Future
Learn

LANCASTER
UNIVERSITY

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

CASS
Corpus Approaches
to Social Science

Corpus Linguistics

1. Introduction to Corpora

What is a corpus?

- a collection of words?
- Is it a *theory* or *methodology* of language?

Why use a corpus?

- Large amounts of data tell us about tendencies and what's normal or typical in real-life language use
- Corpora also reveal instances of very rare or exceptional cases, that we wouldn't get from looking at single texts or introspection.
- Human researchers make mistakes and are slow. Computers are much quicker and more accurate.

Criteria in building a corpus

1. It must be a large body of text.
2. It needs to be representative of language (or a genre of language).
3. Must be in machine-readable form (e.g. txt files on a computer).
4. Acts as a standard reference about what's typical in language.
5. Often annotated with additional linguistic information – e.g. grammatical codes.