

Corpus methods for linguistics: A brief introduction

Ye Tian, Université Paris Diderot

UCL 08 – 09 Nov 2016

Plan

- Day 1 AM: general introduction. What can corpus methods do for linguistics.
- Day 1 PM: the Switchboard corpus: general information, example analysis, and some hands-on practices. You can suggest analyses you want to do.
- Day 2 AM: Linguistics and web-data.
 - Get data from Twitter and Facebook
- Day 2 PM: computational processing (text cleaning, PoS tagging, sentiment analysis, emoji analysis)

Corpus linguistics

- How do people *use* language?
- Corpus: “body” in Latin > using large bodies of naturalistic data to investigate language use.
- Corpus linguistics: an empirical *methodology* rather than a branch of linguistics like syntax or semantics.
- Rely on natural usage data rather than introspection.
- Enables us to quantify linguistic patterns. Most patterns are rarely absolute, mostly gradient.

Corpus linguistics

A corpus linguist is

- *A linguist who tries to understand language, and behind language the mind, by carefully observing extensive natural samples of it and then, with insight and imagination, constructing plausible understandings that encompass and explain those observations.*

Chafe (1992: 96)

Types of corpora

- General (larger) corpus vs. specialized corpus (smaller)
 - General: British National Corpus
(<http://www.phon.ox.ac.uk/AudioBNC>)
 - Corpus of Contemporary American English
<http://corpus.byu.edu/overview.asp>
 - Brown Corpus (http://www.sls.hawaii.edu/bley-vroman/brown_corpus.html)
- Written, spoken or computer-mediated texts
- Examples of written corpora:
 - Google Books Ngram corpus, COCA, the Wikipedia corpus

Types of corpora

- Examples of spoken corpora:
 - Diachronic corpus of present day spoken English (800,000 words)
 - Corpus of spoken professional American English (2million)
 - Michigan Corpus of Academic Spoken English (1.7 million)
 - Switchboard (1.4 million)
 - Corpus of Spontaneous Japanese
- Examples of corpora of computer-mediated-texts:
 - The National University of Singapore's SMS corpus (<http://www.comp.nus.edu.sg/entrepreneurship/innovation/osr/corpus/>)
 - SMS Spam Corpus (<http://www.esi.uem.es/~jmgomez/smsspamcorpus/>)
 - Corpus from twitter data.

Type of corpora

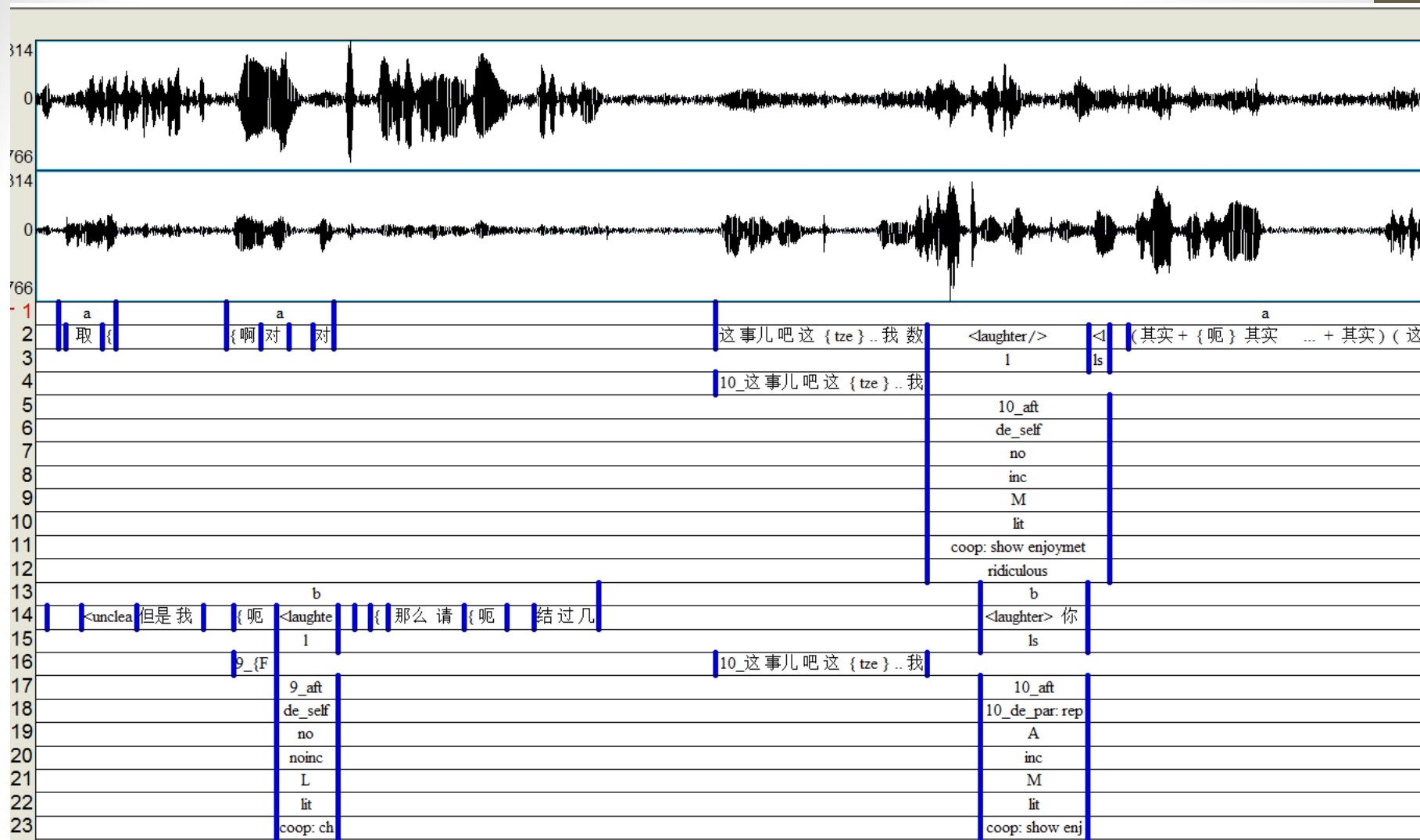
- Multilingual vs. single language corpus
 - The DUEL corpus (<http://www.dsg-bielefeld.de/DUEL/>)
- Specialized corpora
 - Learner corpus (language produced by learners of that language)
 - The Longman Learner Corpus, the International Corpus of Learner English
 - Corpora of language development
CHILDES: <http://childes.psy.cmu.edu/>
 - Multi-modal corpus: gestural corpus
 - Corpora of specific phenomenon
 - Corpus of linguistic and visual metaphor
 - Corpus of speech errors
(<http://www.mpi.nl/resources/data/fromkins-speech-error-database>)

Good places to look

- The linguistic data consortium
- <https://catalog.ldc.upenn.edu/ldc97s62>
- LREC conferences (The international conferences on language resources and Evaluation)
- <http://www.lrec-conf.org/>
- Check out their proceedings. You can often find some surprising and interesting corpora!

Transcription and annotation

- PoS tagging: Part of Speech tagging,
 - Assigns a part-of-speech (e.g. noun, verb, adjective, adverb) to each word.
 - Done manually or automatically (normally a finite-state-model, e.g. a Markove chain model, with probabilities. Leech, 1991).
- Parsing: to identify and label the function of each word and groups of words, according to specific syntactic theories.
 - Parsed corpora are called treebanks
- Written corpora: PoS tagging, parsing, higher level annotation
- Spoken corpora: transcription using the authogriphy of the target language, glossing, translation, IPA transcription, turn/utterance identification (if conversational), PoS tagging, parsing.



Transcription and annotation of disfluency and laughter in Project DUEL

Transcription and annotation

- [What/WP kind/NN] of/IN [experience/NN] do/VBP [you/PRP] ,/, do/VBP [you/PRP] have/VB ,/, then/RB with/IN [child/NN care/NN] ?/.
- (SBARQ (INTJ (UH So)) (, ,) (WHNP-1 (WHNP (WP What) (NN kind)) (PP (IN of) (NP (NP (NN experience)) (PP (-NONE- *ICH*-2))))) (SQ (EDITED (RM (-DFL- \[])) (SQ-UNF (VBP do) (NP-SBJ (PRP you)))) (, ,) (IP (-DFL- \+))) (VBP do) (NP-SBJ (PRP you)) (RS (-DFL- \])) (VP (VB have) (NP (-NONE- *T*-1)) (, ,) (ADVP (RB then)) (PP-2 (IN with) (NP (NN child) (NN care))))) (. ?) (-DFL- E_S))

MAKE USE OF CORPORA DATA

Making use of a corpus

- Search
 - Web-based search engines (e.g. BNCweb, WordSmith Tools, Xaira, Wmatrix, AntConc)
 - Write your own tools
- Quantify and summarize the entire corpus
 - Frequency and distribution (of words, sentences/utterances)
 - Written corpus: sentence length, length and structure of discourse units, etc
 - Spoken corpus: length of utterances, turns, amount of contribution from each conversational participant, etc
 - Collocation
 - Keywords

Keywords

- A keyword is a word which occurs statistically more frequently in one file or corpus, when compared against another comparable or reference corpus.
- Window into language change.

Table 5.2 Some keywords in LOB and FLOB
when compared against each other

| LOB (1961) | FLOB (1991) |
|--------------|---------------|
| COMMONWEALTH | THATCHER |
| MISS | MAJOR |
| MAN | WOMEN |
| THE | OK |
| HE | FUCKING |
| GIRL | AROUND |
| MUST | ET |
| SHALL | PRIVATISATION |
| RHODESIA | MARKET |
| KENYA | BLOODY |

Keywords

- A number of keywords are more indicative of changes in style, which can also ultimately be linked back to social change. For example, the keywords “fucking”, “bloody” and “OK” suggest that written language has become more informal in the 30-year period.

Frequency and distribution of words

- Frequency and distribution of different word classes/ specific words/phrases
 - the most frequent 50 or so words make up about 50% of all tokens.

Table 5.1 Top 10 word frequencies in LOB and FLOB

| LOB (1961) | | | FLOB (1991) | |
|------------|------|----------------|-------------|----------------|
| 1 | THE | 68,379 (6.67%) | THE | 64,813 (6.35%) |
| 2 | OF | 35,769 (3.49%) | OF | 34,147 (3.35%) |
| 3 | AND | 27,932 (2.72%) | AND | 27,292 (2.67%) |
| 4 | TO | 26,907 (2.62%) | TO | 27,058 (2.65%) |
| 5 | A | 23,170 (2.26%) | A | 23,168 (2.27%) |
| 6 | IN | 21,338 (2.08%) | IN | 20,880 (2.05%) |
| 7 | THAT | 11,197 (1.09%) | THAT | 10,481 (1.03%) |
| 8 | IS | 10,995 (1.07%) | IS | 10,923 (1.01%) |
| 9 | WAS | 10,502 (1.02%) | WAS | 10,039 (0.98%) |
| 10 | IT | 10,031 (0.98%) | FOR | 9,344 (0.92%) |

- Identifying neologisms (its appearance and shelf life)

Word senses

- Vossen (1991) studied word senses.
- 67% of 23,800 nouns in the Longman Dictionary of contemporary English contained only one sense, 20% contained two senses, 6.5% contained three senses, 2.5% contained four senses.
- The remaining 600 or so nouns have 5-27 senses (average 7).
- Strong correlation between frequency and degree of polysemy.
- So for a small number of frequent nouns (e.g. line, point, head, place, time, service, thing, body), polysemy is a major feature.
- Very similar pattern found in verbs (55% one sense, 184 verbs have between 5 and 21 senses, averaging 6).

Chunks (phrases, idioms)

- Do we make up new combinations or reuse old ones?
 - The 'open-choice' principle v.s. 'the idiom principle' (Sinclair, 1991:109)
- Altenberg (1991a): over 70% of the words in the London Lund corpus are part of recurrent word combinations.
 - Mostly two to three words in length, some are more than five
- Pawley and Syder (1983): 'habitually-spoken sequences':
 - *Did you have a good trip?*
 - *How are you going to do that?*
 - *How is everyone at home?*
 - *Once you've done that the rest is easy.*
 - *I see what you mean.*
 - *I knew you wouldn't believe me.*
 - *There is nothing you can do about it now.*

Collocation /Concordance

- Investigate context of words.
- A concordance is a list of a word or phrase, with a few words of context either side of it, so we can see at a glance how the word tends to be used.
- “silk” in the Brown Corpus (Kennedy, 1998: 113): colour is most relevant rather than texture or weight.

Mrs Thomas Jordan selected a black taffeta frock made with a skirt of fringed tiers and worn with *crimson silk* slippers.

Mrs Eustis Reilly's *olive-green* street length *silk* taffeta dress was embroidered on the bodice with gold threads and golden sequins and beads.

Her favorite cocktail dress is a Norell, a *black* and *white* organdy and *silk* jersey. a new Turkish Empire embracing 'the union of all Turks throughout Central Asia from Adrianople to the Chinese oases on the *Silk* Trade Route'.

It was amazing how they had herded together for protection: an enormous matriarch in a quilted *silk* wrapper, rising from the breakfast table; a gross boy in his teens, shuffling in from the kitchen with a sandwich in his hands;

Here he sketched, sitting in their flowing gowns of linen and *silk*, young girls not yet twenty, some about to be married, some married a year or two.

Martha Schuyler, old, slow, careful of foot, came down the great staircase, dressed in her best lace-drawn *black silk*, her jeweled shoe buckles held forward.

Finally she had come down; Winston had heard her shaking out the skirt of her new *pink silk* hostess gown.

Dark *gray* sports jacket, lighter *gray* slacks, *pink* flannel shirt, *black silk* necktie.

He was a man in his late forties, with *graying* hair, of medium height; he looked dapper in a lightweight summer suit, *brown silk* tie and *green-tinted* soft collar.

Collocation /Concordance

- Some collocations are more lexicalized than others, but not simply frequency based
 - “no hope” and “pretty good” are less frequent than “the three”, though the former two are recognized as more lexicalized.
 - Kjeller (1984:164): collocational distinctiveness depends on
 - Absolute frequency of occurrence
 - Relative frequency of occurrence
 - Length of sequence
 - Distribution of sequence over texts
 - Distribution of sequence over text categories
 - Structure of sequence

Phonetics, phonology & morphology

- Natural speech provides more variability than citation forms.
- The phonetic forms of natural speech convey also the information structure and pragmatic content.
- Speech corpus allows one to study the variability in phonetic form, but also to study the relation between phonological form and other levels of linguistic structure.
- Speech corpora: search in the Linguistic Data Consortium.
<https://www ldc.upenn.edu/>, and also the resources page of Stanford department of linguistics:
<https://linguistics.stanford.edu/resources/corpora/corpus-inventory>
- Reading:
 - Cole, J., & Hasegawa-Johnson, M. (2012). Corpus phonology with speech resources. *Handbook of laboratory phonology*, 431-440.
 - Harrington, J. (2010). *Phonetic analysis of speech corpora*. John Wiley & Sons.

Semantic studies

- Quantification in use (Kennedy, 1987a): manual analysis of 60,000 words in journalistic sources
 - nearly 15% of the words are used to quantify, in a wide range of syntactic classes (determiners, adjs, nouns, verbs, advs).
 - Suggesting that quantification is important in communication.

Table 3.39 Relative frequencies of use of categories of quantification
(from Kennedy, 1987a: 272)

| Category of quantification | No. of different 'types' in 60,000 words | No. of quantification tokens | % of quantification tokens | % of words in corpus |
|--|--|---------------------------------|-------------------------------|-------------------------|
| A Specific quantities/degrees | | | | |
| 1 Numerical | | | | |
| Cardinals | 190 | 757 | 8.29 | 1.20 |
| Ordinals | 10 | 97 | 1.06 | 0.15 |
| Fractions | 11 | 30 | 0.33 | 0.05 |
| Percentages | 44 | 106 | 1.16 | 0.17 |
| 2 None | 21 | 160 | 1.75 | 0.25 |
| 3 Individual | 39 | 264 | 2.89 | 0.42 |
| 4 Totality | 73 | 440 | 4.82 | 0.70 |
| Total | 388 | 1,854 | 20.30 | 2.93 |
| B Non-specific quantities/degrees | | | | |
| 1 Approximation | 16 | 107 | 1.17 | 0.17 |
| 2 Small quantities/degrees | 75 | 385 | 4.21 | 0.61 |
| 3 Large quantities/degrees | 231 | 1,640 | 17.95 | 2.60 |

Syntactic studies

- NP complexity
- Premodification
- Post-modification
- Subordination
- Complementation
- Topicalization (tough-movement, cleft constructions)
- Ellipsis
- Co-ordination
- Adverbial usage

Syntactic studies - voice

- Voice: Francis and Kucera (1982: 554) studied the proportion of active vs. passive voice predications in the different genres of the Brown Corpus. They found that the passive voice is used more frequently in informative than in imaginative prose.
- Similar results from Svartvik (1966: 155) (figure from Kennedy, 1998)

Table 3.23 The use of passives in different registers
(adapted from Svartvik, 1966: 155)

| Text source | No. of words in sample | Total no. of passives | Passives per 1,000 words |
|-------------|---------------------------|--------------------------|-----------------------------|
| Science | 50,000 | 1,154 | 23.1 |
| News | 45,000 | 709 | 15.8 |
| Arts | 20,000 | 254 | 12.7 |
| Speech | 40,000 | 366 | 9.2 |
| Sports | 30,000 | 269 | 9.0 |
| Novels | 80,000 | 652 | 8.2 |
| Plays | 30,000 | 158 | 5.3 |
| Advertising | 28,000 | 83 | 3.0 |
| Total | 323,000 | 3,645 | 11.3 |

Syntactic studies – postnominal modifiers

- Biber et al. (1994) studied postmodification of nouns in three genres: editorials, fictions and letters (table from Kennedy, 1998)
 - Informal texts (fiction and letters) use less postnominal modification.
 - Prepositional phrases are much more frequent than relative clauses (academic prose uses relative clauses the most, but still less than prepositional phrases)

Table 3.45 Frequencies of different types of postnominal modifiers in three registers (per 1,000 words)
(from Biber et al., 1994: 173)

| Postnominal modifier | Editorials (27 texts) | Fiction (29 texts) | Letters (6 texts) |
|--|--------------------------|-----------------------|----------------------|
| <i>that</i> relative clauses restrictive | 1.8 | 0.7 | 0.5 |
| <i>Wh</i> relative clauses restrictive | 5.4 | 2.1 | 1.1 |
| <i>Wh</i> relative clauses non-restrictive | 1.9 | 2.2 | 0.3 |
| Relative clauses with no relative pronoun | 0.1 | 0.1 | 0.2 |
| Participial postnominal modifiers (past and present) | 4.9 | 1.8 | 0.2 |
| Prepositional phrases as noun modifiers | 38.2 | 15.2 | 16.8 |
| Prepositional phrases as verb modifiers | 44.1 | 56.2 | 45.8 |

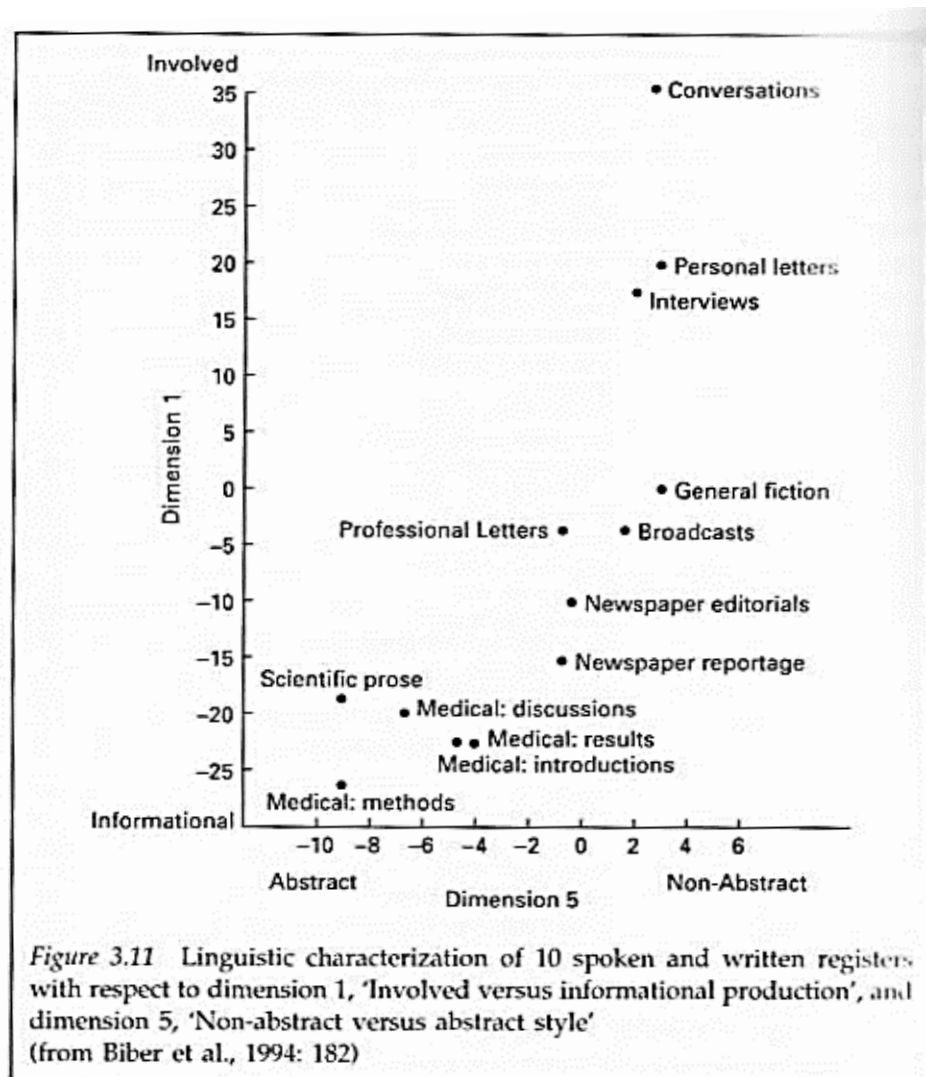
Syntactic studies - clefting

- Collins (1987) studied pseudo-cleft and cleft sentences
 - Non-cleft: *Fred gave Max a ride.*
 - Cleft: *It was Fred who gave Max a ride.*
 - Pseudo-cleft: *What Fred did was give Max a ride.*
- In spoken texts, pseudo-clefts are more than three times as frequent as clefts.
- In written texts, pseudo-clefts : clefts = 1.3 : 1

Pragmatic studies

- Linguistic variation (Biber, 1988)
- Five dimensions:
 - Informational vs. involved production
 - Narrative vs. non-narrative concerns
 - Explicit (situation-independent) vs. situation-dependent reference
 - Overt expression of persuasion
 - Abstract vs. non-abstract style

Pragmatic studies



Biber, (1988),
cited in Kennedy
(1998:190)

Pragmatics – Question Under Discussion

- Corpus with annotated QUD structure (de Kuthy et al., in prep)
- Roberts' (2012): natural discourse in general serves to answer hierarchically ordered Questions under Discussion (QUDs).
- Reflects the information structure of the dynamic context.
- De Kuthy et al. try to annotate QUD structure in German spoken corpus.

Pragmatics and discourse

- Filled pauses and Self Addressed Questions (Tian & Ginzburg, under review)
- SAQs address different problems in different languages
 - most frequently about memory-retrieval in English and Chinese, and about appropriateness in Japanese.
- In relation to filled pauses, British but not American English uses ``um'' to signal a more severe problem than ``uh''.
- Chinese uses different filled pauses to signal the syntactic category of the problem constituent.
- Japanese uses different filled pauses to signal levels of interaction with the interlocutor.

Pragmatics - metaphor

- The VU Amsterdam Metaphor Corpus and the Vismet corpus of visual metaphors:
- <http://metaphorlab.org/metaphor-corpus/>
- Bolognesi, M. (2016): studied semantic similarity between metaphor terms in Flickr tags.
 - Higher degree of similarity between two concepts in visual metaphor than in linguistic metaphor, which doesn't differ substantially from the similarity between two random concepts.
 - Semantically, more situation and entity related features in tags for visual metaphor, while more introspection and taxonomic features for linguistic metaphor.

Corpus of annotated visual metaphor

Total images found: 353

CONTEXT

any

Advertising (Commercial) i

Advertising (Social Campaign) i

Art (Illustration) i

Art (Photograph) i

Art (Graffiti) i

Political Cartoons i

Other i

CONTENT CONCEPTUALIZATION

any

Conventional i

Novel i

EXPRESSION REALIZATION

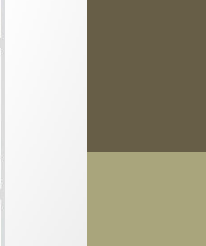







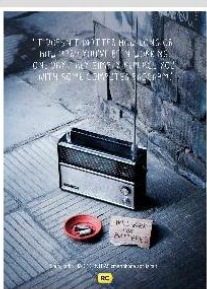













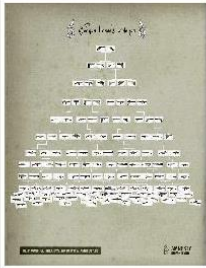


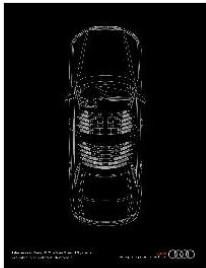


any

Juxtaposition i

Fusion i

Replacement i

HOME PROJECT ANNOTATION SCHEME BROWSE TERMS of USE



Pragmatics – irony and hyperbole

- Burgers (2016). HIP: A Method for Linguistic Hyperbole Identification in Discourse.
- Steen et al. (2010a). *A method for linguistic metaphor identification:*

Other uses of corpora

- Because we are linguists, we are mainly interested in the *language* of the texts/ speech in the corpora
- But one can also be interested in the *content* of the texts/speech - *knowledge-based* analyses
 - What knowledge can we extract from these texts – e.g. automatic processing of bio-medical literature
 - What topics do people talk about
 - What are their opinions on these topics
- Corpus linguistics versus computational linguistics
 - Here is a good read: Speech and language processing: an introduction of natural language processing, computational linguistics, and speech recognition, by Jurafsky and Martin

References

- Baker, P. (2010). Corpus methods in linguistics. *Research methods in linguistics*, 93-116.
- Bolognesi, M. (2016). Modeling Semantic Similarity between Metaphor Terms of Visual vs. Linguistic Metaphors through Flickr Tag Distributions. *Frontiers in Communication*, 1, 9.
- Burgers, C., Brugman, B. C., Renardel de Lavalette, K. Y., & Steen, G. J. (2016). HIP: A method for linguistic hyperbole identification in discourse. *Metaphor and Symbol*, 31(3), 163-178.
- Chafe, W. (1992). The importance of corpus linguistics to understanding the nature of language. In *Directions in corpus linguistics. Proceedings of Nobel Symposium* (Vol. 82, pp. 79-97).
- Cole, J., & Hasegawa-Johnson, M. (2012). Corpus phonology with speech resources. *Handbook of laboratory phonology*, 431-440.
- Francis, W., & Kucera, H. (1982). Frequency analysis of English usage.
- Harrington, J. (2010). *Phonetic analysis of speech corpora*. John Wiley & Sons.
- (Kennedy, 1987a)
- Kennedy, G. (1998, 2014). *An introduction to corpus linguistics*. Routledge.
- Kjeller, G. Some Thoughts on Collocational Distinctiveness in Recent Developments in the Use of Computer Corpora in English Language Research. *Costerus*, 1984, vol. 45, p. 163-171.
- Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford University Press.
- Steen, G. J., Dorst, A. G., Herrmann, J. B., Kaal, A., Krennmayr, T., & Pasma, T. (2010a). *A method for linguistic metaphor identification: From MIP to MIPVU*. Amsterdam, The Netherlands: Benjamins.
- Quirk, R., & Svartvik, J. (1966). *Investigating linguistic acceptability* (No. 54). de Gruyter Mouton.
- Vossen, P. (1991) 'Polysemy and vagueness of meaning description in the Longman Dictional of Contemporary English', in Johansson & Stenstrom (1991): 105-124.