Future
Learn

LANCASTER
UNIVERSITY

E·S·R·C
ECONOMIC
& SOCIAL
RESEARCH
COUNCIL

CASS
Corpus Approaches
to Social Science

Corpus Linguistics

1. Introduction to Corpora

Types of Corpora

1 **Specialised corpus** – e.g.

- genre: the language of newspapers
- time: 2005 to the present day
- place: just texts published in China

2 **General corpus** – needs to be much larger. E.g. The British National Corpus (BNC) has about 100 million words of spoken and written British English:

The BNC

Mode	Text category and description	Number of words
Written 87,284,364 words	“Informative” writing: 8 types: 1) World affairs 2) Leisure 3) Arts 4) Commerce and finance 5) Belief and thought 6) Social science 7) Applied science 8) Natural and pure science	70.9 million
	“Imaginative” writing: 1 type: 9) Fiction	16.4 million
Spoken 10,341,729 words	“Spoken demographic”: informal conversation which has been demographically sampled across the population of the UK	4.2 million
	“Spoken Context governed”: task centered speech recorded at specific locations for specific events, such as business meetings, public talks.	6.1 million

Types of Corpora…

3. **Multilingual corpus** – e.g. English and Spanish. Or American English and Indian English.
4. **Parallel corpus** – e.g. English and Spanish – exactly the same texts translated. E.g. the CRATER corpus.
5. **Learner corpus** – language use created by people learning a particular language. E.g. the International Corpus of Learner English.
6. **Historical or Diachronic corpus** – e.g. Helsinki corpus – 1.5 million words of texts from 700AD to 1700AD.
7. **Monitor corpus** – continually being added to. e.g. the Bank of English.