Corpus linguistics, cognitive linguistics, and psycholinguistics: on their combination and fit

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

Corpus linguistics and linguistic theory: some disagreements

- The relation between corpus linguistics and linguistic theory has traditionally been somewhat problematic, at least from the corpus-linguistic perspective
  - on the one hand, corpus linguists differ with regard to what they think corpus linguistics is
    - a tool, method(ology), methodological approach, discipline, theory, paradigm, framework
  - on the other hand, there are some things that may make corpus linguistics appear less attractive to the (casual or tempted) observer from theoretical linguistics

What is corpus linguistics anyway?

- Some corpus linguists say it's a theory
  - "computer corpus linguistics defines not just a newly emerging methodology for studying language, but a new research enterprise, and in fact a new philosophical approach to the subject" (Leech 1992:106)
  - "a corpus is not merely a tool of linguistic analysis but an important concept in linguistic theory" (Stubbs 1993:2f.)
  - a "pre-application methodology" which possesses "theoretical status" (Tognini-Bonelli 2001:1)
  - "an approach to the description of English with its own theoretical framework", employing the term "corpus theoretical approach" (Mahlberg 2005:2)
  - "a theoretical approach to the study of language" (Teubert 2005:2)

What is corpus linguistics anyway?

- Some corpus linguists say it's a method(ology)
  - McEnery and Wilson (1996)
  - Meyer (2002)
  - Bowker and Pearson (2002)
  - "corpus linguistics is a whole system of methods and principles of how to apply corpora in language studies and teaching/learning, it certainly has a theoretical status. Yet theoretical status is not theory in itself" (McEnery, Xiao, & Tono 2006:7f.)
  - "As a corpus linguist I consider myself primarily a methodologist and CL primarily a methodology, to be applied to whatever theory seems most appropriate for the task at hand" (Hardie 2008)

what is corpus linguistics anyway?

- Some corpus linguists have yet other labels
  - discipline (Aarts 2002, Teubert 2005, Williams 2006)
  - "[c]orpus linguistics is not in itself a method: many different methods are used in processing and analysing corpus data. It is rather an insistence on working only with real language data taken from the discourse in a principled way and compiled into a corpus" (= methodological commitment) (Teubert 2005:4)

It's a theory vs. it isn't ↔ corpus-driven vs. corpus-based

- Whether scholars attribute the status of theory to corpus linguistics or not coincides, to some degree, with where these scholars are on the continuum of
  - corpus-driven linguistics
    - aims to build theory from scratch
    - completely free from pre-corpus theoretical premises
    - "While corpus linguistics may make use of the categories of traditional linguistics, it does not take them for granted. It is the discourse itself, and not a language-external taxonomy of linguistic entities, which will have to provide the categories and classifications that are needed to answer a given research question." (Teubert 2005:3)
    - "Corpus linguists still don't know what a morpheme, a word, a phrase or a pattern is." (Teubert 2009)
  - corpus-based linguistics
    - approaches corpus data from the perspective of moderate corpus-external premises, with the aim of testing and improving such theories
    - uses annotation

My own take on these perspectives

- I
  - consider corpus linguistics "a major methodological paradigm in applied and theoretical linguistics." (Gries 2006:191)
  - agree with Teubert's methodological commitment
- why?
  - a theory whose name is a source of data?
  - a corpus-driven perspective as a theory

A theory whose name is a source of data?

- Aarts is reported as commenting that the term was coined with some hesitation "because we thought (and I still think) that it was not a very good name: it is an odd discipline that is called by the name of its major research tool and data source."
- put differently, I don't accord corpus linguistics the status of a theory just as I don't think there is a linguistic theory called *experimental linguistics* or *self-paced reading time linguistics*
  - even though self-paced reading times may yield results that call into question units/structures/processes assumed in the kind of formal linguistics that (some of) corpus linguistics was a reaction against

Corpus-driven linguistics – does that even exist?

- With maybe very few exceptions, I have yet to see what I consider a truly corpus-driven approach
  - with regard to lexis/grammar
    - a truly corpus-driven approach would require a complete distributional analysis of the corpus to first identify the linguistic units that are manifested in the data
    - Teubert (2009): "Corpus linguists still don't know what a morpheme, a word, a phrase or a pattern is." – but many 'corpus-driven' studies start out from words (e.g., Bill Louw's concordance of *all sorts of*) and traditional (pre-corpus) parts of speech are common
    - "a corpus-driven grammar is not one that is theory-free" (Halliday 2003/2005:174)
    - "applying intuitions when classifying concordances may simply be an implicit annotation process, which unconsciously makes use of preconceived theory", and this implicit annotation is "to all intents and purposes unrecoverable and thus more unreliable than explicit annotation" (Xiao 2009:995)

Corpus-driven linguistics – does that even exist?

- With maybe very few exceptions, I have yet to see what I consider a truly corpus-driven approach
  - with regard to lexis/grammar
    - many corpus-driven studies look at *n*-grams, where *n* is arbitrarily defined as one number (*n*=4 is en vogue), but
      - most do not check whether that number is indeed the best number for all the *n*-grams (one of few exceptions: Biber 2009 checks for 5-grams)
      - most do not check whether it would not indeed be better to have different *n*'s for different *n*-grams
        - *of course*: 2, *in spite of*: 3, *on the one hand*: 4, *as a matter of fact*: 5, *the fact of the matter is*: 6, …
        - and that in spite of the fact that there is corpus work out there exploring such issues (Kita et al. 1992, Mason 2006, Mukherjee & Gries 2009, Brook O'Donnell 2011 …)
      - I don't even mention discontinuous *n*-grams

Corpus-driven linguistics – does that even exist?

- With maybe very few exceptions, I have yet to see what I consider a truly corpus-driven approach
  - with regard to register

    "Register variation can in fact be defined as systematic variation in probabilities; a register is a tendency to select certain combinations of meanings with certain frequencies" (Halliday 1991/2005:66)

  - register distinctions existing in a corpus may not be warranted from a truly bottom-up perspective
    - in terms of 2-gram attractions, acad and news in the BNC Baby are hardly different at all (Gries 2009, cf. also Gries, Newman, Shaoul, & Dilts 2009)
    - in terms of verb preferences of the ditransitive in the ICE-GB, spoken vs. written is as relevant as written:printed vs. written:non-printed (Gries 2011)

Corpus-driven linguistics – does that even exist?

- Xiao (2009) summarizes his own discussion of corpus-driven vs. corpus-based as follows
  - the distinction between the two is overstated
  - the corpus-based approach is better suited to contributing to linguistic theory
- I think that
  - if anything, the distinction is understated, given that **truly** corpus-driven work is a myth at best
  - this conclusion is interesting because in effect it says corpus-driven linguistics, which uses corpus-driven characteristics to argue for corpus linguistics as a theory, is in fact less suited to contributing to linguistic theory than corpus-based linguistics, which often views corpus linguistics as a method(ology) 'only'

Corpus linguistics and linguistic theory: some disagreements

- Some corpus linguists
  - are simply not concerned with the linguistic system that more theoretical linguists may care about – they may
    - use corpora for practical/applied purposes such as lexicography and/or language teaching
    - are not interested in linguistic theory (and I will not be concerned with this perfectly legitimate stance)
  - have rather 'unusual' ideas about potentially relevant neighboring disciplines
  - have rather 'unusual' ways of defending their perspective(s)
  - have rather 'unusual' ideas about the nature of the discipline (above and beyond the above issues)

Corpus linguistics and linguistic theory: some disagreements

- Some corpus linguists have rather 'unusual' ideas about potentially relevant neighboring disciplines
  - Teubert (2008) on the relationship between cognitive linguistics and natural language processing: the latter is the "illegitimate offspring" of the former … ???
  - Mason (2007:2): "Formal approaches […] take for granted a hierarchical (phrase) structure, […]. However, language is not produced in that way, but instead is a linear sequence created in stops and starts. A hierarchical structure thus cannot account for the fact that the beginning of an utterance is already produced before the whole sentence has been completely worked out. Similar issues apply for the reception of language." … an incremental approach to language production and comprehension does by no means require abandoning a largely hierarchical view of language structure!

Corpus linguistics and linguistic theory: some disagreements

- Some corpus linguists have rather 'unusual' ways of defending their perspective(s): a rather radical us-vs.-them ideological warfare rhetoric
  - that uses geographical labels in place of arguments, as when agendas are characterized as "transatlantic" and (implicitly) contrasted with British/Old World corpus linguistics
  - (good) old-fashioned Sinclairian core corpus linguistics vs. those who "piss into" Sinclair's canonical corpus linguistics tent and use corpora in "a seemingly inappropriate, toolbox-like, non-Sinclairian way"
  - "the label *corpus linguistics* has, over the last decade, been hijacked by theoretical linguists of all feathers"
  - Teubert even argues against some software because "it does not matter what kind of strings of information bit are processed. It could be language, but it could also be DNA sequences or the ciphers behind the "3." in the number pi" - as if that wasn't true of any concordancer

Corpus linguistics and linguistic theory: some disagreements

- Some corpus linguists have rather 'unusual' ideas about the nature of the discipline
  - "corpus linguistics looks at phenomena which cannot be explained by recourse to general rules and assumptions"
  - "When linguists come across a sentence such as "The sweetness of this lemon is sublime", their task is […] to look to see if other testimony in the discourse does or does not provide supporting evidence."
  - "Corpus linguistics looks at language from a social perspective. It is not concerned with the psychological aspects of language.", but on the other hand, …
  - "Linguistics is not a science like the natural sciences whose remit is the search for 'truth'. It belongs to the humanities, and as such it is a part of the endeavour to make sense of the human condition. Interpretation, and not verification, is the proper response to the quest for meaning."

A brief interim summary

- All of this must not distract from the facts that corpus linguistics in its present form is a relatively young discipline and has left quite a mark on (theoretical) linguistics
- but corpus linguists can benefit from more interaction with other (neighboring) disciplines
- this is because many corpus linguists take the above delimitation of the field very literally and often develop tools/methods that hardly get validated against anything outside the discourses
  - measures of dispersion
  - measures for collocational strength
  - measures for *n*-grams

Where we should validate more …

- Re validation
  - there are now 20-something measures of dispersion but few **corpus** linguists try to determine which are best in which circumstances (cf. Lyne, Gries for exceptions)
  - there are now 30-something measures of collocational strength but not only do few **corpus** linguists set out to determine which are best in which circumstances (Evert, Wiechmann, Pecina are laudable exceptions),
  - there are now many different ways to generate *n*-grams, but few **corpus** linguists try to determine which result in something that corresponds to something else outside of the narrow confines of the discourses in a corpus (this is true even of Linear Unit Grammar)
  - there are now even corpus linguists who argue for trying different ways to modify measures and pick whatever yields results that intuitively (!) appear best (and then sell that 'functionality')

Why we should validate more …

- And validation is so urgently needed: studies differ with regard to which measures of attraction yield the best results
  - Krug (1998): string frequency; Gries et al. (2005): $p_{\text{Fisher-Yates exact test}}$; Wiechmann (2006): minimum sensitivity; Divjak (in progress): conditional probability
- so, do we really just go on using *MI* (or *t* or …) just because we're supposed to focus on the discourse only and because WordSmith or the WordSketch engine or … make that so easy?
- don't we care that there are psycholinguistic results out there, results that should affect
  - our choice of statistical measures?
  - our interpretation of results in a larger context?

Where to turn to / what to relate to for validation …

- Thus, corpus linguistics would benefit from applying corpus methods outside of corpus linguistics and its discourses proper
  - because that would increase corpus linguistics' visibility in the field of linguistics as a whole and in particular with disciplines that have often independently arrived at similar findings or conclusions
  - because external validation would streamline corpus-linguistic research enterprises
  - because that would in turn improve corpus linguistics: Butler (2004) argued for a greater awareness in corpus linguistics of the need for a more powerful and cognitively valid theory
- this in turn means we need to hook up (more) with theoretical linguistics / other neighboring disciplines … but which theory could we hook up with?

Where to turn to / what to relate to for validation …

- Since I disagree with (nearly!) that Teubert says, let's turn to him for help
  - "For me, corpus linguistics and cognitive linguistics are two complementary, but ultimately irreconcilable paradigms."
  - "Corpus linguistics localises the study of language, once again, firmly and deliberately, in the Geisteswissenschaften, the humanities."
  - "Corpus linguistics looks at language from a social perspective. It is not concerned with the psychological aspects of language."
-  how about a psycholinguistically informed (cognitively-inspired) usage-based linguistics, located, firmly and deliberately, in the social / behavioral sciences?

Additional advantages of that move …

- And, since we're talking about **humanistic** perspective and the ***Geistes****wissenschaften* … isn't illuminating the cognitive system(s) that ultimately give rise to discourse(s) telling us much more about the 'human condition'? and how can we seriously be in the ***Geistes****wissenschaften* if the one thing we *a priori* disregard is *Geist*?
- at some point of time, going psych/cogn is needed: things only enter into discourse when a speaker has processed them and 'decided' to utter them and thereby make them part of the discourse, and the way a hearer processes things is also determined by that hearer's internal structure
- plus, the overlap of notions and interests is already huge

Many things corpus linguists say have immediate cogn./psycholing. relevance

- When corpus linguists talk about token frequencies
  - (theoretical) cognitive linguists become interested because, all other things being equal, token frequencies correlate with
    - degree of entrenchment (Schmid 2000)
    - phonetic reduction and development of new forms (Schuchardt 1885, Fidelholtz 1975, Bybee & Thompson 1997, Bybee & Scheibman's 1999)
    - resistance to morphosyntactic language change (Bybee & Thompson 1997)
  - psycholinguists become interested because, all other things being equal, token frequencies correlate with
    - ease/earliness of acquisition (Casenhiser & Goldberg 2005)
    - lexical decision tasks, word naming, picture naming (Howes and Solomon 1951, Forster & Chambers 1973; re web data, cf. Van Durme et al., in progress)

Many things corpus linguists say have immediate cogn./psycholing. relevance

- When corpus linguists talk about type frequencies
  - (theoretical) cognitive linguists become interested because type frequencies are correlated with (morphological) productivity and language change (Bybee 1985, Albright & Hayes 2003)
  - psycholinguists become interested because type frequencies are correlated with the productivity of, say, constructions in first and second language acquisition

Many things corpus linguists say have immediate cogn./psycholing. relevance

- When corpus linguists not only talk about frequencies, but also about dispersion (which they do too rarely)
  - psycholinguists become interested because
    - dispersion has implications for psycholinguistic experiments (Gries 2010)
    - dispersion has implications for learning/acquisition
      - range can have significant predictive power for processing speed of academic formulae above and beyond raw frequency of occurrence (Simpson & Ellis 2005)

Many things corpus linguists say have immediate cogn./psycholing. relevance

- When corpus linguists argue against a strict separation of syntax and lexis

  "I have always seen lexicogrammar as a unified phenomenon, a single level of wording, of which lexis is the most delicate resolution."
  Halliday (1991/2005:64)

  - (theoretical) cognitive linguists agree
  - many psycholinguists have long assumed a position where both words and syntactic patterns are represented as nodes in an (interactive activation) network where, in production, lexical and syntactic nodes are activated when they fit the particular semantic/pragmatic meaning to be communicated

Many things corpus linguists say have immediate cogn./psycholing. relevance

- When corpus linguists talk about words and patterns and the Idiom Principle's large number of semi-preconstructed phrases that constitute single choices (Sinclair 1991:110)
  - (theoretical) cognitive linguists become interested because it reminds them of Langacker's
    - unit, "a structure that a speaker has mastered quite thoroughly, to the extent that he can employ it in largely automatic fashion, without having to focus his attention specifically on its individual parts for their arrangement"
    - rule-list fallacy: "There is a viable alternative: to include in the grammar both the rules and instantiating expressions. This option allows any valid generalizations to be captured (by means of rules), and while the descriptions it affords may not be maximally economical, they have to be preferred on grounds of psychological accuracy […]. Such units are cognitive entities in their own right whose existence is not reducible to that of the general patterns they instantiate."

Many things corpus linguists say have immediate cogn./psycholing. relevance

- When corpus linguists talk about words and patterns
  - (theoretical) cognitive linguists become interested because Hunston and Francis's patterns are very similar to Goldberg's constructions
    - pattern: "The patterns of a word can be defined as all the words and structures which are regularly associated with the word and contribute to its meaning. A pattern can be identified if a combination of words occurs relatively frequently, if it is dependent on a particular word choice, and if there is a clear meaning associated with it." (Hunston & Francis 2000:37)
    - construction: "Any linguistic pattern is recognized as a construction as long as some aspect of its form or function is not strictly predictable from its component parts or other constructions recognized to exist. In addition, patterns are stored as constructions even if they are fully predictable as long as they occur with sufficient frequency" (Goldberg 2006:5)

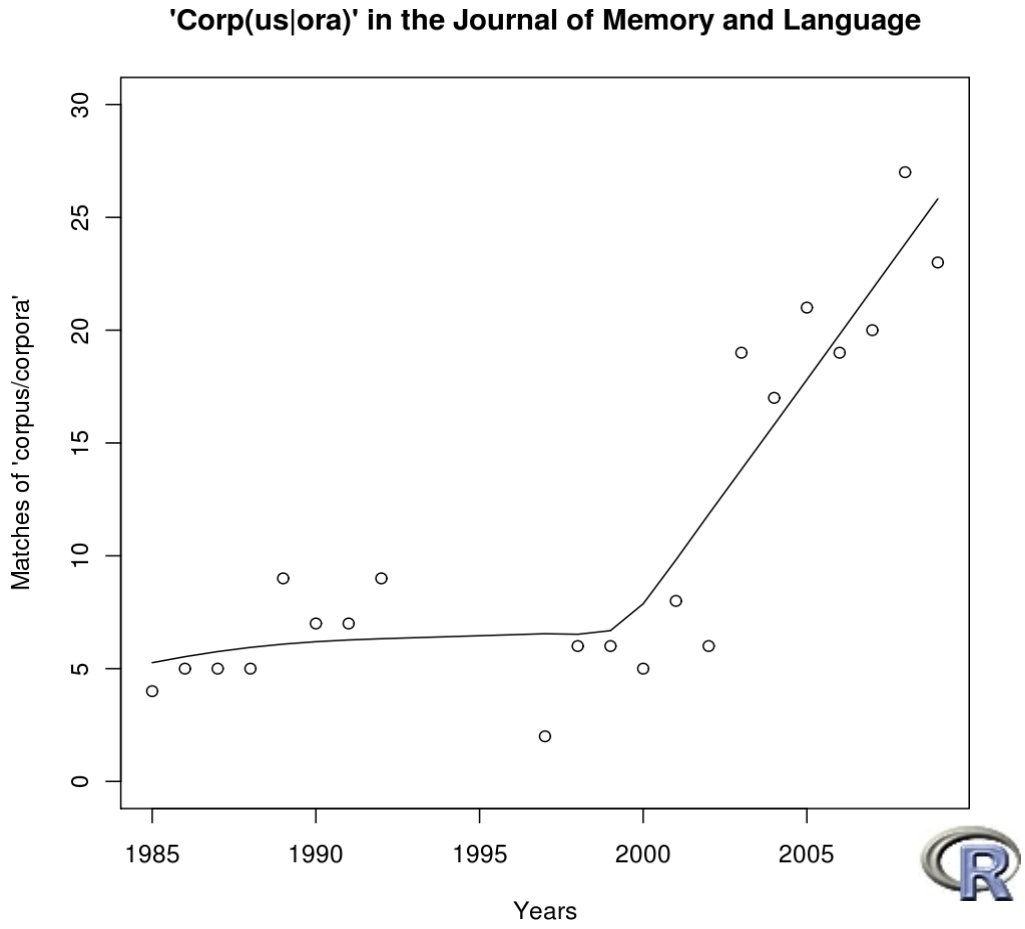
Many things corpus linguists say have immediate cogn./psycholing. relevance

- When corpus linguists talk about concordances, collocations, *n*-grams, colligations – i.e., anything having to do with co-occurrence information
  - psycholinguists become interested because such co-occurrence information
    - helps children discern phonotactic patterns (Saffran et al.)
    - helps children discover word classes (Reddington et al. 1998, Mintz et al.'s 2002)
    - can predict reading times (word-pair frequencies, MacDonald 1993) and gaze duration (bigram probabilities, McDonald, Shillcock, & Brew 2001)
    - helps subjects recognize frequent 4-grams faster (when 1-gram and 2-gram frequency is controlled) (Snider & Arnon 2012)
  - language production and comprehension have been shown to be highly item-specific, which is just another way of saying context-bound (e.g., lexically-specific reduction or priming effects)

Cognitive linguistics / psycholinguistics → ← corpus linguistics

- Re Teubert's focus on the social perspective
  - Geerarts (2003), Croft (2009), and others have been arguing for a cognitive sociolinguistics and the first papers, volumes, and conferences focusing on such issues appear
  - "the function pole in the definition of a construction indeed allows for the incorporation of factors pertaining to social situation, such as, e.g., register" (Goldberg 2003:221)
- re the use of corpus data in cognitive linguistics
  - there have now been several theme sessions on corpora and/or frequency effects in cognitive linguistics at the largest cognitive linguistics conferences
  - psycholinguists use corpora more and more

Cognitive linguistics / psycholinguistics → ← corpus linguistics



Towards a psycholinguistic model

- So, corpus linguists talk about a lot of things that have immediate psycholinguistic and/or cognitive-linguistic relevance
- however, to a considerable degree, it is linguists outside of corpus linguistics that
  - apply our methods, and/or demonstrate their relevance, to notions/data outside of the 'discourses'
  - validate some of the suggestions we've made
- thus, we can benefit from relating to more of what happens in irreconcilably different disciplines
- these disciplines in turn have developed theories and models that would allow us to move
  - from the purely descriptive approach for which corpus linguists are often criticized
  - to explanation, prediction, and the embedding into a larger context, and the kind of psycholinguistic model many of the above studies come with is an exemplar-based approach

Characteristics of exemplar models

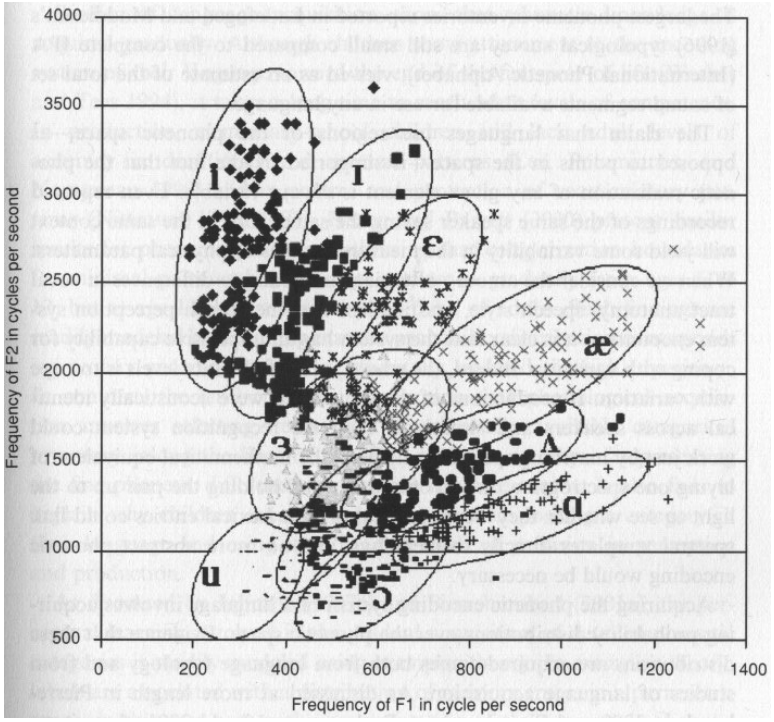
- We have seen above that infants are very good at keeping track of distributional characteristics of the ambient language – but how is that acquired and represented? → exemplar representations
- what are the main assumptions of such models?

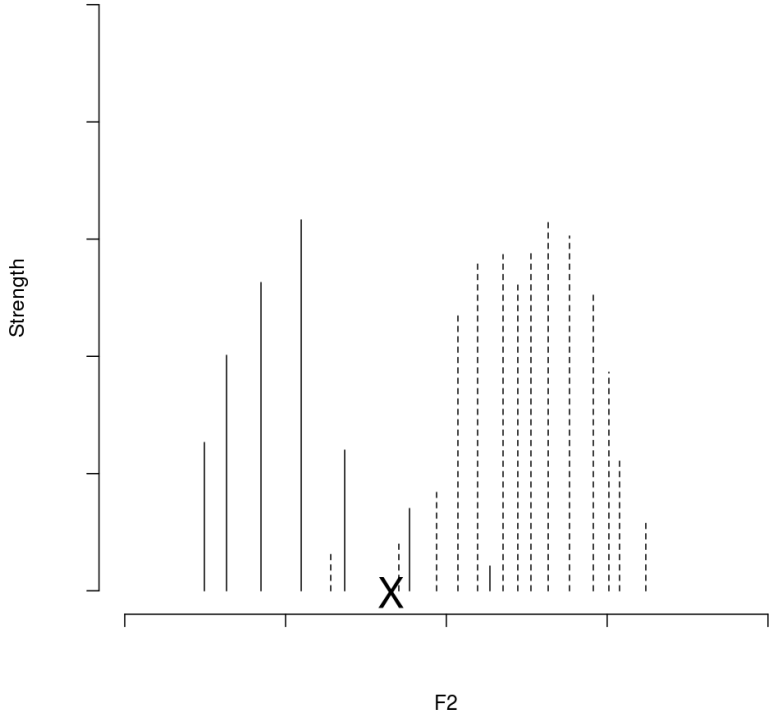
"each instance redefines the system, however infinitesimally, maintaining its present state or shifting its probabilities in one direction or the other" (Halliday 1991/2005:67)
"it is usual that each learning event updates a sta-tistical representation of a category independently of other learning events." (Ellis 2002:147)

Characteristics of exemplar models

- Speakers/listeners remember (aspects of) tokens / exemplars and 'place them' into a multidimensional space / network
- labels (e.g., of phonemes) are "associated with a distribution of memory traces in a parametric space, in this case a cognitive representation of the parametric phonetic space" (Pierrehumbert 2003:185)
- co-occurrence traces involve
  - phonetic, phonological, prosodic, morphemic, lexical, syntactic, … co-occurrence + extra-linguistic aspects: utterance context, socio-linguistic speaker factors, register / genre / mode

Characteristics of exemplar models





Characteristics of exemplar models

- However, speakers/listeners don't remember each token and don't remember everything about each token
  - memories of individual tokens may not always be accessible and memories of aspects of a particular token may not always be accessible
  - "How are learners then able to isolate typical contexts for a particular word? […] the fallibility of human memory: the fact that we normally don't remember things we enounter only once or twice (unless they are particularly striking, or highly significant for personal reasons)." (Dąbrowska 2008:207)
  - aspects may never be stored in long-term memory
  - aspects may decay or may be subject to generalization/abstraction as well as reconstruction
    - "abstraction is an automatic consequence of aggregate activation of high-frequency exemplars, with regression toward central tendencies as numbers of highly similar exemplars increase." (Ellis 2002:153)

Advantages and implications of exemplar models

- Theoretical advantages
  - it explains first language acquisition without recourse to largely untestable parameters etc.
  - we know that s/l store immense amounts of probabilistic information, and the assumption of clouds of remembered exemplars can explain frequency effects well
    - high freqs of occurrence correspond to dense clouds with many different points in very close proximity
    - high freqs of co-occurrence correspond to dense clouds with many different points in very close proximity, but looked at from a different 'dimensional angle'
  - categorization and prototype effects follow from the multidimensional structure of a cloud of exemplars
  - the model can explain how even native speakers of a language can differ considerably in their command of the language and their judgments (cf. Dąbrowska 2009)
  - the model can unproblematically account for register and other contextual effects (as additional dimensions)

Advantages and implications of exemplar models

- Methodological implications
  - towards multifactorial approaches in hypothesis-testing where model selection processes are used to determine which dimensions for which data are available should be retained
    - general(ized) linear models as in, say, studies of alter-nations (Gries 2003), Szmrecsanyi (2005), Bresnan, Cueni, Nikitina, & Baayen (2007), Arppe (2008), Janda, Nesset, & Baayen (2010), etc.
  - towards mixed-effects models, which allow us to model
    - speaker/writer/subject-specific effects
    - lexically-specific effects, etc.
  - towards more bottom-up and/or multivariate exploratory methods to determine which (meaningful) dimensions emerge when the space is compressed and rotated
    - principal components/factor analysis (cf. Biber)
    - cluster analysis, MDS, correspondence analysis, …
  - more comparative register-specific analysis

What I wanted to do …

- I hope I have been able to
  - discuss some of the reasons why theoretical linguists and (especially a particular group of) corpus linguists have so far not yet entered into the kind of fruitful relation that I would like to see more
  - convey my thoughts on why I think that this (only slowly narrowing) gap should be closed at a much faster pace and why esp. that particular group in corpus linguistics is on the wrong track
  - convince you at least in part that much of corpus linguistics is extremely compatible with developments in cognitive construction grammar (of the Goldberg flavor) and some psycholinguistic theories/models, and that these theories can help corpus linguists answer *why*-questions in a much more revealing way than the humanistic hermeneutic-circle meaning-in-discourses-is-negotiated-by-the-community way still upheld by some

Why I wanted to do that …

- For example, is it not better to be able to explain distributions in corpora – e.g., reduced pronunciations of words – with reference to cognitive mechanisms regarding learning, habitualization, and articulatory routines arrived at independently than to what else happens in the discourse?
- for example, is it not better to be able to explain changes in diachronic corpora – e.g., the development of *going to* as a future marker in English – with reference to more generally known effects of automatization as a result of frequency of occurrence than to what else happens in the discourse?

And now what I want you/us to do … ;-)

- So, apart from minor proposals such as
  - maybe we can/should rethink the contrast of 'corpus-driven' and corpus-based linguistics
  - we should definitely rethink the us vs. them hijacking warfare rhetoric
- my main proposal today is for us corpus linguists to assume as the theoretical framework within which to embed our analyses a psycholinguistically informed, (cognitively-inspired) usage-based linguistics
- some have already argued for something similar
  - Miller & Charles's (1991) contextual representation
  - and my favorite: "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." (Hoey 2005:11; see also Taylor's *Mental Corpus*)
- but the major breakthrough has not happened … yet …

Thank you!

http://tinyurl.com/stgries