



Corpus

Stephen Miller
AAC / ÖAW
Texts & Files WS 2005/06
Universität Wien



Definitions / Issues

What is a Corpus?

“A collection of linguistic data, either written texts or a transcription of recorded speech, which can be used as a starting-point of linguistic description or as a means of verifying hypotheses about a language.”

David Crystal, *A Dictionary of Linguistics and Phonetics*, Blackwell, 3rd Edition, 1991.

3

Text and Files WS 2005/06 / Corpus

© AAC / ÖAW 2005

What is *not* a Corpus...

o Distinction made by Leech (1991):

- Archive
 - Unordered collection of data
- Corpus
 - Principled linguistic snapshot of language at a given point in time

4

Text and Files WS 2005/06 / Corpus

© AAC / ÖAW 2005

What is a Corpus?

“A collection of naturally occurring language text, chosen to characterize a state or variety of a language.”

John Sinclair, *Corpus Concordance, Collocation*, OUP, 1991.

5

Text and Files WS 2005/06 / Corpus

© AAC / ÖAW 2005

Chosen how?

“There is no consensus in the community as to the procedures to be followed in corpus design (balanced, opportunistic, statistically sophisticated and defiantly naive approaches all struggle with each other for acceptance) [...]”

C.M. Sperberg-McQueen (1994)

6

Text and Files WS 2005/06 / Corpus

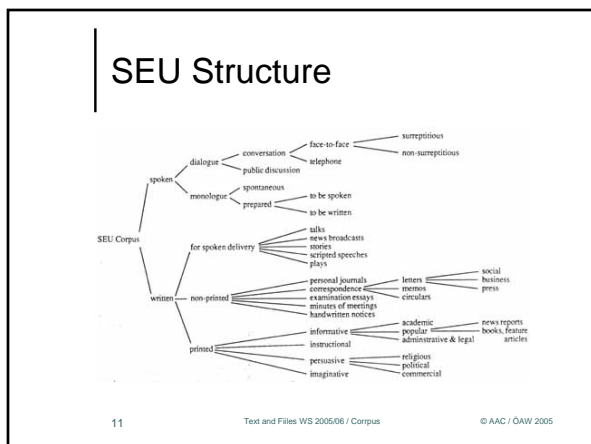
© AAC / ÖAW 2005

Development of Corpora

- ## Phases of Corpora
- Pre-electronic Corpora
 - 1st generation Major Corpora
 - 2nd generation “Mega-Corpora”

Pre-electronic Corpora

- ## Pre-electronic Corpora
- Pre-electronic Corpora
 - Biblical Concordances
 - *Survey of English Usage*
 - UCL
 - 1m words
 - 200 x 5000 words
 - written/spoken



Advent of the Computer

The computer “gives us the ability to comprehend, and to account for, the contents of such corpora in a way which was not dreamed of in the pre-computational era of corpus.”
Leech (1992)

Helsinki Corpus of English Texts

- o [^ TEXT: ALFRED'S PREFACE TO CURA PASTORALIS. KING ALFRED'S WEST-SAXON VERSION OF GREGORY'S PASTORAL CARE, PART I. EARLY ENGLISH TEXT SOCIETY, O.S. 45. ED. H. SWEET. LONDON, 1958 (1871). PP. 3.1 - 9.7^] [^B9.1.1^] <P 3> <R 1> []+DEOS BOC SCEAL TO WIOGORA CEASTRE.]) +alfred kyning hate+d gretan W+arfer+d biscep his wordum luflice & freondlice; <R 2> & +de cy+dan hate +d+at me com swi+de oft on gemynd, hwelce wiotan iu w+aron giond Angelcynn, +ag+der ge godcundra hada ge woruldcundra; <R 4>

19

Text and Files WS 2005/06 / Corpus

© AAC / ÖAW 2005

Corpus Linguistics

Corpus Linguistics

- o The study of language through corpus-based research, but it differs from traditional linguistics in its insistence on the systematic study of authentic examples of language in use.
 - Text linguistics vs Corpus Linguistics
 - Illustration vs Evidence
 - Introspection and Informant Testing vs Observation of Text

21

Text and Files WS 2005/06 / Corpus

© AAC / ÖAW 2005

Methodological Issues

- o Is the Corpus large enough to answer the question?
- o Is the Corpus representative?
 - What were the sampling procedures?
- o Is the transcription (or annotation) reliable?
- o Is the study design sound?

22

Text and Files WS 2005/06 / Corpus

© AAC / ÖAW 2005

Approaches to Corpora

- o Corpus-driven
- o Corpus-based
 - Corpus-formed (alternative term)

23

Text and Files WS 2005/06 / Corpus

© AAC / ÖAW 2005

2nd Generation Corpora

“Mega-Corpora”

- COBUILD
- Bank of English®
- British National Corpus (BNC)
 - 1994 : 100m

25

Text and Files WS 2005/06 / Corpus

© AAC / ÖAW 2005

COBUILD

- Collins Birmingham University International Language Database (COBUILD)
- University of Birmingham (1980)
- Funded by Collins
- Directed by John Sinclair
- *Collins COBUILD English Language Dictionary* (1987), based on the study of the COBUILD corpus.
 - Monolingual learner's dictionary

26

Text and Files WS 2005/06 / Corpus

© AAC / ÖAW 2005

BNC

- British National Corpus (BNC)
 - Developed between 1991–94
 - 100 million words
 - 100,106,008 (to be precise...)
 - 1.5 gigabytes

27

Text and Files WS 2005/06 / Corpus

© AAC / ÖAW 2005

Types of Corpora

DWDS

- Das Ergebnis der Corpuerstellung sind zwei verschiedene Corpora:
- Das **Kerncorpus**: Größe: 100 Millionen Textwörter
 - ausgewogene Textgrundlage
 - rechtlich abgesicherte Nutzungsvereinbarungen
 - annotiert gemäß XML/TEI
- Das **Ergänzungscorpus**: 980 Millionen Textwörter
 - opportunistisch
 - http://www.dwds.de/pages/pages_info/dwds_info.htm

29

Text and Files WS 2005/06 / Corpus

© AAC / ÖAW 2005

Reference / Monitor Corpus

- Reference Corpus
 - fixed
 - publicly-available
- Monitor Corpus
 - open-ended
 - in-house

30

Text and Files WS 2005/06 / Corpus

© AAC / ÖAW 2005

Real Academia Española Corpora

- CREA
 - Corpus de Referencia del Español Actual
- CORDE
 - Corpus Diacrónico del Español

31

Text and Files WS 2005/06 / Corpus

© AAC / ÖAW 2005

Corpus Size

- Growing ever larger...
 - Brown / LOB Corpus 1,000,000
 - British National Corpus 100,000,000
 - Bank of English 450,000,000
 - DWDS 900,000,000
- Constraints
 - Workflow
 - Financial
 - Copyright issues

32

Text and Files WS 2005/06 / Corpus

© AAC / ÖAW 2005

Corpus Exploitation

- NLP
- Speech Processing
- Lexicography
- Language Change
- Stylistics
- TEFL

33

Text and Files WS 2005/06 / Corpus

© AAC / ÖAW 2005

Corpus Typology

Broad Typology

- Native Speaker Learner
- Monolingual Multilingual
- Original Translations
- Synchronic Diachronic
- Plain Annotated

35

Text and Files WS 2005/06 / Corpus

© AAC / ÖAW 2005


Narrow Typology

- Spoken
- Written
- Genre
- Domain
- Register
- Text Types
- Multimedia

36

Text and Files WS 2005/06 / Corpus

© AAC / ÖAW 2005



Corpus Taxonomies

Categories

- Medium
- Design Method
- Language Variables
- Language States
- Plain / annotated

38

Text and Files WS 2005/06 / Corpus

© AAC / OAW 2005

Medium

- Medium
 - Printed
 - Electronic Text
 - Digitized Speech
 - Video
 - Mixed
 - Multimedia

39

Text and Files WS 2005/06 / Corpus

© AAC / OAW 2005

Design Method

- Design Method
 - Balanced
 - Pyramidal
 - Opportunistic

40

Text and Files WS 2005/06 / Corpus

© AAC / OAW 2005

Language Variables

- Language Variables
 - Monolingual Multilingual
 - Original Translations
 - Native Speaker Learner

41

Text and Files WS 2005/06 / Corpus

© AAC / OAW 2005

Language States

- Language States
 - Synchronic Diachronic

42

Text and Files WS 2005/06 / Corpus

© AAC / OAW 2005

Plain / Annotated

- Plain / Annotated
 - Perfectly plain
 - Marked up for formatting attributes
 - e.g. page breaks, paragraphs, font sizes, italics, etc.
 - Annotated with identifying information
 - e.g. edition date, author, genre, register, etc.
 - Annotated for part of speech, syntactic structure, discourse information, etc.

43

Text and Files WS 2005/06 / Corpus

© AAC / ÖAW 2005

National Corpora

Not only in English...

- [CNC](#) (Czech National Corpus)
- [SNC](#) (Slovak National Corpus)
- [IDS](#) (Institut für Deutsche Sprache)
- [FIDA](#) (Corpus of Slovene Language)
- [CNC](#) (Croatian National Corpus)

45

Text and Files WS 2005/06 / Corpus

© AAC / ÖAW 2005

British National Corpus

BNC Definition

- [BNC](#) (British National Corpus)
 - Monolingual
 - Synchronic
 - General
 - Sample

47

Text and Files WS 2005/06 / Corpus

© AAC / ÖAW 2005

Corpus Design

- Sampled (max 45,000 words)
- Synchronic (informative from 1975, imaginative from 1960)
- Not subject-specific
- Monolingual British English
- Spoken and Written

48

Text and Files WS 2005/06 / Corpus

© AAC / ÖAW 2005

BNC Composition

- 90% Written
- 10% Sound

49

Text and Files WS 2005/06 / Corpus

© AAC / ÖAW 2005

Written: Selection Criteria

- Domain
 - 75% *Informative*
 - 8 sub-domains
 - 25% *Imaginative*
- Time
 - 1975 onwards
- Medium
 - 60% Books
 - 25% Periodicals
 - 15% Misc.

50

Text and Files WS 2005/06 / Corpus

© AAC / ÖAW 2005

Spoken: Selection Criteria

- 50% Demographic
 - Spontaneous conversation in a natural setting
- 50% Context-Governed
 - Meetings and events, structured situations
 - 4 sub-domains

51

Text and Files WS 2005/06 / Corpus

© AAC / ÖAW 2005

Links

Gateways

- [Corpus Linguistics on the Internet](#)
- [UCREL](#) (University of Lancaster)
- [Bookmarks for Corpus-based Linguists](#)

53

Text and Files WS 2005/06 / Corpus

© AAC / ÖAW 2005