

Corpus Linguistics (L615)

Available Corpora

Markus Dickinson

Department of Linguistics, Indiana University
Spring 2009

It will help us in several ways to look at a set of available corpora

- ▶ To see how corpora can be categorized
- ▶ To learn about design decisions done in different corpora
- ▶ To become familiar with the range of commonly-reference corpora

The focus here is mostly on English corpora

Corpus-Distributing Organizations

Corpus Linguistics

Available Corpora

Important Corpora

Multilingual corpora

Treebanks

- ▶ Linguistic Data Consortium (LDC)
- ▶ European Language Resources Association (ELRA)

- ▶ **Brown Corpus:** 1 million words of written American English texts from various genres, dating from 1961
- ▶ **Lancaster-Oslo-Bergen (LOB) Corpus:** 1 million words of written British English texts, dating from 1961. Genres are parallel to the Brown Corpus.
 - ▶ FLOB and Frown are 1990s versions of Brown & LOB, out of Freiburg
- ▶ **British National Corpus (BNC):** 100 mio. words of written and spoken language, balanced corpus of current British English
- ▶ **International Corpus of English (ICE):** national or regional varieties of English; one million word collections of contemporary spoken and written English (Great Britain, USA, Australia, South Africa, Canada, Hong Kong, India, etc.)

- ▶ **German National Corpus:** 2.2 bio. words
- ▶ **IPI PAN Polish Corpus:** 300 mio. words
- ▶ **Chinese National Corpus:** 100 mio. words
- ▶ **Czech National Corpus:** 100 mio. words
- ▶ **Hungarian National Corpus:** 80 mio. words
- ▶ **Croatian National Corpus:** 30 mio. words
- ▶ **Hellenic National Corpus:** 20 mio. words
- ▶ **METU Turkish Corpus:** 10 mio. words
- ▶ ...

- ▶ British National Corpus (BNC)
 - ▶ 90% written, 10% spoken
 - ▶ Represents as wide a range of modern British English as possible
- ▶ American National Corpus (ANC)
 - ▶ Same model as the BNC
 - ▶ Will have about 100 million words

Some examples:

- ▶ Guangzhou Petroleum English Corpus: petrochemical domain
- ▶ HKUST Computer Science Corpus: undergraduate textbooks in CS
- ▶ Corpus of Professional Spoken American English (CPSA)
- ▶ Michigan Corpus of Academic Spoken English (MICASE): <http://quod.lib.umich.edu/m/micase/>

- ▶ London-Lund Corpus (LLC): spoken British English from 1960s to mid-1970s
- ▶ Spoken English Corpus (SEC): spoken British English from 1980s, mainly radio broadcasts
- ▶ Cambridge and Nottingham Corpus of Discourse in English (CANCODE)
 - ▶ Corpus coded with the relationship between speakers: intimates, casual acquaintances, colleagues at work, strangers
- ▶ Wellington Corpus of Spoken New Zealand English (WSC): 1988-1994
 - ▶ 75% informal speech/dialogue (more private material than most)

Synchronic and Diachronic corpora

Synchronic corpora often compare regional varieties

- ▶ ICE, comparisons of Brown & LOB, etc.
- ▶ Longman Spoken American Corpus, Survey of English Dialects (SED)
- ▶ Need detailed speaker information

Diachronic corpora usually cover a wide range of time periods

- ▶ Corpus of English Dialogues
- ▶ Helsinki Dialect Corpus
- ▶ Helsinki Diachronic Corpus of English Texts
 - ▶ Old, Middle, & Early Modern English
- ▶ ARCHER corpus: A Representative Corpus of Historical English Registers
 - ▶ 1650-1990, divided into 50-year periods

Learner corpora collect the language of second language (L2) learners

- ▶ *Developmental corpora* (e.g., CHILDES) are for L1 language

Some examples:

- ▶ International Corpus of Learner English (ICLE)
- ▶ Cambridge Learner Corpus part of the Cambridge International Corpus (CIC)
- ▶ Longman Learners' Corpus
- ▶ Standard Speaking Test (SST) Corpus
- ▶ Chinese Learner English Corpus (CLEC)
- ▶ HKUST(Hong Kong University of Science and Technology) Corpus of Learner English

See <http://jones.ling.indiana.edu/wiki/LearnerCorpora>

Monitor Corpora (theory)

Monitor corpora continue to grow

- ▶ Ensures larger corpus size and allows for large individual sample sizes
- ▶ Often only admit new material which has new features not already in corpus
- ▶ Used to track changes across different periods of time
 - ▶ Monitor corpora could be a series of static corpora

Disadvantages:

- ▶ No attempt to balance the corpus
- ▶ Text availability can become an issue (e.g., copyrights)
- ▶ Confusing to indicate specific corpus version
- ▶ Cannot easily compare results run on corpora of different sizes

- ▶ Bank of English (BoE)
- ▶ Global English Monitor Corpus
 - ▶ Collection of newspapers in English
 - ▶ Monitors language use and semantic change in English across US, Britain, Australia, Pakistan, & South Africa

Multilingual corpora are corpora with multiple languages (two or more)

- ▶ Gain new insights, as compared to monolingual corpora
- ▶ Highlight language-specific, typological, or cultural features
- ▶ Useful for lexicography

Three types of multilingual corpora:

- ▶ Type A: Source texts plus translations (e.g., Hansards)
- ▶ Type B: Monolingual subcorpora designed with the same sampling technique
- ▶ Type C: Combination of A&B (e.g., EMILLE)

Parallel corpus (for us) is Type A, and **comparable corpus** is Type B

- ▶ Corpora with different varieties of the same language (e.g., Brown, LOB) are **comparative corpora**

Parallel corpora can be uni- or multi-directional

- ▶ i.e., there could be translations in either or both directions
- ▶ be on the watch out for “translationese”

An important step is to *align* corpus units

- ▶ at the level of the text, section, paragraph, sentence, and/or word
- ▶ often useful to have a separate alignment file with pointers to, e.g., word IDs

Basic methods for doing sentence alignment automatically:

- ▶ statistical: based on sentence length, in terms of words or characters
- ▶ lexical/rule-based: exploit morpho-syntactic information to align
 - ▶ often more accurate, but slower, than statistical approaches
- ▶ hybrid: integrate linguistic knowledge into a probabilistic system

Fairly accurate for sentence alignment of European language pairs

- ▶ **MULTEXT-East**: for Bulgarian, Croatian, Czech, English, Estonian, Hungarian, Lithuanian, Resian, Romanian, Russian, Slovene, and Serbian. For most languages: Orwell's 1984.
- ▶ **Hansard Corpus**: from the official records (Hansards) of the 36th Canadian Parliament [1997-2000], 3 mio. words
- ▶ **Europarl**: extracted from the proceedings of the European Parliament; includes versions in 11 European languages: Romanic (French, Italian, Spanish, Portuguese), Germanic (English, Dutch, German, Danish, Swedish), Greek and Finnish. Ca. 20 mio. words.

Verbmobil Example

- ▶ e102ach1_109_JLF_420000: well , I guess it depends on how we are going to go to the branch to do the our business .
- ▶ e102ach1_109_JLF_420000_D: also , ich denke , es kommt darauf an , wie wir zur Zweigstelle kommen , um die , unsere Arbeit zu machen .
- ▶ e102ach2_110_SNC_420000: mhm .
- ▶ e102ach2_110_SNC_420000_D: mhm .
- ▶ e102ach1_111_JLF_420000: will we take the train or will we drive I am not I don't know how to get there .
- ▶ e102ach1_111_JLF_420000_D: nehmen wir den Zug oder fahren wir , ich bin nicht , ich wei"s nicht , wie man da hinkommt .
- ▶ e102ach2_112_SNC_420000: yeah , me neither . maybe we should # maybe we should # take a train .
- ▶ e102ach2_112_SNC_420000_D: ja , ich auch nicht . vielleicht sollten wir # # einen Zug nehmen .

English:

<CHAPTER ID=1> Approval of the Minutes of the previous sitting <SPEAKER ID=1 NAME="President"> The Minutes of yesterday 's sitting have been distributed . <P> Are there any comments ? <P> (The Minutes were approved) <SPEAKER ID=2 LANGUAGE="FR" NAME="Wurtz"> Mr President , as you know , today is the eleventh World Press Freedom Day . Many of our fellow Members would certainly wish to take this opportunity once again to state their solidarity with this struggle , a struggle , furthermore , which is proving successful , because , according to the Reporters sans Frontiers association , fewer journalists are being imprisoned and fewer media outlets are being censured than a year ago . <P>

German:

<CHAPTER ID=1> Genehmigung des Protokolls der vorangegangenen Sitzung <SPEAKER ID=1 NAME="Der Pr<E4>sident"> Das Protokoll der gestrigen Sitzung wurde verteilt . <P> Gibt es Einw<E4>nde ? <P> (Das Parlament genehmigt das Protokoll .) <SPEAKER ID=2 LANGUAGE="FR" NAME="Wurtz"> Herr Pr<E4>sident , bekanntlich begehen wir heute zum elften Mal den internationalen Tag der Pressefreiheit . Zahlreichen Kolleginnen und Kollegen ist es sicher Herzenssache , bei dieser Gelegenheit ihre Solidarit<E4>t mit diesem Kampf zu bekunden , der im <DC>brigen bereits Fr<FC>chte getragen hat , denn nach Informationen der Vereinigung Reporter ohne Grenzen sind heute weniger Journalisten in Haft und unterliegen weniger Medien der Zensur als noch vor einem Jahr . <P>

- ▶ Association for Computational Linguistics (ACL) wiki:
<http://aclweb.org/aclwiki>
- ▶ at Linguist List: <http://linguistlist.org/sp/Texts.html>
- ▶ Stanford list:
<http://www-nlp.stanford.edu/links/statnlp.html#Corpora>
- ▶ Tübingen list: <http://www.sfb441.uni-tuebingen.de/c1/corpora-engl.html>
- ...

Syntactically Annotated Corpora: Treebanks

Corpus Linguistics

Available Corpora

Important Corpora

Multilingual corpora

Treebanks

English:

- ▶ Penn Treebank
- ▶ BLLIP Treebank
- ▶ The Penn-Helsinki Parsed Corpus of Middle English
- ▶ Susanne Corpus and Christine Project
- ▶ International Corpus of English ICE (British)
- ▶ Lancaster Treebank
- ▶ The Redwoods HPSG Treebank

- ▶ Basque
 - ▶ Eus3LB project
- ▶ Bulgarian
 - ▶ HPSG-based Syntactic Treebank of Bulgarian (BulTreeBank)
- ▶ Catalan
 - ▶ CAT3LB project
- ▶ Chinese
 - ▶ The Chinese Treebank Project
- ▶ Czech
 - ▶ Prague Dependency Treebank

TreebankProjects (2)

Corpus Linguistics

Available Corpora

Important Corpora

Multilingual corpora

Treebanks

- ▶ Danish
 - ▶ Danish Dependency Treebank
- ▶ Dutch
 - ▶ The Alpino Treebank
- ▶ French
 - ▶ Project TALANA
- ▶ German
 - ▶ NeGra Project - NeGra Corpus
 - ▶ Project TIGER
 - ▶ Verbmobil Treebank of Spoken German (TüBa-D/S)
 - ▶ The Tübingen Treebank of Written German (TüBa-D/Z)

Trebanks Projects (3)

- ▶ Italian
 - ▶ Turin University Treebank TUT
 - ▶ Italian Syntactic-Semantic Treebank
- ▶ Portuguese
 - ▶ The Floresta Sinta(c)tica project
- ▶ Slovene
 - ▶ Slovene Dependency Treebank
- ▶ Swedish
 - ▶ Swedish Treebank
- ▶ Turkish
 - ▶ METU treebank