5. Corpus Linguistics (CL)

5.1 Definition

corpus1 = body or collection of written or spoken material upon which linguistic analysis is based (structuralism)

corpus2 = machine-readable “representative”, i.e. stratified “model” i.e. more than a text collection!

for computer-based language analysis:
corpus-informed (language awareness/ELT)
corpus-based vs. -driven (research)

5.1.2 Tools corpus analysis software:
WordSmith (with ICAME CD)
Sara (with BNC in TUC Bib)
AntConc downloadable free from http://www.antlab.sci.waseda.ac.jp/antconc_index.html

5.1.3 CL: methodology or theory?

• pro methodology:
  - CL is not parallel to phonology, syntax, lexicology or pragmatics (core linguistics)
  - CL is not restricted to any linguistic level (can be used to address phonological, syntactic, pragmatic etc. questions, (as is sociolinguistics \( \rightarrow \) ?))
  - “corpus” no reference to area of linguistic investigation (vs. sociolinguistics, psycholinguistics, computational linguistics, etc.)
  - methodologies adopted from social sciences inform sociolinguistics, but are not a theory in themselves (participant observation, interviews, etc.)

• pro theory:
  - CL has a particular outlook on language
  - rules of language are usage-based, not normative (as in prescriptive grammars)
  - linguistic change occurs when speakers use L for communication
  - CL introduced new methods and principles which have theoretical status \( \rightarrow \) theory

  -- pro methodology e.g. combined with SFL
5.1.4 The corpus-based vs. the intuition-based approach

- descriptions of English strongly biased (personal views)
- normative description do not take variation into consideration
- authenticity of invented examples can be questionable
- introspection-based results not verifiable, e.g. “I get myself a soft drink.” (pop, soda, coke considered incorrect)
- personal opinion is reflecting idiolect, not real speech
- professor’s shoeboxes
- Jespersen kept thousands of notes of real English from literary texts
- First to use them as authentic examples in his grammar
- corpus data can address preferences / tendencies; quantitative questions in general (frequency effects etc.)
- Improved reliability of corpus-based over intuition-based approach
- BUT: corpus-based approach not suitable for all research questions → approaches should be seen as complementary rather than exclusive

5.1.5 Reasons

for the popularity of corpus linguistics, esp. among non-native speakers

because it combines a qualitative and quantitative perspective

- offers citations used as real language samples of language usage
- provides a view beyond individual experience
- rules out individual salience
- computer processable

output: - concordances (KWIC-key word in context)
- collocates (milk gets sour, butter rancid, eggs addled = groceries spoil differently)
- relative vs. absolute frequencies → “normalise” = per 1 M. words

5.2 Corpus-based vs. corpus-driven approaches

1) type of corpus data
   - representativeness (c-driven: large corpora balance themselves; c-based: belief in natural balance unwarranted)
   - corpus size (c-driven: the larger the better; BUT: generally only analyse every nth instance – unclear how this is different from c-based methodology)
2) attitudes towards existing theories and intuitions
   - corpus annotation (c-driven: strong objections)
     - c-driven: tabula rasa ideal → no preconceived idea as to results; BUT: in reality, c-driven approaches make use of traditional categories such as word classes, etc. without defining them (which is a de-facto annotation ...)
   - c-based: typically start out from a theoretical issue / problem and use corpus data to illustrate / solve it
3) research focus
   - c-driven: holistic approach to language description (since such notions as pragmatics or syntax demand a theory)
   - c-based: focus on individual levels of linguistics
4) corpus-based approach by no means as radical as corpus-driven approach
   - c-driven approach “claims to be a new paradigm within which a whole language can be described”

5.3 Developments in corpus compilation

5.3.1 50 years of corpus history

(forerunners 1950s American structuralists, e.g. Harris)
1959 Quirk: Survey of English Usage (SEU)
1,000,000 words written/spoken 1953-1987
• London-Lund corpus of spoken English
1963/64 Francis/Kucera: Brown Corpus
1M of written American English from 1961
• 1M written BrE (Lancaster-Oslo/Bergen Corpus) from 1961
1980 · Cobuild Corpus (Birmingham, Sinclair) → Bank of English
1990 · International Corpus of English (ICE):
   - UK, US, CA, AU/NZ, EA (UK/TZ), ZA, HK, SG, IN, PH, etc.
1990 · International Corpus of Learner English (ICLE)
1990 - 1993 British National Corpus 100M (10M spoken)
• 1990 Freiburg Corpora: FLOB and Frown from 1991 etc. (parallel LOB/Brown)
for recent language change
• since 1998 www as “corpus” (WebCorp, WebPhraseCount)

5.3.2 Corpora on the history and variation of English

Kortmann (2005: 36)
5.3.3 Reference corpora on the WWW


5.4 Corpus compilation principles

5.4.1 Corpus types

large and stratified:
- British National Corpus (BNC)
  90 M written/10 M spoken, demographic/context-governed from 1991-94
  http://www.natcorp.ox.ac.uk/
- American / Australian National Corpus being compiled now (problematic; ANC 20 M)
- "national corpora", e.g. ICE http://www.ucl.ac.uk/english-usage/ice/
  e.g. ICE-East Africa, ICE-Canada (parallel corpora)
- genre/domain specific corpora
  e.g. SPACE (Specialised & Popular Academic English), Trains (dialogue corpus)
- translation corpora
  e.g. EU corpus
- English-German Translation Corpus
  "quick and dirty"/ad-hoc corpora, e.g. for translation problems, not translating

5.4.2 Representativeness / Balance in corpus design

Leech: representative = findings can be generalised (What)
Biber: representative findings show the same degree and extent of variability as the total population (How)
a representative corpus must contain different text types / genres / registers
possibility of constructing a monitor corpus, depending on the view on a corpus (dynamic vs. static)
in a general corpus, balance and sampling are responsible for achieving representativeness
acceptable level of balance also depends on intended uses (e.g. specialized corpora)
N.B.: "any claim of corpus balance is largely an act of faith rather than a statement of fact as, at present, there is no reliable scientific measure of corpus balance" (McEnery, Xiao & Tono 2006: 16)
researchers often adopt earlier corpus construction procedures (primarily BNC)
→ National Corpora in Australia, US, etc.

5.4.3 Sampling

sample = scaled-down version of a larger population
sampling units: e.g. a book, newspaper, periodical
sampling frame: the list of sampling units actually used in the corpus compilation (e.g. all books available in one particular library)
target population: group to be represented in the corpus
sampling techniques:
- simple random sampling: sampling units are numbered, elements are chosen based on a list of random numbers; problem: rare types / genres may not be selected
- stratified random sampling: divides population into groups (strata), samples each stratum at random
sample size:
- use full texts or text chunks from written sources?
- if chunks, where from (initial, middle, end chunks)?
→ again, these should be balanced (no either-or)
proportion & number of samples from each category

5.4.4 Annotation / mark up

Types of mark-up:
- structural markup: descriptive information about the texts
- "metalinguage" = structure of electronic documents (e.g. structure of conversations, categorizing parts of speech, segmenting of spoken or written text, marking of overlapping speech)
- bibliographic information about written text (genre, number of words, tagger which assign part-of-speech), ethnographic information about. Individuals in spoken texts (e.g. age, gender, social class, region; usually very limited)
- part of speech markup: part of speech designation (e.g. noun, verb); produced by software program called tagger (e.g. CLAWS; 95% accuracy)
- grammatical markup: parses grammatical structures (e.g. phrases, clauses); produced by software program called parser (usually 70 - 80% accuracy)
- always manual checks necessary

5.5 Corpora in Language Planning


Academic English Corpora:
- Michigan Corpus of Academic Spoken English (MICASE)
- Michigan Corpus of Upper-Level Student Papers (MICUSP)
- British Academic Spoken / Written English (BASE / BAWE)
cf. ChemCorpus of student / academic writing (theses)
5.4.5 Annotation procedures

taggers
• assign part-of-speech designations to each word in a sentence
• first tagging programme 1971 by Greene and Rubin
• out of this programme developing of CLAWS tagset by University of Lancaster
    (still widely used in its updated form)
• according to Leech (1997: 25-6) tags should strive for:
  - conciseness: tags should be as short as possible
  - perspicuity: tags should be as readable as possible
  - analysability: tags should have order and hierarchy above more specific tags

taggers are of two types:
• rule-based: based on rules of grammar written into the tagger -- e.g. EngCG-2
• probabilistic: based on statistical likelihood that a given tag will occur in a given context; can be trained on corpora
  -- the larger the tagset, the greater the accuracy of tagging

5.5 Corpus search strategies

3.5.1 Pattern types: investigating context
  collocation = the appearance of one particular word form in certain distance of another particular word forms
  in different meanings can have different collocates

colligation = the appearance of one particular word form in a particular grammatical structure

connotation = the semantic environment,
  can have positive or negative value ("semantic prosody")
  e.g. happen, cause, attempt, try, fail

collostruction analysis has 3 methods (WikiPedia):
  - collexeme analysis, to measure the degree of attraction/rejection of a lemma to a slot in one particular construction;
  - distinctive collexeme analysis, to measure the preference of a lemma to one particular construction over another, functionally similar construction; multiple distinctive collexeme analysis extends this approach to more than two alternative constructions;
  - covarying collexeme analysis, to measure the degree of attraction of lemmas in one slot of a construction to lemmas in another slot of the same construction.

http://en.wikipedia.org/wiki/Collostructional_analysis (15/12/13)

5.5.2 Types of frequency: exploring vocabulary

absolute vs. relative frequency of a word form
(standard) deviation from mean frequency of word forms

5.5.3 Corpus research examples

• How frequent is a particular morphological form/grammatical structure?
• Which particular structures have particular locations in texts?

Corpus tasks have degrees of complexity

• relevance of tagging:
  - parts-of-speech (POS), e.g. CLAWS tagging for LOB
    (http://ucrel.lancs.ac.uk/claws/trial.html)
  - semantic, semantic web/web 3.0

5.6 Corpus applications

5.6.1 Computational linguistics

5.6.2 Lexicography

5.6.3 Academic Writing

Corpus compilation
• variables: genres/text types, (sub-)discipline, gender, L1/MT, ...
• Corpus analysis --> interpretation
  - research hypotheses confirmed/refuted
  - research hypotheses developed

Corpus search strategies
• lexical hypotheses developed
• lexical hypotheses confirmed/refuted

Corpus applications
• part-of-speech tagging + syntactic annotation/treebanks + semantic web

Language analysis: an interdisciplinary field dealing with the statistical and/or rule-based modelling of natural language from a computational perspective
• not limited to any particular field of linguistics
• traditionally, performed by computer scientists who had specialized in the application of computers to the processing of a natural language
• often grouped under artificial intelligence today, but that has older applications (1950s) as well:
  - language analysis: tagging, parsing, annotation (5.6.1)
  - machine translation: SYSTRAN (5.6.2)
  - text processing: spell checkers, style checkers, automatic text production (abstraction/summarisation), Q&A systems
  - speech recognition and synthesis (telephone/communication systems)
  - others: expert/dialogue systems, CALL, etc.

Semantic tagging: semantic web

Tim Berners-Lee has described the semantic web as a component of "Web 3.0"
5.6.2 The corpus revolution in lexicography: word-watching → corpus compilation/analysis

today all dictionaries are based on large-scale corpora, esp. the BNC
- new lexical entries are found
- existing lexical entries are enriched by additional information extracted via corpus analysis (e.g. most common forms, connotation, etc.)
- important aspects of word meaning and grammar are highlighted,
  which were simply never noticed by linguists who had no data to work with
- word frequency analysis is used for annotating lexical entries
- collocational information is collected, organized, and presented (e.g. idiom identification)
- (domain specific) knowledge is extracted
- lexical items unlikely to be found in dictionary sources are extracted (e.g. proper nouns)
- real examples showing how central and typical features of English are used are provided
- paradigmatic- and syntagmatic-driven semantic clustering is performed

5.6.3 Academic Writing
Matrix of genre types in Academic Writing

Problem

Strand

Classification

ChemCorpus set-up by genre, degree programme and specialisation

<table>
<thead>
<tr>
<th>genre</th>
<th>specialisation</th>
<th>number of texts</th>
<th>average length</th>
<th>total words</th>
</tr>
</thead>
<tbody>
<tr>
<td>fixed Mag paper</td>
<td>language-linguistics</td>
<td>70</td>
<td>4,200</td>
<td>0.3 Mill</td>
</tr>
<tr>
<td>MSc thesis</td>
<td>language-linguistics</td>
<td>25</td>
<td>35,600</td>
<td>0.6 Mill</td>
</tr>
<tr>
<td>culture-literature</td>
<td>11</td>
<td>30,000</td>
<td>0.4 Mill</td>
<td></td>
</tr>
<tr>
<td>Total Mag paper</td>
<td>106</td>
<td>1.3 Mill</td>
<td></td>
<td></td>
</tr>
<tr>
<td>term paper B4</td>
<td>language-linguistics</td>
<td>100</td>
<td>4,200</td>
<td>0.5 Mill</td>
</tr>
<tr>
<td>culture-literature</td>
<td>100</td>
<td>4,700</td>
<td>0.5 Mill</td>
<td></td>
</tr>
<tr>
<td>project report</td>
<td>(submitted)</td>
<td>120</td>
<td>4,000</td>
<td>0.5 Mill</td>
</tr>
<tr>
<td>B4 thesis</td>
<td>language-linguistics</td>
<td>80</td>
<td>12,000</td>
<td>1 Mill</td>
</tr>
<tr>
<td>culture-literature</td>
<td>80</td>
<td>16,000</td>
<td>1 Mill</td>
<td></td>
</tr>
<tr>
<td>term paper M4</td>
<td>language-linguistics</td>
<td>80</td>
<td>4,500</td>
<td>0.5 Mill</td>
</tr>
<tr>
<td>culture-literature</td>
<td>80</td>
<td>6,000</td>
<td>0.5 Mill</td>
<td></td>
</tr>
<tr>
<td>M4 thesis</td>
<td>language-linguistics</td>
<td>40</td>
<td>25,000</td>
<td>1 Mill</td>
</tr>
<tr>
<td>culture-literature</td>
<td>40</td>
<td>25,000</td>
<td>1 Mill</td>
<td></td>
</tr>
<tr>
<td>Total B-M4</td>
<td>720</td>
<td>0.5 Mill</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Functional linkers by L1 (per one million words) (Albrecht 2015: 76, Fig. 5)
Functional categories by gender (per one million words)  
(Albrecht 2015: 75, Fig. 4)

Usage of epistemic adverbs in BA and MA theses per 10,000 words  
(Beyer 2015: 93, Fig. 3)

Sentence subject in may-clauses per 100,000 words  
(Küchler 2015: 109)

Figure 2: Frequency of may in the ChemCorpus TP (N=399)  
Figure 3: Frequency of may in the SYSU Corpus (N=901)

- This may be one reason for the implementation of the New Economic Policy (NEP) in 1971.  
- Graduation is concerned with the resources by which the force or intensity of an utterance may be raised or lowered. (CC13FMAT_3)  
- They may lack time or knowledge ... (CC13FMATP_83)  
- Maybe, it is possible to think that both systems may appear to be irrelevant to each other. (CTP12FBALJR_23)

Sentence subject in will-clauses per 100,000 words  
(Küchler 2015: 114, Fig. 10)

Meanings of will per 100,000 words  
(Küchler 2015: 113, Fig. 9)