Corpus: Some key terms

1

CASS: Briefings

CASS
Corpus Approaches to Social Science

CASS: Briefings

Published 2013 (2018 revision)
by
The ESRC Centre for Corpus Approaches to Social Science (CASS), Lancaster University, UK

CASS Centre Director
Tony McEnery

Series Editing & Design
Mark McGlashan

Contents

About CASS...

The ESRC funded Centre for Corpus Approaches to Social Science (CASS) is a research centre based at Lancaster University which aims to bring the methods and benefits of the corpus approach to other disciplines.

From the Centre Director

Prof. Tony McEnery

The corpus approach harnesses the power of computers to allow analysts to work to produce machine aided analyses of large bodies of language data - so-called corpora. Computers allow us to do this on a scale and with a depth that would typically defy analysis by hand and eye alone. In doing so, we gain unprecedented insights into the use and manipulation of language in society. The centre's work is generating such insights into a range of important social issues like climate change, hate crime and education. This series of briefings aims to spread the social impact and benefits of the work being done by the centre and, in so doing, encourage others to use our methods in future.

What is Corpus Linguistics?

Corpus linguistics, broadly, is a collection of methods for studying language. It begins with collecting a large set of language data - a *corpus* - which is made usable by computers. **Corpora** (the plural of **corpus**) are usually so large that it would be impossible to analyse them by hand, so software packages (often called **concordancers**) are used in order to study them. It is also important that a corpus is built using data well matched to a *research question* it is built to investigate. To investigate language use in an academic context, for example, it would be appropriate for one to collect data from academic contexts such as academic journals or lectures. Collecting data from the sports pages of a tabloid newspaper would make much less sense.

Software

A number of software packages are available with varying functionalities and price tags. Some pieces of software can be downloaded and used for free, others cost money or are available only online but have built-in **reference corpora**. This table an idea of the variety of software currently available:

Name	Where?	Online	Download	Reference Corpora	Fee	Platform(s)
#LancsBox	http://corpora.lancs.ac.uk/lancsbox	stats online	✔	✔	Free	PC, MAC, Linux
AntConc	www.laurenceanthony.net/software/antconc/		✔		Free (donation requested)	Mac/PC
COCA	http://corpus.byu.edu/coca/	✔		✔	Free (after registration)	N/A
CQPweb	www.cqpweb.lancs.ac.uk	✔		✔	Free (after registration)	N/A
Sketch Engine	www.sketchengine.co.uk	✔		✔	£50 (single user)	N/A
WMatrix	http://ucrel.lancs.ac.uk/wmatrix/	✔			Free (after registration)	N/A
Wordsmith	www.lexically.net/wordsmith		✔		£59.50 (single user)	PC

Annotation

Codes used within a corpus that add information about things such as, for example, grammatical category. Also refers to the process of adding such information to a corpus.

Balance

A property of a corpus (or, more precisely, a **sampling frame**). A corpus is said to be balanced if the relative sizes of each of its subsections have been chosen with the aim of adequately representing the range of language that exists in the population of texts being sampled (see also, **sample**).

Colligation

More generally, colligation is co-occurrence between grammatical categories (e.g. verbs colligate with adverbs) but can also mean a co-occurrence relationship between a word and a grammatical category.

Collocation

A co-occurrence relationship between words or phrases. Words are said to *collocate* with one another if one is more likely to occur in the presence of the other than elsewhere.

Comparability

Two corpora or subcorpora are said to be *comparable* if their sampling frames are similar or identical.

Concordance

A display of every instance of a specified word or other search term in a corpus, together with a given amount of preceding and following context for each result or 'hit'.

Concordancer

A computer program that can produce a **concordance** from a specified **text** or **corpus**. Modern concordance software can also facilitate more advanced analyses.

Corpus

From the Latin for 'body' (plural *corpora*), a corpus is a body of language representative of a particular variety of language or genre which is collected and stored in electronic form for analysis using **concordance** software.

Corpus construction

The process of designing a corpus, collecting texts, **encoding** the corpus, assembling and storing the **metadata**, marking up (see **markup**) the texts where necessary and possibly adding linguistic **annotation**.

Corpus-based

Where **corpora** are used to test preformed hypotheses or exemplify existing linguistic theories. Can mean either:

(a) any approach to language that uses corpus data and methods.

(b) an approach to linguistics that uses corpus methods but does not subscribe to **corpus-driven** principles.

Corpus-driven

An inductive process where corpora are investigated from the bottom up and patterns found therein are used to explain linguistic regularities and exceptions of the language variety/genre exemplified by those corpora.

Diachronic

Diachronic corpora sample (see **sampling frame**) texts across a span of time or from different periods in time in order to study the changes in the use of language over time. Compare: **synchronic**.

Encoding

The process of representing the structure of a **text** using **markup language** and **annotation**.

Frequency list

A list of all the items of a given **type** in a corpus (e.g. all words, all nouns, all four-word sequences) together with a count of how often each occurs.

Key word in context (KWIC)

A way of displaying a **node** word or search term in relation to its context within a **text**. This usually means the node is displayed centrally in a table with co-text displayed in columns to its left and right. Here, 'key word' means 'search term' and is distinguished from **keyword**.

Keyword

A word that is more frequent in a **text** or corpus under study than it is in some (larger) **reference corpus**. Differences between corpora in how the word being studied occurs will be statistically significant (see, **statistical significance**) for it to be a keyword.

Lemma

A group of words related to the same base word differing only by inflection. For example, *walked*, *walking*, and *walks* are all part of the verb lemma WALK.

Lemmatisation

A form of **annotation** where every **token** is labelled to indicate its **lemma**.

Lexis

The words and other meaningful units (such as idioms) in a language. The lexis or vocabulary of a language in usually viewed as being stored in a kind of mental dictionary, the *lexicon*.

Markup

Codes inserted into a corpus file to indicate features of the original text other than the actual words of the **text**. In a written text, for example, markup might include paragraph breaks, omitted pictures, and other aspects of layout.

Markup language

A system or standard for incorporating **markup** (and, sometimes, **annotation** and **metadata**) into a file of machine-readable text. The standard markup language today is **XML**.

Metadata

The texts that makeup a corpus are the data. Metadata is data *about* that data - it gives information about things such as the author, publication date, and title for a written text.

Monitor corpus

A corpus that grows continually, with new texts being added over time so that the dataset continues to represent the most recent state of the language as well as earlier periods.

Node

In the study of **collocation** - and when looking at a **key word in context (KWIC)** - the node word is the word whose co-occurrence patterns are being studied.

Reference corpus

A corpus which, rather than being representative of a particular language variety, attempts to represent the general nature of a language by using a **sampling frame** emphasing **representativeness**.

Representativeness

A representative corpus is one sampled (see, **sample**) in such a way that it contains all the types of **text**, in the correct proportions, that are needed to make the contents of the corpus an accurate reflection of the whole of the language or variety of language that it samples (also see: **balance**).

Sample

A single text, or extract of a text, collected for the purpose of adding it to a corpus. The word sample may also be used in its statistical sense by corpus linguists. In this latter sense, it means groups of cases taken from a population that will, hopefully, represent that population such that findings from the sample can be generalised to the population. In another sense, corpus is a sample of language.

Sample corpus

A corpus that aims for **balance** and **representativeness** within a specified **sampling frame**.

Sampling frame

A definition, or set of instructions, for the samples (see: **sample**) to be included in a corpus. A sampling frame specifies how samples are to be chosen from the population of text, what types of texts are to be chosen, the time they come from and other such features. The number and length of the samples may also be specified.

Significance test

A mathematical procedure to determine the **statistical significance** of a result.

Statistical significance

A quantitative result is considered statistically significant if there is a low probability (usually lower than 5%) that the figures extracted from the data are simply the result of chance. A variety of statistical procedures can be used to test statistical significance.

Synchronic

Relating to the study of language or languages as they exist at a particular moment in time, without reference to how they might change over time (compare: **diachronic**). A synchronic corpus contains texts drawn from a single period - typically the present or very recent past.

Tagging

An informal term for **annotation**, especially forms of annotation that assign an analysis to every word in a corpus (such as part-of-speech or semantic tagging).

Text

As a count noun: a text is any artefact containing language usage - typically a written document or a recorded and/or transcribed spoken text. As a non-count noun: collected discourse, on any scale.

Token

Any single, particular instance of an individual word in a **text** or corpus. Compare: **lemma**, **type**.

Type

(a) A single particular wordform. Any difference of form (e.g. spelling) makes a word a different **type**. All **tokens** comprising the same characters are considered to be examples of the same type.

(b) Can also be used when discussing *text types*.

Type-token ratio

A measure of vocabulary diversity in a corpus, equal to the total number of **types** divided by the total number of **tokens**. The closer the ratio is to 1 (or 100%), the more varied the vocabulary is. This statistic is not comparable between corpora of different sizes.

XML

A **markup** language which is the contemporary standard for use in corpora as well as for a range of data-transmission purposes on the Internet. In XML, tags are indicated by <angle> <brackets>.

Part of our aim at CASS is to make Corpus Linguistics accessible, which is why we have created our **free online FutureLearn course**. With the course, we aim to demonstrate that corpus approaches can offer researchers from all disciplines unique, valuable insights into the use and manipulation of language in society. We provide you need to start 'doing' Corpus Linguistics yourself.

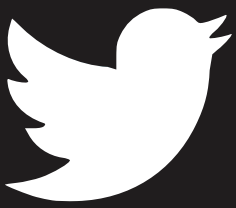
This briefing should act as an introduction and companion to the course where you will begin to apply the concepts and methods mentioned here in a practical way relevant to your field of interest.

The course is free, can be done from home, and comes with a whole range of content and support from world-leading scholars in the field of Corpus Linguistics. For more, visit:

futurelearn.com/courses/corpus-linguistics

For more about CASS and our freely available resources, please visit:

cass.lancs.ac.uk



CASS

@CorpusSocialSci

CASS: Briefings is a series of short, quick reads on the work being done at the ESRC/CASS research centre at Lancaster University, UK. Commissioning work from internationally recognised academics in the field of Corpus Linguistics, CASS: Briefings set out to make cutting edge research easily accessible, providing a good introduction to the variety of vital and exciting research going on in the area of Corpus Linguistics.