FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

PHILOSOPHISCHE FAKULTÄT
UND FACHBEREICH THEOLOGIE

EMLex Unit B1.5
An introduction to corpus linguistics

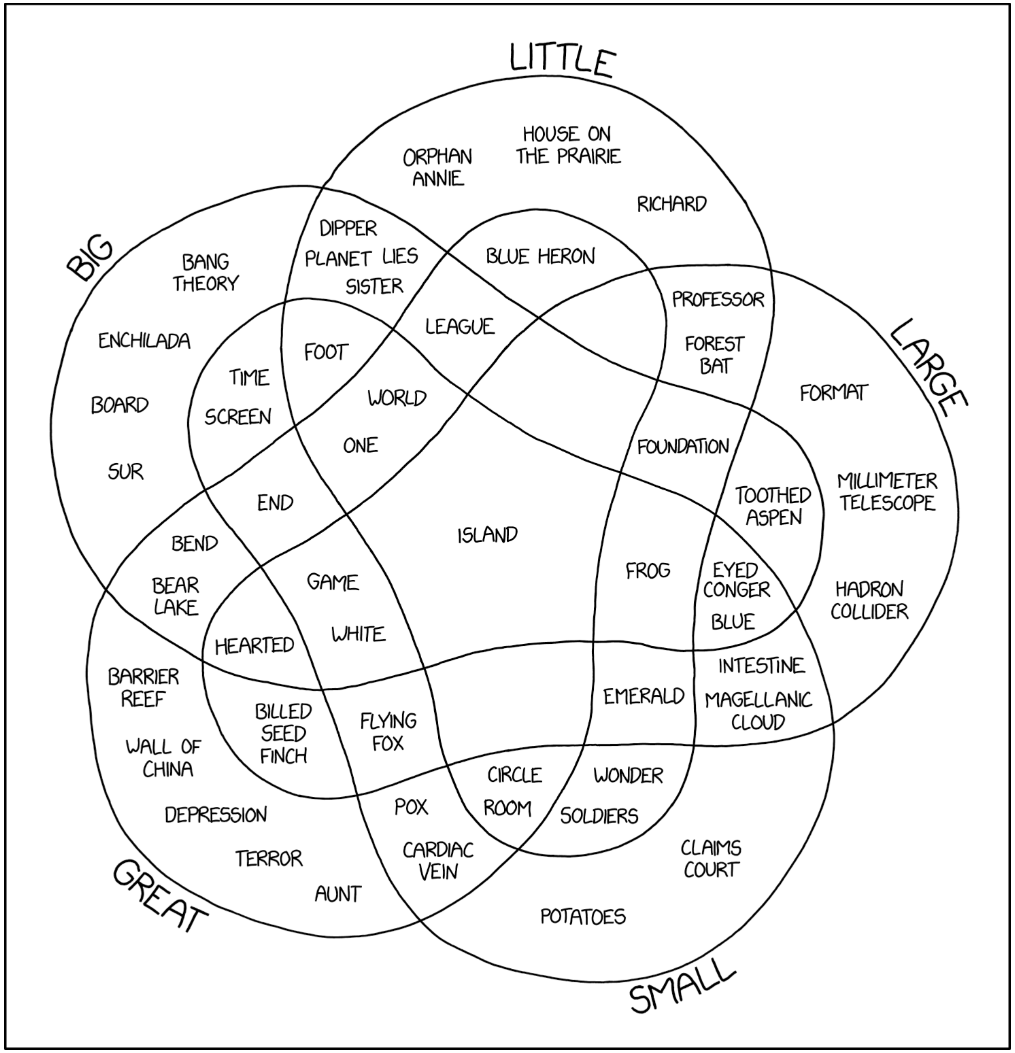
Stefan Evert | FAU Erlangen-Nürnberg

www.linguistik.fau.de | www.stefan-evert.de

What is corpus linguistics?

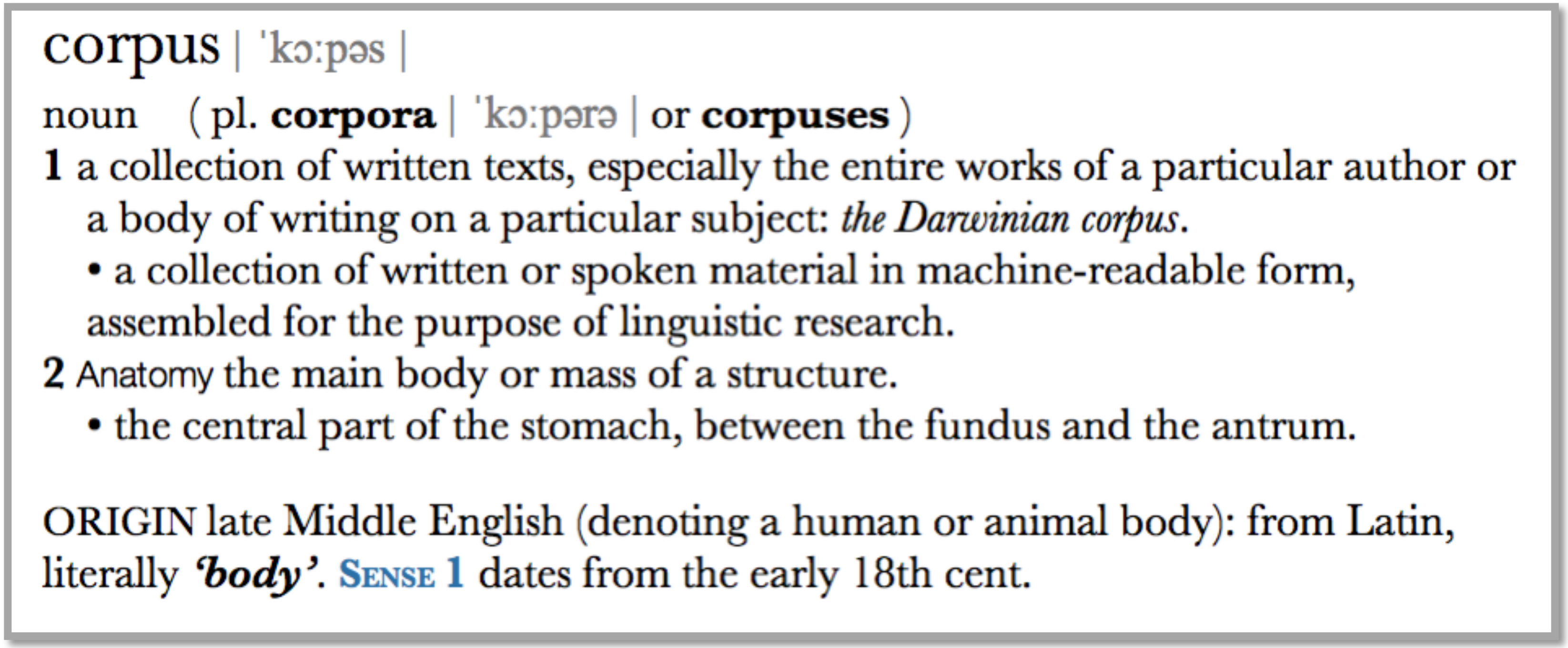
In a nutshell …

The empirical
scientific study
of language
based on
authentic samples
of language use



https://xkcd.com/2122/

What is a corpus?



corpus | ˈkɔːpəs |

noun (pl. **corpora** | ˈkɔːpərə | or **corpuses**)

1 a collection of written texts, especially the entire works of a particular author or a body of writing on a particular subject: *the Darwinian corpus*.

• a collection of written or spoken material in machine-readable form, assembled for the purpose of linguistic research.

2 Anatomy the main body or mass of a structure.

• the central part of the stomach, between the fundus and the antrum.

ORIGIN late Middle English (denoting a human or animal body): from Latin, literally ***'body'***. **SENSE 1** dates from the early 18th cent.

- Corpus as electronic text collection (archive)
  - wide sense: often used in computational linguistics / NLP
- Corpus as representative sample of language
  - narrow sense: dominant in corpus linguistics

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 (in M = million running words)
  - small & clean vs. large & messy

History of corpus linguistics

- First corpus-based quantitative studies in late 19th century
- Orthography and frequency lists
  - Kaeding (1897): German frequency dictionary based on corpus of approx. 11 million words (completely manual analysis!)
- Lexicography
  - Murray: several million index cards for OED (1879–1828)
- Language acquisition
  - first longitudinal studies ca. 1876–1926 (parent diaries)
  - large cross-sectional studies ca. 1927–1957
- Foreign language teaching
  - basic vocabulary, vocabulary levels, collocations (e.g. Palmer 1933)
- Strucuralist language documentation
  - Boas (1940), Firth (1930–1955), …

History of corpus linguistics

- Since 1950: Humanities Computing (➞ Digital Humanities)
  - Index Thomisticus by Robert Busa & IBM (1949–)
- 1950–1960: Mechanolinguistics (Juilland: contrastive corpora) ➞ quantitative / mathematical linguistics (Harris 1968)
- 1960–1980: Corpus linguistics as European counter-movement against mainstream of generative linguistics
  - 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), COBUILD
    - principles of collocation & colligation, "trust the text"
- Since 1990: CL established as subdiscipline of linguistics

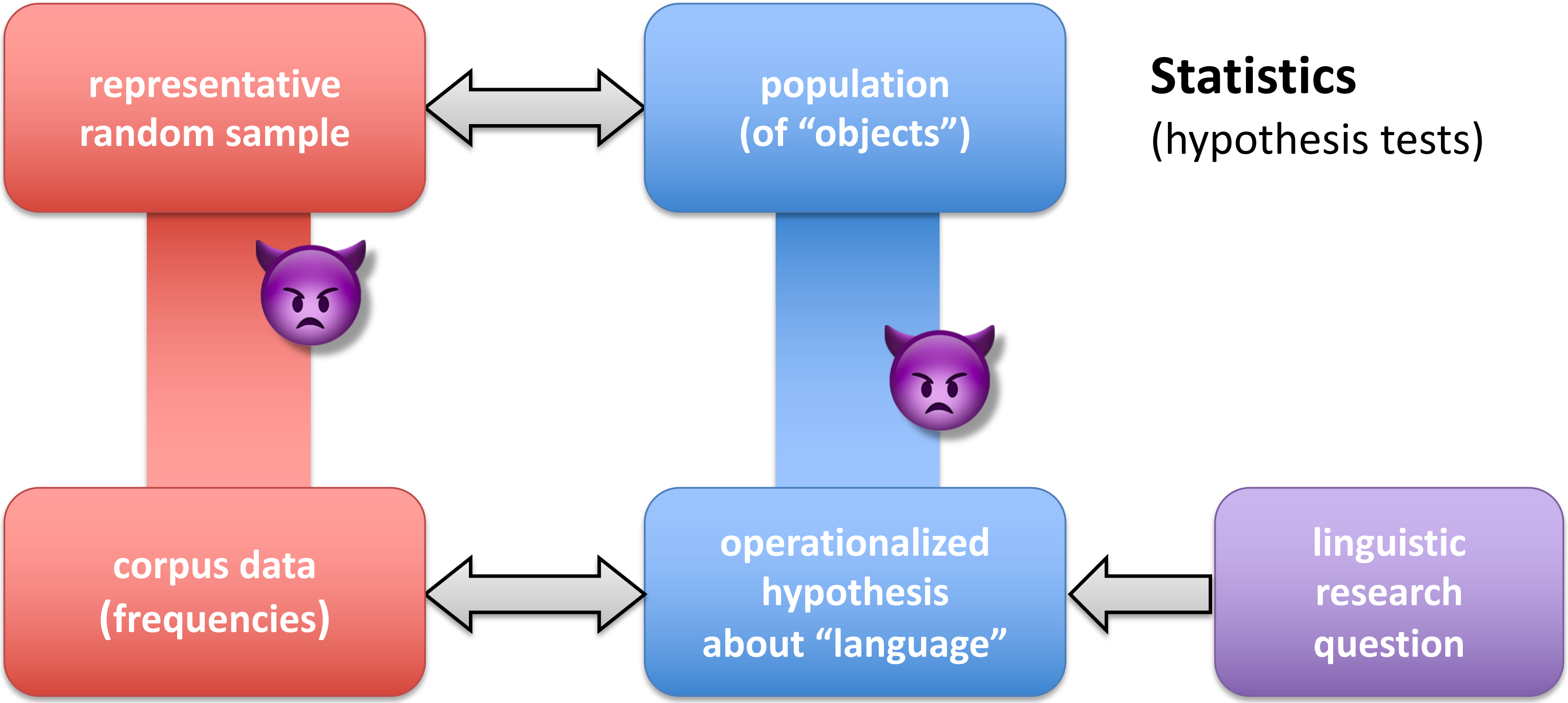
Applications of corpus linguistics

- Dialectology & contrastive linguistics
- Historical linguistics & language change
- Language description (e.g. endangered languages)
- Language teaching, language acquisition, CALL
- Lexical semantics
- Language variation, register studies (→ Biber's MDA)
- Lexicography & lexicology
- Morphology (→ quantitative productivity)
- Phonology (esp. studies of phonological variation)
- Pragmatics & discourse analysis (→ rhetoric, ideology, politics, …)
- Sociolinguistics (e.g. gender studies)
- Stylometrics & literary studies
- Syntax & grammar
- Translation studies (→ “translationese”, CAT)

Goals of corpus design

- Representativeness
  - a corpus should be representative of the (sub-)language to be studied
  - full representativeness difficult to achieve
  - must at least be balanced
    (= good coverage of different registers, speakers, …)
- Comparability
  - corpus studies often build on frequency comparison
  - prerequisite: comparable corpora

Why representativeness?


representative random sample
population (of “objects”)
corpus data (frequencies)
operationalized hypothesis about “language”
linguistic research question

Statistics
(hypothesis tests)

Representativeness & sampling

- Design criteria → sampling frame
  - defines the population from which sample is taken
  - "A sampling frame is an operational definition of the population, an itemized listing of population members from which a representative corpus can be chosen." (Biber 1993, 244)
- Specification of sampling frame
  - mode (spoken/written), genre / text type, domain, …
  - publication date, region, medium, target audience, …
  - properties of speakers: sex, age, dialect, social class, …
- Balance: include texts from all (combns of) categories

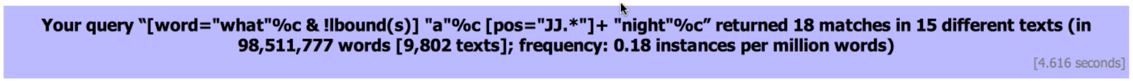
The stages of a corpus study

1. Operationalization
   – hypothesis, definition of population (*sampling frame*)
2. Corpus compilation
   – selection of texts (from sampling frame), digitization / format conversion
   – collection of metadata, legal & ethical issues
3. Representation format
   – standards: Unicode, XML, TEI, XCES, …
   – important for archiving and data exchange
4. Linguistic annotation
   – manual annotation, GUI, annotator agreement
   – automatic annotation with NLP tools for larger corpora

The stages of a corpus study

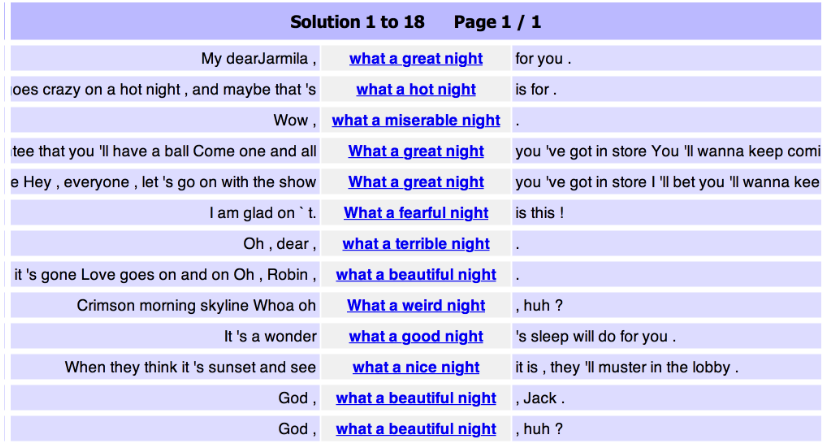
5. Indexing & search

- search for keyword, phrase, linguistic pattern,
- view results as concordance ("kwic" = keyword in context)
- efficient search usually 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 techniques, data analysis, visualisation

7. Interpretation

Pre-compiled corpora

- Corpus compilation is expensive & time-consuming
  - 100M words hardly feasible for individual researcher
- Stages 1 – 4 can be skipped if a pre-compiled corpus is used for the study
  - often involves taking a subset of the corpus (= subcorpus) that matches desired sampling frame
- Additional linguistic annotation may be needed
- Issues: accessibility, licensing conditions, fees

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 German, English, French and Italian
  - other Web as Corpus projects cover additional languages

Corpora for lexicography: English

- British National Corpus — 100 M
  http://www.natcorp.ox.ac.uk/
  - BNC v2 in progress, with texts from around 2015
- Movie subtitles (Erlangen: 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 for lexicography: Other languages

- Few reference corpora available (similar to BNC)
  - American National Corpus aborted at 15 M words
    http://www.anc.org/
  - German DeReKo (28 G words) and DWDS (balanced core 100 M, extension 2.5 G) only with limited Web access
    https://cosmas2.ids-mannheim.de/cosmas2-web/ http://www.dwds.de/ressourcen/korpora/
  - 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 for lexicography: 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 for lexicography: 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 14 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
- Crawl your own (specialized) corpus with BootCaT
  http://bootcat.dipintra.it/

Textbooks & handbooks

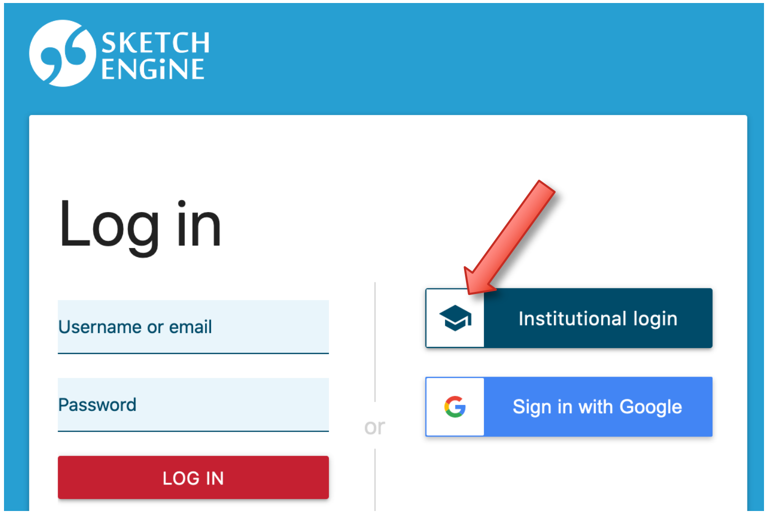
- McEnery, Tony and Hardie, Andrew (2012). *Corpus Linguistics: Method, Theory and Practice*. Cambridge University Press, Cambridge.
- McEnery, Tony; Xiao, Richard; Tono, Yukio (2006). *Corpus-Based Language Studies: An advanced resource book*. Routledge, London and New York.
- Hoffmann, Sebastian *et al.* (2008). *Corpus Linguistics with BNCweb – a Practical Guide*, vol. 6 of English Corpus Linguistics. Peter Lang, Frankfurt.
- Lüdeling, Anke and Kytö, Merja (eds.) (2008). *Corpus Linguistics. An International Handbook*. HSK 29. Walter de Gruyter, Berlin, New York.
- Gouws, Rufus H.; Heid, Ulrich; Schweickard, Wolfgang; Wiegand, Herbert Ernst (eds.) (2013). *Dictionaries. An International Encyclopedia of Lexicography. Supplementary volume: Recent Developments with Focus on Electronic and Computational Lexicography*, HSK 5.4. De Gruyter Mouton, Berlin. Chapters XVIII + XIV.

CL research community

- Important conferences
  - Corpus Linguistics (Lancaster / Birmingham)
  - ICAME
  - 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 portals
  - David Lee's bookmarks: http://tiny.cc/corpora
  - Linguistics Web (S. Bartsch): http://www.linguisticsweb.org/

Sketch Engine

- https://auth.sketchengine.eu/
  - large selection of corpora in many languages available
  - upload / create own corpora up to 1 M words
- Free non-commercial access (until 03/2022)
  - select Institutional Login
  - enter home university account & password on SSO login page
  - agree to share your information
  - may need to create SkE account
- User guide: https://www.sketchengine.eu/user-guide/


SKETCH ENGINE
Log in
Username or email
Password
LOG IN
or
Institutional login
Sign in with Google