HS Corpus Linguistics / Korpuslinguistik

2. Fundamentals

Prof. Dr. Stephanie Evert
Chair of Computational Corpus Linguistics
www.linguistik.uni-erlangen.de

FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

PHILOSOPHISCHE FAKULTÄT
UND FACHBEREICH THEOLOGIE

Pre-history of corpus linguistics

- First corpus-based quantitative studies in late 19th century
- Orthography and frequency lists
  - Käding (1897): German frequency dictionary based on corpus of ca. 11 million words (manual work!)
- Language acquisition
  - first longitudinal studies ca. 1876–1926 (parent diaries)
  - large cross-sectional studies ca. 1927–1957
- Lexicography
  - Murray: several million index cards for OED (1879–1928)
- Foreign language teaching
  - basic vocabulary, vocabulary levels, collocations (e.g. Palmer 1933)
- Strucuralist language documentation
  - Boas (1940), J.R. Firth (1930–1955), …
- Comparative philology
  - Eaton (1940): semantic frequency lists for English, French, German, Spanish

Chomsky (1957)

- Rationalism (introspection) vs. empiricism (data-oriented)
  - corpus linguistics: empirical description of language patterns
  - Chomsky: explanatory theory, must be cognitively plausible
- Competence vs. performance
  - Chomsky: corpus reflects speaker performance (with mistakes), empirical frequency data irrelevant for language competence
  - counter-argument: *armchair linguistics* based on invented examples
- Representativity
  - human speakers can produce infinitely many well-formed utterances; even a large corpus only contains a small subset → not representative
  - Chomsky: "Any natural corpus will be skewed. Some sentences won't occur because they are obvious, others because they are false, …"
- Learnability: the poverty of stimulus argument
  - corpus data insufficient for language acquisition: no negative examples

Recent history of corpus linguistics

- Since 1950: Humanities Computing (→ Digital Humanities)
- 1950–1960: Mechanolinguistics (Juilland: contrastive corpora)
  → quantitative / mathematical linguistics (Harris 1968)
- 1960–1980: Corpus linguistics as European counter-movement against mainstream of generative linguistics (→ Chomsky)
- Corpus-based grammars (e.g. Quirk/Greenbaum)
  - Survey of English Usage (SEU) since 1960
  - Brown Corpus 1961–1963
- British contextualism (Firth & Sinclair)
  - Firth (1957) building on Malinowski & Jones, Sinclair (1991), computational lexicography (COBUILD)
  - principles of collocation & colligation, "trust the text"
- Since 1990: Corpus linguistics established as subdiscipline of linguistics, begins to accept methodological innovations from other subdisciplines
- Since 2010: Corpus evidence has become an essential part of linguistic research

Applications of corpus linguistics (I)

- Dialectology & contrastive linguistics
- Historical linguistics & language change
- Language description (e.g. endangered languages)
- Language teaching, language acquisition, CALL
- Language variation, register studies (→ Biber's MDA)
- Lexical semantics (e.g. semantic prosodies)
- Lexicography & lexicology (→ COBUILD, collocation dictionaries)
- Morphology (→ quantitative productivity)
- Phonology (esp. studies of phonological variation)

See McEnery, Xiao & Tono (2006, Unit A10)
for examples and references

Applications of corpus linguistics (II)

- Pragmatics & discourse analysis (→ rhetoric, ideology, politics)
- Psycholinguistics & psychology (e.g. frequency & association norms)
- Sociolinguistics (e.g. gender studies, language ideology, power relations)
- Stylometry & literary studies (e.g. authorship attribution, literary stylistics)
- Syntax & grammar (→ Quirk/Greenbaum, LGSWE)
- Theoretical linguistics (→ validation of theoretical predictions)
- Translation studies (→ “translationese”, CAT)
- Usage-based linguistics (→ cognitive linguistics, construction grammar)

See McEnery, Xiao & Tono (2006, Unit A10)
for examples and references

The stages of a corpus study

1. Operationalization
   - research question → quantitative hypothesis, definition of population (*sampling frame*)

2. Corpus compilation
   - selection of texts (from sampling frame), digitization / format conversion
   - collection of metadata, legal & ethical issues
   - shortcut: reuse existing corpus (often by selecting a suitable subcorpus)

3. Linguistic annotation
   - manual annotation, GUI, annotator agreement
   - automatic annotation with NLP tools for larger corpora

4. Representation format
   - standards: Unicode, XML, TEI, XCES, … important for archiving and data exchange
   - efficient binary index formats (e.g. CWB) for corpus search & quantitative analysis

The stages of a corpus study

FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG
PHILOSOPHISCHE FAKULTÄT
UND FACHBEREICH THEOLOGIE

5. Indexing & search

- search for keyword, phrase, linguistic pattern, …
- view results as concordance (“kwic” = keyword in context)
- analysis = grouping & structuring of concordance in order to identify recurrent patterns
- efficient search based on binary index format

Your query “[word="what"%c & !lbound(s)] "a"%c [pos="JJ.*"]+ "night"%c” returned 18 matches in 15 different texts (in 98,511,777 words [9,802 texts]; frequency: 0.18 instances per million words)

[4.616 seconds]

Solution 1 to 18 Page 1 / 1		
My dearJarmila ,	what a great night	for you .
oes crazy on a hot night , and maybe that 's	what a hot night	is for .
Wow ,	what a miserable night	.
tee that you 'll have a ball Come one and all	What a great night	you 've got in store You 'll wanna keep comi
e Hey , everyone , let 's go on with the show	What a great night	you 've got in store I 'll bet you 'll wanna kee
I am glad on ` t.	What a fearful night	is this !
Oh , dear ,	what a terrible night	.
it 's gone Love goes on and on Oh , Robin ,	what a beautiful night	.
Crimson morning skyline Whoa oh	What a weird night	, huh ?
It 's a wonder	what a good night	's sleep will do for you .
When they think it 's sunset and see	what a nice night	it is , they 'll muster in the lobby .
God ,	what a beautiful night	, Jack .
God ,	what a beautiful night	, huh ?

6. Quantitative analysis

- many insights based on systematic analysis of frequency data (esp. for large corpora)
- frequency comparison, keywords, co-occurrence
- statistical hypothesis tests, data analysis, visualisation

7. Interpretation

Types of corpora

- written vs. spoken vs. multimodal/multi-media
- reference corpus vs. specialized corpus
- synchronic vs. diachronic (discrete, continuous)
- closed corpus vs. monitor corpus
- monolingual vs. multilingual (parallel, comparable)
- unannotated (raw text) vs. annotated
  - metadata = information about texts & speakers/authors
  - linguistic annotation = systematically coded interpretation
- corpus size: small & clean vs. large & messy
  - measured in M = million (or G = billion) running words

Some corpora everybody should know

- Brown Corpus (Francis & Kucera 1964)
  - American English, written (edited), texts published in 1961
  - 500 samples @ 2000 words from 15 text genres (*categories*)
- Brown Family
  - Brown (AmE, 1961), LOB (BrE, 1961) – Frown (AmE, 1991), FLOB (BrE, 1991) – BLOB (BrE, 1931), BE2006 (BrE, 2006)
- Penn Treebank (Marcus, Santorini & Marcinkiewicz, 1993)
  - ca. 3 million words of AmE with syntactic analyses (*parse trees*)
- British National Corpus (Aston & Burnard 1998)
  - British English, 90% written / 10% spoken, collected ca. 1991
  - approx. 100 million words in 4048 files (= texts / collections)
- Web as Corpus: WaCky (Baroni et al. 2009)
  - ca. 2 billion words of text from automatically crawled Web pages for each of DE, EN, FR, IT
  - many other Web as Corpus projects: larger corpora, additional languages (Arachnea, COW, SkE 10^{10})

Recommended textbooks

- McEnery, Tony and Wilson, Andrew (2001). *Corpus Linguistics*. Edinburgh University Press, 2nd ed.
- McEnery, Tony; Xiao, Richard; Tono, Yukio (2006). *Corpus-Based Language Studies: An advanced resource book*. Routledge, London/New York. https://www.lancaster.ac.uk/fass/projects/corpus/ZJU/xCBLS/CBLS.htm
- McEnery, Tony and Hardie, Andrew (2012). *Corpus Linguistics: Method, Theory and Practice*. Cambridge University Press, Cambridge.
- Lüdeling, Anke and Kytö, Merja (eds.) (2008). *Corpus Linguistics. An International Handbook*. Mouton de Gruyter, Berlin.
- Kennedy, Graeme D. (1998). *An Introduction to Corpus Linguistics*. Longman (Pearson Education Ltd), London and New York.
- Hoffmann, Sebastian *et al.* (2008). *Corpus Linguistics with BNCweb – a Practical Guide*, vol. 6 of English Corpus Linguistics. Peter Lang, Frankfurt.
- Lemnitzer, Lothar and Zinsmeister, Heike (2015). *Korpuslinguistik: Eine Einführung*. Narr, Tübingen, 3rd edition.

Research community

- Important conferences
  - Corpus Linguistics (Lancaster / Birmingham / UK)
  - ICAME = International Computer Archive of Modern and Medieval English
  - AACL = American Association for Corpus Linguistics
  - CILC = International Conference on Corpus Linguistics

- Scientific journals
  - *International Journal of Corpus Linguistics* (IJCL)
  - *Corpora*
  - *ICAME Journal*
  - *Corpus Linguistics and Linguistic Theory* (CLLT)

- Web resources
  - David Lee's bookmarks (maintained by Martin Weisser): http://tiny.cc/corpora (navigation: CBL Links)
  - Linguistics Web (Sabine Bartsch): http://www.linguisticsweb.org/

Presentation topics

- Presentation: summary & discussion of 1 research paper
  - or in some cases 2 short papers
  - combine with small corpus study / partial replication of the paper
- Term paper (MA only, ca. 15 pages)
  - usually based on presentation, but find at least 1 additional relevant paper
  - extension of the corpus study
  - take feedback from your seminar presentation into account

- Complete list of topics available in StudOn
  - please select at least 3 topics you are interested in and note down their codes

- Assignment of topics & presentation schedule on Friday, May 20th