HS Corpus Linguistics / Korpuslinguistik

5. Representation formats & corpus queries

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

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

Catching up: Overview of existing corpora

FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

PHILOSOPHISCHE FAKULTÄT
UND FACHBEREICH THEOLOGIE

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})

Corpora: English

- British National Corpus — 100 M
  http://www.natcorp.ox.ac.uk/
  - BNC v2 in progress, with texts from around 2015
- Movie subtitles (DESC) — 90 M
- Gigaword newspaper corpus — 4 G
  - current: 5th edition (2011) / 2nd edition ca. 2 G words
    https://catalog.ldc.upenn.edu/LDC2011T07
- New York Times Annotated — 1.2 G
  https://catalog.ldc.upenn.edu/LDC2008T19
  - articles from 1987–2007 with manual categorization
- Corpus of Contemporary AmE (COCA) — 440 M
  http://corpus.byu.edu/coca/
  - only limited access via BYU Web interface
- Wackypedia (English Wikipedia of 2009) — 1 G
  http://wacky.sslmit.unibo.it/doku.php?id=corpora

Corpora: Other languages

- Few reference corpora available (similar to BNC)
  - American National Corpus aborted at 15 M words
    http://www.anc.org/
  - German DeReKo (53 G words) and DWDS (balanced core 100 M, extension 28 G)
    https://cosmas2.ids-mannheim.de/cosmas2-web/ http://www.dwds.de/ressourcen/korpora/
    only with limited Web access
  - Frantext only paid & limited Web access (ca. 200 M words)
    http://www.frantext.fr/
  - Hungarian National Corpus (ca. 100 M words)
    http://corpus.nytud.hu/mnsz/index_eng.html
  - Corpus Brasileiro (ca. 1 G words)
    http://corpusbrasileiro.pucsp.br/cb/Inicial.html
  - most w/o substantial amounts of spoken language

- Newspaper corpora difficult to acquire
  - LexisNexis does not allow systematic download & analysis
    http://www.lexisnexis.com/
  - newspaper publishers often ask steep prices

Corpora: Parallel corpora

- EuroParl debates of the EU Parliament — 10 – 60 M
  http://diates.lingfil.uu.se/Europarl.php
  - parallel corpus with translations into 21 EU languages
  - aligned at sentence level
- OpenSubtitles 2016 — up to 2.5 G
  http://diates.lingfil.uu.se/OpenSubtitles2016.php
  - parallel corpus of movie subtitles in 60 languages
- Parallel Web corpus (linguatools) — ca. 200 M
  http://linguatools.org/tools/corpora/webcrawl-parallel-corpus-german-english-2015/

Corpora: Web corpora

- WaCky (Web as Corpus kool ynitiative) — ca. 2 G
  http://wacky.sslmit.unibo.it/doku.php?id=corpora
  - first publicly available Web corpora (EN, DE, FR, IT)
- Aranea collection — 1 G
  http://sketch.juls.savba.sk/aranea_about/
  - Web corpora in 12 languages
- Corpora from the Web (COW) — 5 – 20 G
  http://corporafromtheweb.org/
  - up-to-date Web corpora in DE, EN, FR, ES, NL, SV
- USENET newsgroup corpus — ca. 7 G
  http://www.psych.ualberta.ca/~westburylab/downloads/usenetcorpus.download.html
  - newsgroup postings from 2005–2011
- Global Web-based English (GloWbE) — ca. 2 G
  http://corpus.byu.edu/glowbe/
  - onle limited Web access via BYU
- TenTen corpus family (≥ 10^{10} tokens in many languages) — up to 36 G
  https://www.sketchengine.eu/documentation/tenten-corpora/
  - only accessible in commercial Sketch Engine
- Crawl your own (specialized) corpus with BootCaT
  http://bootcat.dipintra.it/

Corpus queries

FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

PHILOSOPHISCHE FAKULTÄT
UND FACHBEREICH THEOLOGIE

CQPweb

- https://corpora.linguistik.uni-erlangen.de/cqpweb/
  - Login: `studentX` (1 … 15)
  - Password: `erlangen`

CEQL*

- Background information
  - Hardie (2012); Evert & Hardie (2011)
  - http://cwb.sourceforge.net/

- Documentation: YouTube tutorial videos
  https://www.youtube.com/user/CorpusWorkbench

*http://cwb.sourceforge.net/ceql.php

Other Web UIs @ FAU

- BNCweb
  https://corpora.linguistik.uni-erlangen.de/bncweb/
  CEQL*
  - Login: `studentX` (1 … 15)
  - Password: `erlangen`
  - for use with textbook *Corpus Linguistics with BNCweb – a Practical Guide* (Hoffmann et al. 2008)

- EuroParl debates
  https://corpora.linguistik.uni-erlangen.de/demos/CQP/Europarl/
  CEQL*

- HGC German Newspapers
  https://corpora.linguistik.uni-erlangen.de/demos/auth/HGC/
  CEQL*
  - Login: `demo`
  - Password: `demo`
  - annotated with morphological information

*http://cwb.sourceforge.net/ceql.php

Other Web interfaces using the same CWB technology

- OPUS collection of parallel corpora
  http://diates.lingfil.uu.se/
- Leeds IntelliText (multilingual, Web corpora)
  http://corpus.leeds.ac.uk/itweb/htdocs/Query.html
- BFSU CQPweb (Chinese & English corpora at BFSU)
  http://111.200.194.212/cqp/ http://www.bfsu-corpus.org/channels/corpus
- Linguateca AC/DC (Portuguese)
  http://www.linguateca.pt/ACDC/
- Hungarian National Corpus
  http://corpus.nytud.hu/mnsz/index_eng.html
- Corpus del Español Actual (Spanish)
  http://spanishfn.org/tools/cea/english
- Varitext (French)
  http://syrah.uni-koeln.de/varitext
- Spraakbanken (Swedish)
  http://spraakbanken.gu.se/parole/
- KorpusDK (Danish)
  http://ordnet.dk/korpusdk/
- Georgetown University CQPweb (some free corpora)
  https://corpling.uis.georgetown.edu/cqp/

Other Web interfaces using the same CWB technology

- TSCorpus (Turkish)
  http://tscorpus.com/
- CORIS/CODIS (Italian)
  http://corpora.ficlit.unibo.it/
- SSLMIT La Repubblica (Italian newspapers)
  http://dev.sslmit.unibo.it/corpora/corpus.php?path=&name=Repubblica
- BwanaNet (Catalan, Spanish, English)
  http://bwananet.iula.upf.edu/
- PolMine (German political corpora)
  http://polmine.sowi.uni-due.de/cwb/
- Perugia Corpus (Italian)
  https://www.unistrapg.it/cqpweb/
- CorpusEye (several languages, few free corpora)
  http://corp.hum.sdu.dk/
- CorpusWiki initative (multilingual, still very small)
  http://www.corpuswiki.org/

Further Web interfaces

- **BYU Corpora** (by Mark Davies)
  https://www.english-corpora.org
  - COCA, COHA, Soap Operas, GloWbE, TIME, Spanish, Portuguese, …
- Google Web 1T 5-Grams (n-gram database)
  http://corpora.linguistik.uni-erlangen.de/cgi-bin/demos/Web1T5/Web1T5_freq.perl
  - search n-gram tables, pre-computed (quasi-)collocation
  - NetSpeak offers a nicer Web interface to the database
    http://www.netspeak.org/
- Google Books Ngram Viewer (info)
  https://books.google.com/ngrams/ https://books.google.com/ngrams/info
  - visualize frequency changes over time (words, phrases)
- Linguee: English, German, French
  http://www.linguee.de/ http://www.linguee.com/ http://www.linguee.fr/
  - Web-crawled parallel corpora for many language pairs
  - useful to find possible translations (but *caveat emptor*)
- Treebank.info (automatically parsed corpora)
  http://treebank.info/
- Commercial **Sketch Engine** platform
  https://www.sketchengine.eu/
  - many large & small corpora in different languages
  - free access for master students in EMLex (MA Lexikographie)

Simple query syntax

- Most Web interfaces offer a “simple” query syntax
  - simply type a word or phrase
  - limited support for wildcards
- In this course: CEQL syntax
  - relatively powerful simple query language
  - supported by BNCweb, CQPweb and a few other UIs
- Tutorial & documentation
  - **Ch. 6** of Hoffmann, Sebastian *et al.* (2008). *Corpus Linguistics with BNCweb – a Practical Guide*, vol. 6 of English Corpus Linguistics. Peter Lang, Frankfurt.
  - official documentation: https://cwb.sourceforge.io/ceql.php
  - CQPweb simple query manual
    https://cqpweb.lancs.ac.uk/doc/cqpweb-simple-syntax-help.pdf

CEQL quickstart

• `speak`	matches specific word form
• `{speak}`	matches all inflected forms
• `at the end of the day`	specific phrase
• `is n't it \?`	tokenization rules & escapes
• `*able`	suffix *-able*
• `+able`	without the word *able*
• `light_JJ`	the adjective *light*
• `Mr _N*`	person (male)
• `[Mr,Mrs] _N*`	person (male or female)
• `Mr _N* {be} _J*`	what is said about the person
• `Mr (_N*)+ {be} (_RB)? _J*`	

CEQL quickstart

- `Smith:C` — turn off case folding
- `deja:d vu:d` — ignore diacritics
- `\D` — number (one or more digits)
- `\u\u\u\u:C` — acronym (4 uppercase letters, e.g. *YMCA*)
- `\u\L:C` — starts with uppercase letter
- `take * off` — optional word
- `take ++*** off` — between 2 and 5 words
- `in ( _JJ* )? time` — optional adjective
- `Mr (_N*)+ {be} (_RB)? _J*` — what is said about a person (refined query)
- `his ( _JJS | most _JJ )* _N*` — alternatives
- `<s> but` — start of sentence
- `<ne_type=PERSON> (+)+ </ne_type>`

CQPweb & CEQL practice

- What are the most frequent *uber-* words?
- Search for your favourite topic (one or more lemmas)
- In which year and newspaper is it most frequent?
- Carry out a collocation analysis for this topic
- Find different kinds of numbers and acronyms
- Can you identify predications like *austerity is good*?
- Find different types of named entities
- What are the typical patterns of headlines? (`<title>` …)

CQP query syntax

- Formal query notation
  - based on regular expression at multiple levels
  - allows precise specification of search pattern
  - much more flexible and powerful than CEQL syntax
- Supported by all CWB-based Web interfaces!
- Tutorial & documentation
  - **Ch. 12** of Hoffmann, Sebastian *et al.* (2008). *Corpus Linguistics with BNCweb – a Practical Guide*, vol. 6 of English Corpus Linguistics. Peter Lang, Frankfurt.
  - CQP Query Language Tutorial (online version)
    http://cwb.sourceforge.net/files/CQP_Tutorial.pdf http://cwb.sourceforge.net/files/CWB_Encoding_Tutorial/

CQP queries: single tokens

- Quoted regexp matches surface form of a token
  - `"(over|under)\w+"` or `'(over|under)\w+'`
  - duplicate embedded quotes: `""""` matches "

- Append flags for case/diacritic-insensitive search
  - `"deja"%c` … case-insensitive
  - `"deja"%d` … ignore diacritics
  - `"deja"%cd` … both
  - `"?"%l` … literal string (no metacharacters)

Regular expressions

- Regular expressions (regexp) are a sophisticated formal wildcard notation from computer science, used to describe patterns of characters or other elements
- Fundamental building blocks of regular expressions
  - `(…)?` optional element (0 or 1)
  - `(…)*` any number of repeats, incl. 0 (Kleene star)
  - `(…)+` at least one repetition
  - `(…|…|…)` alternatives
  - nesting of such elements makes regexp very powerful
- CEQL uses regexp notation over *tokens*
  - for optional tokens, repetitions and alternatives
- CQP & full-text search use regexp notation over individual characters (letters, digits, punctuation, …)
  - CQP also uses regexp notation over tokens (→ later)
- Different regexp "flavours": CQP supports PCRE
  - POSIX, PCRE = Perl-compatible regexp, Python, Oniguruma, …

PCRE regular expressions
PCRE = Perl-compatible regular expressions

- `(…)?` = optional (0 or 1)
- `(…)*` = any number of repeats (0 or more)
- `(…)+` = at least one repeat (1 or more)
- `(…){3}` = exactly 3
- `(…){2,4}` = between 2 and 4
  - applies to single character if parentheses are omitted
- `(…|…|…)` = alternatives (matches exactly one)
- `.` = any character (matchall)
  - esp.: `.?` (optional character), `.*` (arbitrary string), `.+`
- escapes: `\.` = ., `\*` = *, `\?` = ?, `\+` = +, …

PCRE regular expressions
PCRE = Perl-compatible regular expressions

- `[aeiou]` = character class (matches exactly one)
  - `[a-z]` = `[abc … z]` and `[A-Z]` = `[ABC … Z]`
  - `[0-9]` = `[0123456789]`
- `[^aeiou]` = everything(!) except `[aeiou]`
- escape sequences:
  - `\w` = letters, digits and _ (word character)
  - `\s` = any single whitespace (blank, TAB, newline, …)
  - `\d` = digit
  - `\pL` = letter, `\p{Ll}` = lowercase, `\p{Lu}` = uppercase
  - `\pN` = digit, `\p{Cyrillic}` = cyrillic letter, …
    - see https://www.pcre.org/original/doc/html/pcrepattern.html#SEC5

CQP queries: single tokens

- Search token annotation with attribute-regexp pair:
  - `[lemma = "(over|under)\w+_ADJ"]` (BNC)
  - `[pos = "AJS"]` … superlatives (BNC)
  - `"deja"%cd` is shorthand for `[word = "deja"%cd]`

- Combine constraints with Boolean operators:
  - operators: `&` (and), `|` (or), `!` (not), `!=` (doesn't match)

  - `[(word="can"%c) & (pos!="VM.*")]`
  - same as: `[(word="can"%c) & !(pos="VM.*")]`

token description

- All examples for BNCweb with CLAWS tagset

CQP queries: token sequences

- CQP queries are regular expressions over token descriptions (`[…]`)
  - `"in" [pos="AJ.*"]? [hw="year"]` … optional
  - `"in" [pos="AJ.*"]+ [hw="year"]` … one or more
  - `"in" [pos="AJ.*"]{2} [hw="year"]` … exactly two
  - `([pos="AJS"] | "most"%c [pos="AJ0"])` … either

- Skipping arbitrary tokens
  - `[]` … matchall (any token)
  - `"dog" []{0,4} "cat"` … within 5-token span
  - `"dog" []{0,4} "cat" within s` … must not cross a sentence boundary (s-attribute)

CQP queries: s-attributes

- XML tags match start/end of s-attribute regions
  - `<head> "UK"` … as first word of heading
  - `"UK" </head>` … as last word of heading
  - `<head>` … doesn't match anything (0 tokens)
  - `<mw> []* </mw>` … paired tags match entire region

- Search within a region:
  - `"Twain" within quote;`
  - `[pos="NN.*"] :: match.mw_pos = "PRP";`
    … add “global constraint” to check s-attribute annotation
  - pre-defined anchors: `match`, `matchend`, `target` (`@`)

CQP queries: token sequences

- Repetition operators and alternatives can be nested to search for complex lexicogrammatical patterns:

```
([pos="AJS"] | "most"%c [pos="AJ0"])
(
  "(and|\,)"%c
  ([pos="AJS"] | "most"%c [pos="AJ0"])
)+
[pos = "NN.*"]
```

What does this query do?

- Matching strategy defaults to non-greedy
  - `"ho"%c (","  "ho"%c)+` … always matches *ho, ho*
  - `(?longest) "ho"%c (","  "ho"%c)+`
    … recent CQP versions support inline modifier at start of query

CQP query practice

- Find noun compounds / names with 4+ components
  - What are the longest compounds/names in the BNC?
- Find bare nouns (e.g. *went to school*)
- Find co-occurrences of *coffee* and *drink* (5-word span)

- Find verb-object combinations (active voice)
  - design flexible pattern for matching noun phrases
  - don't forget about phrasal verbs and adverbs

- What are the typical patterns of headlines?
  - Does your query account for all headlines in the BNC?

- Can you find inflected forms of verbs (≠ base form)?
  - hint: `normalize(word, "c")` → lowercased word form