Corpus data and experimental data: examples and applications

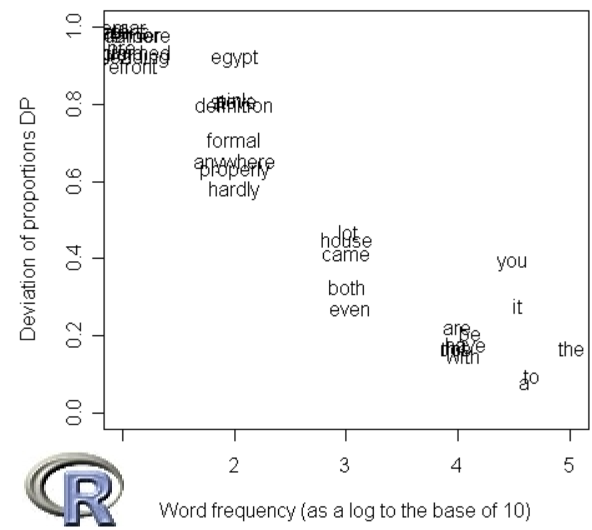
Stefan Th. Gries
Department of Linguistics
University of California, Santa Barbara
http://tinyurl.com/stgries

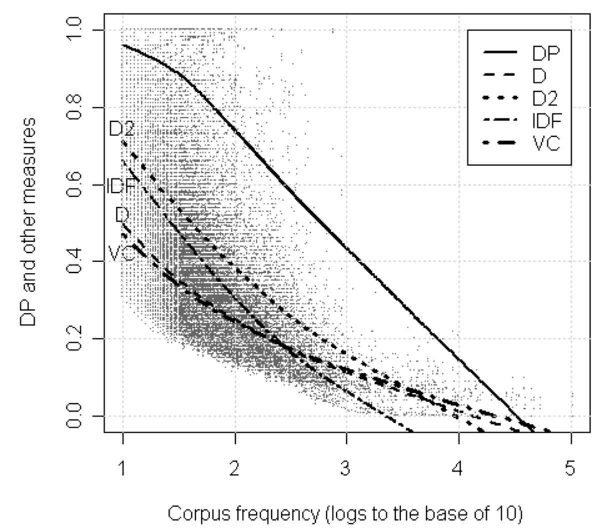
After all my telling what to do, now I will even tell you (more about) why ;-)

- During the last 9 talks, I have told you to do a lot of things
  - e.g., don't just use frequencies – add dispersion measures (or adjusted frequencies) to your data
  - e.g., don't just use probabilities of co-occurrence – use association measures (such as p_{FYE} or ΔP) instead
  - e.g., don't just use co-occurrence frequencies or isolated examples to describe the semantics of synonyms, antonyms, and polysemous items – use Behavioral Profiles
  - e.g., don't rely on introspective data to, say, predict speaker behavior – use multifactorial models instead
- I have sometimes alluded to experimental evidence for the recommended methods – in this talk, I will discuss several kinds of experimental evidence in more detail

Recap: dispersion to make frequencies more precise

- Earlier, I discussed the risks that come with frequency data that do not also take dispersion into account
- I proposed a measure ≈$0 \leq DP \leq 1$ that can serve to put frequencies into a better perspective
- *DP* has many attractive properties
  - handles differently large corpus parts
  - easy to understand: difference of %s
  - can handle frequencies of occurrence and co-occurrence
  - sensitive: does not return extreme values too quickly
    not too sensitive: does not overpenalize zeros and does not react to low expected frequencies

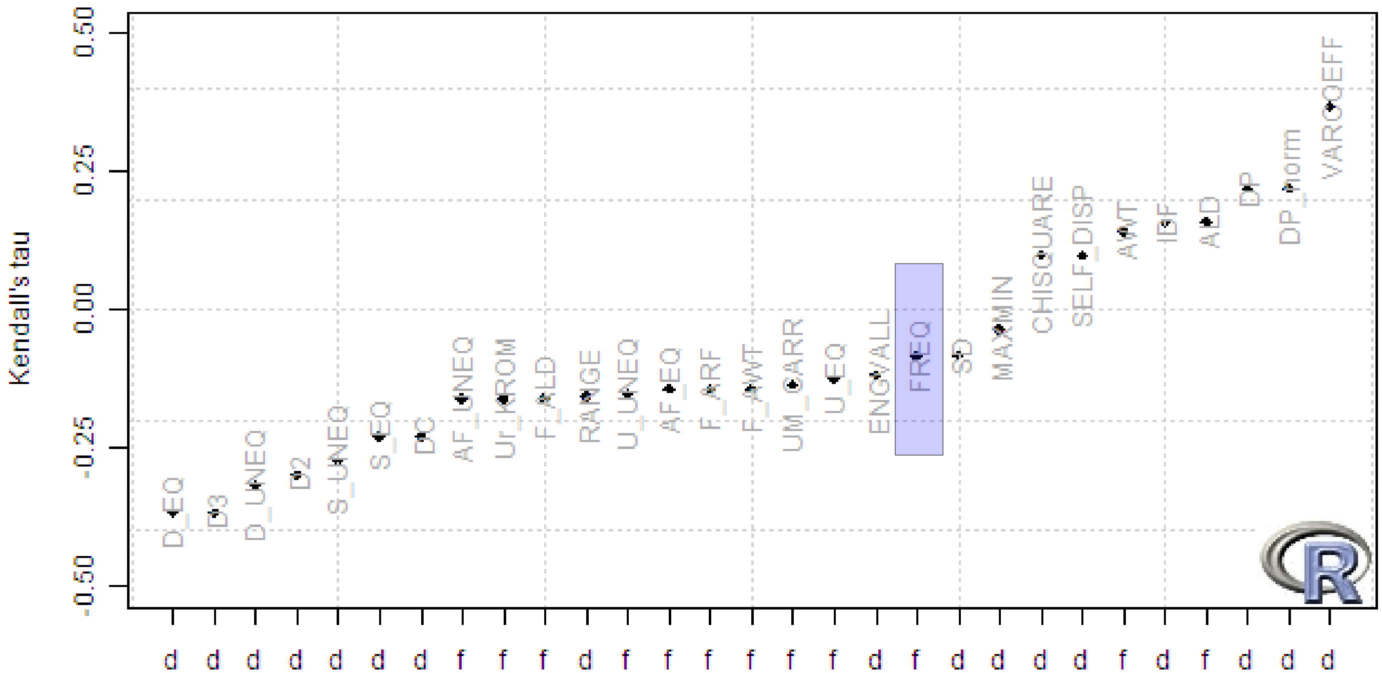




What dispersion measures buy us

- This is not just corpus-linguistic playing with numbers
  - Ellis & Simpson-Vlach (2005) and Ellis et al. (2007) show that a dispersion measure (range) has significant predictive power above and beyond raw frequency
  - Gries (2010) shows that some dispersion measures correlate more highly with
    - response time latencies from Balota & Spieler (1998) than raw frequencies
    - lexical decision task times from Baayen (2008)
- "given a certain number of exposures to a stimulus […], learning is always better when exposures or training trials are distributed over several sessions than when they are massed into one session." (Ambridge et al. 2006: 175)
- thus, there is good experimental reason to augment frequencies with dispersion measures

What dispersion measures buy us



Recap: collostructions to measure verb-constructions associations better

- Earlier, I discussed the advantages of using collostructional analysis (CA) to study the association of words to constructional slots
- I already mentioned a few studies that showed expe-rimentally that CA is often better than the use of just frequencies/probabilities of co-occurrence
  - Gries, Hampe, & Schönefeld (2005): sentence completions are predicted better by p_{FYE} than by frequency
  - Wiechmann (2008): p_{FYE} is the best unproblematic measure to predict eye-tracking data from Kennison (2001)
  - Gries, Hampe, & Schönefeld (2010): self-paced reading times are predicted better by p_{FYE} than by frequency
- but if the logic underlying CA is correct, association effects should also be observable for advanced learners

A test case with advanced learners of English

- Target of study: *to-* vs. *ing*-complementation
  - *People began to make strenuous efforts*
    People began making strenuous efforts
- this alternation
  - is often tricky for learners (because of the overall semantic similarity but occasional differences)
    - *Sheila tried to bribe the jailor*
      Sheila tried bribing the jailor
    - *I remembered to fill out the form*
      I remembered filling out the form
  - is characterized by strong lexical associations
  - has not been studied much from an SLA perspective
- sequence of methods
  - corpus analysis of *to* vs. *ing* based on the ICE-GB
  - questionnaire experiment that combines
    - an acceptability judgment task
    - a sentence completion task

Methodology

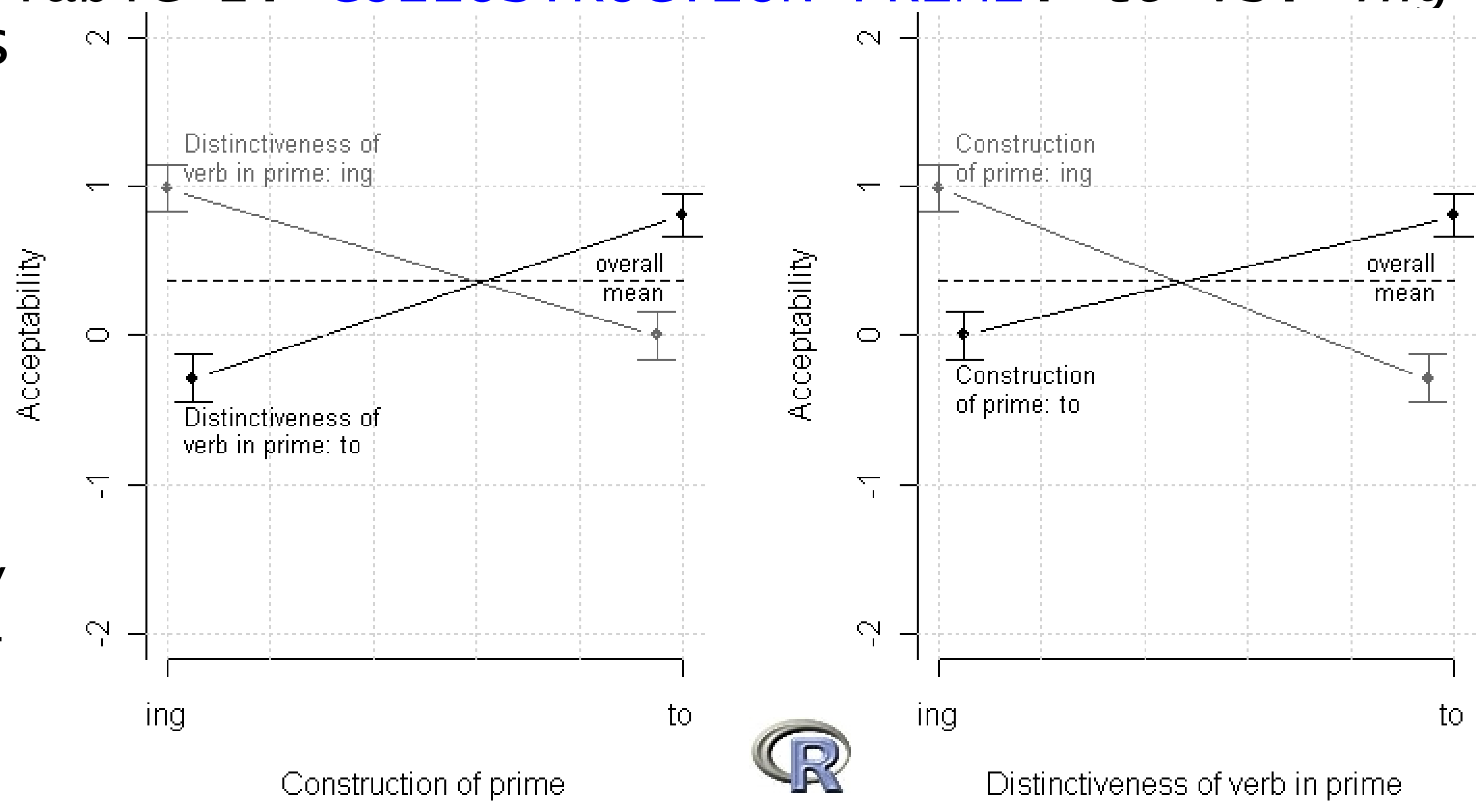
- Corpus analysis with distinctive collexeme analysis
  - verbs associated with *to*:
    want (55.67), *try* (22.44), *wish* (5.39), *manage* (4.77), *seek* (4.35), *tend* (4.06), *intend* (3.67), *attempt* (3.19), *hope* (3.19) *fail* (3.09), *like* (3.03), *refuse* (2.98), …
  - verbs associated with *ing*:
    keep (76.45), *start* (35.23), *stop* (29.45), *avoid* (11.87), *end* (11.87), *enjoy* (11.87), *mind* (11.87), *remember* (10.14), *go* (7.99), *consider* (5.45), …
- experiment

  PRIME Sally tried to open the door. **RATING** ____
  TARGET John started ________________.

  - 12 experimental items (6 completions + 6 ratings) + 24 filler items
  - acceptability judgments on a scale from -3 to +3

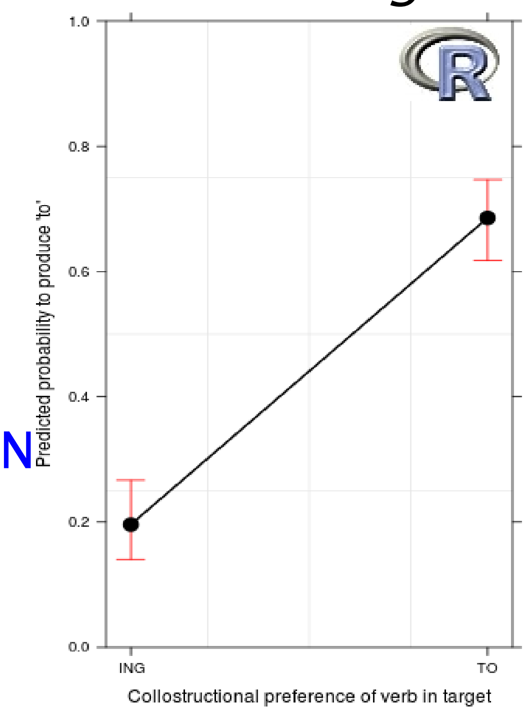
Results from the acceptability judgments

- We obtained 556 ratings from 94 subjects and analysed the data with a linear model
  - dependent variable: RATING
  - independent variable 1: CONSTRUCTION PRIME: *to* vs. *ing*
  - independent variable 2: COLLOSTRUCTION PRIME: *to* vs. *ing*
- overall results
  - the model is significant: $F_{3,\ 552}$=15.15, p<0.001
  - the effect is very weak: R^2=0.07
  - but the interaction strongly confirms collostructions



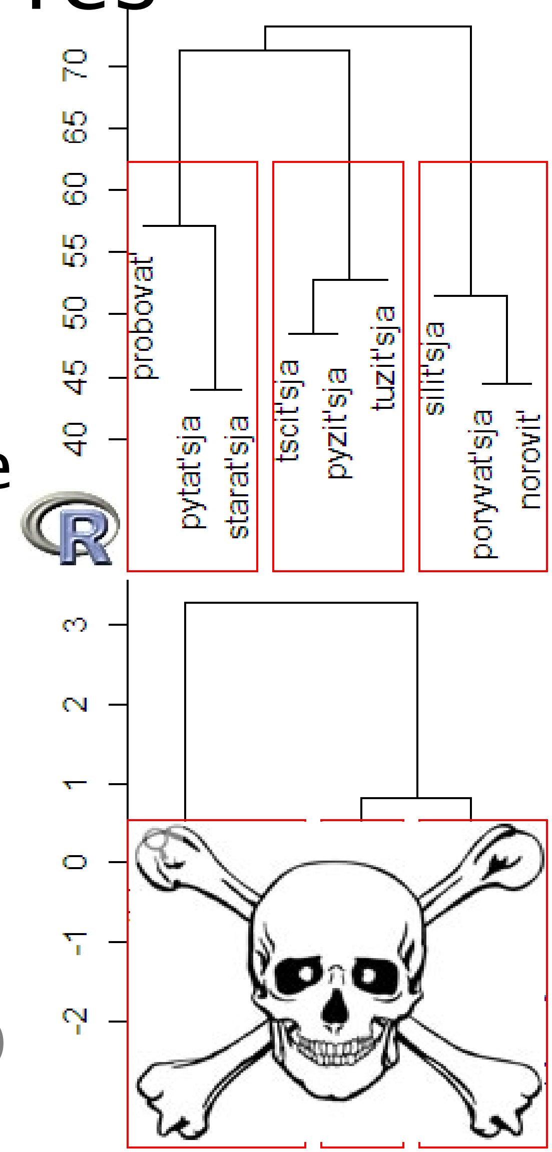
One result from the sentence completions

- We obtained 560 completions from 94 subjects, in-cluding 176 *to*- + 193 *ing*-constructions, and analysed them with a logistic regression
  - dependent variable: COMPLETION: *to* vs. *ing*
  - indep. variable 1: CONSTRUCTION PRIME: *to* vs. *ing*
  - indep. variable 2: COLLOSTRUCTION PRIME: *to* vs. *ing*
  - indep. variable 3: COLLOSTRUCTION FRAGMENT: *to* vs. *ing*
  - …
- overall results
  - the model is significant: LL X^2=128.46, *df*=5, *p*<0.0001
  - the predictive accuracy is good: R^2=0.39, *C*=0.82
  - the collostructional preference of the verb in the target fragment (COLLOSTRUCTION PRIME) is the strongest predictor (OR=9) and supports collostructions



Recap: Behavioral profiles

- Earlier, I discussed the advantages of Behavioral Profiling
- I already mentioned that the cluster analyses and post-hoc analyses of BP were quite revealing and versatile
- the question of course now is, is there any independent, not to say converging evidence, to support the clusters and make them more than correlations in corpus data?
- after all, a cluster analysis will always generate some tree whatever nonsense it is fed …
- some (experimental) validation is indispensable (cf. Divjak & Gries 2008)



An experimental validation of BP using a sorting task

- Students from a Moscow CompSci and Econ Dept were given instructions to sort 9 sentences that only differed with regard to the verb meaning 'to try'
  - into *n* groups of similar sentences
  - into 3 groups of similar sentences
  - into 3 groups of 3 similar sentences each
- but how do we evaluate such data?
- how do we compare this with a cluster diagram?
- two approaches
  - with a newly developed evaluation metric
  - with a comparison of dendrograms

    (I will focus only on the first sorting task, the results for all others are virtually identical)

The evaluation metric (theory)

- Step 1: generate a co-classification matrix that states for each verb how often it was put into one group with every other verb
- step 2: compute the Pearson residuals for every cell in the table to identify deviations
  - (obs-exp)/sqrt(exp)
- step 3: mark the highest Pearson residuals in every row
  - if a target verb's highest Pearson residual was observed for a verb from the same cluster (in the corpus analysis), score 1 point
  - otherwise, score 0 points

The evaluation metric (practice)

- Step 1

	noro	*pory*	*sil*	*prob*	*pyt*	*star*	*pyz*	*tschi*	*tuz*
noro		*a*	*b*	*c*	*d*	…	…	…	…
pory	*a*		*f*	*g*	*h*	…	…	…	…
sil	*b*	*f*		*k*	*l*	…	…	…	…
…	…	…	…		…	…	…	…	…

- step 2

	noro	*pory*	*sil*	*prob*	*pyt*	*star*	*pyz*	*tschi*	*tuz*
noro		5.7	-2.27	-1.5	-2.12	-2.18	-2.56	-0.75	-2.63
pory	5.7		-3.22	-1.45	-1	-0.54	-3.04	-1.59	-3.36
sil	-2.27	-3.22		-1.67	-2.25	-1.84	1.73	0.15	2.74
prob	-1.5	-1.45	-1.67		3.77	1.32	-2.93	-2.9	-3
pyt	-2.12	-1	-2.25	3.77		3.22	-3.26	-2.97	-3.32
star	-2.18	-0.54	-1.84	1.32	3.22		-2.32	-2.73	-2.64
pyz	-2.56	-3.04	1.73	-2.93	-3.26	-2.32		0.19	4.39
tschi	-0.75	-1.59	-0.15	-2.9	-2.97	-2.73	0.19		0.36
tuz	-2.63	-3.36	2.74	-3	-3.32	-2.64	4.39	0.36	

- step 3: 8 points …
- … but what kind of a result is this? there is not immediately available expected distribution → step 4

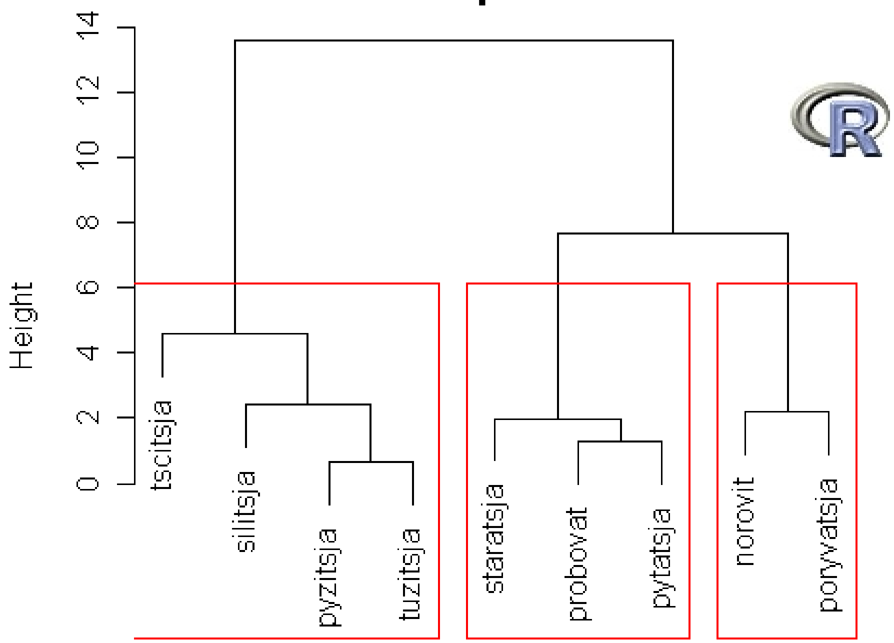
The evaluation metric: inference

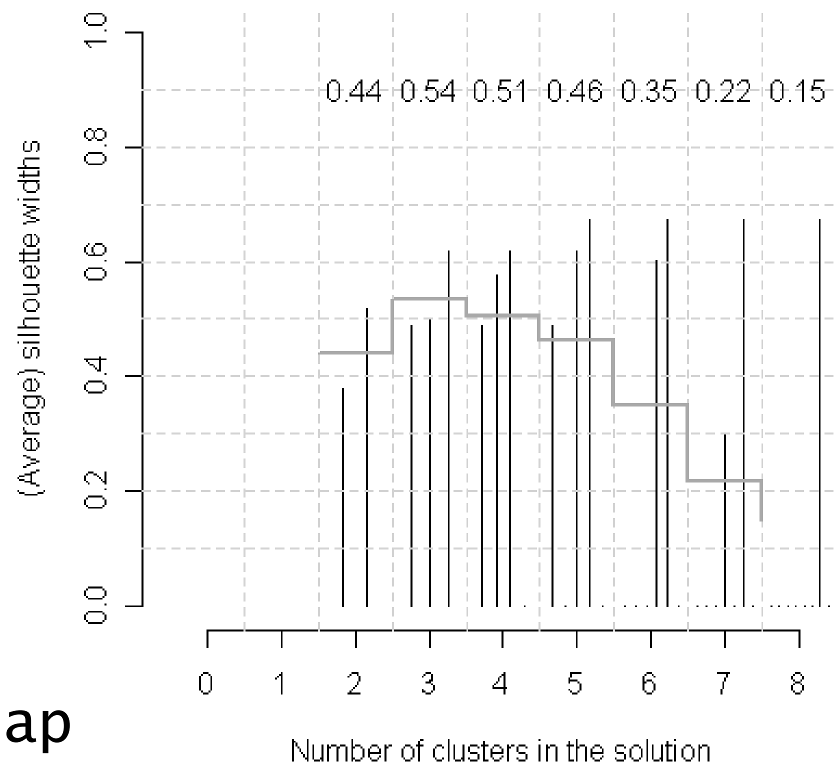
- Step 4
  - the minimal obtainable value is 0
  - the maximal obtainable value is 9
  - the expected score is 2.25 (9 Vs scoring ¼ on average)
  - Monte Carlo simulation: we generated a vector with all possible scores {1,1,0,0,0,0,0,0} and sampled one value from it with replacement 9 times and added the values up
  - we did that 100,000 times
  - we counted how often we obtained our sample result of 8 as a sum or even more
  - 12 out of 100,000 times, i.e. p=0.00012
  - quantiles of the simulation data

Quantile	0.005	0.010	0.025	0.050	0.500	0.950	0.975	0.990	0.999
Σ	0	0	0	0	2	4	5	6	6

Comparison of dendrograms

- We computed a cluster analysis on the sorting data (with the same parameters as for the corpus data)





Fowlkes & Mallows (1983)
B_k=0.74 (0≤B_k≤1): good overlap

- both kinds of analyzing the sorting data result in a clear and significant confirmation of the corpus-based BP cluster analysis

Recap: multifactorial models are indispensable

- Earlier, I discussed how multifactorial modeling is often the most useful approach to study data (esp. if those data are complex)
- however, with the exception of some newer develop-ments (NDL or Bayesian networks), the math under-lying regression models is hardly cognitively realistic
- thus, it would be good if there was a way to determine whether what they predict
  - does not just have a good classification accuracy when it comes to the corpus data from which the model was derived
  - but also predicts experimental behavior
- we have seen some examples above with regard to verb-construction associations – the following will consider prototypicality of construction exemplars

The design of the corpus part of the study

- Target of study: the dative alternation in English
  - *John gave Mary the book* ditransitive
    John gave the book to Mary prepositional dative
- the dative alternation is affected by a large number of interconnected factors
- Gries (2003) coded
  - whether the VP denotes transfer
  - animacy of patient and recipient
  - NP type of patient and recipient
  - definiteness of patient and recipient
  - length of patient and recipient
  - times of preceding mention of patient and recipient
  - distance to last mention of patient and recipient
- two main questions (at the time)
  - can the constructional choice be predicted?
  - can prototypical instances of the two constructions be identified?

Findings from the corpus analysis

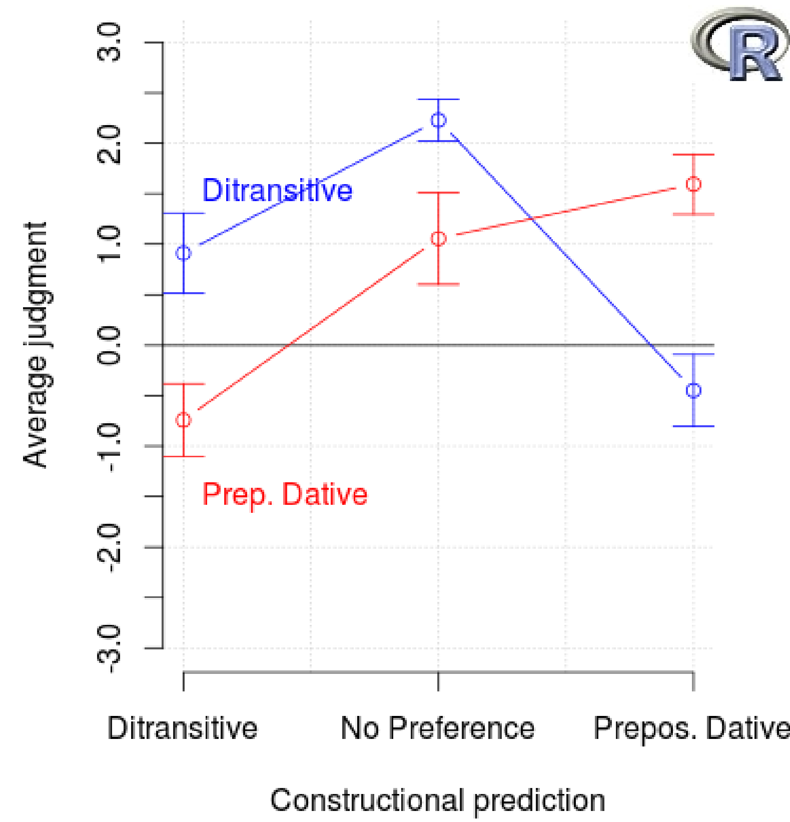
- A linear discriminant analysis shows the constructional choices can be predicted well
  - the model is significant: X^2=112.12, *df*=30, p<0.001
  - canonical *R*=0.821, classification accuracy=88.9%
- how does the model predict constructional choices?
- it uses a discriminant score
  - if that score > 0, the model predicts ditr
  - if that score < 0, the model predicts prep
  - the further away the score of a sentences is from 0, …
  - the more that sentence has the characteristics typical for one construction, …
  - and the more certain is the prediction
- prototypes for
  - ditr.: *going round beer festivals gave me the idea …*
  - prep.: *[X, Y, and Z] gave a new impetus both to the study of these themes and to action upon them*

Follow-up acceptability judgment experiment

- From this, it follows that the sentences with the most extreme scores should embody the prototypes, and speakers should strongly disprefer these sentences in the opposite construction
- experimental design
  - independent variable 1: PREDICTION: I picked
    - 2 sentences predicted to be highly typical of ditr
    - 2 sentences predicted to be highly typical of prep
    - 2 sentences predicted to accept both constructions
  - independent variable 2: CONSTRUCTION: each sentence was provided in its original construction or the opposite
  - dependent variable: JUDGMENT (ranging from -3 to +3)
  - 36 native speakers of English
  - plus the usual experimental controls
- prediction
  - the speakers should like stimuli when they are presented in the structure that the corpus analysis predicted to be preferred

Results of the experiment

- The result of a linear model is quite clear
  - the model is significant: $F_{5,\ 173}$=12.22, p<0.0001
  - the effect is intermediately strong: adj. R^2=0.24
  - the predicted interaction is
    - the strongest effect
    - exactly as predicted
      - when the corpus model predicts ditr, then
        - ditr is liked
        - prep is not
      - when the corpus model predicts prep, then
        - prep is liked
        - ditr is not
      - when the corpus model predicts both, both are liked
- the multifactorial corpus model receives very strong support



To sum up

- For many of the tools or methodological proposals made in the course of this week, supportive experimental evidence has been presented
- ideally, we would always try to seek this type of converging evidence
  - from experiments for corpus data
  - from corpus data for experiments
  - with different methodologies and data sets within each of these two types of data
- this is a lot of work and not without its own problems (cf. Arppe et al. 2011), but it ensures replicable progress with regard to our analysis of (hopefully) falsifiable hypotheses
- and that in turn is the only guarantee that cognitive linguistics will evolve further as a truly empirical and interdisciplinary science

Thank you!

http://tinyurl.com/stgries