CORPUS LINGUISTICS

as a new achievement in modern lexicography

Points to discuss

- Corpus Linguistics and What it Does
- History
- Corpus Lexicography
- Corpora in Lexical Studies

Corpus

A collection of linguistic data

(usually contained in a computer database)

used for research, scholarship, and teaching

Corpus Linguistics

a study of language and a method of linguistic analysis which uses a collection of natural or “real word” texts known as corpus

What Corpus Linguistics Does

- Gives an access to naturalistic linguistic information
- Facilitates linguistic research
- Enables the study of wider patterns and collocation of words
- Allows analysis of multiple parameters at the same time
- Facilitates the study of the second language

What Corpus Linguistics Does Not

- Does not explain why
- Does not represent the entire language

1600

“A Table Alphabetical of Hard Words”

by Robert Cawdrey, 1604

considered to be the first monolingual English dictionary ever made

Robert Cawdrey

- concise definition of each word
- synonym or explanatory phrase
- fixed form of many of the difficult words

1700

“A Dictionary of the English Language in Which the Words are

Johnson’s difficulties

- Selecting words
- Orthography
- Pronunciation
- Etymology and derivation

Johnson’s difficulties

- Analogy
- Syntax
- Phraseology
- Interpretation
- Distribution

New dictionary

- 1857 - appointment of the committee for collection of words that were not in the dictionary
- 1858 – decision on creation a new dictionary

Main aims of the project

- to record every word that can be found in English from about the year 1000
- to exhibit the history of each from a selection of quotations from the whole range of English writings

Levels of selection

- the range of documents which were to be read
- the choice of the reader according to what they felt was significant

OED

- 1884 – (after 24 years) the first instalment of the dictionary that covers part of the letter A
- 1900 – (after 16 years) four and a half volume of dictionary was published until the letter H
- 1928 – the final section of the dictionary was issued

Rules

- word to be explained
- pronunciation and accent
- various forms assumed by the word
- its principal grammatical inflexions

Rules

- etymon of the word
- cognate forms in kindred languages
- meanings which are logically deduced from the etymology, and arranged to show the common thread or threads which unite them together

Creating a dictionary

The methods employed by Johnson were still relevant to lexicographers and were the main steps to be taken in making a dictionary before corpus linguistics was introduced in dictionary making

1960s

Brown Corpus - million word corpus of written text from 500 reading passages made out of a survey of English usage conducted by two universities, University of London and the Brown University Corpus in Providence

This corpus was the first corpus to employ a computer in its making

1982

British version of the corpus, the **LOB corpus** (Lancaster-Oslo-and Bergen) was compiled by Hofland and Johansson

1988

The International Corpus of English

one-million-word corpora by Sidney Greenbaumin

The unique feature of this corpus is that it samples more spoken language(60%) than its written counterpart (40%)

early 1990s

British National Corpus (BNC)

containing 100 million words (1980-1993)

The compilers: Oxford University Press, Longman, Chambers, the British Library, Oxford University and Lancaster University

BNC

The aim – to provide a balanced corpus that represents British English

The corpus includes 10% spoken language and 90% written language, which comprises of 25% fiction and 75% non-fiction

BNC and Brown corpus

BNC took samples from a longer piece of text between 40,000 and 50,000 words

Advantages

- large sample of authentic spoken and written text as a source
- citation comes from real-life discourse
- real contexts that provide accurate, well-defined lexical meanings in the definition

Advantages

- rich information available for words that have many invariant meanings and tend to be overlooked in the previous dictionary practice
- information on word frequency can also be obtained

1987

Collins COBUILD series of English Language Dictionary

the first dictionary to be founded wholly on corpus

by John Sinclair

Definition

junk:

“You can use junk to refer to old and second-hand goods that people buy and collect”

Modern corpus linguistics

- fast machines
- sufficient storage

Modern corpus linguistics

- part of speech
- prosodic intonation
- proper names
- bilingual parallel corpora

Modern corpus linguistics

1980-1986

completely new set of techniques for language observation, analysis, and recording

Modern corpus linguistics

One of the most prominent uses
of a corpus in recent years is
as a resource for lexicography

Modern corpus linguistics

collections of texts

that are stored and accessed electronically

are usually larger than the paper-based collections

Corpus of British English

The main structural features:

- A classification into genres of printed texts (15)
- A large number (500) of fairly short extracts (2000 words), giving a total of around one million words
- A close to random selection of extracts within genres

Collections available

- Association for Computational Linguistics/ Data Collection Initiative (ACL/DCI)
- European Corpus Initiative (ECI)
- British National Corpus (BNC)

Collections available

- Linguistic Data Consortium (LDC)
- Consortium for Lexical Research (CLR)
- Electronic Dictionary Research (EDR)

The use of corpora

changed dictionaries in a way that it has stressed on

- frequency
- collocation and phraseology
- variation
- lexis in grammar
- authenticity

Corpus lexicography

the process of compiling or revising a dictionary based on texts (of written and/or spoken language) collected in an electronic format

Corpus lexicography

- John Sinclair - british linguist, the founder of the COBUILD project at the University of Birmingham, the initiator of the first strictly corpus-based dictionary of general language (*Collins COBUILD English Language Dictionary*)
- Britain was the site of the first corpus-based collocation dictionaries

Corpus lexicography

Le Robert & Collins English-French Dictionary edited by B.T.S. Atkins with Valerie Grundy and Marie-Helene Correard's **Oxford-Hachette Dictionary** which covers the same language pair

The use of (monolingual) corpora lead to a remarkably greater number of multiword translation units (collocations, set phrases) and to context profiles

Corpus lexicography

Wörter und Wortgebrauch in Ost und West

by Manfred W. Hellmann (1992)

the only German example of that era, using the corpus for lemma selection rather than semantic description

Corpus lexicography

Schlüsselwörter der Wendezeit

by Dieter Herberg, Doris Steffens and Elke Tellenbach (1997)

2010

the Oxford English Corpus "contains over 2 billion words of real 21st century English. It is not only size that matters, though: it is the size of the corpus coupled with the careful selection and development of its contents . . ."

Corpora in Lexical Studies

- Difference
- The possibility to produce through text-processing software contexts for the totality of words in the corpus ordered alphabetically with frequency counts attached

Corpora in Lexical Studies

The main difference – the superior processing capacity of machines to sort the data for human interpretation

Corpora in Lexical Studies

quiver and *quake*

Corpora in Lexical Studies

corpus data contains a rich amount of textual information:

- regional variety
- date
- part-of-speech tags etc
- author
- genre

Corpora in Lexical Studies

The open-ended monitor corpus enables lexicographers:

- to keep on top of new words entering the language
- existing words changing their meanings
- the balance of their use according to genre

Corpora in Lexical Studies

- Finite corpora have an important role in the area of quantification
- It is possible to rapidly produce reliable frequency counts and to subdivide these areas across various dimensions according to the varieties of language in which a word is used

Corpora in Lexical Studies

phrases and collocations can be treated more systematically than was previously possible

Notable English language corpora

Brown University Standard Corpus of Present-Day American English

compiled in the 1960s by Henry Kucera and W. Nelson Francis at Brown University, Providence, Rhode Island as a general corpus (text collection) in the field of corpus linguistics

Notable English language corpora

British National Corpus (BNC)

100-million-word text corpus of samples of written and spoken English from a wide range of sources

The corpus covers British English of the late 20th century from a wide variety of genres with the intention that it be a representative sample of spoken and written British English of that time

Notable English language corpora

American National Corpus (ANC)

Text corpus of American English containing 22 million words written and spoken data produced since 1990

The ANC may at some point of time include a range of genres comparable to the British National Corpus. It is annotated for part of speech and lemma, shallow parse, and named entities

Notable English language corpora

Corpus of Contemporary American English (COCA)

The largest freely-available corpus of English, and the only large and balanced corpus of American English. The corpus was created by Mark Davies of Brigham Young University, and it is used by tens of thousands of users every month

The corpus contains more than 450 million words of text and is equally divided among spoken, fiction, popular magazines, newspapers, and academic texts

Notable English language corpora

International Corpus of English (ICE)

set of corpora representing varieties of English from around the world. Over twenty countries or groups of countries where English is the first language or an official second language are included

Notable English language corpora

International Corpus of English (ICE)

ICE corpora contain 60% (600,000 words) of orthographically transcribed spoken English

The father of the project, Sidney Greenbaum, insisted on the primacy of the spoken word. This emphasis on word-for-word transcription marks out ICE from many other corpora, including those containing, e.g. parliamentary or legal paraphrases