Types of corpora and some famous (English) examples

**Balanced, representative**

Texts selected in pre-defined proportions to mirror a particular language or language variety.

*Examples:*

- **Brown family** [http://khnt.aksis.uib.no/icame/manuals/>
  - Written. 1 m words, 15 text categories, 500 texts a 2,000 words
  - • Indian (Kolhapur), NZ (Wellington), Australian (ACE)

- **BNC: British National Corpus** [http://www.natcorp.ox.ac.uk>]
  - 100 m words, 10% spoken
  - Carefully composed to be ‘balanced’
  - Oxford access at [https://ota.oerc.ox.ac.uk//bncweb-cgi/BNCquery.pl](https://ota.oerc.ox.ac.uk/bncweb-cgi/BNCquery.pl)

**Monitor**

New texts added by and by to ‘monitor’ language change.

*Examples:*

  - Written and spokne, much newspaper/media language
  - Different varieties and text categories
  - Part can be searched online

- **COCA: Corpus of Contemporary American English** [http://www.americancorpus.org/>]
  - currently 385 mwd
  - 5 genres (one spo) a 4 mwd from 1990 – 2008
  - Searchable online

**Parallel (translation)**

Same texts in two (or more) languages

*Examples:*

- **OPUS** open source parallel corpus [http://urd.let.rug.nl/tiedeman/OPUS/>]
  - Access to aligned corpora, mainly EU texts.
  - Unknown size.

- **ENPC English-Norwegian Parallel Corpus** [http://www.hf.uio.no/ilos/forskning/forskningsprosjekter/enpc/>]
  - Originally English and Norwegian originals with Norwegian and English translations, now also German, Dutch, Portuguese.
  - 50 text extracts in each direction, fiction and non-fiction

**Comparable**

Similar texts in two, or more, languages or language varieties

*Examples:*

- **ICE: International Corpus of English** [http://www.ucl.ac.uk/english-usage/ice/>]
– Different varieties of English (British, Irish, East African, etc)
– 50% spoken
– Some freely available

**ICLE – International Corpus of Learner English**
<http://www.fltr.ucl.ac.be/FLTR/GERM/ETAN/CECL/Cec/Projects/Icle/icle.htm>
– Material produced by language learners in different countries

### Diachronic
Include texts from different (consecutive) periods, preferable comparable ones

**Examples:**

**Helsinki Corpus of English Texts**<http://khnt.hit.uib.no/icame/manuals/HC/>
– Old, Middle, and Early Modern English
– Text samples
– Ca 1,5 mwd

**COCA Corpus of Contemporary American English**<http://www.americancorpus.org/>
– Monitor corpus
– Material from 1990-2008 (and growing)
– Five genres of similar size, 20% spoken
– Part of Brigham Young University corpus collection (Mark Davies)

**Time Magazine**
– Part of Brigham Young University corpus collection (Mark Davies)
– Complete text from Times Magazine searchable online by decade

### Specialized
Include a specific type of text

**Examples:**

**Air Traffic Control Speech corpus**
– 50 sessions of real-time simulation exercises

**Lampeter Corpus of Early Modern English Tracts**
<http://khnt.hit.uib.no/icame/manuals/LAMPETER/LAMPHOME.HTM>
– Historical, written
– Tracts published between 1640 and 1740
– Six domains, ten decades
– 120 different texts, c. 1.1 million words

**USE – Uppsala Student English corpus**<http://www.engelska.uu.se/use.html>
– Ca 1500 essays by Swedish learners of English

### Multi-media
Include multi-media material (for example video recordings and transcriptions)

**Example:**

**SACODEYL**<http://www.um.es/sacodeyl/>
– Interviews with teenagers
– Seven languages
– Available as video, sound and orthographic transcription