



# Corpora Galore

Stephen Miller

AAC / ÖAW

***Text & Files*** WS 2004



# 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.

# What is *not* a Corpus...

- Distinction made by Leech (1991):
  - Archive
    - Unordered collection of data
  - Corpus
    - Principled linguistic snapshot of language at a given point in time

# 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.

# 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)



# 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

# SEU Structure



# 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)



Major Corpora

# Major Corpora

- Brown Corpus
  - Brown Corpus of Standard American English
- LOB Corpus
  - Lancaster-Oslo/Bergen Corpus

# Brown Corpus

- Brown Corpus:
  - 1 million words
  - 500 x 2000 words
    - 15 text categories
  - Written American English
  - 1961

# LOB Corpus

- Lancaster-Oslo/Bergen Corpus
  - Same design /date as Brown Corpus
  - But built from British English

# LOB (Untagged)

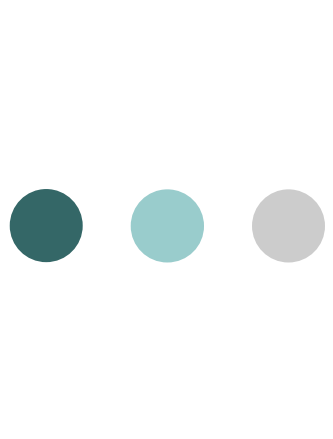
A01 1 **\*\*[001 TEXT A01\*\*]**  
A01 2 **\*<\*7STOP ELECTING LIFE PEERS\*\*\*>**  
A01 3 **\*<\*4By TREVOR WILLIAMS\*>**  
A01 4 |^A \*0MOVE to stop \0Mr. Gaitskell from nominating any more Labour  
A01 5 life Peers is to be made at a meeting of Labour {0M P}s tomorrow.  
A01 6 |^\0Mr. Michael Foot has put down a resolution on the subject and  
A01 7 he is to be backed by \0Mr. Will Griffiths, {0M P} for Manchester  
A01 8 Exchange.  
A01 9 |^Though they may gather some Left-wing support, a large majority  
A01 10 of Labour {0M P}s are likely to turn down the Foot-Griffiths  
A01 11 resolution.  
A01 12 **\*<\*7'ABOLISH LORDS\*\*\*>**  
A01 13 |^\*0\0Mr. Foot's line will be that as Labour {0M P}s opposed the  
A01 14 Government Bill which brought life peers into existence, they should  
A01 15 not now put forward nominees.  
A01 16 |^He believes that the House of Lords should be abolished and that  
A01 17 Labour should not take any steps which would appear to **\*\*prop up\*\*** an  
A01 18 out-dated institution.  
A01 19 |^Since 1958, 13 Labour life Peers and Peeresses have been created.  
A01 20 |^Most Labour sentiment would still favour the abolition of the  
A01 21 House of Lords, but while it remains Labour has to have an adequate  
A01 22 number of members

# LOB (Tagged)

A01 2 ^ '\*' '\*' stop\_VB electing\_VBG life\_NN peers\_NNS '\*' '\*' .\_.  
A01 3 ^ by\_IN Trevor\_NP Williams\_NP .\_.  
A01 4 ^ a\_AT move\_NN to\_TO stop\_VB \0Mr\_NPT Gaitskell\_NP from\_IN  
A01 4 nominating\_VBG any\_DTI more\_AP labour\_NN  
A01 5 life\_NN peers\_NNS is\_BEZ to\_TO be\_BE made\_VBN at\_IN a\_AT meeting\_NN  
A01 5 of\_IN labour\_NN \0MPs\_NPTS tomorrow\_NR .\_.  
A01 6 ^ \0Mr\_NPT Michael\_NP Foot\_NP has\_HVZ put\_VBN down\_RP a\_AT  
A01 6 resolution\_NN on\_IN the\_ATI subject\_NN and\_CC  
A01 7 he\_PP3A is\_BEZ to\_TO be\_BE backed\_VBN by\_IN \0Mr\_NPT Will\_NP  
A01 7 Griffiths\_NP ,\_, \0MP\_NPT for\_IN Manchester\_NP  
A01 8 Exchange\_NP .\_.  
A01 9 ^ though\_CS they\_PP3AS may\_MD gather\_VB some\_DTI left-wing\_JJB  
A01 9 support\_NN ,\_, a\_AT large\_JJ majority\_NN  
A01 10 of\_IN labour\_NN \0MPs\_NPTS are\_BER likely\_JJ to\_TO turn\_VB down\_RP  
A01 10 the\_ATI Foot-Griffiths\_NP  
A01 11 resolution\_NN .\_.  
A01 12 ^ '\*' '\*' abolish\_VB Lords\_NPTS '\*' '\*' .\_.  
A01 13 ^ \0Mr\_NPT Foot's\_NP\$ line\_NN will\_MD be\_BE that\_CS as\_CS labour\_NN  
A01 13 \0MPs\_NPTS opposed\_VBD the\_ATI  
A01 14 government\_NN bill\_NN which\_WDTR brought\_VBD life\_NN peers\_NNS into\_IN  
A01 14 existence\_NN ,\_, they\_PP3AS should\_MD  
A01 15 not\_XNOT now\_RN put\_VB forward\_RB nominees\_NNS .\_.  
A01 16 ^ he\_PP3A believes\_VBZ that\_CS the\_ATI House\_NPL of\_IN Lords\_NPTS  
A01 16 should\_MD be\_BE abolished\_VBN and\_CC that\_CS  
A01 17 labour\_NN should\_MD not\_XNOT take\_VB any\_DTI steps\_NNS which\_WDTR  
A01 17 would\_MD appear\_VB to\_TO '\*' '\*' prop\_VB up\_RP '\*' '\*' an\_AT  
A01 18 out-dated\_JJ institution\_NN .\_.

# Helsinki Corpus of English Texts

- [^ 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>



# Corpus Linguistics

# Corpus Linguistics

- 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

# Methodological Issues

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

# Approaches to Corpora

- Corpus-driven
- Corpus-based
  - Corpus-formed



# 2nd Generation Corpora

# “Mega-Corpora”

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

# 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

# BNC

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



# Types of Coprora

# 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](http://www.dwds.de/pages/pages_info/dwds_info.htm)

# Reference / Monitor Corpus

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

# Real Academia Española

## Corpora

- CREA

- Corpus de Referencia del Español Actual

- CORDE

- Corpus Diacrónico del Español

# Corpus Size

- Growing ever larger...

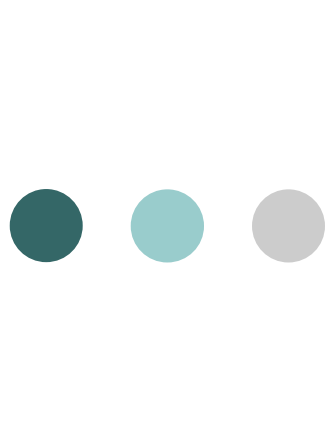
● Brown / LOB Corpus	1,000,000
● British National Corpus	100,000,000
● Bank of English	450,000,000
● DWDS	900,000,000

- Constraints

- Workflow
- Financial
- Copyright issues

# Corpus Exploitation

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



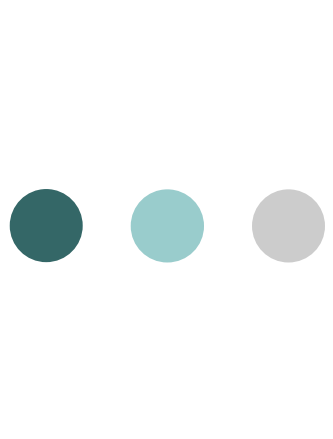
# Corpus Typology

# Broad Typology

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

# Narrow Typology

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



# Corpus Taxonomies

# Categories

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

# Medium

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

# Design Method

- Design Method
  - Balanced
  - Pyramidal
  - Opportunistic

# Language Variables

- Language Variables

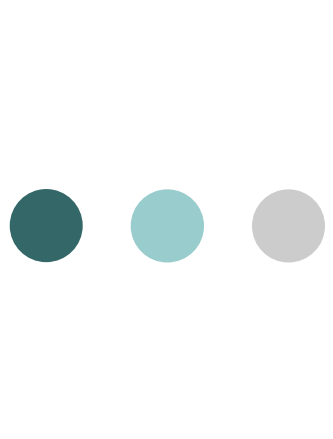
- Monolingual                      Multilingual
- Original                              Translations
- Native Speaker                      Learner

# Language States

- Language States
  - Synchronic Diachronic

# 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.



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)



# British National Corpus

# BNC Definition

- BNC (British National Corpus)
  - Monolingual
  - Synchronic
  - General
  - Sample

# Corpus Design

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

# BNC Composition

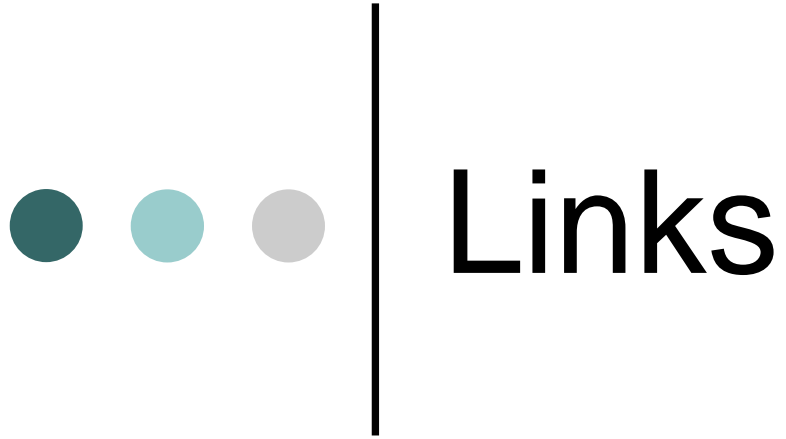
- 90% Written
- 10% Sound

# Written: Selection Criteria

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

# Spoken: Selection Criteria

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



Links

# Gateways

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