Corpus-based cognitive semantics: behavioral profiles for polysemy, synonymy, and antonymy

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

Some questions (cognitive) semanticists face

- Polysemy
  - how to determine whether two usage events are identical or sufficiently similar to be considered a single sense
  - how to determine the degree of similarity of different senses
  - how to determine how/where to connect a sense to others in the network
  - how to determine the prototypicality of (a) sense(s)
- (near) synonymy
  - in the above, replace *sense* with *word*
  - what are the differences (in meaning/construal/…) between near synonyms
  - what is the functional relation between near synonyms in a semantic domain

Avenues of research pursued by (cognitive) semanticists 1

- Approaches not using empirical data
  - e.g., Lakoff and others' full-specification approach: minimal perceived differences between usage events constitute different senses
- partially empirical approaches
  - e.g., Tyler & Evans' (2001) principled polysemy approach
    - empirically testable distributional assumptions
    - additional meaning components and distributional features
    - lexical choices regarding patterns of modification and/or complementation
- some problems with these approaches
  - what is the ontological status of the proposed networks?
  - not all fine-grained sense distinctions are supported by data
  - frequent use of artificial/decontextualized examples

Avenues of research pursued by (cognitive) semanticists 2

- Empirical approaches
  - Schmid (1993): corpus data on *begin* and *start* in the LOB corpus
  - Sandra & Rice's (1995) / Rice's (1996): sorting, judgments, sentence generation
  - Raukko (1999, 2003): sentence generation, paraphrasing, prototypicality judgments
  - Kishner & Gibbs (1996), Gibbs & Matlock (2001)
    - collocate analysis (R1)
    - colligations / syntactic patterns
    - correlating senses and syntactic patterns
    - "our findings suggest the need to incorporate information about […] lexico-grammatical constructions in drawing links between different senses of a polysemous word"
  - e.g., Glynn and QLVL group with their correspondence analysis approach

Corpus-based approaches to lexical semantics and the main assumption

- The domain of linguistics that has probably been studied most with corpora is lexical semantics
- the main assumption underlying all this work is that distributional characteristics of an item reveal many of its semantic and functional properties and purposes (and for once, let's not quote Firth here)
  - Harris (1970:785f.): "[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."
  - Bolinger (1968:127): "a difference in syntactic form always spells a difference in meaning"
  - Cruse (1986:1): "the semantic properties of a lexical item are fully reflected in appropriate aspects of the relations it contracts with actual and potential contexts"

Corpus-based information in lexical semantics

- This kind of logic has been applied especially fruitfully in the domain of synonymy, where contextual information of two kinds has been particularly useful and revealing
  - collocational information: what are the words that are modified by
    - *strong* and *powerful* (Church et al. 1991)
    - *absolutely*, *completely*, and *entirely* (Partington 1998)
    - *big*, *large*, and *great* (Biber et al. 1998)
    - \w+*ic* and \w+*ical* (Gries 2003)
  - syntactic information: what are the preferred grammatical associations of
    - *quake* and *quiver* (Atkins & Levin 1995)
    - *little* vs. *small* or *begin* vs. *start* (Biber et al. 1998)
    - causative *get* and *have* (Gilquin 2003)
    - several Finnish verbs meaning 'think' (Arppe & Järvikivi 2007 and Arppe 2008)

Some fine-grained analytical approaches in corpus semantics

- Atkins (1987)
  - collocation in a L7-R7 window
  - POS characteristics of the head word
  - ID tag: collocation/colligation correlating with a sense
- Hanks (1996)
  - collocations/colligations
  - sense triangulation: correlating colloates in different clause roles
  - behavioral profiles: the set of complementation patterns of a word → the semantics of a verb are determined by the totality of its complementation patterns
- some problems with these approaches
  - not much evidence for the predictive power of ID tags is offered
  - methodology lacks quantitative sophistication
- → Behavioral Profiles (BPs)

From this previous work …
to the present approach

- Problems of many previous approaches
  - they focus on individual pairs of synonyms and antonyms only and usually do not take larger sets of synonymous/antonymous words into consideration
  - they focus only on the base forms of the words
  - they focus only on collocational aspects or only on syntactic patterning, but do not combine lexical and syntactic distributional characteristics
    - exceptions from a cognitive-linguistic perspective: Schmid (1993), Kishner & Gibbs (1996), Gibbs & Matlock (2001)
    - exceptions from a corpus-linguistic perspective: Atkins (1987), Hanks (1996), Arppe & Järvikivi (2007)
  - they do not analyze their distributional data in the most revealing way but rather restrict themselves to observed frequencies of co-occurrence of a lexical item and a collocate or a syntactic pattern
  - they do not integrate their corpus-based findings into a theoretical account

From this previous work … to the present approach

- This approach reported on today tries to extend previous work in several respects: it
  - is geared towards allowing explorations of larger sets of synonymous and/or antonymous words or senses of polysemous words
  - not only allows but, in a sense, encourages the study of different word forms
  - includes a much larger range of distributional characteristics than just collocations and colligations
  - involves statistical analysis in various ways to get the most out of the large amount of distributional information corpus data offer

Behavioral profiling: overview

- The Behavioral Profiling approach involves four steps (cf. Gries 2003, 2010, Gries & Divjak 2010)
  - retrieval of (a representative sample of) all instances of a word's lemma from a corpus in their context (usually at least the complete utterance or sentence)
  - (so far largely) semi-manual analysis and annotation of many properties of the use of the word forms (ID tags)
    - morphological, syntactic, semantic, … characteristics (other characteristics could of course also be included)
  - co-occurrence tables indicating how often in % each ID tag level is attested with each word/sense (behavioral profile)
  - evaluation of the table by means of
    - descriptive summary statistics: counts of what is attested
    - correlational methods: pairwise distributional similarity
    - cluster analyses to identify structure(s) in the sets of words/senses explored

      (various extensions follow-up analyses are possible)

Overview of applications (with examples of extensions)

Phenomenon	Method
polysemy	
English *run*	frequencies, correlations
English *get*	clustering w/ *p*-values
(near) synonymy	
within one L1: Russian 'to try'	silh. widths, *t*-values, *z*-scores
between two L1s: phasal verbs in English & Russian	pairwise diffs w/ snake plots
between an L1 and its L2 variant: English *can*/*may*	logistic regression
(near) synonymy and antonymy	
English SIZE adjectives	pairwise diffs w/ snake plots

Example 1: the polysemy of *run* (steps 1 and 2)

- Data:
  - 815 instances of the verb lemma *run* were retrieved from the ICE-GB and the Brown corpus
- annotation (252 ID tag levels)
  - morphological ID tags
    - tense, aspect, voice
  - syntactic ID tags
    - transitivity, clause type, sentence type
  - semantic ID tags
    - semantic roles of subjects, objects, complements
    - *run*'s sense: 'fast pedestrian motion', 'manage', 'location/extension', 'function', 'flee', 'be a candidate', 'operate', …
  - lexical collocates in the same clause

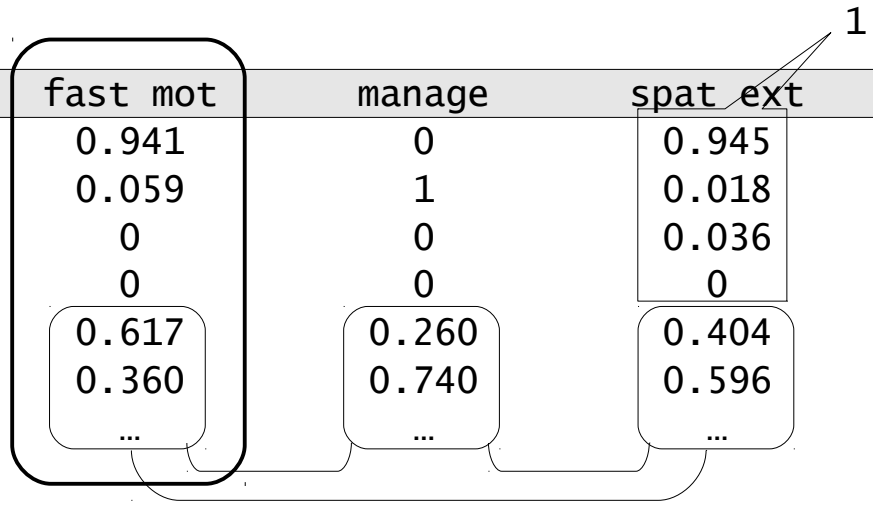
Example 1: the polysemy of *run* (steps 2 and 3)

- Step 2: the data were annotated for a variety of characteristics

Sense	VForm1	VForm2	ClType	SubjectType	ObjectType	Transitivity
fast mot	*run*	base	main	count	NA	intrans
function	*runs*	3PersSg	subord	non-count	quantity	monotrans
manage	*running*	progress	subord	count	NA	intrans
…	…	…	…	…	…	…

- Step 3: with the script BP 1.0, this table was converted into a co-occurrence table of Behavioral Profile vectors of %s

ID tag	ID tag level	fast mot	manage	spat ext	function	…
VForm2	intrans	0.941	0	0.945	0.979	…
	monotrans	0.059	1	0.018	0	…
	copula	0	0	0.036	0.021	…
	other	0	0	0	0	
ClType	main	0.617	0.260	0.404	0.408	…
	subord	0.360	0.740	0.596	0.551	…
…	…	…	…	…	…	…



Example 1: the polysemy of *run* (steps 4: evaluation)

- Frequencies of ID tag combinations: which senses to lump or split
  - in Croft (1998:169)
    - *Jack ate lunch with Jill* → 'dine' + comitative = attested
    - *Jack ate pizza with Jill* → 'consume' + comitative ≠ attested
    - the meanings of 'dine' and 'consume' are different senses
  - for *run*, this means lumping these senses
    - *and we ran back* [$_{goal}$ *to the car*]
    - *Durkin and Calhoun came running* [$_{source}$ *from the post*]
    - *I once ran* [$_{source}$ *from the Archive studio*] [$_{goal}$ *to the Start The Week studio*]
  - for *run*, this means splitting these 'escape' senses
    - *If Adelia had felt about someone as Henrietta felt about Charles, would she have run away* [$_{comitative}$ *with him*]*?* → 'to move away to engage in a romantic relation'
    - *He wanted to know if my father had beaten me or my mother had run away* [$_{source}$ *from home*] → to move away from something undesirable

Example 1: the polysemy of *run* (steps 4: evaluation)

- Correlations: where to connect senses in a network
  - do we connect the 'to escape' senses of *run*
    - to 'fast pedestrian motion'?
      - because 'fast pedestrian motion' is the prototypical sense
      - because escaping typically involves fast pedestrian motion
    - to 'motion' or 'fast motion'?
      - because then the connection involves the most general link
  - why not measure the correlations between different senses' behavioral profiles (cf. Biber 1993, McDonald 1997, Hare et al. 2003)
    - the correlations of all senses (Pearson's *r*) to each other range from 0.38 to 0.92
    - most different senses: *their cups were already running over without us* and *He ran his eye along the roof copings*
    - most similar senses: 'escape' and 'fast pedestrian motion', which should therefore be connected

Overview of applications (with examples of extensions)

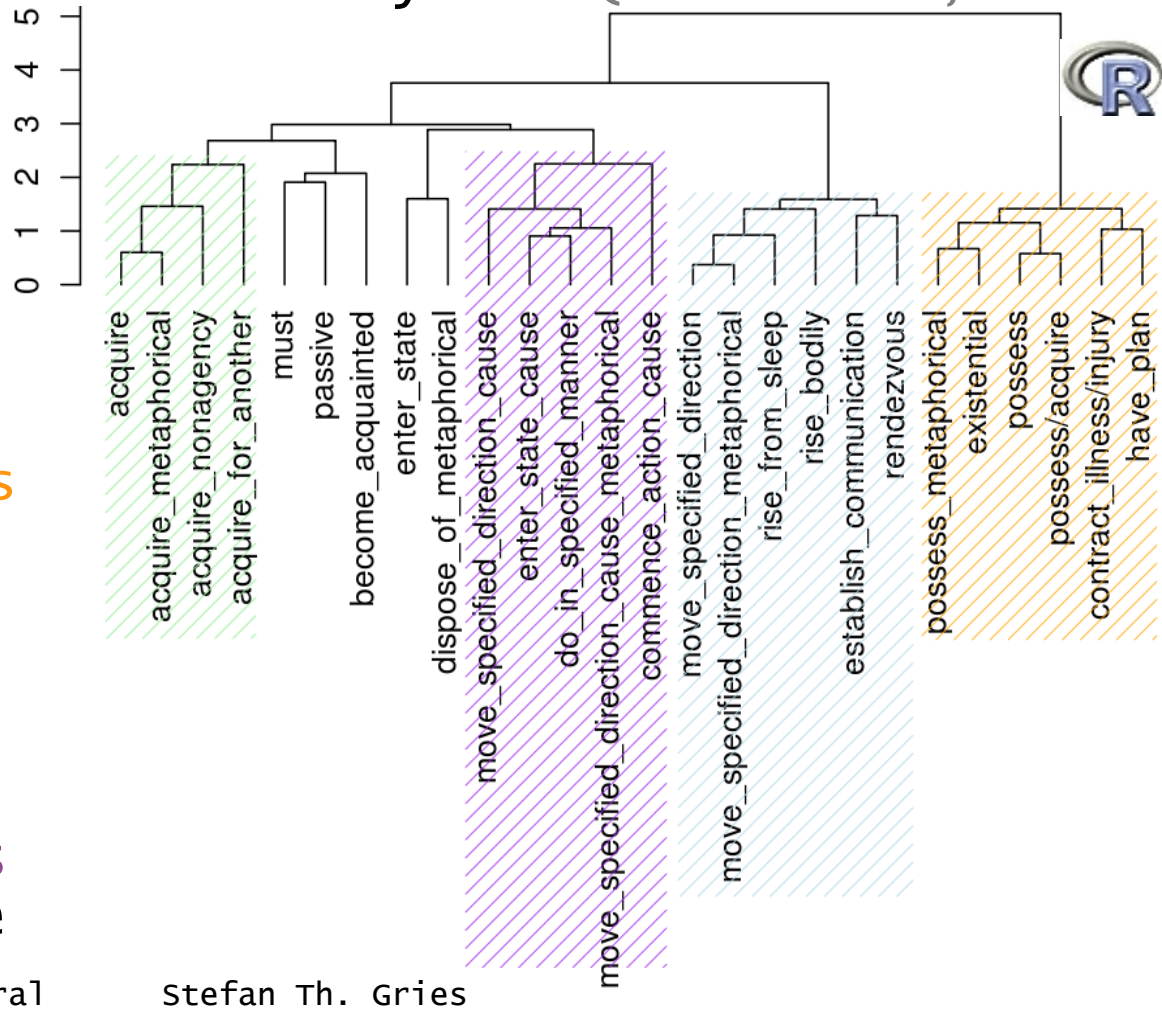
Phenomenon	Method
polysemy	
English *run*	frequencies, correlations
English *get*	clustering w/ *p*-values
(near) synonymy	
within one L1: Russian 'to try'	silh. widths, *t*-values, *z*-scores
between two L1s: phasal verbs in English & Russian	pairwise diffs w/ snake plots
between an L1 and its L2 variant: English *can*/*may*	logistic regression
(near) synonymy and antonymy	
English SIZE adjectives	pairwise diffs w/ snake plots

Example 2: the polysemy of *get* (steps 1, 2, and 3)

- Data
  - 600 instances of the verb lemma *get* were retrieved from the ICE-GB
- annotation (54 ID tag levels)
  - morphological ID tags
    - tense, aspect, voice
  - syntactic ID tags
    - transitivity, clause type, clause function,
  - semantic ID tags
    - abstractness of sense: abstract vs. concrete
    - *get*'s senses: 'acquire', 'stable possession', '(cause) movement in specified direction (concrete)', '(cause) movement in specified direction (metaph.)', '(cause) entering state', …
- with the script BP 1.0, the raw data table was converted into a co-occurrence table of BP vectors

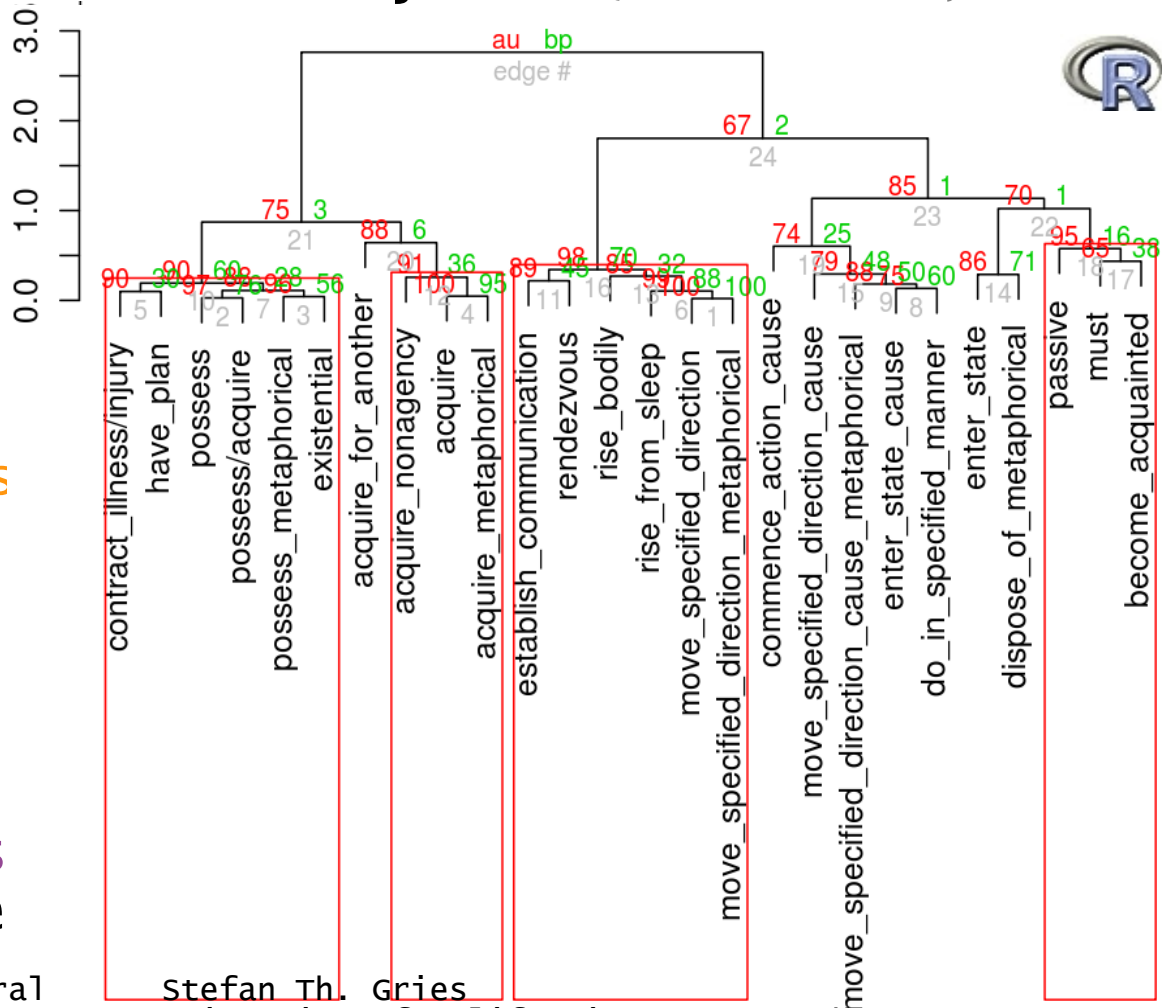
Example 2: the polysemy of *get* (step 4)

- Evaluation: 26 senses occurring 5+ times were entered into a hier. cluster analysis (Canberra, Ward)
  - several interpretable clusters emerge
    - various 'acquire' senses
    - various causative meta-phorical motion senses
    - various metaphorical motion senses
    - various possession senses
- which of the clusters reach some level of significance? → multiscale bootstrap resampling
  - clusters other than this are supported, as is the 'more grammatical one'



Example 2: the polysemy of *get* (step 4)

- Evaluation: 26 senses occurring 5+ times were entered into a hier. cluster analysis (Canberra, Ward)
  - several interpretable clusters emerge
    - various 'acquire' senses
    - various causative meta-phorical motion senses
    - various metaphorical motion senses
    - various possession senses
- which of the clusters reach some level of significance? → multiscale bootstrap resampling
  - clusters other than this are supported, as is the 'more grammatical one'



Overview of applications (with examples of extensions)

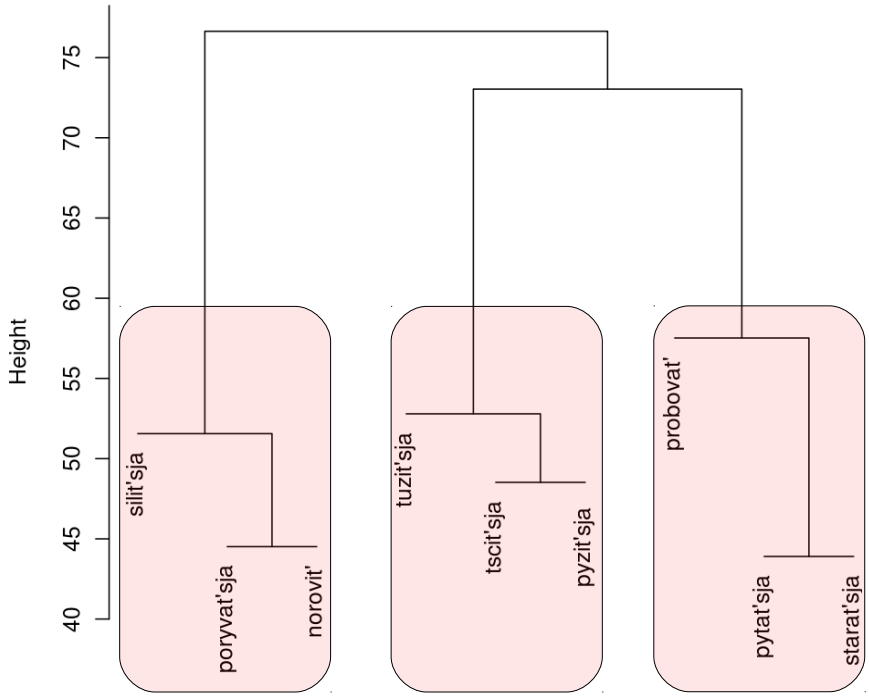
Phenomenon	Method
polysemy	
English *run*	frequencies, correlations
English *get*	clustering w/ *p*-values
(near) synonymy	
within one L1: Russian 'to try'	silh. widths, *t*-values, *z*-scores
between two L1s: phasal verbs in English & Russian	pairwise diffs w/ snake plots
between an L1 and its L2 variant: English *can*/*may*	logistic regression
(near) synonymy and antonymy	
English SIZE adjectives	pairwise diffs w/ snake plots

Example 3: Russian verbs meaning 'to try' (steps 1, 2, and 3)

- Data
  - 1585 instances of nine verbs meaning 'to try' from various sources (Amsterdam corpus, RNC, WWW)
- annotation (87 ID tag levels)
  - morphological ID tags
    - tense, aspect, mode
  - syntactic ID tags
    - clause type, sentence type, subject structure
  - syntactico-semantic ID tags
    - adverbial specification and particles: duration, repetition, intensity, futility
    - negation
    - connectors
    - semantic roles of subjects and infinitives
- with the script BP 1.0, the raw data table was converted into a co-occurrence table of BP vectors

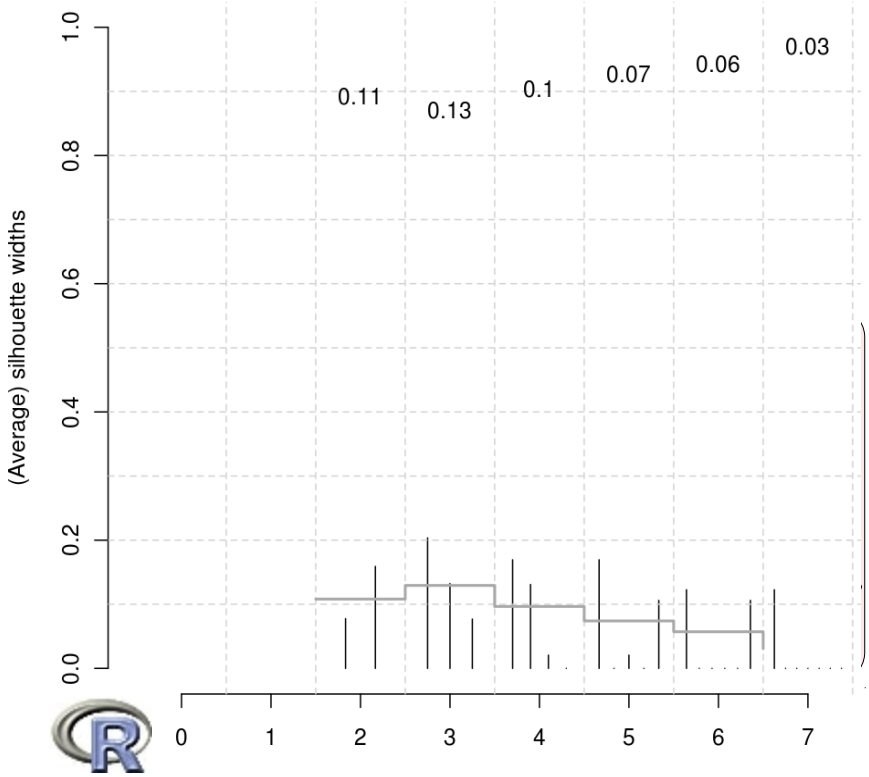
Example 3: Russian verbs meaning 'to try' (step 4, evaluation)

- Evaluation
  - the nine verbs were entered into a hierarchical cluster analysis (Canberra, Ward)
    - it seems there are 3 or 4 clusters – but how to decide?
    - silhouette widths: a statistic that compares
      - the similarity of an element to the other elements of the same cluster to
      - the similarity of an element to the most similar cluster to which it does not belong
      - on the basis of the average of all elements' silhouette widths
    - in this case, assuming 3 clusters is the best interpretation
  - but what do they represent?



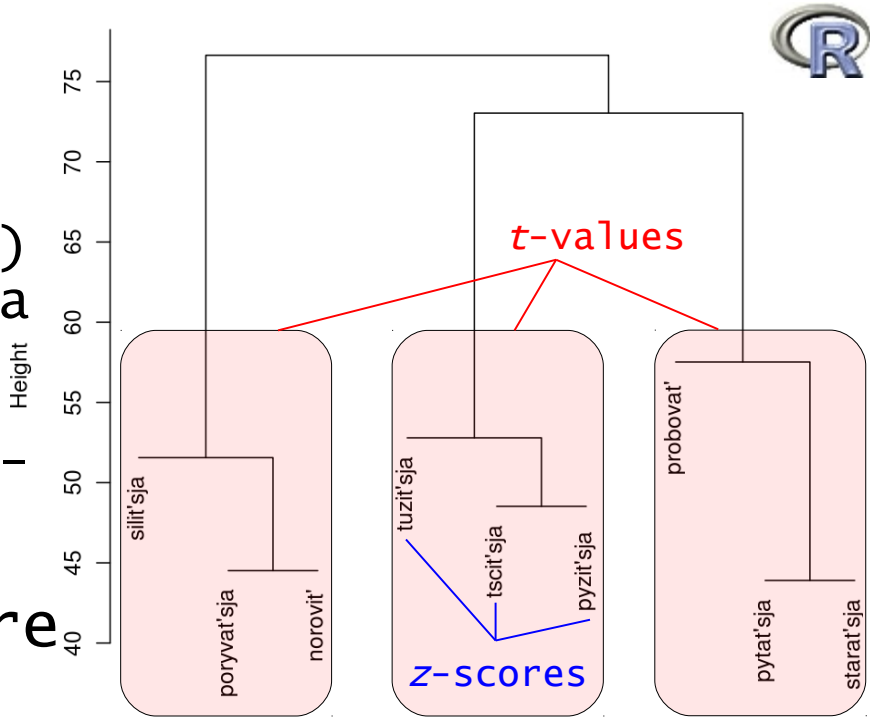
Example 3: Russian verbs meaning 'to try' (step 4, evaluation)

- Evaluation
  - the nine verbs were entered into a hierarchical cluster analysis (Canberra, Ward)
    - it seems there are 3 or 4 clusters – but how to decide?
    - silhouette widths: a statistic that compares
      - the similarity of an element to the other elements of the same cluster to
      - the similarity of an element to the most similar cluster to which it does not belong
      - on the basis of the average of all elements' silhouette widths
    - in this case, assuming 3 clusters is the best interpretation
  - but what do they represent?



Example 3: Russian verbs meaning 'to try' (step 4, interpretation)

- 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



Overview of applications (with examples of extensions)

Phenomenon	Method
polysemy	
English *run*	frequencies, correlations
English *get*	clustering w/ *p*-values
(near) synonymy	
within one L1: Russian 'to try'	silh. widths, *t*-values, *z*-scores
between two L1s: phasal verbs in English & Russian	pairwise diffs w/ snake plots
between an L1 and its L2 variant: English *can*/*may*	logistic regression
(near) synonymy and antonymy	
English SIZE adjectives	pairwise diffs w/ snake plots

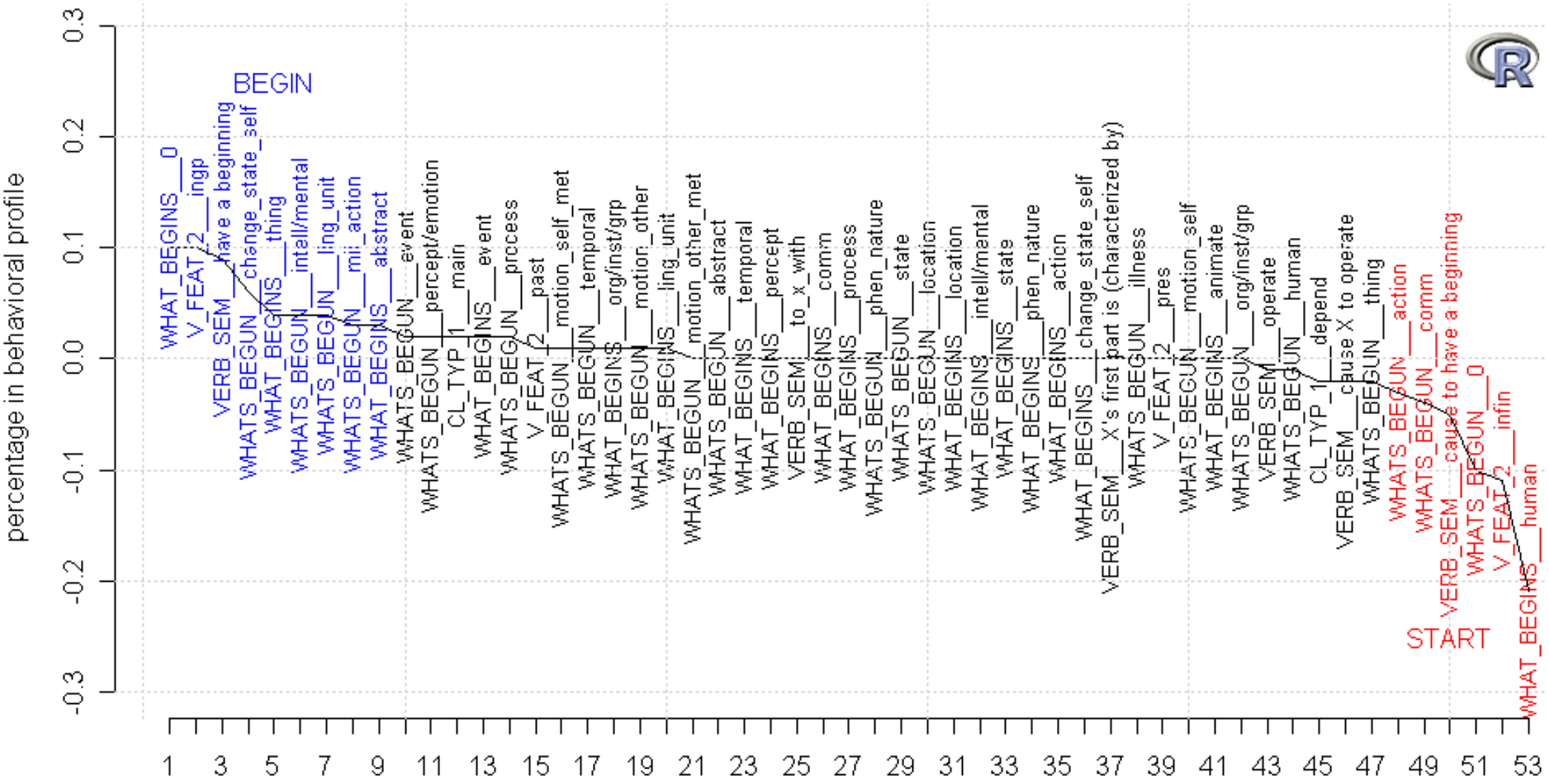
Example 4: contrastive phasal verbs (steps 1, 2, and 3)

- Data
  - 298/531 instances of the lemmas *begin/start* (ICE-GB)
  - 321, 173, and 156 instances of the lemmas *načinat'/ načat'*, *načinat'sja/ načat'sja* and *stat'* from journalese (Uppsala Corpus)
- annotation (73 ID tag levels)
  - morphological ID tags
    - tense, aspect, voice, mood, person
  - syntactic ID tags
    - complement type: N, verbal inf. (and V-*ing* for English)
    - clause type, sentence type
  - semantic ID tags
    - semantic roles of beginner and beginnee: abstract, action, human, perception/emotion, mental, temporal, …
    - sense: '(cause to) have a beginning', '(cause to) operate', 'first part characterized by'
- with the script BP 1.0, the raw data table was converted into a co-occurrence table of BP vectors

Example 4: contrastive phasal verbs (step 4a, English)

- Evaluation: since the BP vectors are percentages of co-occurrence, subtracting them from each other provides pairwise preferences for two words/senses
  - *begin*: main clauses, progressive, when nothing that is expressed or a concrete object begins to initiate a change of state of itself or something abstract (events, processes, percepts)
  - *start*: transitively in subordinate clauses, when a human instigator causes an action (particularly communicative actions) or, less so, causes a concrete object to operate ('prototype': *he started the bike*)

Example 4: contrastive phasal verbs (step 4a, English)



Example 4: contrastive phasal verbs (step 4b, Russian)

- Evaluation: since the BP vectors are percentages of co-occurrence, subtracting them from each other provides pairwise preferences for two words/senses
  - *načinat'/ načat'*: imperfective, non-finite or present tense, expresses situations with clear source and clear starting point, typically abstract concepts or changes of states (instigated by unknown entities or nature)
  - *stat'*: perfective, indicative and past tense, general actions or communicative activities instigated by humans

Example 4: contrastive phasal verbs (step 4c, Russian)

- Evaluation: since the BP vectors are percentages of co-occurrence, subtracting them from each other provides pairwise preferences for two words/senses
  - *načinat'sja/ načat'sja*: expresses no explicit beginner and only 'have a beginning' (of processes, events, and other temporals), typically in main clauses
  - *načinat'/ načat'*: present tense in dependent clauses, wide range of senses, actions and changes of states instigated by humans and groups/institutions

A brief cross-linguistic comparison

- In a cluster analysis of all verbs, the languages are neatly separated, but …
- across languages, *begin* is most similar to *načinat'/načat'* and *start* is somewhat similar to *stat'*
  - *načinat'/načat'/begin* prefer zero and more abstract beginners
  - *start/stat'* prefer past tense and similar beginnees (actions, communications, mental activities)
  - *begin/stat'* highlight the view into the state after the onset of the action
- the prototypes for each (set of) verb(s) revolve around different sets of characteristics
  - 12 of the 15 most distinctive ID tags for *begin/start* involve beginners & beginnees (*begin*'s abstract processes and *start*'s concrete actions by humans)
  - for the Russian verbs, most distinctive ID tags involve lexical preferences & aspectual properties of the verbs
  .

Overview of applications (with examples of extensions)

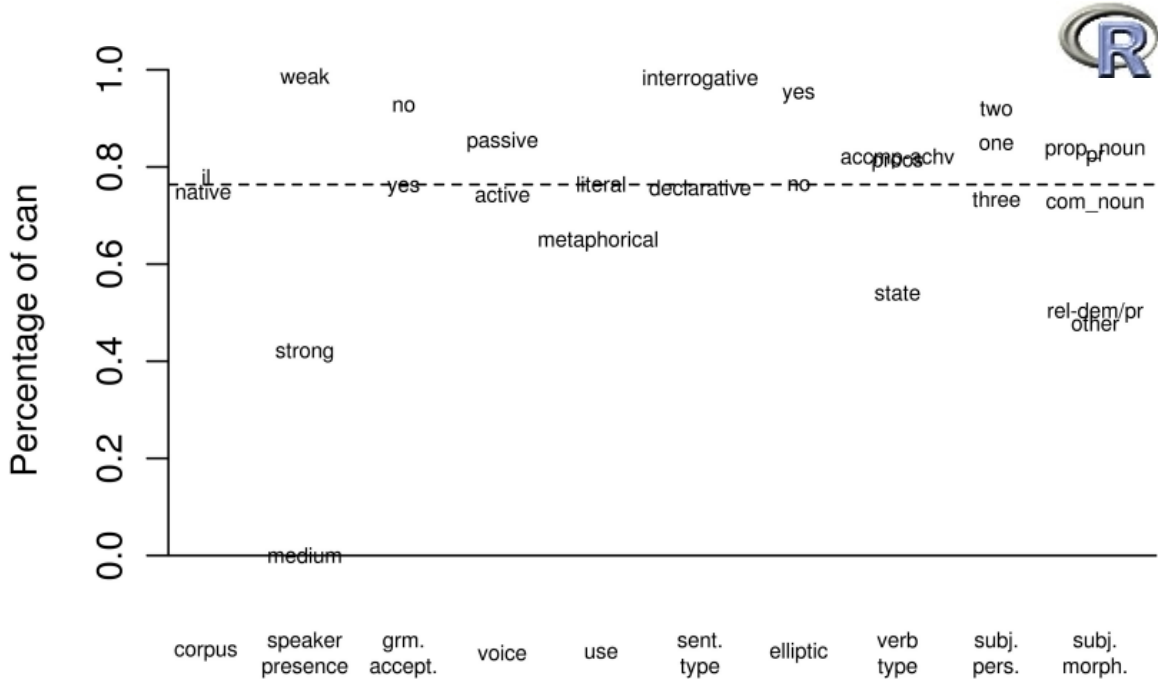
Phenomenon	Method
polysemy	
English *run*	frequencies, correlations
English *get*	clustering w/ *p*-values
(near) synonymy	
within one L1: Russian 'to try'	silh. widths, *t*-values, *z*-scores
between two L1s: phasal verbs in English & Russian	pairwise diffs w/ snake plots
between an L1 and its L2 variant: English *can/may*	logistic regression
(near) synonymy and antonymy	
English SIZE adjectives	pairwise diffs w/ snake plots

Example 5: *can/may/pouvoir* in L1 and L2 (steps 1 and 2)

- Data
  - 1322 and 466 instances of *can* and *may* from LOCNESS (L1)
  - 1290 and 366 instances of *can* and *may* from ICLE-FR (L2)
  - 265 instances of *pouvoir* from CODIF (L1)
- annotation (22 ID tags)
  - morphological ID tags
    - verb form, subject person, subject number, subject type, voice, aspect, mood
  - syntactic ID tags
    - sentence type, clause type, negation
  - semantic ID tags
    - modal senses, modal use, speaker presence, verb semantics, subject referent animacy (type)
  - corpus: native English, learner English, native French
  - acceptability: yes vs. no
- with the script BP 1.0, the raw data table was converted into a co-occurrence table of BP vectors

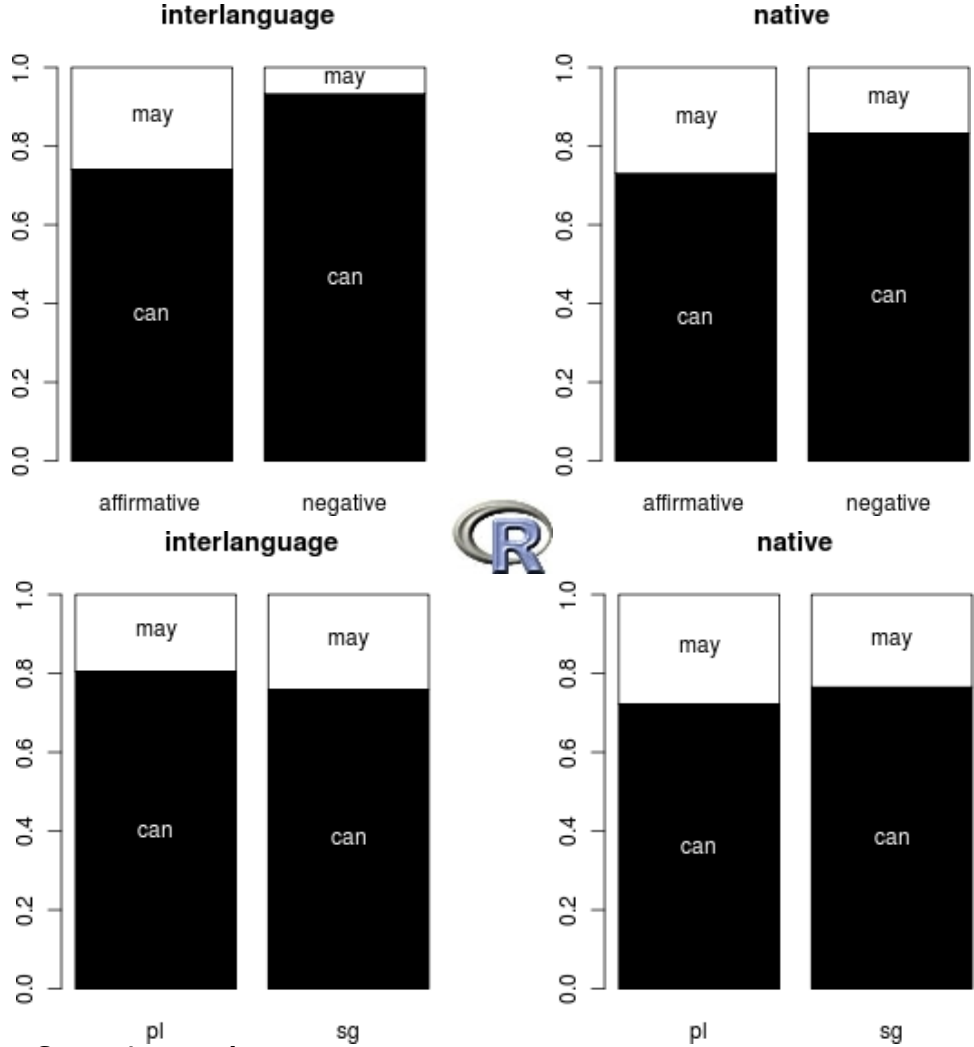
Example 5: *can/may/pouvoir* in L1 and L2 (step 4)

- The data on *can* and *may* were entered into a logistic regression with a subsequent model selection process
  - dependent variable: *can*. vs. *may*
  - independent variables: all above annotation columns plus their interactions with CORPUS; this allowed us to test whether some ID tags operate differently in L1 and L2
- result (after ns predictors were eliminated)
  - highly significant correlation: R^2=0.955; p<0.001; classification accuracy: 99%
  - several main effects were significant, but …
  - there were also several significant interactions



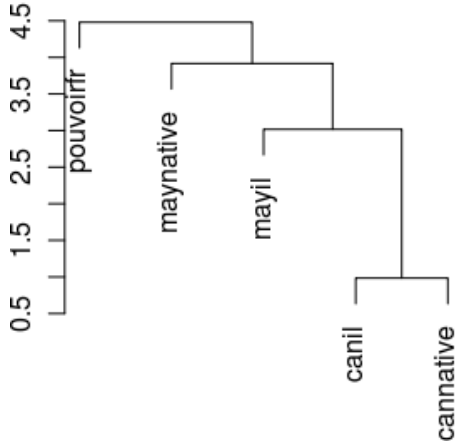
Example 5: *can/may/pouvoir* in L1 and L2 (step 4)

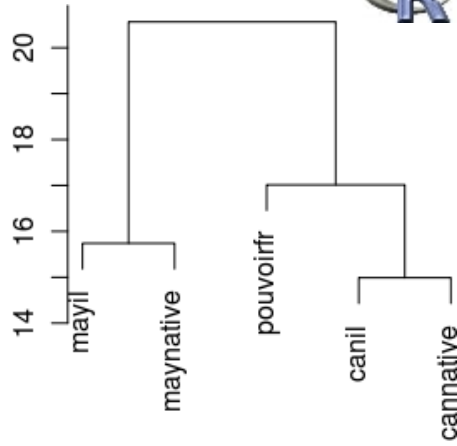
- The discussion of two interactions must suffice here
  - CORPUS × NEGATION
    - all speakers prefer *can* in negated clauses, but L2 speakers do so more strongly
    - one way to explain this: the complexity principle: in complex environments, speakers make more explicit/default choices
  - CORPUS × SUBJ-NUMBER
    - native speakers use *can* more often with singular subjects, learners behave the other way round
    - the complexity principle would again be compatible with that finding



A brief cross-linguistic comparison

- In a clustering of all five modals based on all ID tags, the expected groupings were found, but when the ID tags are split up, the results change
  - syntactically, *can* and *may* stick together, but semantically, *pouvoir* is closer to *can* than *may* is
  - how exactly and why is that?
    - in complex environments, the learners resort to the more frequent 'default' verb
    - learners overuse *can* with animate subjects and underuse it with time/place verb semantics





- both the clustering and the regression results indicate straightforwardly where the largest and most significant results between L1 and L2 use are

Overview of applications (with examples of extensions)

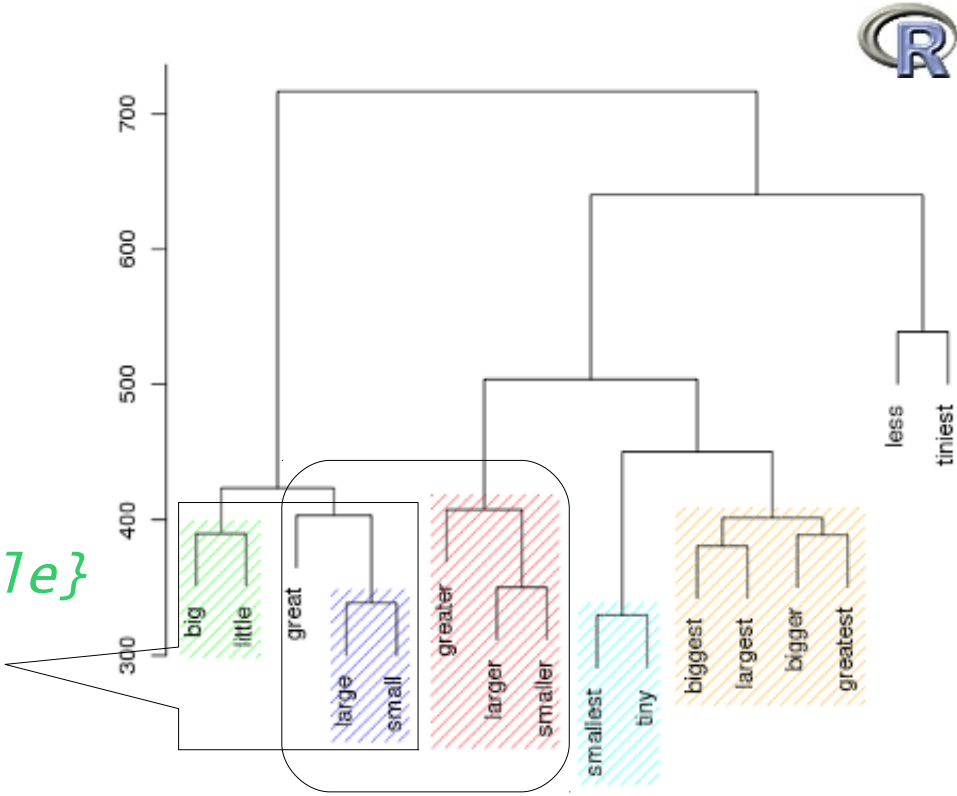
Phenomenon	Method
polysemy	
English *run*	frequencies, correlations
English *get*	clustering w/ *p*-values
(near) synonymy	
within one L1: Russian 'to try'	silh. widths, *t*-values, *z*-scores
between two L1s: phasal verbs in English & Russian	pairwise diffs w/ snake plots
between an L1 and its L2 variant: English *can*/*may*	logistic regression
(near) synonymy and antonymy	
English SIZE adjectives	pairwise diffs w/ snake plots

Example 6: English size adjectives (steps 1, 2, and 3)

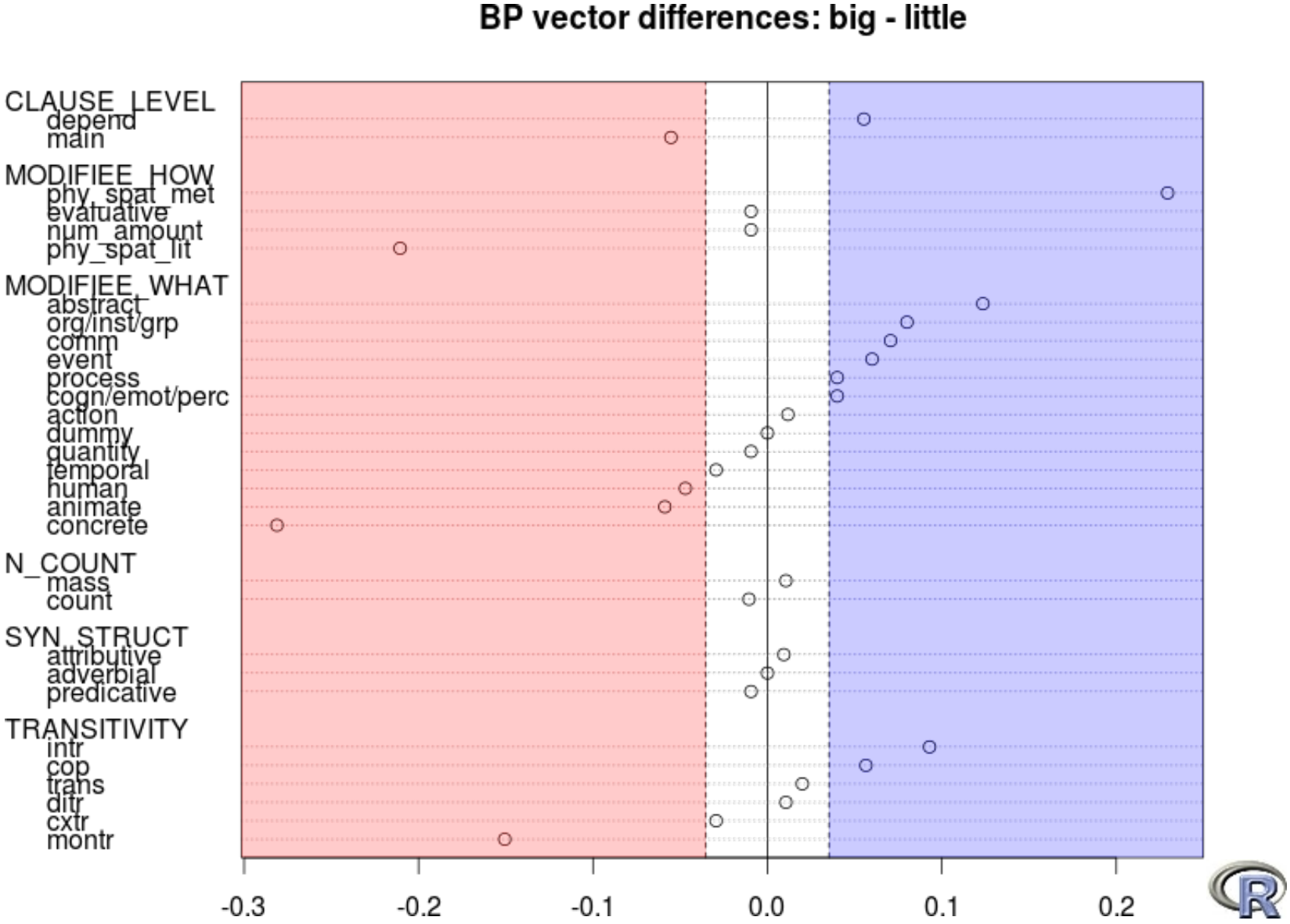
- Data
  - we retrieved 362/409/609 instances of *big/large/great* plus 250/409/34 instances of *little/small/tiny* (plus comparatives and superlatives) from the ICE-GB
- annotation (539 ID tag levels)
  - morphological ID tags
    - tense, voice
  - syntactic ID tags
    - attributive/predicative use, transitivity of main verb, clause type, clause function
  - semantic ID tags
    - countability, animacy, abstractness, and semantic type of the modified noun plus how the noun is modified (literally vs. metaphorically vs. quantitatively vs. evaluatively)
- with the script BP 1.0, the raw data table was converted into a co-occurrence table of BP vectors

Example 6: English size adjectives (step 4)

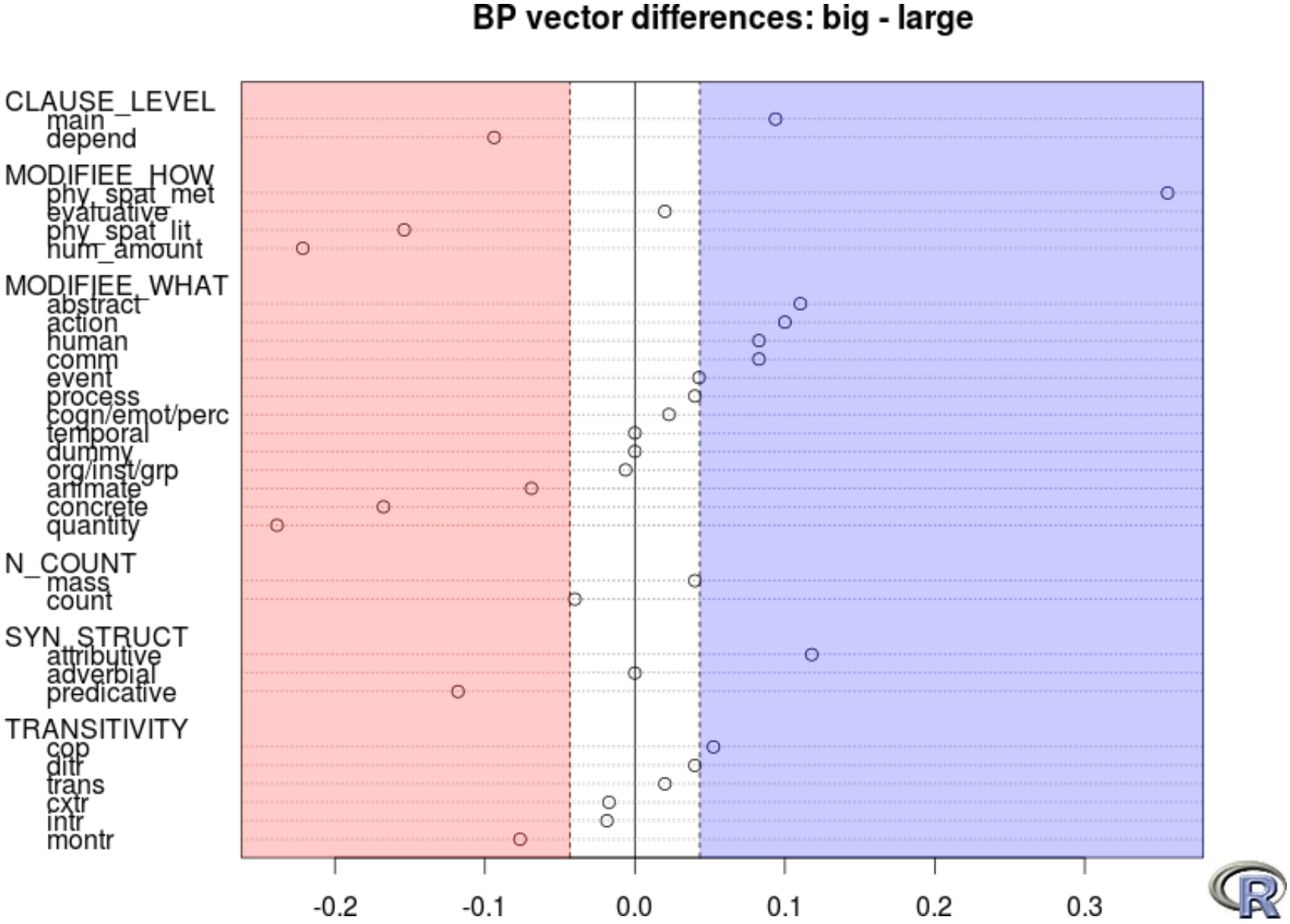
- The adjective forms were entered into a hier. cluster analysis (Canberra, Ward)
- several parameters give rise to the structure of the tree
  - oppositeness of meaning
  - sameness of meaning
  - morphological form
- starting from the bottom
  - synonymy: *smallest = tiny*
  - (canonical!) antonymy: *{big little}* *{large small}*
  - morphological form: the leftmost cluster contains only base forms
  - (canonical!) antonymy, synonymy, and morphological form: *{{larger smaller} greater}*
  - synonymy and morphological form: *{biggest largest bigger greatest}*



Revised plot: *little* vs. *big*



Revised plot: *large* vs. *big*



Post hoc comparisons: summary

- Comparing *little* vs. *big*
  - *little*
    - used literally/physically
    - concrete things, in particular humans and other animates
    - attributively
  - *big*
    - used metaphorically
    - abstract things, events, or organizations/institutions
    - predicatively
- comparing *large* vs. *big*
  - *large*
    - count nouns
    - quantities, but also organizations/institutions and animates (not humans)
  - *big*
    - non-count nouns
    - abstract nouns, but also humans and actions

Interim summary

- The corpus-based Behavioral Profile approach alone achieves what several different studies have shown
  - *smallest = tiny* (cf. Deese's 1964 rating study)
  - *big* and *little* as well as *large* and *small* are canonical antonyms (cf. most previous studies, but also Jones et al. 2007, which did not associate *big* and *little* well)
  - morphologically clean clusters reflect subjects' preference to respond to a stimulus with a morphologically identical form (cf. Ervin-Tripp 1970)
- these findings were largely obtained with all ID tags as well as just the syntactic or semantic ones

How and why does that work, and where does that leave corpus linguistics?

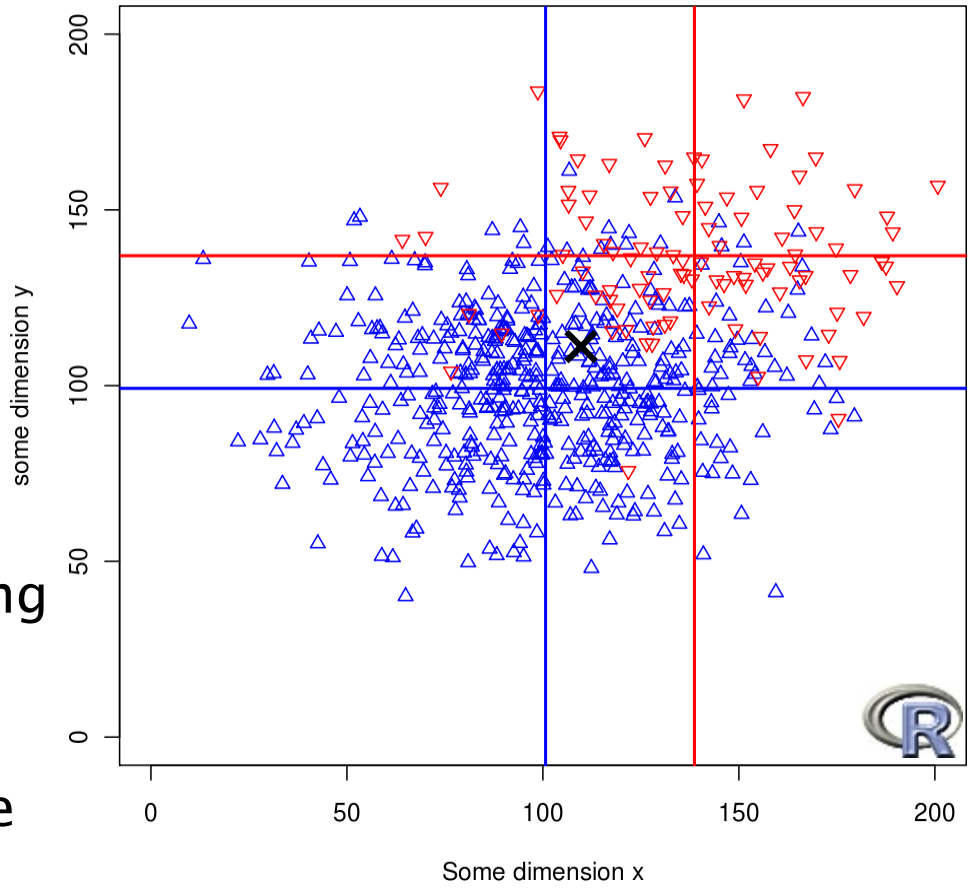
- Apparently, this approach does something, works, but
  - why does it work?
  - into what larger theoretical context can this approach be embedded?
  - and what does this say about corpus linguistics?
- the first two questions can and should be answered together – the answer to the third one follows from that
- in a nutshell
  - 1 and 2: it works because the BP approach taps into frequency information that's at the heart of contemporary exemplar-based / usage-based models
  - 3: this supports the view that corpus linguistics is not a theory on its own, but a method that is intimately connected, and contributing, to linguistic theory in general and cognitive/psycholinguistic theories in particular

Exemplar-based / usage-based models: main assumptions

- What are the main assumptions of such models?
  - each time a speaker processes a particular token/exemplar E, (aspects of) E is/are 'placed' in a multidimensional space/network
    - Pierrehumbert (2003:185): phonemes, for instance, are "associated with a distribution of memory traces in a parametric space, in this case a cognitive representation of the parametric phonetic space"
    - Halliday (2005:67): "each instance redefines the system, however infinitesimally, maintaining its present state or shifting its probabilities in one direction or the other"
  - such distributional characteristics of E involve
    - phonetic, phonological, prosodic characteristics of E
    - morphological and syntactic characteristics of E
    - semantics and discourse-pragmatic characteristics of E
    - sociolinguistic characteristics of E
    - co-occurrence information of all aspects of E
      - linguistic aspects
      - extra-linguistic aspects (e.g., utterance contexts)

Exemplar-based / usage-based models: learning, memory, and categorization

- What does learning, memory, and categorization look like in such an approach?
  - if E is close enough in multidimensional space to a cloud of already memorized exemplars (i.e., sufficiently similar to a category), then
    - E will be 'added' into the multidimensional space according to its characteristics
    - E will thereby strengthen the category formed by the already memorized exemplars to a degree proportional to
      - its similarity to the cloud of already memorized exemplars
      - the homogeneity of the cloud of already memorized exemplars



Exemplar-based / usage-based models: learning, memory, and categorization

- What does learning, memory, and categorization look like in such an approach?
  - that is, speakers have very rich memory representations of events, but …
  - speakers do not remember each exemplar and everything about it: (aspects of) memories of individual exemplars may not be accessible because they
    - were not noticed
    - decay over time
    - may be subject to generalization/abstraction as well as reconstruction (Ellis 2002:153; Langacker 2009)
  - note: fallible memory in fact implicitly facilitates the identification of typical contexts (Ellis 2002:153: "abstraction is an automatic consequence of aggregate activation of high-frequency exemplars, with regression toward central tendencies as numbers of highly similar exemplars increase.")

Does this relate to corpus linguistics at all?

- Hoey (2005:11)
  - "the mind has a mental concordance of every word it has encountered, a concordance that has been richly glossed for social, physical, discoursal, generic and interpersonal context […] all kinds of patterns, including collocational patterns"
- Miller & Charles (1991) on contextual representations
  - "knowledge of how that word is used"
  - "some abstraction or generalisation derived from the contexts that have been encountered"
  - "a mental representation of the contexts in which the word occurs, a representation that includes all of the syntactic, semantic, pragmatic, and stylistic information required to use the word appropriately."
  - "Similarly, the contextual representation of a word is not an actual linguistic context but an abstraction of information in the set of natural linguistic contexts in which a word occurs.

Does this relate to corpus linguistics at all?

- Does this relate to corpus linguistics? very much so
  - Miller & Charles on how associations between words arise: "a consequence of frequently perceiving and using these words together in the same syntactic structures"
  - so, BP can be useful because
    - it involves relative frequency vectors
    - coupled with a discriminant analysis or logistic regression, it involves very fine-grained information on co-occurrence frequencies
  - note
    - the latter approach is more precise than the former BP approach proper because it involves data on a case-by-case basis and co-occurrence frequencies
    - the former approach is better-suited for more coarse-grained studies that don't require individual case data or data with many levels of the dependent word/sense variable and/or fewer data points

What I wanted to do

- I hope I have been able to
  - explain and exemplify a particular corpus-linguistic method: Behavioral Profiles
  - show how widely applicable this method is
    - polysemy, synonymy, antonymy
    - within a language, between languages (L1s and L2s)
  - show how extendable this method is by adding various ways of follow-up exploration
  - hint at how well this method is supported experimentally
  - discuss why it works as well as it does: it is highly compatible with several recent theoretical developments in cognitive and psycholinguistics …
    - … which happen to be just as compatible with corpus linguistics as a methodology in general
    - … which underscore corpus linguistics's affinity to empirical social sciences

Thank you!

http://tinyurl.com/stgries