Bottom-up methods in cognitive and corpus linguistics: on letting the data decide

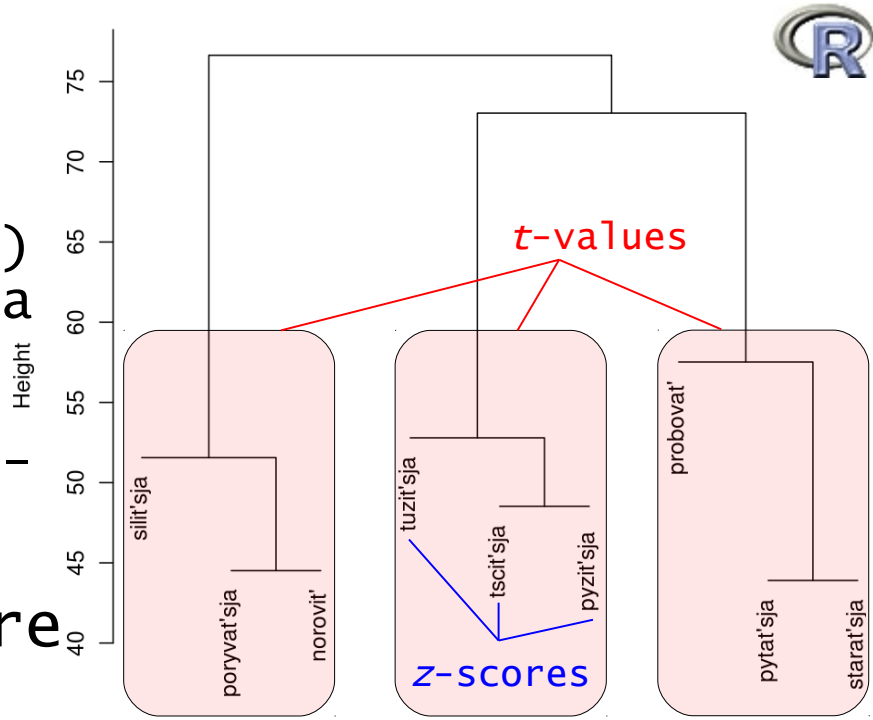
Stefan Th. Gries
Department of Linguistics
University of California, Santa Barbara
http://tinyurl.com/stgries

Why the field has to become more statistical (in two directions) …

- More and more corpus-linguistic studies are based on
  - increasingly larger (samples from) corpora
  - increasingly complex (samples from) corpora
    - complex in terms of both composition and annotation
  - temporally- or otherwise ordered corpora
- these developments often lead to large multi-dimensional data sets whose size and complexity defies
  - mere eye-balling of the data
  - introspective analysis of the data
- therefore, statistical tools are becoming more important and more frequently used
  - sometimes, statistical applications are used in an exploratory / hypothesis-generating way
    - (which will be the topic of this talk)
  - sometimes, statistical applications are used in a hypothesis-testing way
    - (which will be the topic of the next talk)

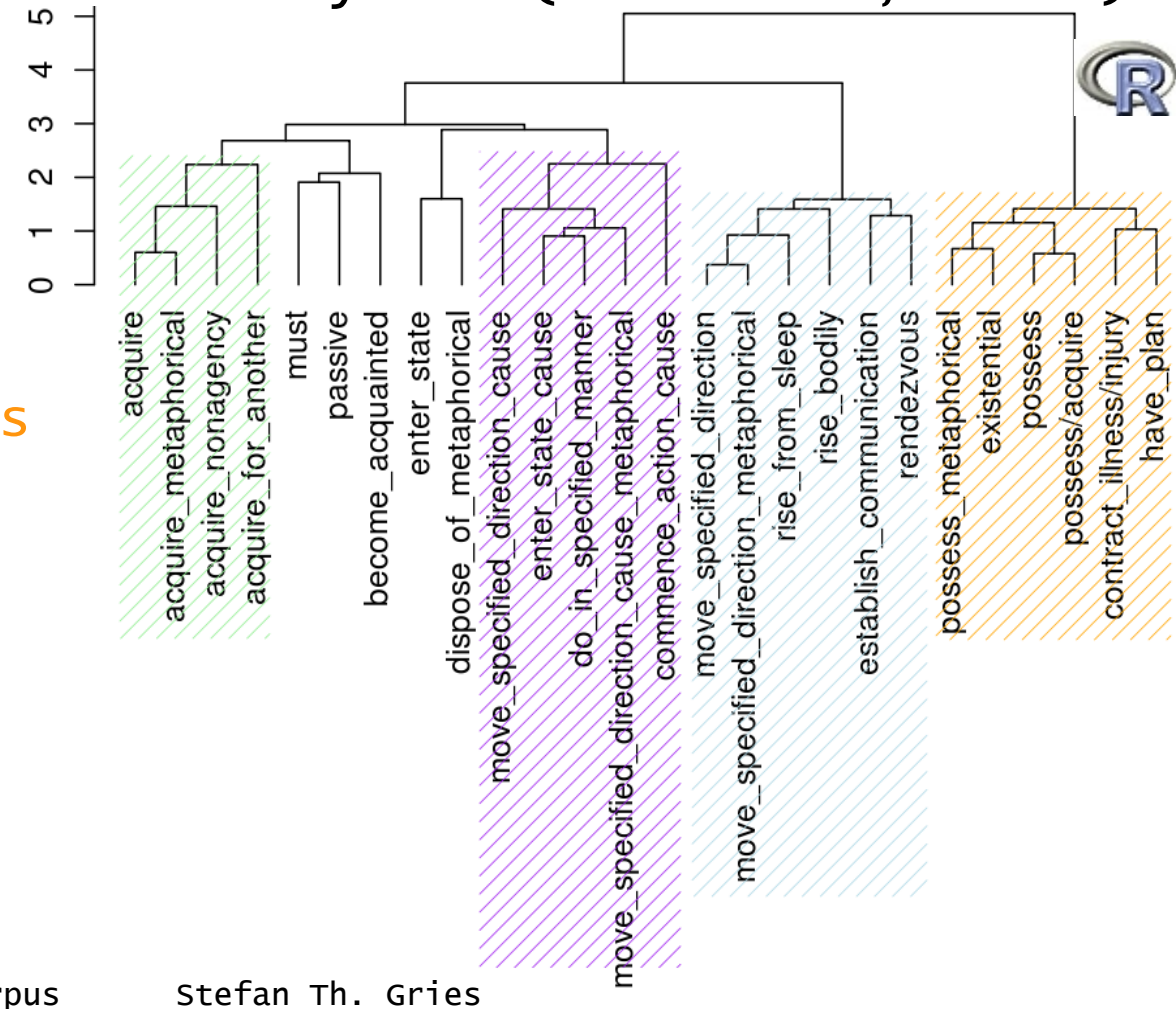
Recall the clustering of near synonyms?

- Interpretation: what do the clusters represent? why are they the way they are?
  - *t*-values: which ID tags 'load high' on which cluster
  - z-scores: which ID tags load high on which verb
    - {*poryvat'sja norovit' silit'sja*}
      - semantics: inanimate subjects physical-motion verbs, uncontrollable, repeated actions
    - {*pyzit'sja tuzit'sja tschit'sja*}
      - semantics: inanimate subj, (fig.) physical-motion verbs affecting a second entity, high vainness
    - {*probovat' pytat'sja starat'sja*}
      - semantics: animate subj. were exhorted to undertake attempt and perform it at reduced intensity
- = compatible with some, but more precise than, previous work



Recall the clustering of senses?

- Evaluation: 26 senses occurring 5+ times were entered into a hier. cluster analysis (Canberra, Ward)
  - several interpretable clusters emerge
    - various 'acquire' senses
    - various causative metaphorical motion senses
    - various metaphorical motion senses
    - various possession senses



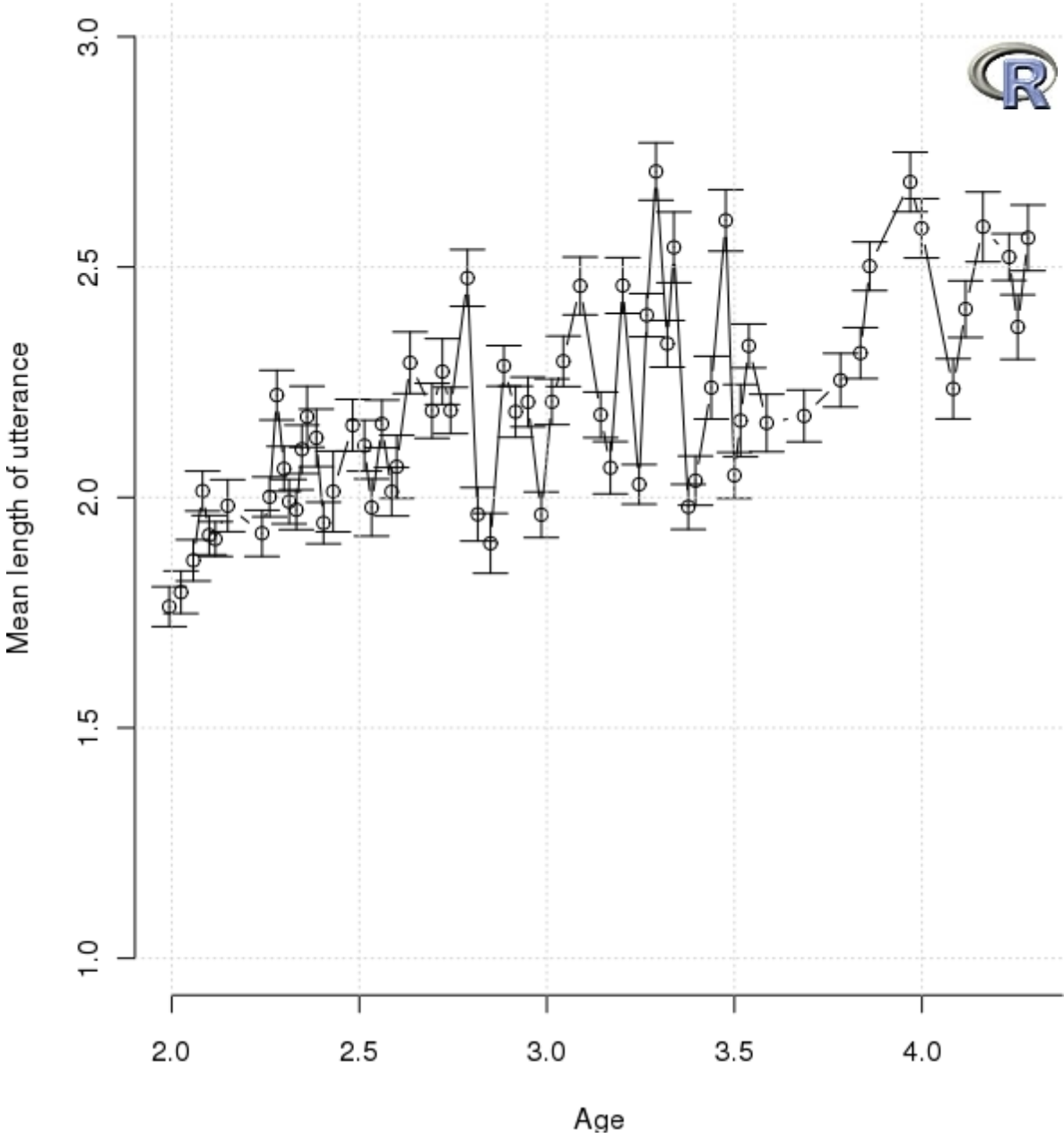
More and more corpus studies on temporal developments …

- With the growing availability of corpora, more and more studies in contemporary corpus linguistics use data from corpora containing sequentially-ordered data
- prominent fields of application
  - historical linguistics
  - language acquisition
- nearly all of these studies include comparisons of how the frequencies of linguistic expressions change over time

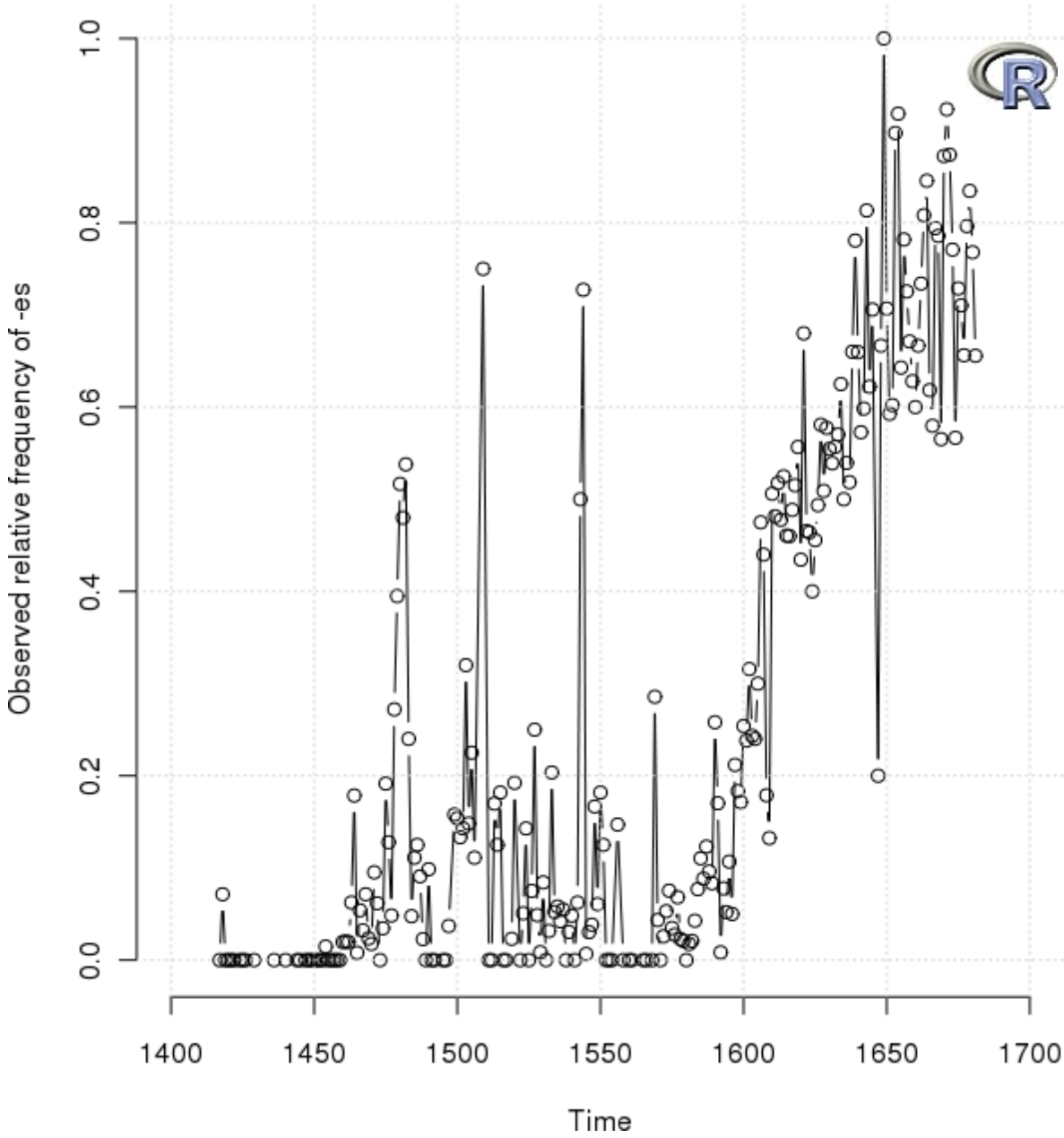
… lead to a variety of pressing methodological questions

- However, with few exceptions many larger resources have become available only recently …
- … which also means that there is as yet little if any consensus as to how sequentially-ordered data can be studied most revealingly and objectively
  - how do we identify trends across the whole range or parts of the data?
  - how do we separate seemingly meaningful developments from seemingly accidental/arbitrary fluctuations?
  - how do we identify groups, i.e. coherent chunks?
  - how do we identify outliers or surprising data points?

Example 1: MLUs

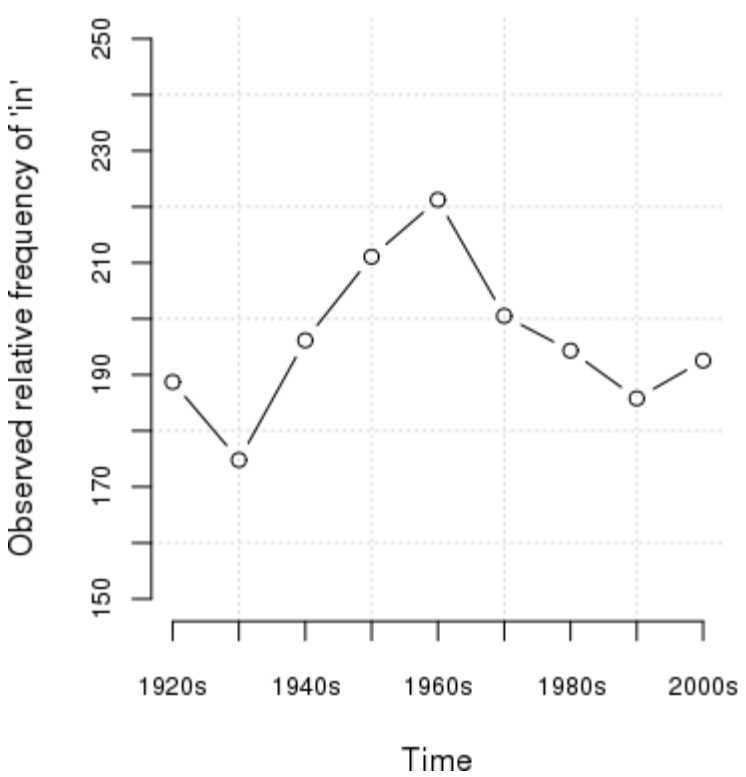


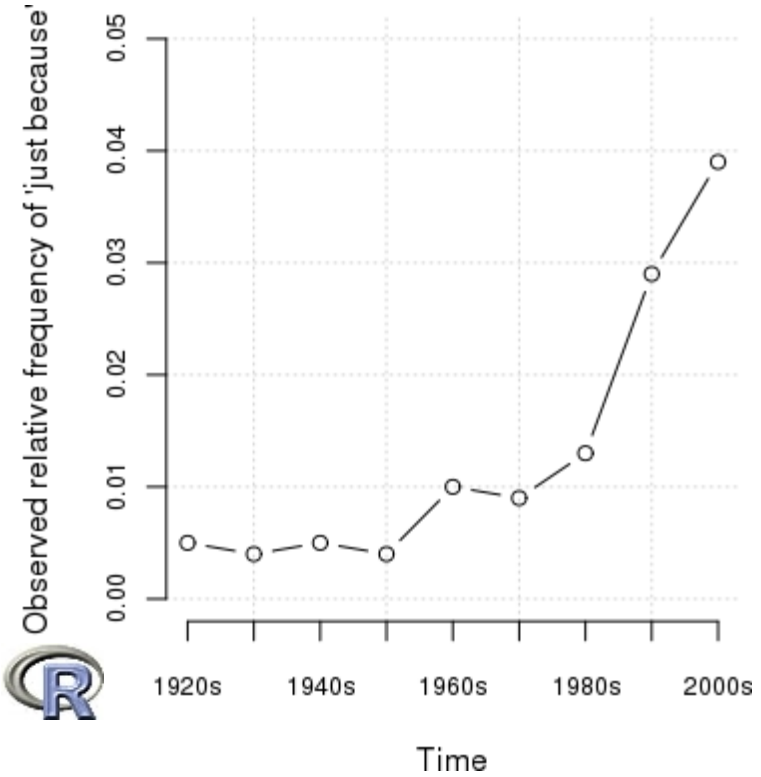
Example 2: *-(e)th → -(e)s*



Identifying trends can actually be easy: the frequencies of *in* and *just because*

- Hilpert & Gries retrieved the normalized frequencies of occurrence of *in* and *just because* from the TIME corpus (<http://corpus.byu.edu/time/>)
  - 106 million words
  - nine time periods (ranging from 1923 to 2006)





Is there an overall trend? the rank correlation

- Question: is there one overall trend in the developments?
- paraphrase: is there a correlation between the successive time periods and the corresponding frequencies?

Time period	Freq of *in*	Freq of *just because*
1920s	188.72	0.005
1930s	174.79	0.004
1940s	196.15	0.005
1950s	211.07	0.004
1960s	221.24	0.010
1970s	200.52	0.009
1980s	194.32	0.013
1990s	185.78	0.029
2000s	192.54	0.039

- while Pearson's *r* is used most frequently, it presupposes normally-distributed interval data and is sensitive to outliers – better choice: Kendall's τ
  - Kendall's τ for *in*: 0 ($p_{\text{two-tailed}}$=1)
  - Kendall's τ for *just because*: 0.857 ($p_{\text{two-tailed}}$=0.001)

Interim conclusions

- The correlational approach showed that there is a significant positive correlation between the time period and, say, the frequency of *just because* …
- … but the correlational approach implicitly treats the data as one 'homogeneous' set …
- … and maybe the data do not constitute one homogeneous set
- however, one may not have any hypotheses whatsoever
  - about how many sets/trends – i.e., linguistically different time periods – there are: 2, 3, 4, …?
  - about how long each of these linguistically different time periods is
- thus, data sets like these might call for exploratory / bottom-up tools

Multiple trends: the acquisition of tense/aspect in Russian

- An example involving sequential data from Russian first language acquisition
- children's acquisition of aspect is often characterized by a strong tense-aspect correlation
  - present tense and imperfective aspect
  - past tense and perfective aspect
    (the aspect hypothesis)
- empiricist/cognitive approaches explain the above distribution with reference to the child's initially inflexible formation of islands/prototypes
- this tight correlation gets relaxed as the child learns that past events can in fact be viewed as incomplete

The correlation of tense and aspect

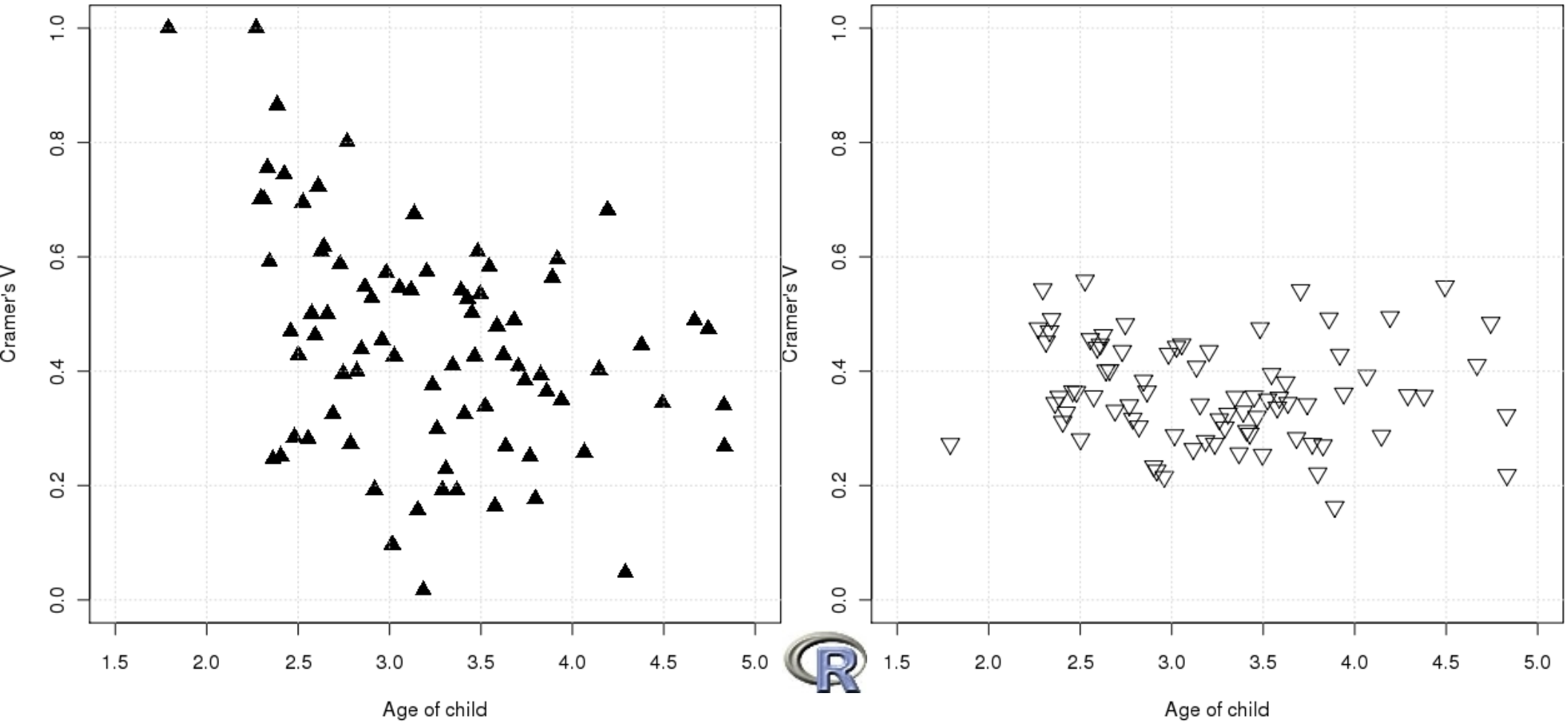
- The following data were analyzed
  - 80 recordings from a Russian child from the Stoll corpus of Russian acquisition
  - 6,796 utterances with verbs by the child
  - 31,687 utterances with verbs by all caretakers
- for each recording, each verb was coded for its tense and its aspect

observed	Non-past	Past	Totals
Imperfective	60	20	80
Perfective	20	28	48
Totals	80	48	128

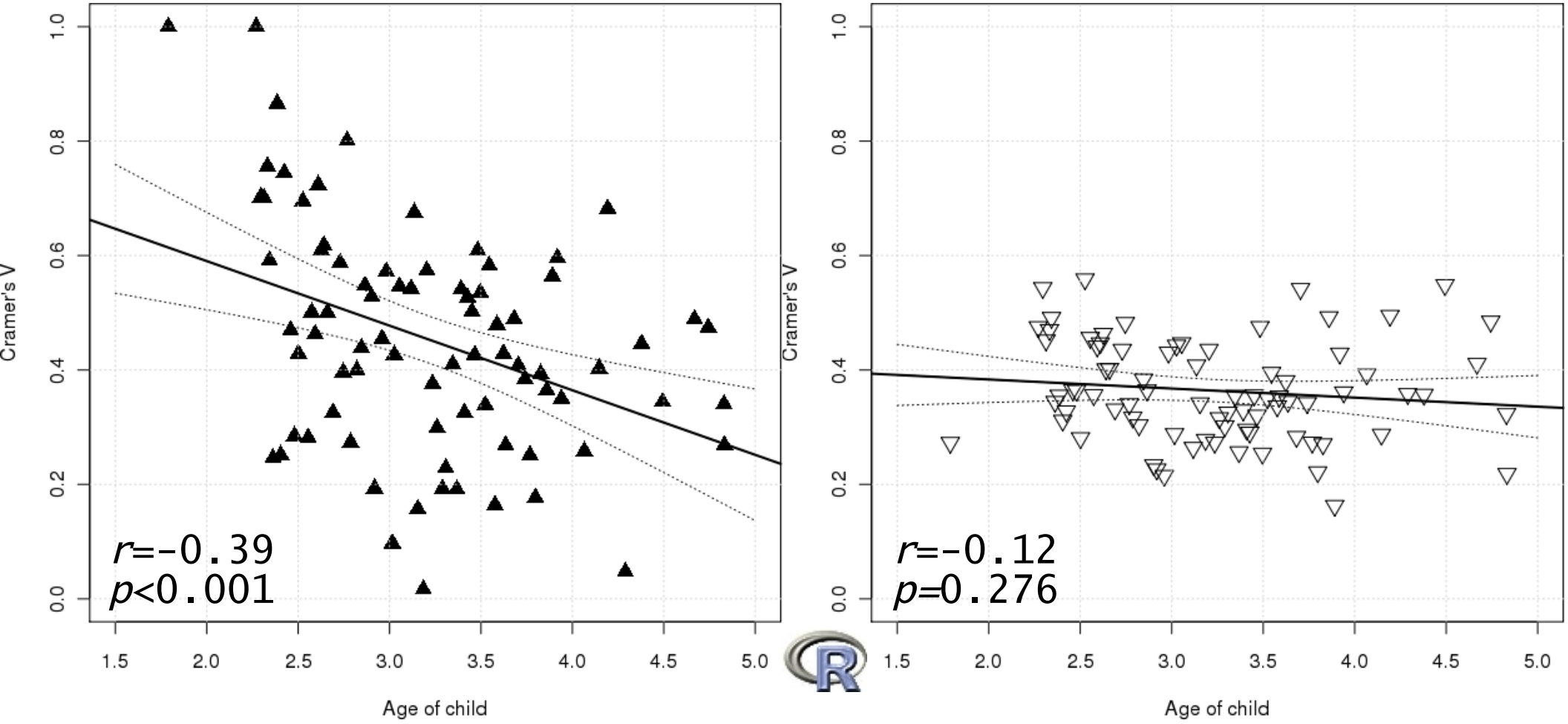
expected	Non-past	Past	Totals
Imperfective	50	30	80
Perfective	30	18	48
Totals	80	48	128

- X^2=14.22; Cramer's V=0.33 ($0 \leq V \leq 1$) and scatterplots of the results of all recordings

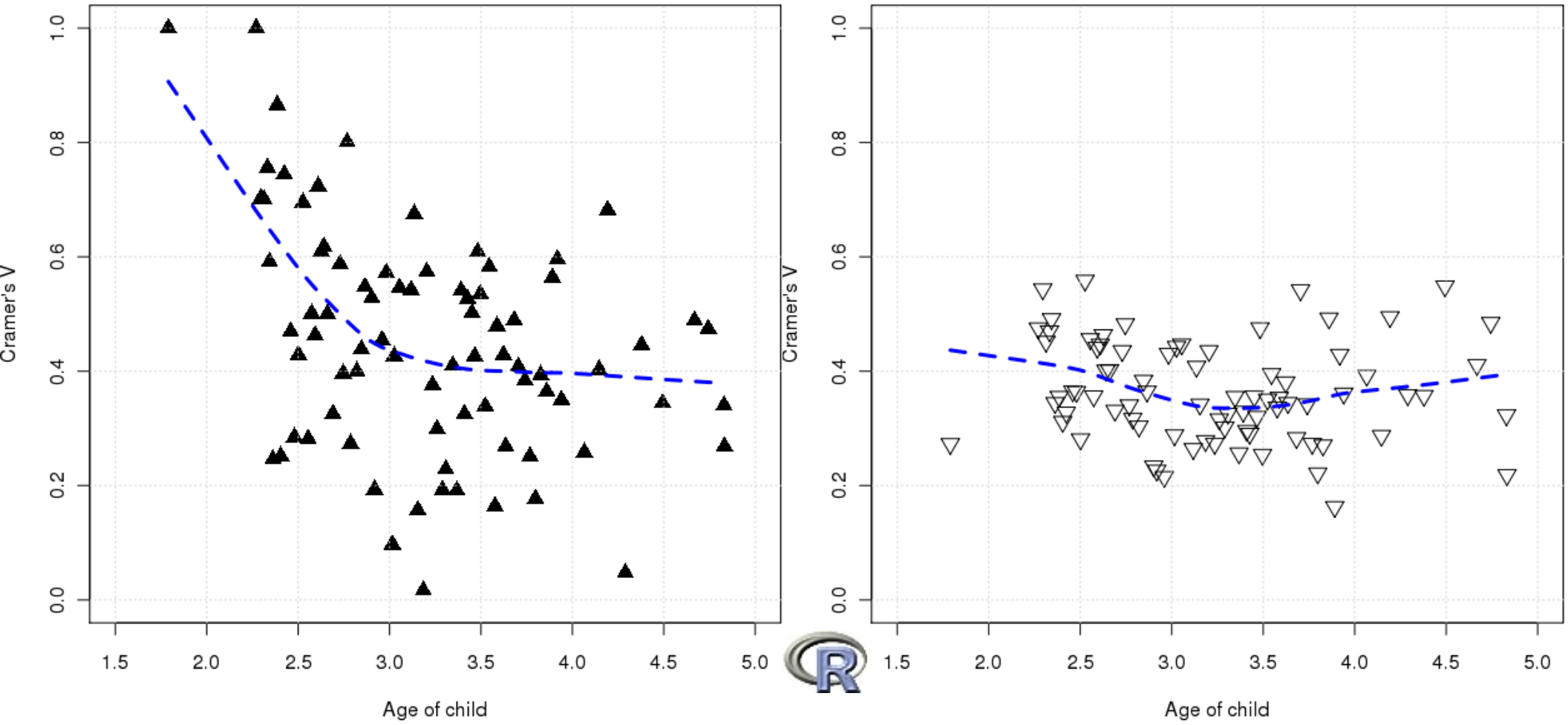
Cramer's *Vs* for the child (left) and the caretakers (right)



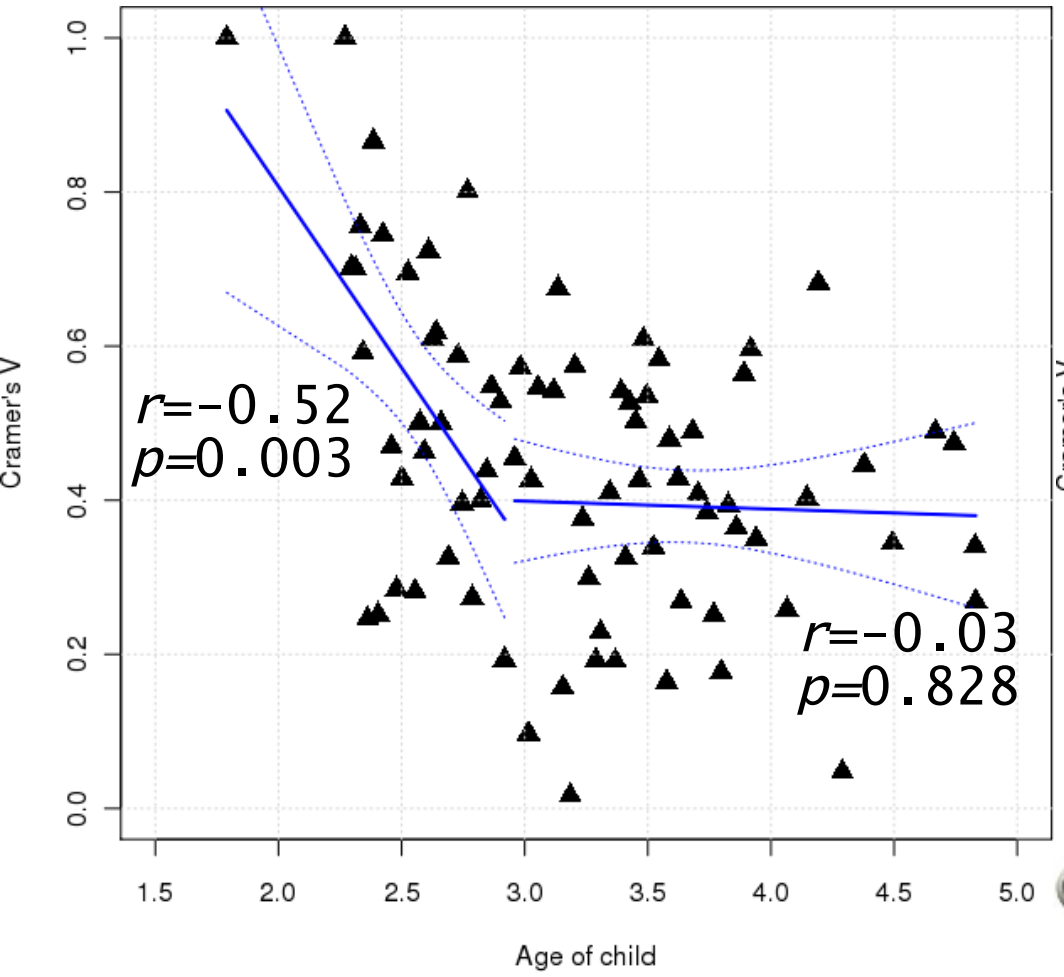
Regressions for the child (left) and the caretakers (right)

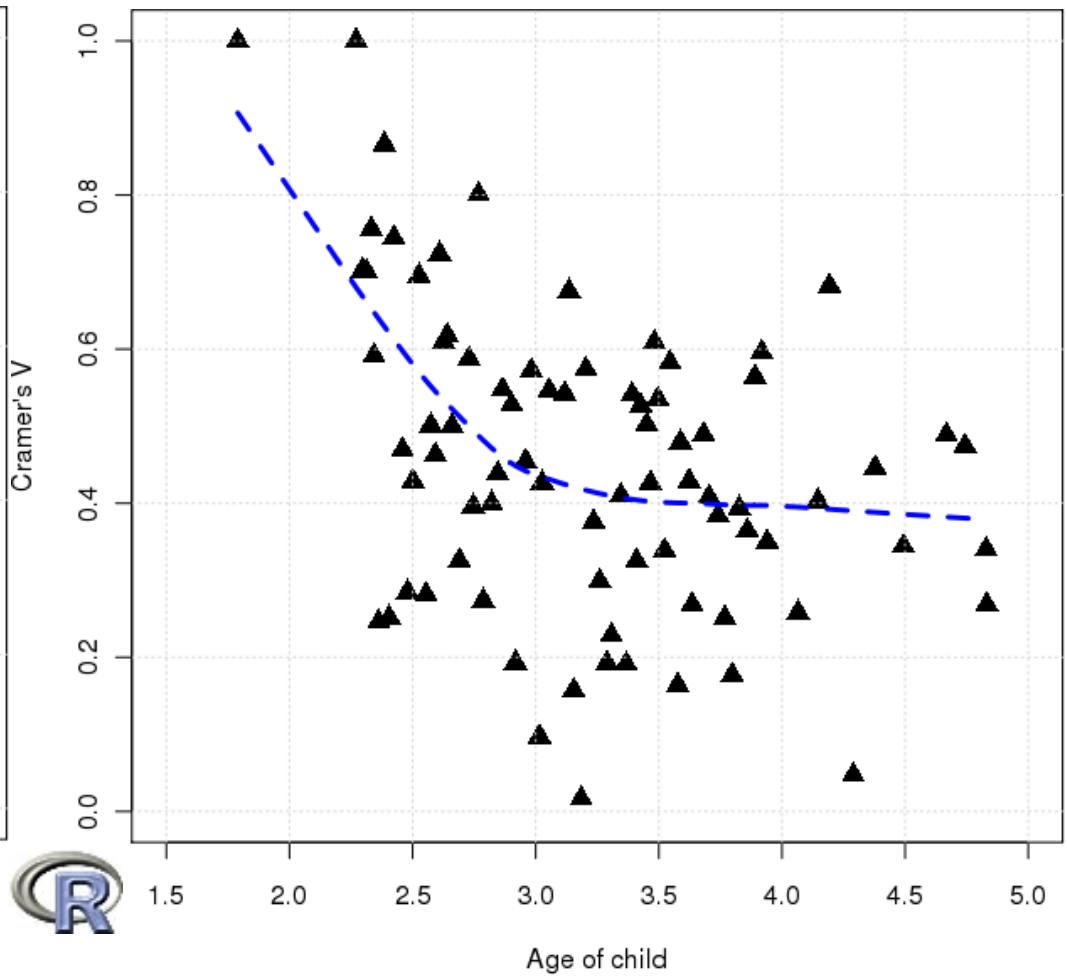


Smoothers for the child (left) and the caretakers (right)



Regression with breakpoints for the child (left)

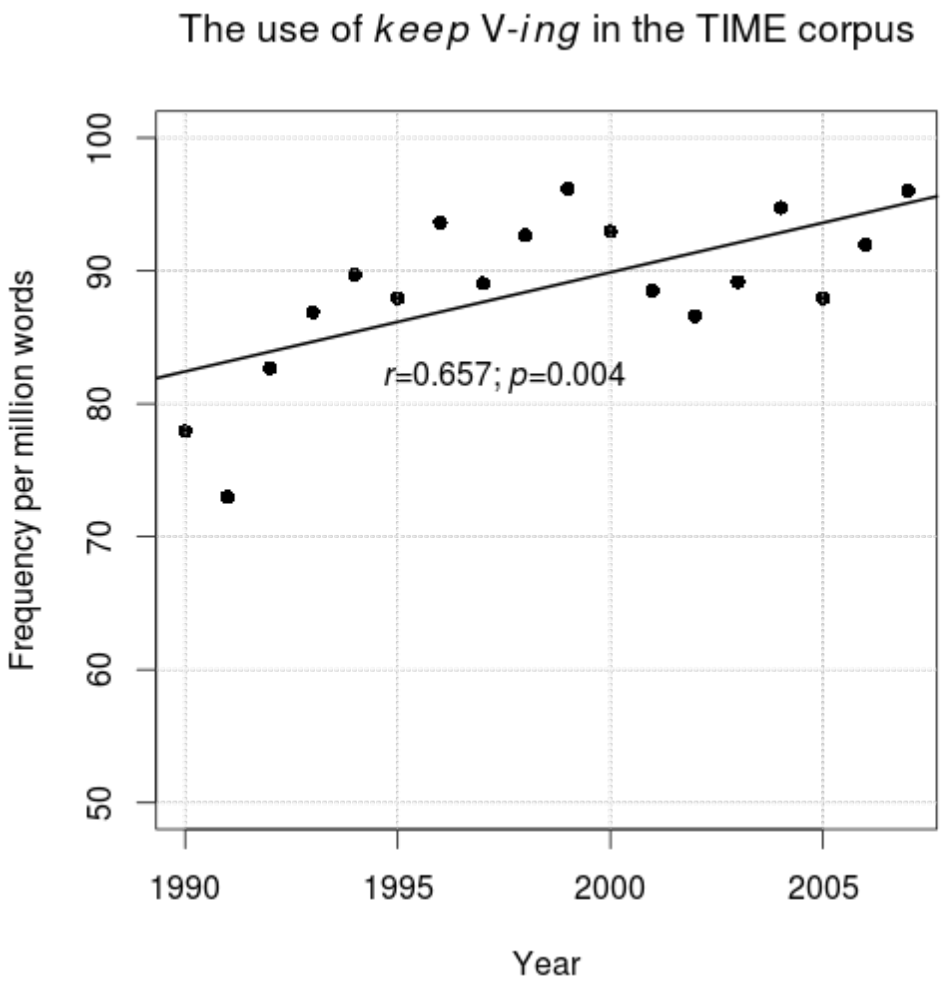


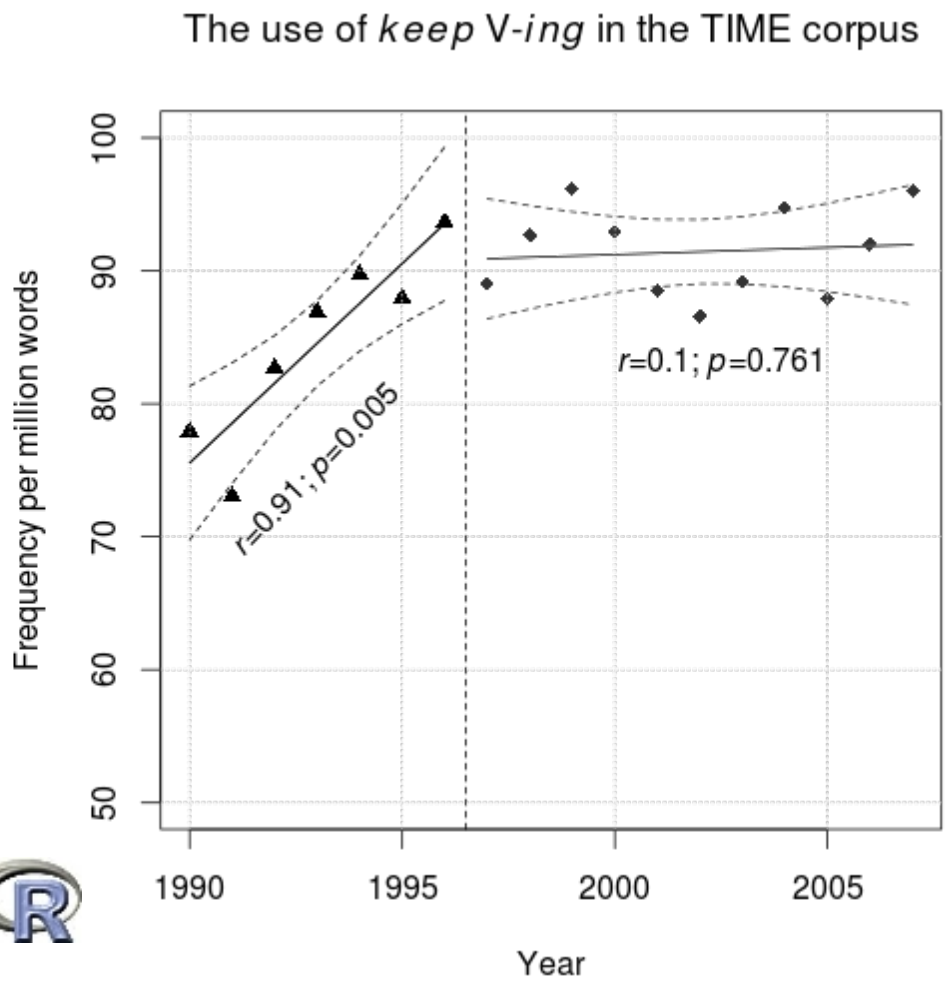


Interim conclusions

- The data
  - provide support for the aspect hypothesis (and the distributional bias hypothesis)
  - show a differential pattern indicating two distinct age groups
    - the former is characterized by rapid and highly significant adaptation towards a tense-aspect correlation characteristic of adult language use
    - the latter is characterized by a tense-aspect correlation that does not differ from non-developing adults anymore
  - but simple plotting does not provide that information … and neither does the standard kind of correlation/ regression methods
  - but locally-weighted robust smoothing techniques and regressions with breakpoints reveal the patterns immediately

A similar case in 'historical' data





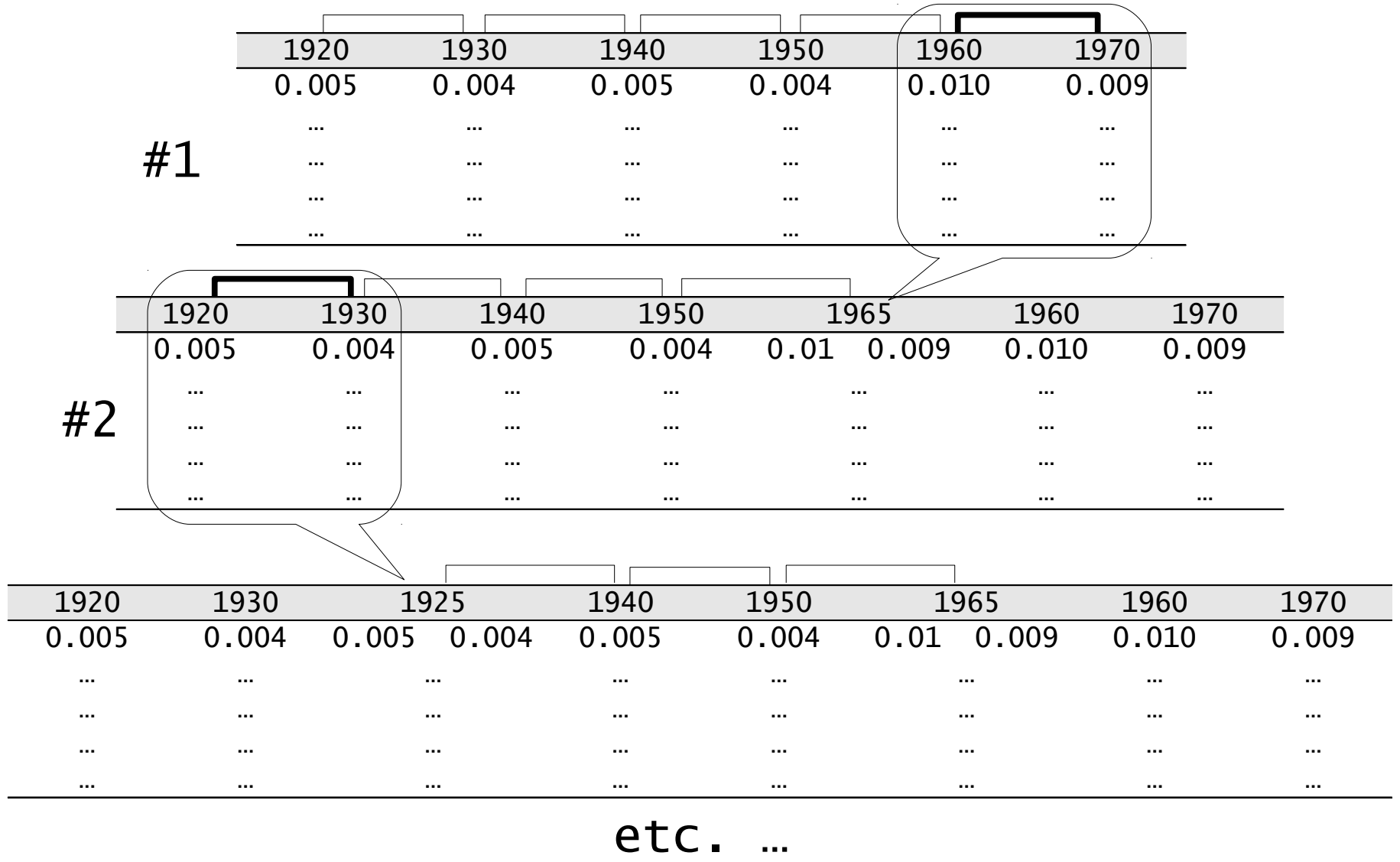
Detecting different kinds of structures: variability-based neighbor clustering

- Knowing there is an overall trend, or several trends, is often not enough because the data may exhibit substructures that are not best characterized as trends, plus it is not always obvious how many different trends to assume a priori
- detecting such structure is often done with clustering approaches …
- but such approaches cannot be applied here, mainly because amalgamations of temporally disparate recordings are not appropriate
- the solution: variability-based neighbor clustering
- VNC is a recursive algorithm whose main distinct characteristic is that it takes temporal ordering into account: it can *not* cluster temporally non-adjacent recordings/files

Variability-based neighbor clustering: the algorithm

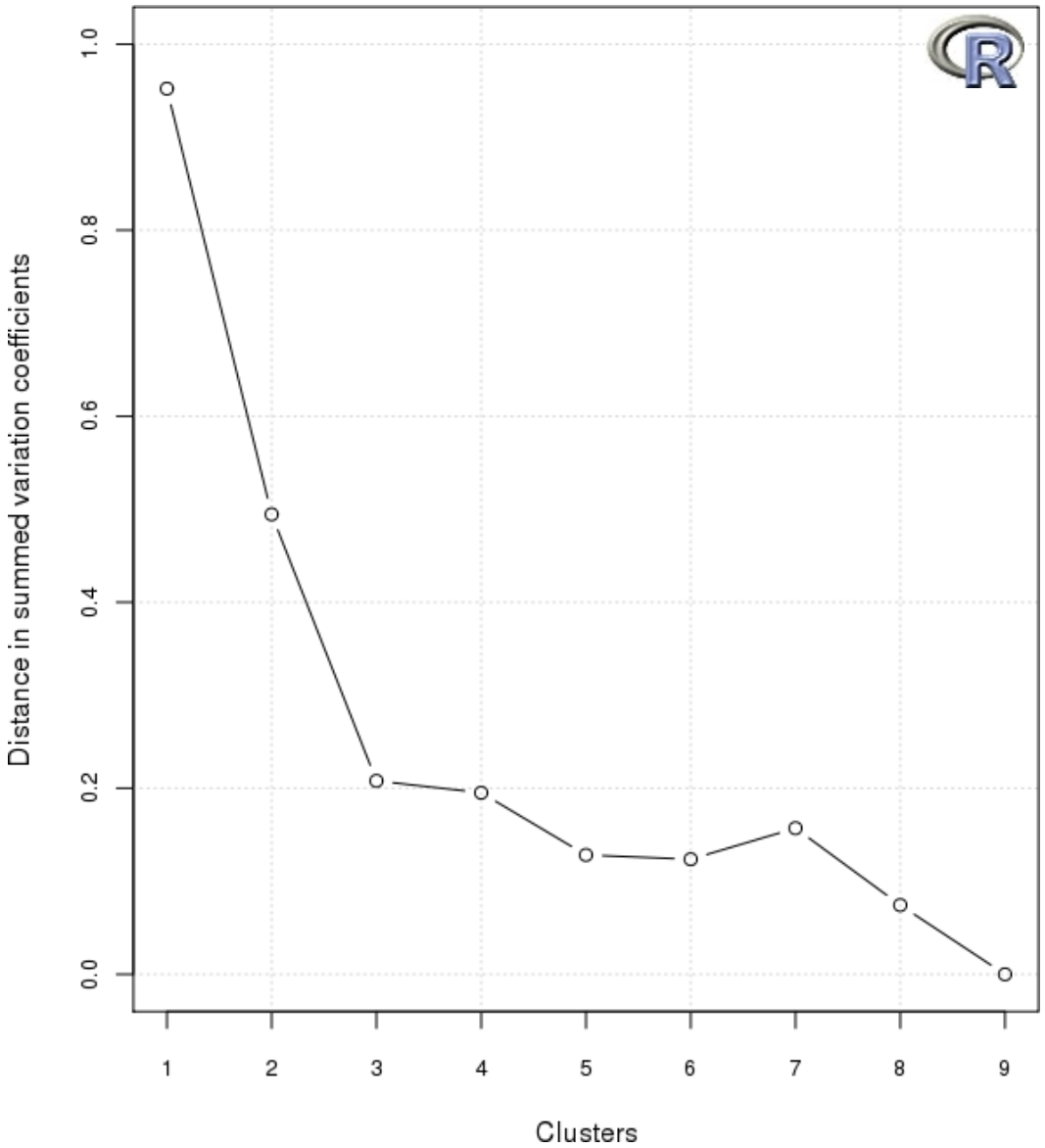
- The pseudocode
  - repeat
  - for all but the last recording date
  - access the measure (frequency) from recording x
  - access the measure (frequency) from recording x+1
  - compute & store a measure of their variability
  - identify & store the smallest measure of variability
  - merge the two recordings whose measure of variability is smallest
  - by concatenating the measures of the recordings, and
  - by renaming the recording (to the weighted mean of the original recordings)
  - until there is just one (huge!) recording left

VNC: the algorithm

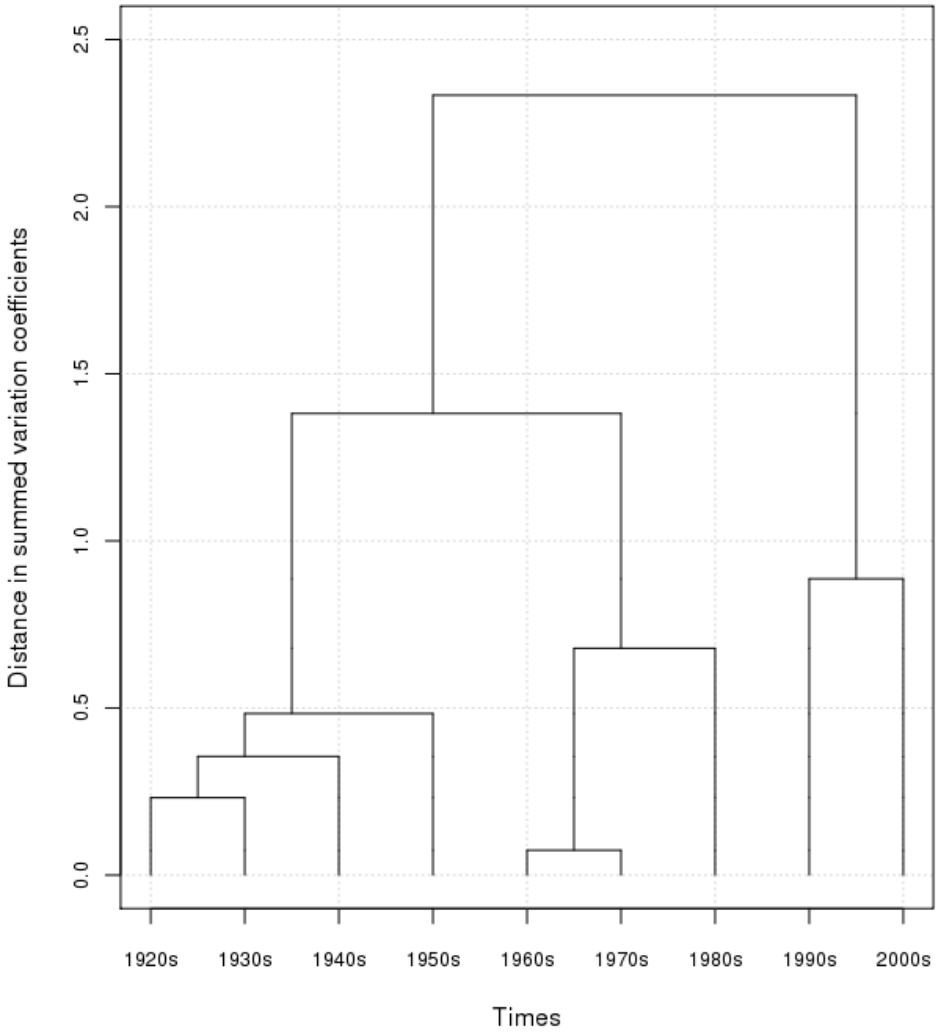


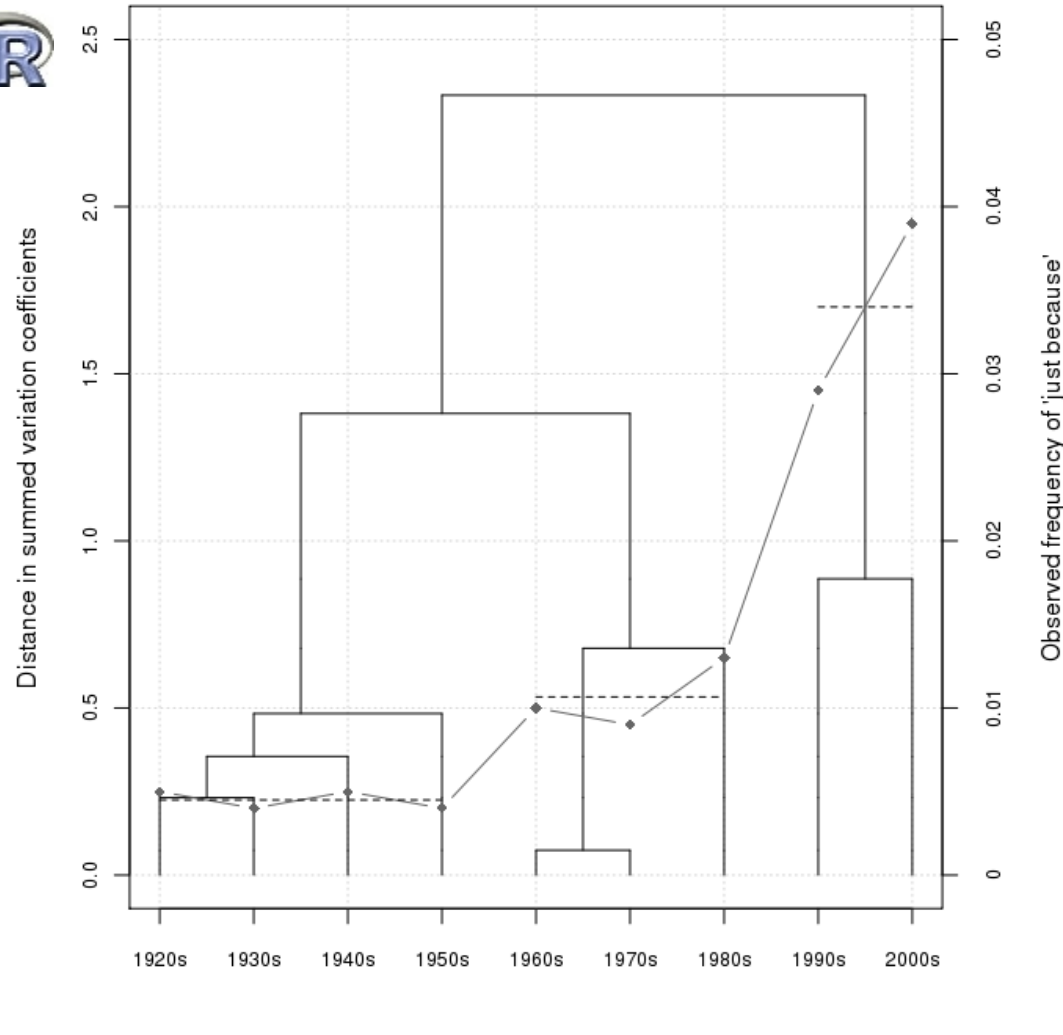
etc. …

VNC (*just because*): how many groups?



VNC (*just because*): what are they?

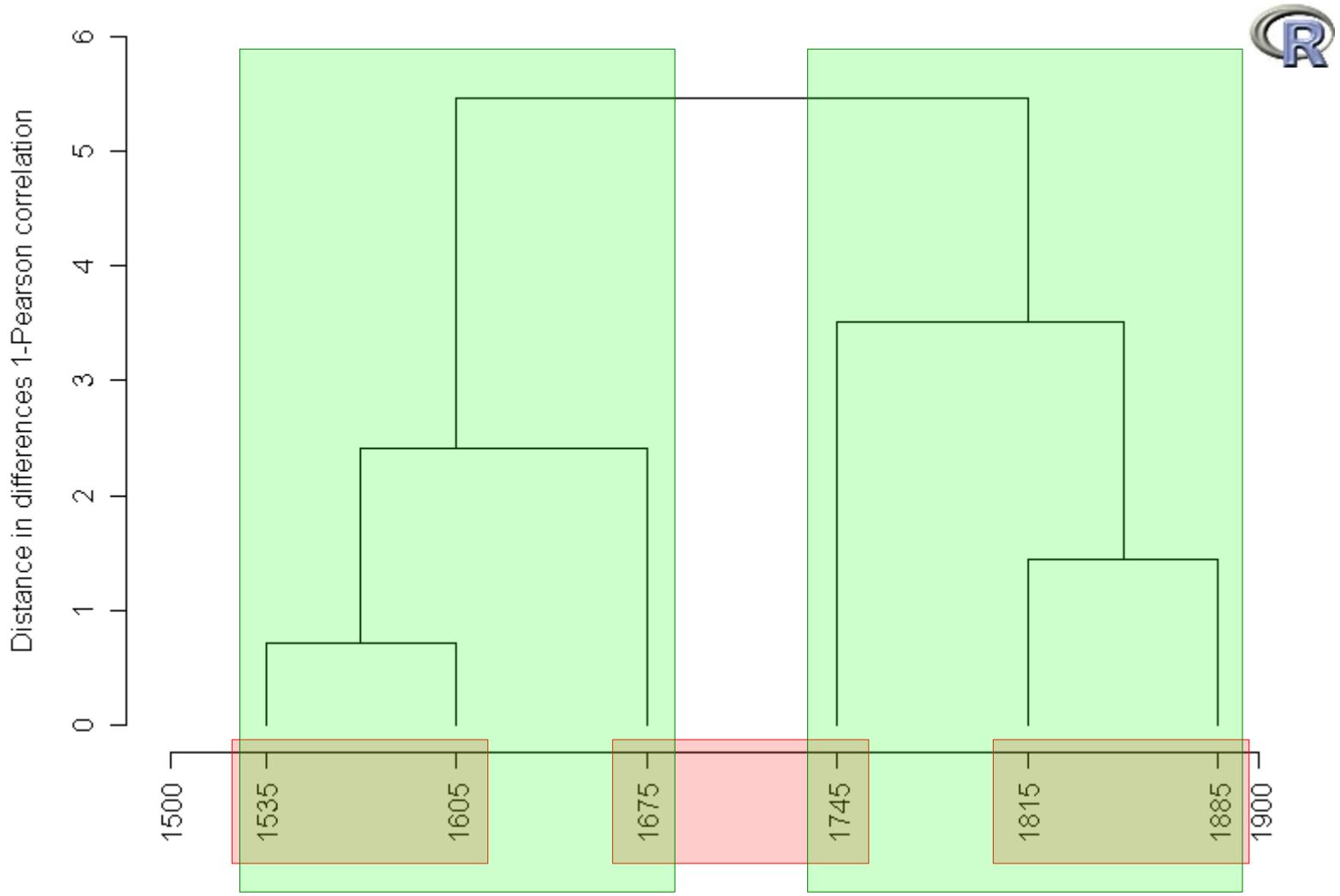




Interim conclusions

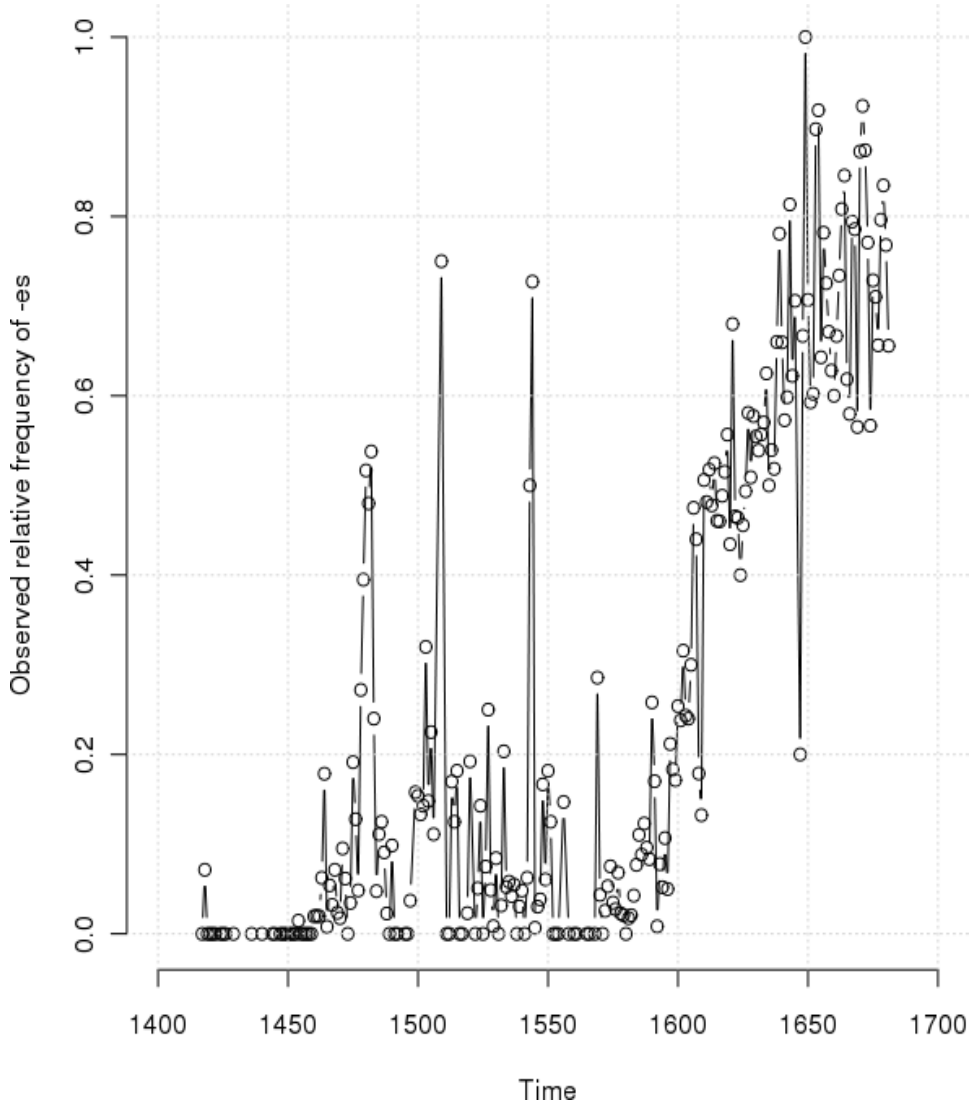
- The correlational approach showed that there is a significant positive correlation between the time period and, say, the frequency of *just because* …
- … but the correlational approach implicitly treats the data as one 'homogeneous' set …
- … and maybe the data do not constitute one homogeneous set
- one possible exploratory approach, VNC, suggests
  - there is indeed a development such that *just because* is used more frequently over time
  - but that development seems to come in three stages and two differently steep trends
- note, once one has decided to use VNC, these results are arrived at completely objectively
- thus, VNC can sometimes help rectify decisions that researchers have arrived at incorrectly …

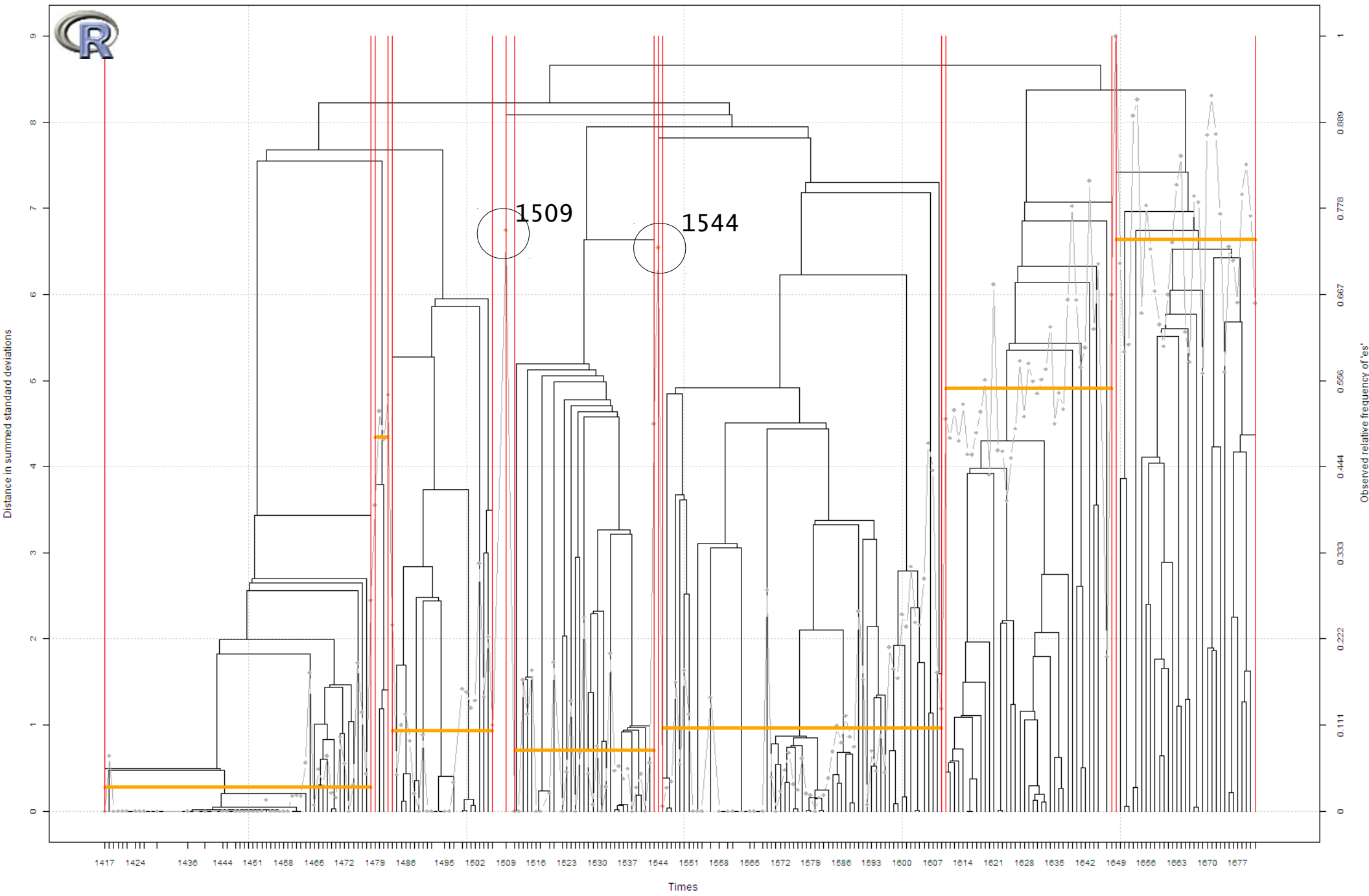
Variability-based neighbor clustering

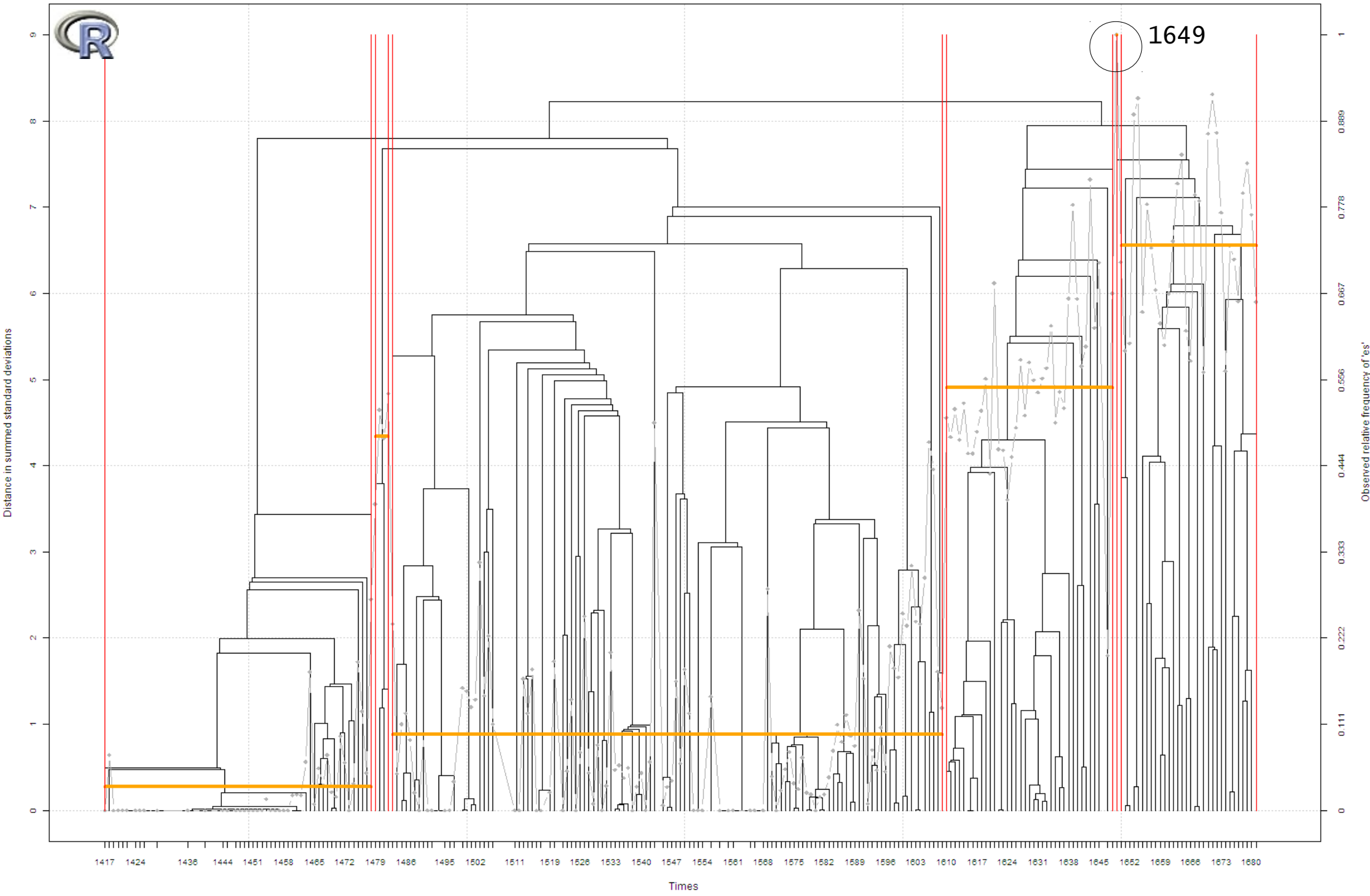


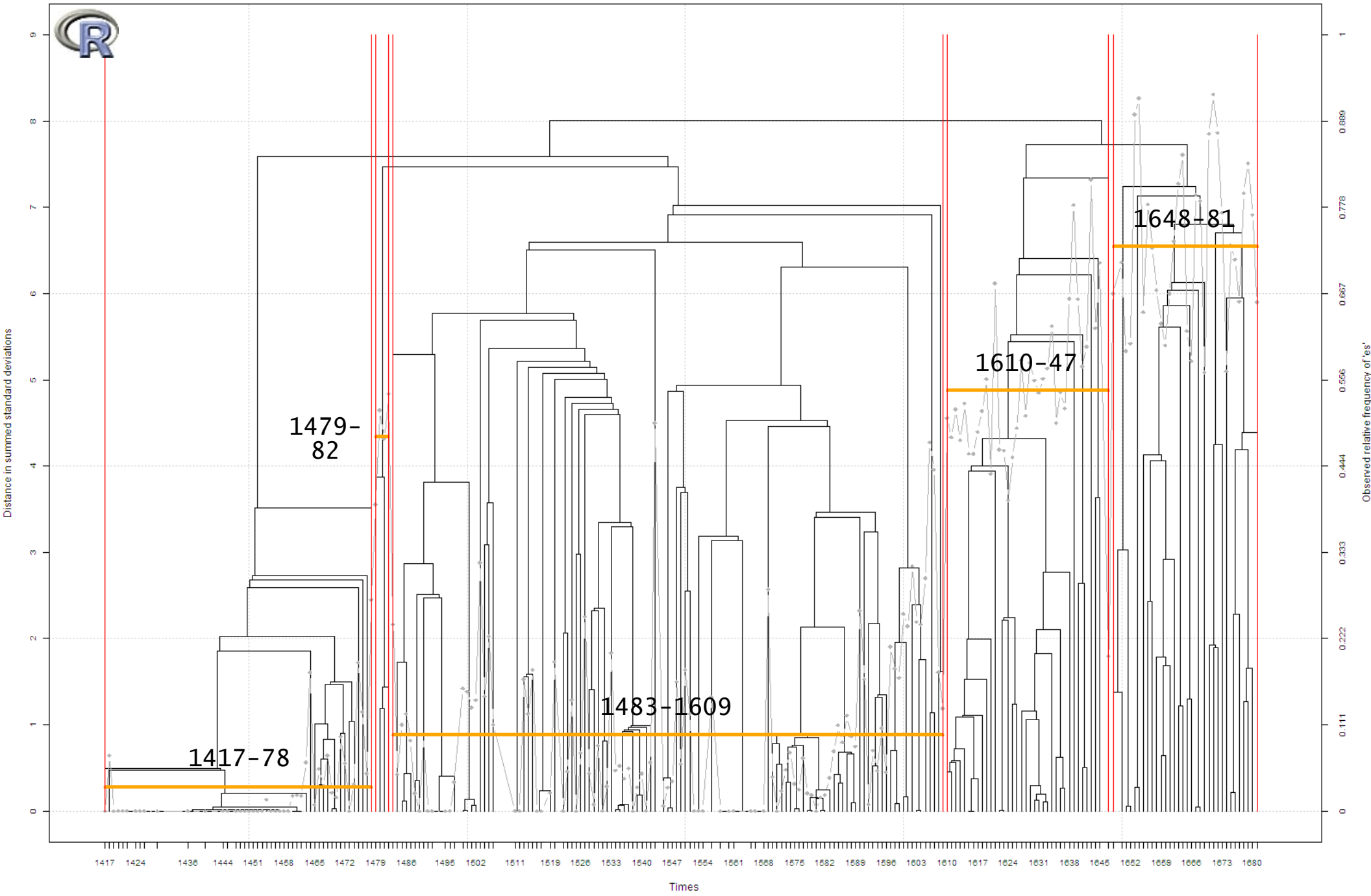
-(e)th and *-(e)s* in the CEEC

- We retrieved from the CEEC
  - ≈13,100 cases of *-(e)th*
  - ≈7,500 cases of *-(e)s*
  - in 233 time periods
- when the proportions of *-(e)s* are plotted against time,
  - there is an overall increasing trend …
  - … which is interrupted by several outliers



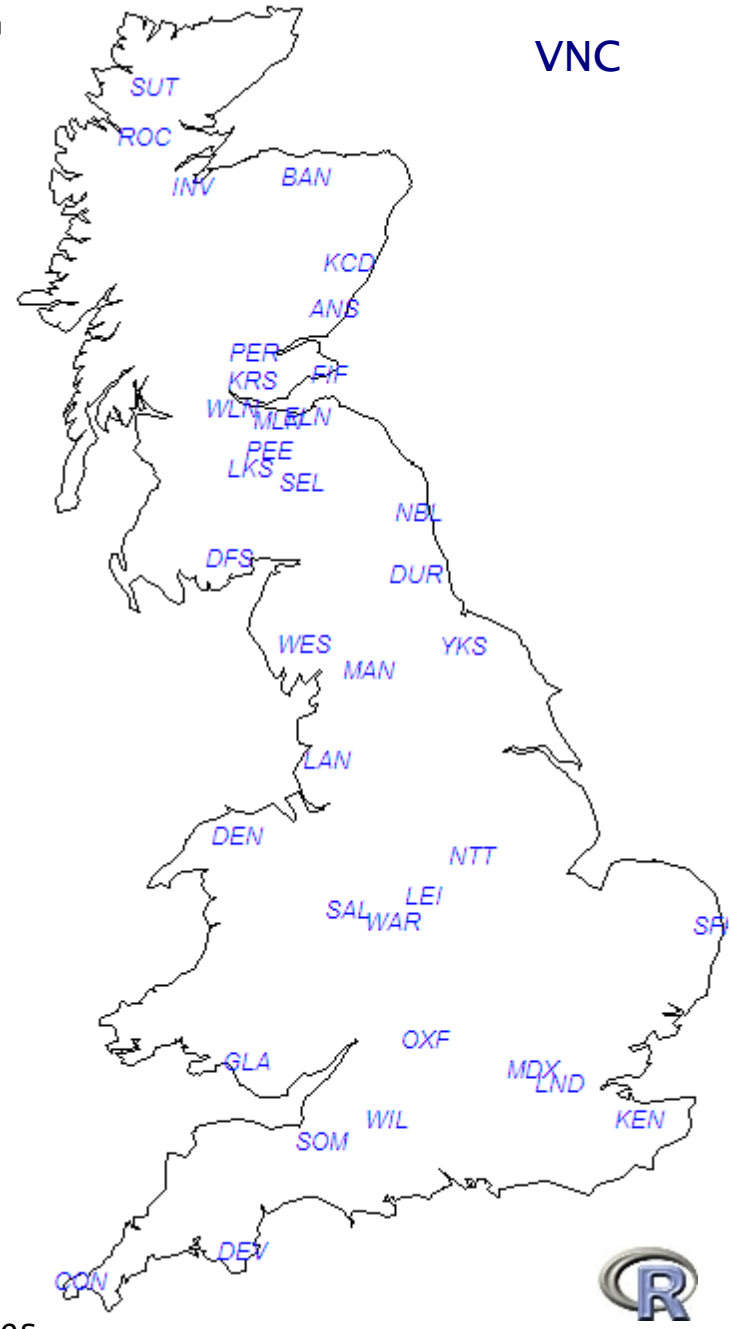






Variability-based neighbor clustering (2 dimensions)

- The application of VNC applied to very messy temporal data resulted in
  - a principled and replicable way to determine/discard outliers
  - a bottom-up way to identify temporal stages (with a degree of predictive power that exceeds that of the original data, cf. the next talk)
- this can in fact be applied to all kinds of data with internal structures
  - be they temporal
    - diachronic data
    - language acquisition/learning data (as shown above)
  - be they geographical …



Many other tools are available

- Other exploratory methods that have been used in cognitively-inspired corpus linguistics
  - multidimensional scaling (MDS)
    - a method that tries to express the similarity of entities by plotting them into a usually two-dimensional plane
  - (multiple) correspondence analysis (MCA)
    - basically a principal component analysis on frequency data
    - has been used as an alternative to Behavioral Profiles
    - for example, Glynn (2010) applies MCA to the verb *bother*
      - 650 examples from both AmE and BrE are annotated for a variety of characteristics
      - the MCA reveals semantic differences between two kinds of syntactic patterns, an agentive and a predicative *bother*
    - for example, Levshina (in progress) applies MCA to discover structure in the semantic field of seating furniture
      - she annotates different words for pieces of furniture as represented in online furniture catalogs for characte-ristics such as 'ab-/presence of armrests', 'use of upholstery', etc.

Thank you!

http://tinyurl.com/stgries