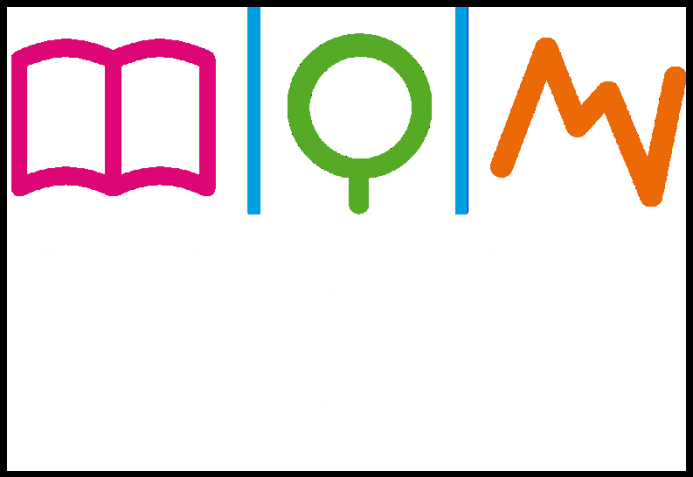


CZECH NATIONAL CORPUS

Introduction to Text Corpora and Their Applications

History and development of language corpora

Lucie Chlumská, Ph.D.

lucie.chlumska@korpus.cz

OUTLINE:

1. LECTURE

- where did the ‘corpus’ come from in the first place?
- pre-electronic corpora and the rise of computers
- 1st generation corpora
- 2nd generation corpora and beyond

2. SEMINAR

- reading (W. Teubert): the nature of language
- what is a language...? what is a discourse...?

LECTURE

The origins of corpus and corpus linguistics

Corpus is...

http://www.oed.com/view/Entry/41873?redirectedFrom=corpus&

Is corpus just a text archive?

Corpus is:

- assembled with particular purposes in mind and often *representative* of some language or text type (Leech 1992)
- machine-readable and usually *annotated*
- a *sample* of language

Corpus linguistics

Is corpus linguistics a discipline?

CL:

- does not concentrate on one single aspect of language (such as syntax, psycholinguistics or sociolinguistics etc.)
- focuses on large amounts of authentic data and strives to observe patterns in language (in the broadest sense of the word)
- works with frequencies, co-occurrences, collocationsetc.
- focuses on meaning and context rather than grammar (unlike Chomsky)

Pre-electronic corpora

The beginnings

First attempts to collect data similar to corpora (before 1960s) were made in the following areas:

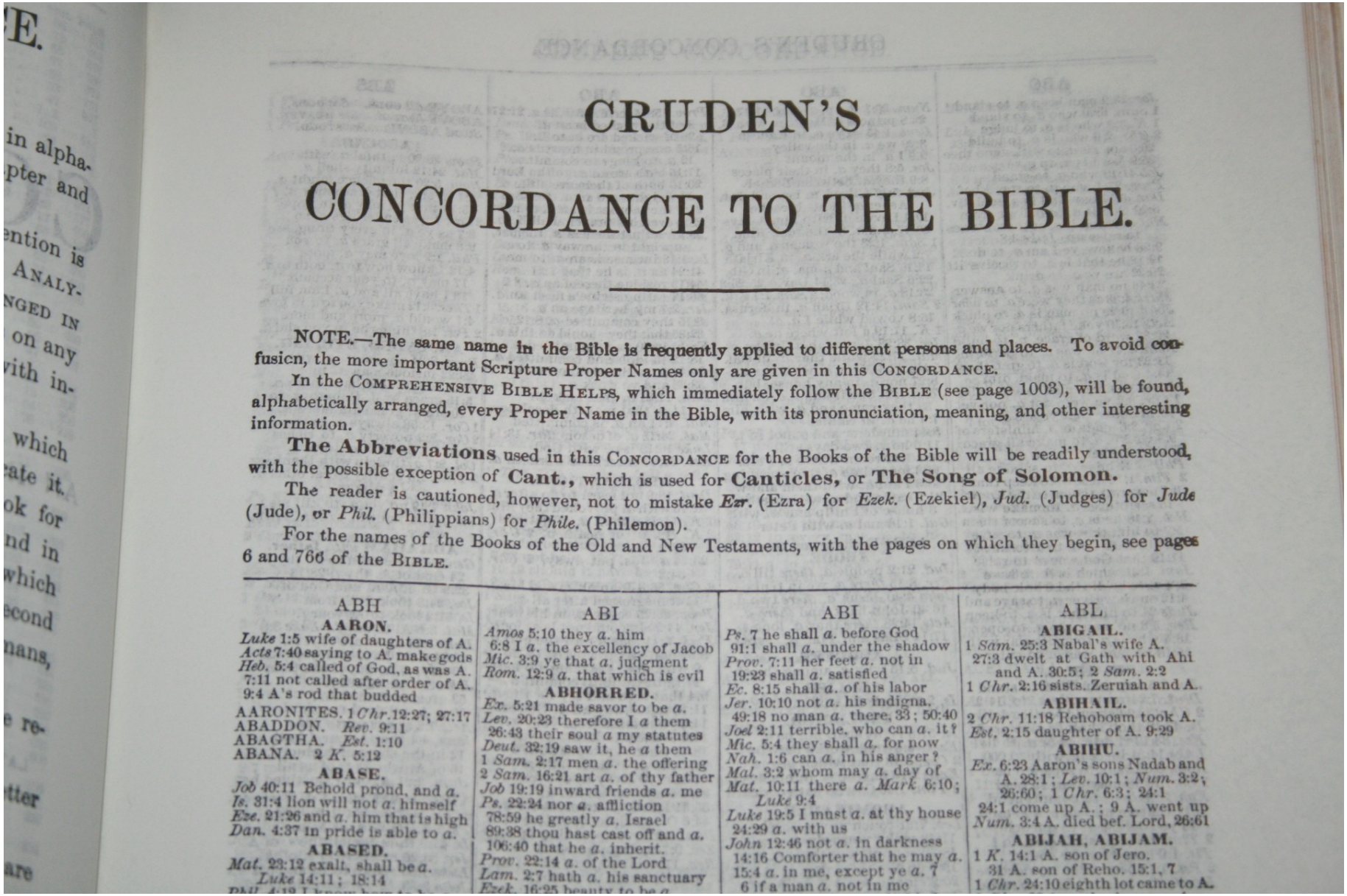
- biblical and literary studies
- lexicography
- dialect studies
- language education studies
- grammatical studies

Biblical and literary studies

- Bible as a corpus on which to base commentaries or criticism, in the form of **word lists and concordances**

- in 18th century, Alexander Cruden, a London bookseller, proofreader and morals campaigner, produced the most famous of these for the Authorized (King James) Version of Bible
  - *Cruden's Concordance*, first published in 1736, 42 editions before 1879!
  - it included concordances not only for the main content words in the Bible, but also some function words and phrases (*how*, *you*, *he*, *once*, *how much*, *all the nations* etc.)

Biblical and literary studies



E.

in alpha-
pter and

ention is
ANALY-
NGED IN
on any
with in-

which
ate it.
ok for
nd in
which
econd
nans,

e re-

tter

are

CRUDEN'S

CONCORDANCE TO THE BIBLE.

NOTE.—The same name in the Bible is frequently applied to different persons and places. To avoid confusion, the more important Scripture Proper Names only are given in this CONCORDANCE.

In the COMPREHENSIVE BIBLE HELPS, which immediately follow the BIBLE (see page 1003), will be found, alphabetically arranged, every Proper Name in the Bible, with its pronunciation, meaning, and other interesting information.

The Abbreviations used in this CONCORDANCE for the Books of the Bible will be readily understood, with the possible exception of **Cant.**, which is used for **Canticles**, or **The Song of Solomon.**

The reader is cautioned, however, not to mistake ***Ezr.*** (Ezra) for *Ezek.* (Ezekiel), *Jud.* (Judges) for *Jude* (Jude), or *Phil.* (Philippians) for *Phile.* (Philemon).

For the names of the Books of the Old and New Testaments, with the pages on which they begin, see pages 6 and 766 of the BIBLE.

ABH

AARON.
Luke 1:5 wife of daughters of A.
Acts 7:40 saying to A. make gods
Heb. 5:4 called of God, as was A.
7:11 not called after order of A.
9:4 A's rod that budded
AARONITES. 1 *Chr.* 12:27; 27:17
ABADDON. *Rev.* 9:11
ABAGTHA. *Est.* 1:10
ABANA. 2 *K.* 5:12

ABASE.
Job 40:11 Behold proud, and *a.*
Is. 31:4 lion will not *a.* himself
Eze. 21:26 and *a.* him that is high
Dan. 4:37 in pride is able to *a.*

ABASED.
Mat. 23:12 exalt, shall be *a.*
Luke 14:11; 18:14

ABI

Amos 5:10 they *a.* him
6:8 I *a.* the excellency of Jacob
Mic. 3:9 ye that *a.* judgment
Rom. 12:9 *a.* that which is evil

ABHORRED.
Ex. 5:21 made savor to be *a.*
Lev. 20:23 therefore I *a* them
26:43 their soul *a* my statutes
Deut. 32:19 saw it, he *a* them
1 *Sam.* 2:17 men *a.* the offering
2 *Sam.* 16:21 art *a.* of thy father
Job 19:19 inward friends *a.* me
Ps. 22:24 nor *a.* affliction
78:59 he greatly *a.* Israel
89:38 thou hast cast off and *a.*
106:40 that he *a.* inherit.
Prov. 22:14 *a.* of the Lord
Lam. 2:7 hath *a.* his sanctuary
Ezek. 16:25 beauty to be *a.*

ABI

Ps. 7 he shall *a.* before God
91:1 shall *a.* under the shadow
Prov. 7:11 her feet *a.* not in
19:23 shall *a.* satisfied
Ec. 8:15 shall *a.* of his labor
Jer. 10:10 not *a.* his indigna.
49:18 no man *a.* there, 33; 50:40
Joel 2:11 terrible, who can *a.* it?
Mic. 5:4 they shall *a.* for now
Nah. 1:6 can *a.* in his anger?
Mal. 3:2 whom may *a.* day of
Mat. 10:11 there *a.* *Mark* 6:10;
Luke 9:4
Luke 19:5 I must *a.* at thy house
24:29 *a.* with us
John 12:46 not *a.* in darkness
14:16 Comforter that he may *a.*
15:4 *a.* in me, except ye *a.* 7
6 if a man *a.* not in me

ABL

ABIGAIL.
1 *Sam.* 25:3 Nabal's wife A.
27:3 dwelt at Gath with Ahi
and A. 30:5; 2 *Sam.* 2:2
1 *Chr.* 2:16 sists. Zeruiah and A.

ABIHAIL.
2 *Chr.* 11:18 Rehoboam took A.
Est. 2:15 daughter of A. 9:29

ABIHU.
Ex. 6:23 Aaron's sons Nadab and
A. 28:1; *Lev.* 10:1; *Num.* 3:2;
26:60; 1 *Chr.* 6:3; 24:1
24:1 come up A.; 9 A. went up
Num. 3:4 A. died bef. Lord, 26:61

ABIJAH, ABIJAM.
1 *K.* 14:1 A. son of Jero.
31 A. son of Reho. 15:1, 7
1 *Chr.* 24:10 eighth lot came to A.

CZECH NATIONAL CORPUS

Lexicography

- as early as 17th century!
- Samuel Johnson recorded on slips of paper a large corpus of sentences from ‘writers of the first reputation’ to illustrate meanings and uses of English words in his *Dictionary of the English Language*
  - Johnson worked with 6 assistants to assemble over 150,000 illustrative citations for the app. 40,000 headword entries
- similarly, *Oxford English Dictionary* (*OED*) also corpus-based
  - twelfth and final volume published in 1928
  - 71 years of sustained work on a corpus of the canon of mainly literary written English from about AD 1000
  - 2,000 volunteer readers collected about five million citations amounting to 50 million words to illustrate 414,825 entries

Lexicography

- Compiling the *OED*...

- How it began 1857: The Philological Society of London calls for a new English Dictionary
- More work than they thought 1884: Five years into a proposed ten-year project, the editors reach *ant*
- One step at a time 1884-1928: The Dictionary is published in fascicles
- Keeping it current 1933-1986: Supplements to the *OED*
- Making it modern 1980s: The Supplements are integrated with the *OED* to produce its Second Edition
- Into the electronic age 1992: The first CD-ROM version of the *OED* is published
- The future has begun The present: The *OED* is now being fully revised, with new material published in parts online

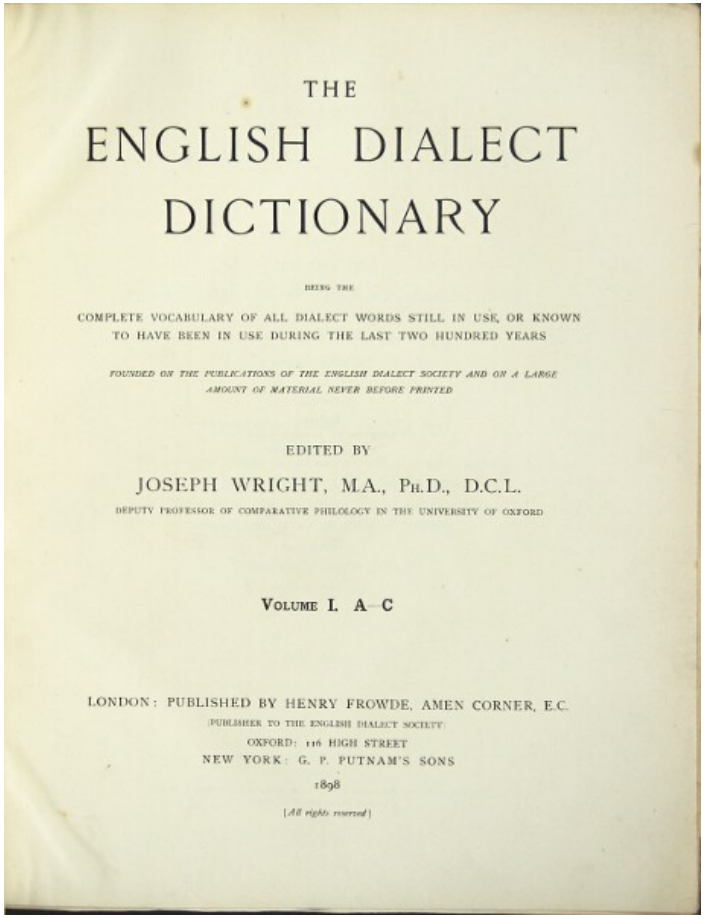
Lexicography

- parallel to the work on the second edition of *OED* in the latter part of 19th century, another great corpus of citations was being assembled to support the third edition of Noah Webster's *An American Dictionary of the English Language*

  - in 1961, the third edition of Webster's *New International Dictionary* had available a corpus of over 10 million citation slips
  - probably the last major English dictionary to be completed without an electronic database...

Dialect

- interest in linguistic variation in regional dialects in 19th century
  - the *English Dialect Dictionary* (Wright, 1898–1905)
  - *The Existing Phonology of English Dialects* (A. Ellis, 1889)



THE

ENGLISH DIALECT DICTIONARY

BEING THE

COMPLETE VOCABULARY OF ALL DIALECT WORDS STILL IN USE, OR KNOWN TO HAVE BEEN IN USE DURING THE LAST TWO HUNDRED YEARS

FOUNDED ON THE PUBLICATIONS OF THE ENGLISH DIALECT SOCIETY AND ON A LARGE AMOUNT OF MATERIAL NEVER BEFORE PRINTED

EDITED BY

JOSEPH WRIGHT, M.A., PH.D., D.C.L.

DEPUTY PROFESSOR OF COMPARATIVE PHILOLOGY IN THE UNIVERSITY OF OXFORD

VOLUME I. A–C

LONDON: PUBLISHED BY HENRY FROWDE, AMEN CORNER, E.C.

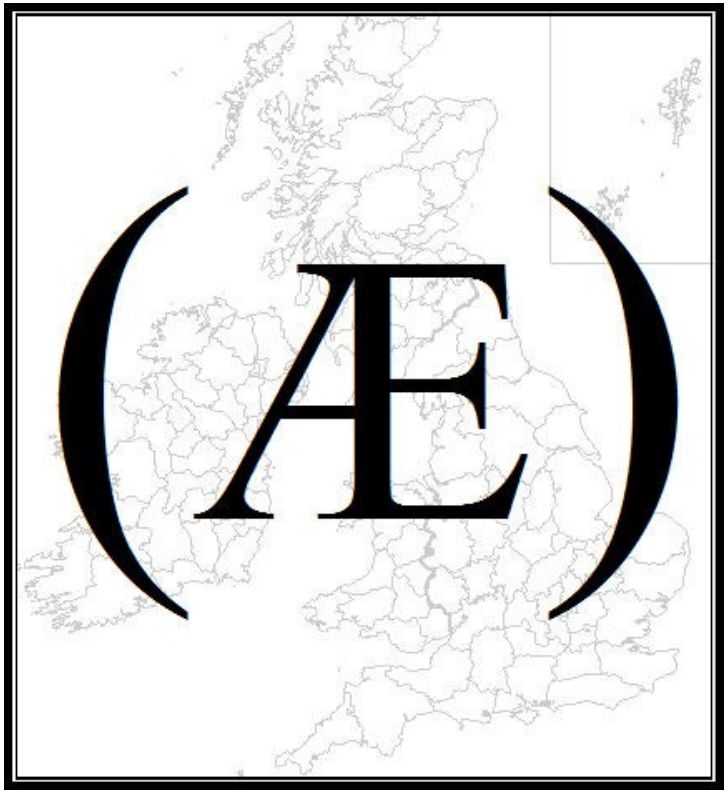
PUBLISHER TO THE ENGLISH DIALECT SOCIETY

OXFORD: 116 HIGH STREET

NEW YORK: G. P. PUTNAM'S SONS

1898

[*All rights reserved*]



Language education

- some of the most influential corpus-based research in the 1st half of 20th century had a **pedagogical purpose**
  - Edward Thorndike (1921) compiled a corpus of 4.5 million words from 41 sources to make a frequency list (for better curricula materials for teaching literacy to native speakers of English in the US)
    - ¾ Bible and classic fiction, ¼ letters, newspapers, school readers
  - a corpus to gather statistical information of German words and letters in order to improve the stenographers' tranining
    - J. W. Kaeding needed help of 5,000 assistants over a period of years to process the corpus of 11 million words he used in 1890s for his analysis

Grammar

- Otto Jespersen, Danish professor, is said to have his villa filled with shoeboxes containing hundreds of thousands of paper slips with examples of interesting English sentences
  - monumental work *A Modern English Grammar on Historical Principles* (1909–49)

- Charles C. Fries used a corpus of letters written to the US government by persons of different educational and social backgrounds to demonstrate social class differences in usage in his *American English Grammar* (1940)
- Later in *The Structure of English* (1952) he used a 250,000-word corpus of recorded telephone conversations
- he analyzed all his corpora manually...

Grammar

- the most important pre-electronic corpus was the *Survey of English Usage Corpus* (*SEU*) by Randolph Quirk et al. in 1968
- it marked a transition between earlier non-computerized corpus-based description and the rise of corpus linguistics
- founded in 1959 by Quirk, the SEU aimed to collect 200 samples (each about 5,000 words) representative of both written and spoken language > corpus of 1 million words
  - SEU Corpus contains texts produced between 1953 and 1987, originally available in the form of paper slips filed at the University College London
  - there was a slip for every word in the corpus, containing 17 lines of text plus a mark-up (grammatical features, prosody...)
  - basis for the *A Comprehensive Grammar of the English Language* (1985)

The history of electronic corpora

Three stages of electronic corpora building

1. 1960–1980 : learning how to build and maintain corpora up to a million words; no material available in electronic form
2. 1980–2000:
   - the 1980s: decade of the scanner
   - the 1990s, or the *First Serendipity*, text becomes available as the by-product of computer type-setting
3. the new milleninum, or the *Second Serendipity*, texts that never existed as hard copies become available in unlimited quantities from the internet

The first electronic corpora

- the Brown corpus
  - compiled in the 1960s at Brown University by Nelson Francis and Henry Kučera
  - a million words of American English from documents published in 1961
  - 500 samples of about 2,000 words
  - a broad range of categories (informative prose: 374 samples, imaginative prose: 126 samples)

The first electronic corpora

- the LOB corpus
  - 1970–1978 at the Lancaster and Oslo University
  - a comparable counterpart to the American Brown corpus
  - a million words of British English from documents published in 1961
  - 500 samples of about 2,000 words
  - better technology and tools (sentence-initial markers etc.)

- FROWN and FLOB: Freiburg University corpora from 1992

The second generation

- the BNC
  - 1991–1994 (BNC XML Edition from 2007)
  - 100-million word collection of samples of written and spoken English
  - monolingual, synchronic, general and sample corpus
  - 4,054 texts
  - 90 % written part, 10 % spoken part
  - spoken language: demographic and context-governed

Modern corpora

- larger, better annotated & lemmatized etc.
- monitor corpora
  - COCA (1990–2015): 520 million tokens
  - NOW (2010–yesterday): 5.1 billion+
- web corpora
  - include texts from the internet: WaC, Aranea, EnTenTen, SYN, Gigafida etc.

Size of corpora and related resources

- 1st generation = small (*1-5 million*): Brown corpus and Brown family corpora (LOB, Frown, FLOB)
- 2nd generation = moderately-sized (*100 million*): BNC, SYN2000
- larger, more up-to-date corpora (*cca 450 million*): COCA
- extremely large text archives (billions): Google Books
- the Web as Corpus

But does size matter?

Frequency of different phenomena

Table 1.2 *Frequency of different phenomena in COCA, BNC, and Brown (numbers explained in detail in Sections 3.1–3.5)*

		COCA (450 m)	BNC (100 m)	Brown (1 m)
1	Lexical: individual	(See discussion in Section 3.1 above)		
2	Lexical: word lists	100,705	43,758	3,956
3	Morphology: substrings	*-ousness* 112 *-ism* 512	*-ousness* 25 *-ism* 278	*-ousness* 1 *-ism* 6
4	Morphology: compare	*prove{n/d}* 2,616 + 3,001 *sincere* 85 + 65	*prove{n/d}* 82 + 1,169 *sincere* 11 + 12	*prove{n/d}* 3 + 7 *sincere* 1 + 0
5	Syntax: high frequency	modals 5,794k perfects 1,837k *be* passives 2,900k	modals 1,421k perfects 446k *be* passives 890k	modals 14k perfects 4k *be* passives 10k
6	Syntax: low frequency	*love* 12,178 + 5,393 *hate* 3,968 + 1,773 *for* 931	*love* 1,192 + 351 *hate* 389 + 475 *for* 103	*love* 10 + 2 *hate* 8 + 2 *for* 0
7	Phraseology: words	*true feelings* 654 *naked eye* 175	*true feelings* 148 *naked eye* 53	*true feelings* 2 *naked eye* 0
8	Phraseology: constructions	*way* 251v : 15,868t *into* 275v : 2,160t	*way* 83v : 3,533t *into* 111v : 358t	*way* 15v : 44t *into* 6v : 6t
9	Semantics: collocates	*riddle* (n) 57 *nibble* (v) 96 *crumbled* (j) 33 *serenely* (r) 24	*riddle* (n) 0 *nibble* (v) 13 *crumbled* (j) 1 *serenely* (r) 4	*riddle* (n) 0 *nibble* (v) 0 *crumbled* (j) 0 *serenely* (r) 0
10	Semantics: prosody	*budge* (v) 1,427 *cause* (v) 1,344	*budge* (v) 164 *cause* (v) 358	*budge* (v) 3 *cause* (v) 0

Let's talk language!

SEMINAR

Reading

- common reading:

Teubert, W. (2004). Corpus linguistics: a different look at language. In M. A. K. Halliday, W. Teubert, C. Yallop and A. Čermáková, *Lexicology and Corpus Linguistics,* pp 96-112. NY: Continuum.

- additional reading recommendations:

http://www.grsampson.net/ARsy.pdf

http://www.vox.com/the-big-idea/2016/9/14/12910180/noam-chomsky-tom-wolfe-linguist

https://aeon.co/essays/the-evidence-is-in-there-is-no-language-instinct

Discussion

- What is a language?
- What is meaning? How can we discover a meaning of a word?
- What do we understand under the term *discourse*?
- Can we say that: what cannot be found in a corpus, does not exist in language?
- Why did Noam Chomsky object against corpora and corpus linguistics?
- What is and is not corpus linguistics good for?
- What are the main differences between a dictionary and a corpus?
- What was first, written or spoken language?
- What is a word?

https://www.futurelearn.com/courses/corpus-linguistics?lr=4