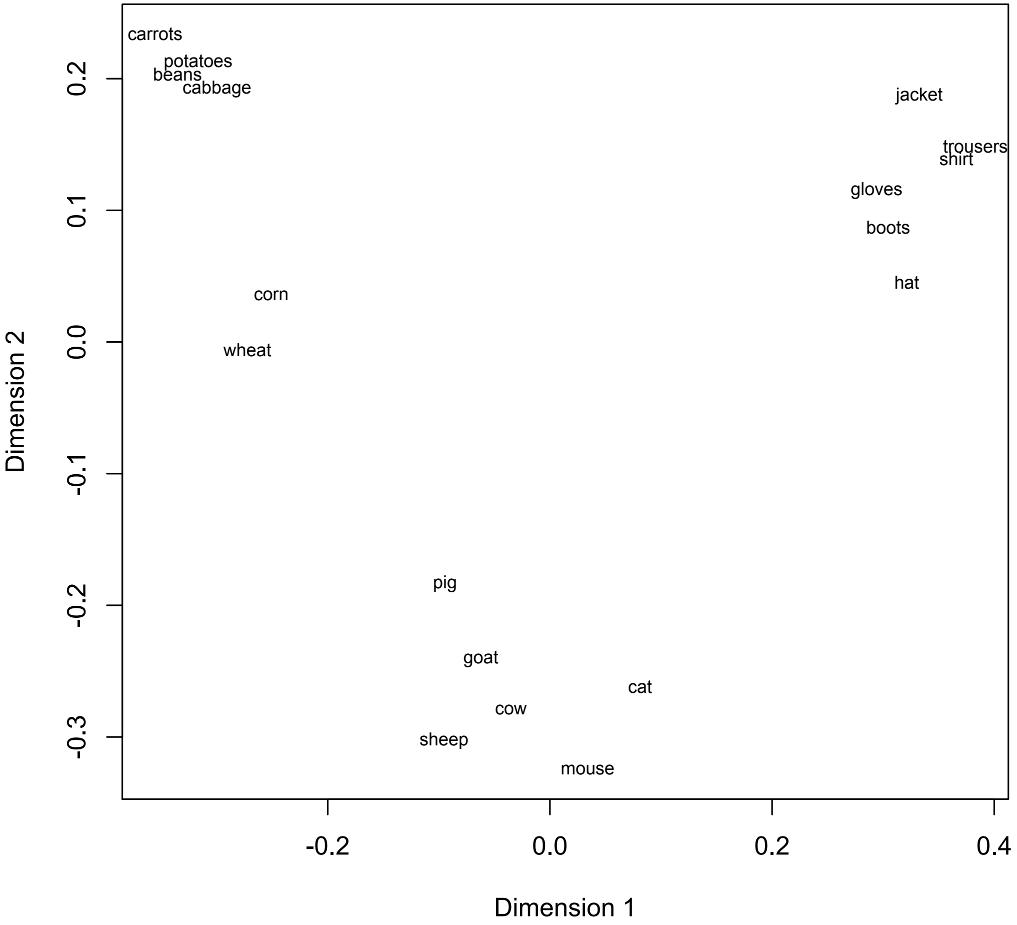
gloves

goat

carrots

cat

corn


carrots
potatoes
beans
cabbage
jacket
trousers
shirt
gloves
boots
hat
corn
wheat
pig
goat
cat
cow
sheep
mouse
0.2
0.1
0.0
-0.1
-0.2
-0.3
Dimension 2
-0.2
0.0
0.2
0.4
Dimension 1

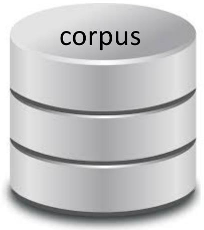
corpus

much hope there . He has his	goat	clinic on Fridays . I hope you
established . Within the mountain	goat	community , this leads to continual
is to get a dead mountain	goat	down from a mountain -- it simply
driving a donkey laden with	goat	fodder . As he passed our party
he wouldn't be a tethered	goat	from choice . He went into
was in good spirits . The	goat	had been sacrificed at the shrine
her a tart or because her pet	goat	had gone missing ; she always
usually from a combination of	goat	hair , cotton and jute , and
And have that lecherous old	goat	hanging round my door

corpus

much hope there . He has his	goat	clinic on Fridays . I hope you
established . Within the mountain	goat	community , this leads to continual
is to get a dead mountain	goat	down from a mountain -- it simply
driving a donkey laden with	goat	fodder . As he passed our party
he wouldn't be a tethered	goat	from choice . He went into
was in good spirits . The	goat	had been sacrificed at the shrine
her a tart or because her pet	goat	had gone missing ; she always
usually from a combination of	goat	hair , cotton and jute , and
And have that lecherous old	goat	hanging round my door

stop words



hope goat clinic Fridays hope

established mountain goat community leads continual

dead mountain goat mountain simply

driving donkey laden goat fodder passed party

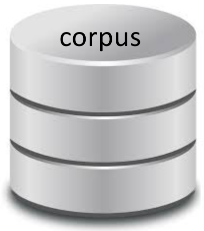
tethered goat choice

good spirits goat sacrificed shrine

tart pet goat missing always

usually combination goat hair cotton jute

lecherous old goat hanging round door



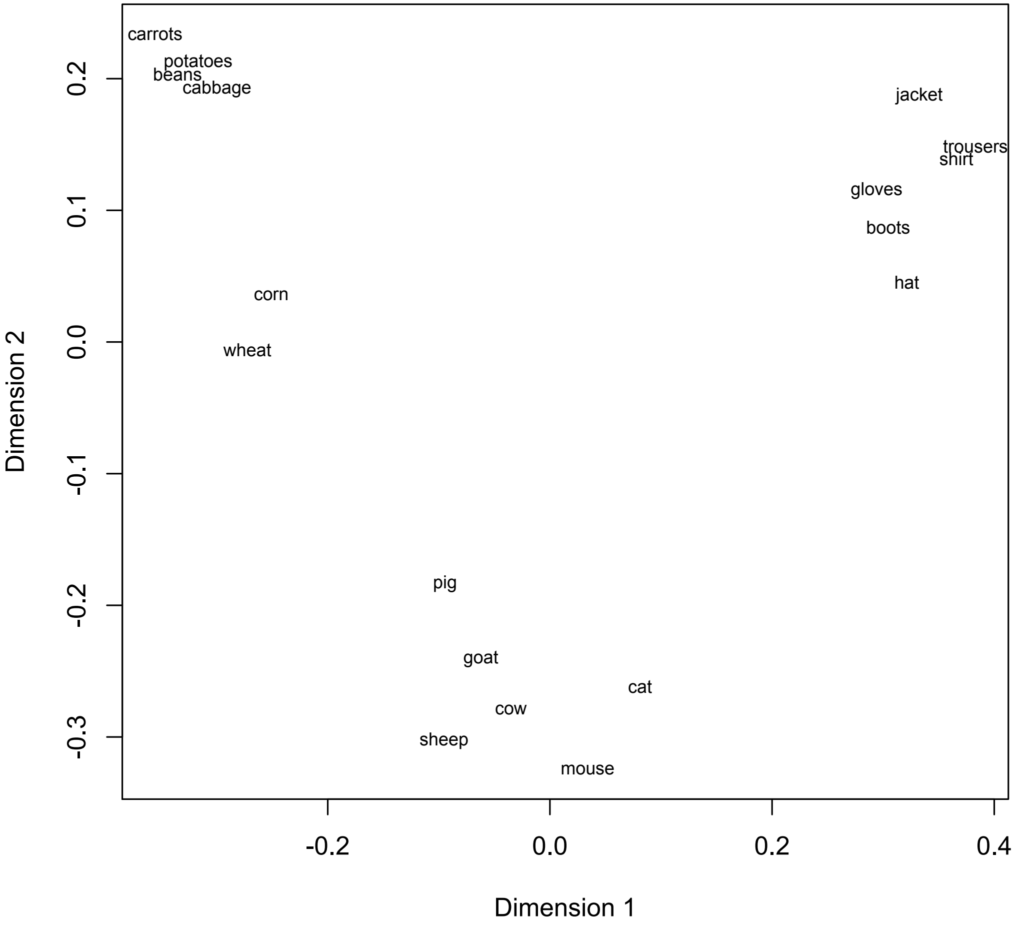
goat
goat
goat
goat
goat
goat
goat
goat
goat

CONTEXT ITEM	FREQUENCY
mountain	48
goat	32
milk	30
cheese	20
sheep	13
meat	9
horns	8
antibodies	8
black	8
gets	7
hens	7
eat	6
tiger	6
head	6
hand	6

CONTEXT ITEM	FREQUENCY
milk	119
cow	80
mad	39
stupid	38
disease	34
silly	28
parsley	26
sheep	21
calf	18
per	17
sacred	17
say	16
little	16
dairy	15
bull	14

CONTEXT ITEM	FREQUENCY
pig	84
wild	27
head	24
pigs	23
iron	20
says	17
farm	16
meat	14
farmer	14
food	13
fact	13
dog	13
thought	12
prices	12
pot	12

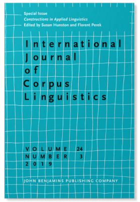
CONTEXT ITEM	FREQUENCY
shirt	150
pair	133
jacket	123
white	108
black	107
trousers	80
wearing	67
shoes	60
grey	58
blue	55
wore	54
boots	51
dressed	48
wear	48
cotton	44


carrots
potatoes
beans
cabbage
jacket
trousers
shirt
gloves
boots
hat
corn
wheat
pig
goat
cat
cow
sheep
mouse
0.2
0.1
0.0
-0.1
-0.2
-0.3
Dimension 2
-0.2
0.0
0.2
0.4
Dimension 1

the *hell* construction

- That scared the hell out of me.
- They beat the hell out of that poor guy.
- Leave it to Patrick to take a simple issue and complicate the hell out of it.

- data from COHA
- 362 tokens with 105 verb types



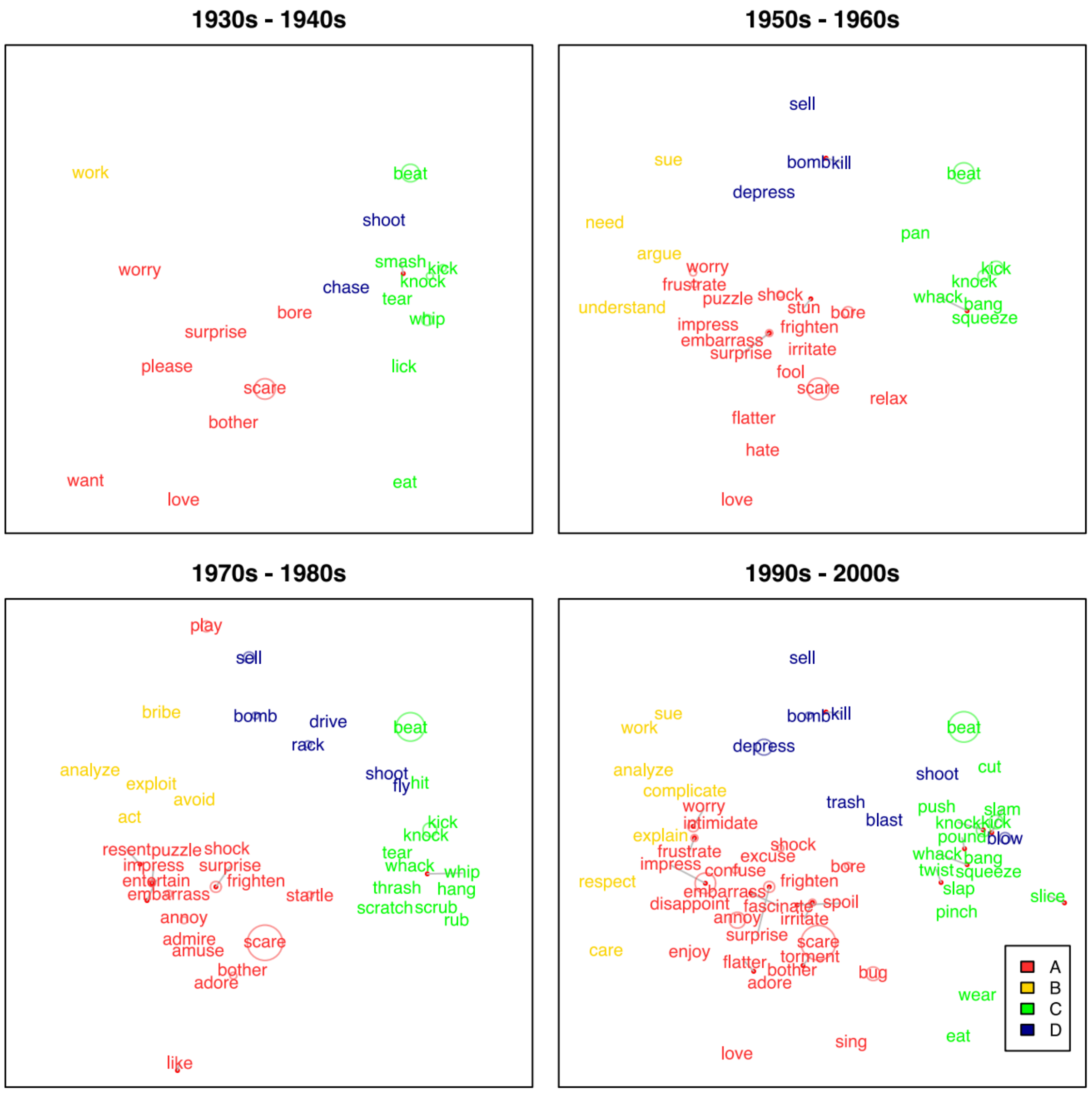
Special Issue
Constructions in Applied Linguistics
Edited by Susan Hunston and Florent Perek

International
Journal
of
Corpus
Linguistics

VOLUME 24
NUMBER 3
2019

JOHN BENJAMINS PUBLISHING COMPANY

Perek & Hilpert (2017) *IJCL* 22/4

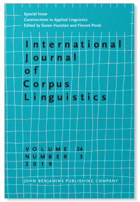


1930s - 1940s

1950s - 1960s

1970s - 1980s

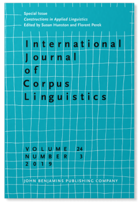
1990s - 2000s



Perek & Hilpert (2017) *IJCL* 22/4

VNC with distributional information

- Each decade in the COHA contains examples of the *hell* construction with several verb types.
- For each verb type, a collocate vector was created.
- All collocate vectors of a given decade were combined by averaging into a single ‘decade vector’.
- VNC was applied to a sequence of decade vectors.



Special Issue
Constructions in Applied Linguistics
Edited by Susan Hunston and Florent Perek

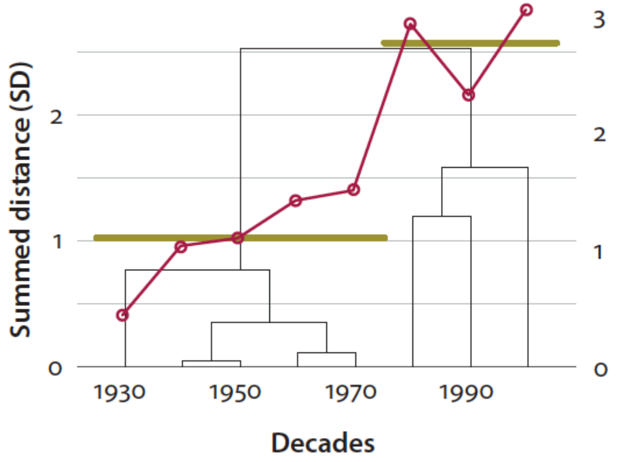
International
Journal
of
Corpus
Linguistics

VOLUME 24
NUMBER 3
2019

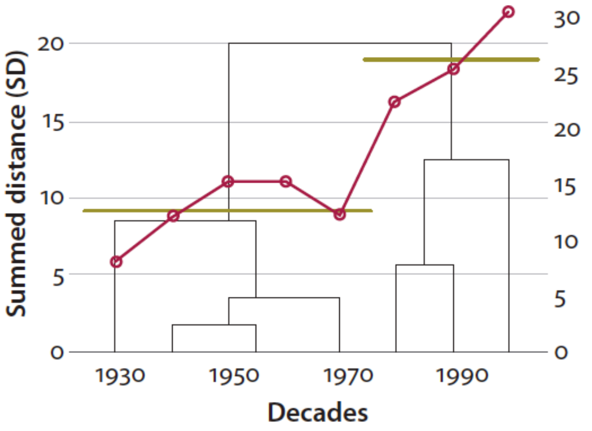
JOHN BENJAMINS PUBLISHING COMPANY

Perek & Hilpert (2017) *IJCL* 22/4

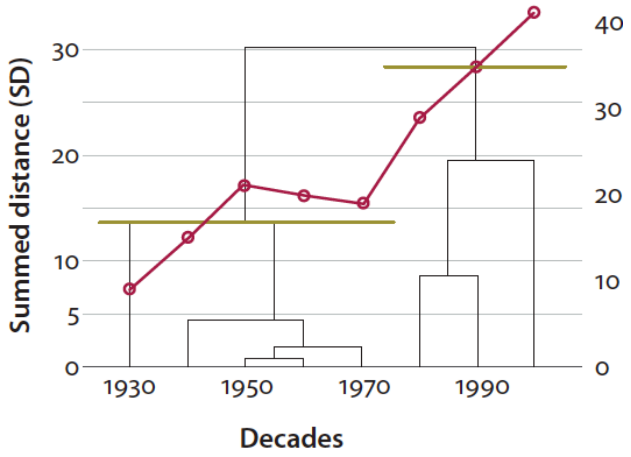
Token frequency (per million words)



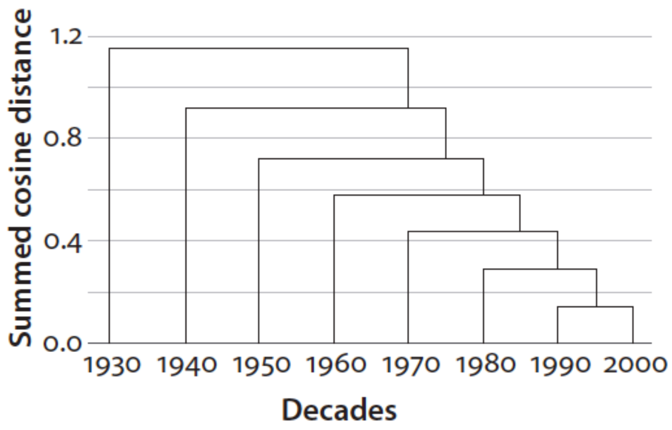
Hapax legomena

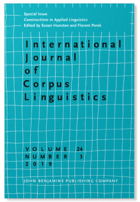


Type frequency



VNC dendrogram



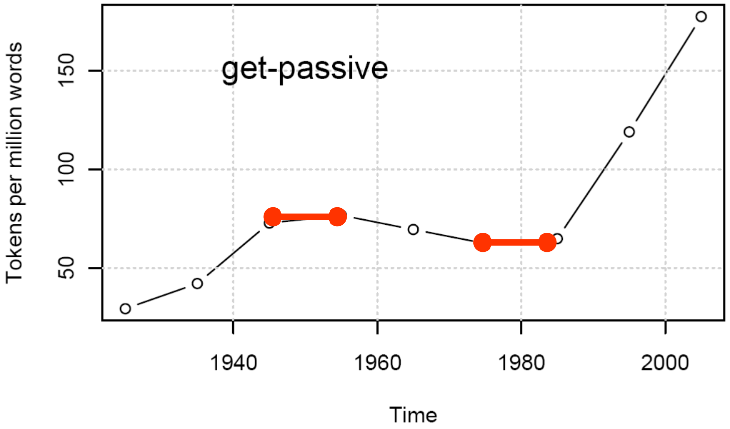


Perek & Hilpert (2017) *IJCL* 22/4

Conclusions

Variability-based Neighbor Clustering

- find the two closest historically adjacent neighbors
- merge them and take the mean value
- now find again the two closest neighbors
- merge them and take the mean value
- ...
- until all periods are merged



- Upsides:
  - VNC can find stages in a data-driven, bottom-up way
  - identifying stages is useful, sometimes necessary, for the decription of changes
  - VNC can be used for the detection of outlier data points
  - finding different stages for two forms can show that they are indeed different constructions

- Downsides:
  - clustering does not provide divine truths: results reflect the similarity measure used in the input
  - determining stages is usually just a first step in an analysis

martin.hilpert@unine.ch

UNIVERSITE
DE
NEUCHATEL