What words are most useful for learners of English? Introducing the *New General Service List*

7

CASS: Briefings

CASS
Corpus Approaches to Social Science

CASS: Briefings

Published 2015
by

The ESRC Centre for Corpus Approaches to Social Science (CASS), Lancaster University, UK

Research
Vaclav Brezina
Dana Gablasova

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

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.

Prof. Tony McEnery

Background and outline

Learning vocabulary is a complex process in which the learner needs to acquire both the form and a variety of meanings of a given vocabulary item (Nation, 2001; Carter, 2012). For the beginner, the main question, of course, is where to start. Arguably, the best starting point would be to look at general words. These are words that occur frequently in different texts across different genres. For example, the most frequently and widely used noun in English is *time*. We speak and write about *time* all the time. For learners of English, general words are crucial because they form the core of the language; without them speakers find it very difficult to communicate.

General vocabulary lists can assist in the process of learning words by providing common vocabulary items. Although there are a number of general vocabulary lists available, the by far most influential and widely-used both in pedagogy and vocabulary research is Michael West's *General Service List (GSL),* published in 1953. However, as a guideline for vocabulary learning it has long been out of date and requires revision. For example, the *GSL* includes a number of items that are no longer in general use (e.g. *gay* [=happy], *cart, shilling, servant, footman, milkmaid* and *telegraph*) and excludes newer items (e.g. *television, computer* and *internet*).

In response to the problems identified with West's *GSL*, we decided to investigate the core English vocabulary with very large language corpora using current corpus linguistic technology. We identified useful words occurring frequently in a number of different contexts in English. The result of this investigation is the *New General Service List* (*new-GSL*) that contains approximately 2,500 words. The full study has been published in the *Applied Linguistics* journal (Brezina & Gablasova, 2013).

Research question

Is there a core general vocabulary in the English language?

Method

This study offers a bottom-up, quantitative approach to the development of a *New General Service List* (*new-GSL)* by means of examining frequent general words across four language corpora (*LOB, BNC, BE06* and *EnTenTen12*) of the total size of over 12 billion running words (see table below).

Corpora	*LOB*	*BNC*	*BE06*	*EnTenTen12*
Tokens	1 million	100 million	1 million	12 billion
Period	1961	1990s	2005-7	2012
Variety of English	British	British	British	International
Spoken component	NO	YES (10%)	NO	NO
Sample size	2k words of each text	40-50k words of each text	2k words of each text	whole documents included
No. of texts	500	4,049	500	21.55 million
Sampled text-types	15 genres of writing	Fiction (20%) and non-fiction (70%) writing and speech (10%)	15 genres of writing	A wide range of internet texts

The four corpora were selected to represent a variety of corpus sizes (from one million to over 12 billion tokens) and approaches to representativeness and sampling (from small samples to whole documents).

Key findings

The study brought strong evidence about the stability of the core English vocabulary across a variety of language corpora including different written and spoken contexts. We examined the overlap between the 3,000 most common vocabulary items in wordlists based on the four corpora described above.

We identified substantial correspondence between the wordlists in terms of the number of shared items (71%) as well as the distribution of the words in the wordlists. The final product, the *new-GSL*, consists of a total of 2,496 words. It is divided into a lexical base (2,118 items) and a lexical innovations part (378 items).

- Examples of words that are stable across all corpora (lexical base): *say, make, time, year, know, then, now, good, use, people, way, new, think, look*…
- Examples of words that occur frequently in current English texts (lexical innovations): *Internet, website, online, email, medium, phone, key* (as an adjective), *technology, guy, kid, environment*…

In its present form, the *new-GSL* can be used both for lexical research and development of teaching materials.

100 new-GSL entries in alphabetical order

Overleaf is an example of 100 *new-GSL* items presented in alphabetical order. In brackets after each word, the information about the word class and the rank of the word in the *new-GSL* is presented. The rank shows the relative importance of the word based on its frequency and distribution in English texts.

The following colour coding and font types are used to distinguish words in different frequency bands and to highlight lexical innovations:

FIRST 500 WORDS; 500 – 1000 words; 1001-2500; *lexical innovation*

A (x, 5)
A LOT/LOTS (OF) (x, 369)
abandon (v, 1881)
ability (n, 800)
ABLE (adj, 252)
accident (n, 1426)
accommodation (n, 2342)
accompany (v, 1499)
ACCORDING TO (con, 410)
account (n, 508)
account (v, 2141)
accurate (adj, 2043)
accuse (v, 2028)
achieve (v, 569)
achievement (n, 1858)
acid (n, 2460)
acknowledge (v, 1887)
acquire (v, 1369)
ACROSS (con, 372)
act (n, 676)
act (v, 762)
ABOUT (adv, 151)
ABOUT (con, 60)
above (adj, 2259)
above (con, 509)
absence (n, 1428)
absolute (adj, 2348)
address (v, 1080)
adjust (v, 2388)
adjustment (n, 2489)
administration (n, 1625)
admit (v, 795)
adopt (v, 1227)
adult (n, 1204)
advance (n, 1491)
advantage (n, 851)
adventure (n, 2132)
advice (n, 935)
advise (v, 1469)
affair (n, 1115)
affect (v, 661)
afford (v, 1600)
afraid (adj, 1878)
AFTER (con, 90)
afternoon (n, 1153)
AGAIN (adv, 142)
absolutely (adv, 1390)
abuse (n, 1965)
academic (adj, 1971)
accept (v, 541)
acceptable (adj, 2442)
access (n, 871)
aid (n, 1495)
aim (n, 1404)
aim (v, 1166)
air (n, 523)
airport (n, 1992)
album (n, 1896)
alcohol (n, 1663)
alive (adj, 1951)
ALL (adv, pron, 147)
ALL (x, 39)
ALLOW (v, 212)
ALMOST (adv, 234)
alone (adv, 663)
along (adv, 1678)
along (avp, 1521)
ALONG (con, 481)
alongside (con, 2091)
ALREADY (adv, 218)

ACTION (n, 406)
active (adj, 1179)
ACTIVITY (n, 449)
actor (n, 2283)
actual (adj, 1104)
ACTUALLY (adv, 391)
ad (n, 2073)
ADD (v, 276)
addition (n, 888)
additional (adj, 1183)
address (n, 1522)
AGAINST (con, 162)
AGE (n, 321)
age (v, 1341)
agency (n, 913)
agenda (n, 2417)
agent (n, 1366)
AGO (adv, 411)
AGREE (v, 436)
agreement (n, 1000)
ahead (adv, 1022)
ALSO (adv, 68)
alternative (adj, 2295)
alternative (n, 1697)
ALTHOUGH (con, 197)
ALWAYS (adv, 163)
AMONG (con, 340)
amongst (con, 2055)
amount (n, 535)
analyse (v, 2169)
[...]

Key

n...noun;
v...verb;
adj...adjective;
adv...adverb;
avp...adverbial particle in phrasal verbs;
con...preposition or conjunction;
pron...pronoun;
x...determiner, quantifier or particle

More information

For more information please visit the Lancaster University vocabulary website at **corpora.lancs.ac.uk/vocab**

or read the original article on which this CASS briefing is based:

Brezina, V. & Gablasova, D. 2013. Is There a Core General Vocabulary? Introducing the *New General Service List, Applied linguistics* (advanced access).

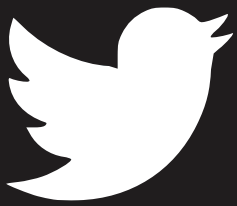
Part of our aim at CASS is to make Corpus Linguistics accessible, which is why we have created our **free online FutureLearn course**. With the course, we aim to demonstrate that corpus approaches can offer researchers from all disciplines unique, valuable insights into the use and manipulation of language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