Corpus Search Basics, Using SketchEngine

Doug Arnold
doug@essex.ac.uk

June 8, 2016

1 Introduction

- By 'corpora' we mean collections of text, 'electronic corpora' are corpora that can be accessed on a computer
- the singular of 'corpora' is 'corpus' (it just means 'body' in Latin – in this context, a body of text)
- There are lots of ways of accessing corpora
- Sketch Engine `https://the.sketchengine.co.uk` provides a relatively cheap and very easy method

2 Getting Started

- go to `https://the.sketchengine.co.uk/login/`
- You can register as a user yourself:
  - you can get a free 30 day trial
  - there are some corpora that are open to everyone: `https://the.sketchengine.co.uk/open/` (you see some adverts when you use this)
  - for an academic individual, an account costs about 6 euro per month, see: `https://www.sketchengine.co.uk/prices/`
- my login name is: dougarnold02
- I'll tell you my password:
- You can search existing corpora, or create your own (I am only interested in the former)

3 A Simple Search

- simple search for a word
- try 'enjoy' – even in the Brown Corpus, we get more than 120 hits
- the results are displayed as a Key Word in Context (KWIC) 'concordance'.
- look at the options, you can 'save', change 'view options', 'sort', 'filter', . . .
- try these out.
- there will be things you don't understand (and don't need to understand)

Incidentally, when you try a new search, **don't** use the browser 'back' button, click on the 'Search' menu item again, or use the search box at top left.

4 Moving on

- look again at the basic search, you can:
  - change the query type
  - limit the text type
  - change 'context' settings

5 More information

- word list (more later)

- word sketch (more later)
- thesaurus
- sketch diff (more later)
- corpus info
- my jobs
- user guide

6 Look at the Frequency Information

Exercise 6.1 Look at the frequency information with *enjoy*. What do you think it means?

7 Lemmas

A *lemma* is a 'basic word' — roughly speaking, a word without inflectional affixes. Thus, *take*,*takes*, *took*, *taken*, are forms of the lemma `take`. The idea is similar to the idea of a dictionary headword or a citation form. Of course, one can lemmatize in different ways. For example, do the noun *take* and the verb *take* share a lemma or not?

Exercise 7.1 If we search for *enjoy* we actually get results for all the forms *enjoys*, *enjoyed*, *enjoying*, *enjoyed*. But this does not happen with the forms of *take*. Why?

Do a lemma search for *take*, and see which form is the most frequent.

Exercise 7.2 Check in SketchEngine whether the verb and noun *take* share the same lemma.

8 Exercises

Exercise 8.1 How big is the Brown Corpus, how big is the British National Corpus?

Exercise 8.2 Are there any Japanese corpora available in Sketch Engine?

Exercise 8.3 Basic searches are a good way of looking for information about words, especially unusual words, or fixed phrases.

Do queries for the following. Can you work out what these words mean?

(1) a. *zilch*
b. *headway*
c. *snook*

Exercise 8.4 Some words you will not find in the BNC include the following (they have come into the language only in the last few years). Do a google search to see how common they are. There is an obvious lesson here in how to use corpora (what?).

(2) a. *snowclone*
b. *chav*

Exercise 8.5 It is also easy to search for fixed phrases, some of these are interesting because they basically defy the normal conventions of English grammar:

(3) a. *per diem*
b. *by and large*

Some others are fixed, but though, they do not seem grammatically very odd:

(4) a. *by dint of*

b. *higgledy piggledy* (there are actually **two** instances of this in the BNC, but you will find only one, to find the other, search just for *higgledy*; we will see how to avoid this problem later).

c. *a dab hand (at)*

What variation can you find here (e.g. does *dint* occur anywhere else?)

Exercise 8.6 I am pretty sure that I never say *'if and when'* — I think I say *'when and if'* (as in *I'll do that when and if I can*). In this, I think I am typical, and I would even wonder if *'if and when'* is acceptable. Similarly with ?*if or when* vs. *'when or if'*. Am I right about this? (a) What do you think? (b) What does a corpus search show?

You can use a 'phrase search' for this.

Exercise 8.7 In discussions of modal verbs, it is common to assert that English does not allow more than one modal verb at a time. Thus, **could might* is predicted to be ungrammatical (and in fact it does not occur in the BNC).

What about the sequence *can might*? Do you find any instances of this? Are any of them real counter-examples?

Exercise 8.8 I was once faced with a class who refused to believe you could say things like (5)

(5) They might have been being taken

They were wrong, of course, but they were Native English speakers and I could not convince them. Nowadays, I would do a corpus search to challenge them:

(6) a. *might have been being* (no results)
b. *have been being* (some results)
c. *has been being* (some results)

I'm pretty sure this would have convinced them.

You can't use a lemma search here (I don't think). We will look at ways you can search for alternatives like *have* and *has* at the same time later.

Exercise 8.9 In discussions of modal verbs, it is common to assert that English does not allow more than one modal verb at a time. Thus, **could might* is predicted to be ungrammatical (and in fact it does not occur in the BNC).

What about the sequence *can might*? Do you find any instances of this? Are any of them real counter-examples?

Exercise 8.10 Using *enjoy* try two of the alternatives to search: Word Sketch, and Thesaurus (leave Word List for another time).

Exercise 8.11 Compare *enjoy* and *love* with 'Sketch diff'. Try to work out what you are looking at.

Exercise 8.12 Have a look at the Sketch Engine documentation, see what sorts of help is available.

My personal view is that there is not much point in becoming an expert Sketch Engine user. It is good to have a general idea of what is possible, and a grasp of some of the basic search techniques, but in general you can (re-)learn what you need when you need it.

8.1 Some General Discussion Points

Exercise 8.13 Why, and to what extent, are corpora useful in (a) lexicography, (b) language teaching/learning, (c) sociolinguistics, (d) descriptive/theoretical linguistics?

Exercise 8.14 Look at one of corpora provided in Sketch Engine, and evaluate it from the point of view of one of these sub-disciplines.

Exercise 8.15 Discuss the advantages and disadvantages of using corpus data vs. informant judgements for various kinds of investigation.

9 What are corpora good for?

Investigating (a) Use/Usage (vs Structure) ; and (b) Variation/Variability.

Variation of linguistic features in relation to other linguistic features, or non-linguistic features (text type, speaker/hearer gender, etc).

Variation among texts (how are face-to-face conversations different from telephone conversations?)

What are the alternatives? (i) work with informants (possibly oneself) accessing native speaker judgements; (ii) observation, and work with non-digitized recordings or corpora; (iii) experimental work specifically designed to tease out particular distinctions (e.g. psycholinguistic work).

- Psycholinguistic style experiment is not always appropriate, or even possible; and as regards normal grammar, one is still faced with the problem of infering facts about competence from observations about performance.
- Work with non-digitized material is of necessity small scale, and hence limited.
- Intuition is a poor guide to *usage*; usage has been generally neglected in mainstream, Chomskian, theoretical linguistics, where the focus is on *competence*, but it is important:
  - In Lexicography — how a word is used is can tell us important, non-obvious, things about its meaning (see below: *great* vs *big* vs *large*);
  - Sometimes we have no access to native speaker intuitions: e.g. with dead languages, child-language, or sublanguages — usage is all we can study;
  - In Sociolinguistics, intuitions about variation across situations, registers, etc. are non-existent or unreliable (or anyway, in need of checking): how does the language used in tutorials differ from that in lectures? How do doctors speak to patients?
  - Similarly, in other areas of grammar, intuitions may reflect prescriptive bias, or just be misconcieved/misguided (and susceptible to correction in the face of attested examples: *They might have been being deceived*).
  - In syllabus design for language teaching, information about relative frequency of constructions (in particular situations or text types) might guide the order in which they are taught (e.g. Prog vs Perf vs Simple aspect).

See Meyer (2002, Ch1).

10 Limitations

Corpus investigation is not always useful, appropriate, or even possible:

- Not all grammatical features are susceptible to investigation, because they are too hard to detect (e.g. parasitic gaps: (7))

  (7) a. Which papers did you photocopy [] before reading [] ?
  b. *Which papers did you photocopy your essay before reading [] ?
  c. Which papers did you photocopy [] before reading the introductions?
- The fact that grammatical feature X occurs in a corpus does not entail it is grammatical (it might be a mistake) — unless it occurs often, of course.
- Nothing at all can be infered from the fact that feature X does not occur . . .
- . . . and corpora are *always incomplete*
- What does frequency have to do with (e.g.) grammaticality? How can we draw inferences from *Parole* to *Langage* (Saussure), from *Performance* to *Competence* (Chomsky), from *E-Language* to *I-Language* (Chomsky)?
- Corpora provide undeniable facts, but do not indicate their significance, and not all facts are *interesting*.

  A conversation between a theoretical linguist (T) and a corpus linguist (C):
  C: Why should I think what you tell me is true?

T: Why should I think what you tell me is interesting?
(adapted from Fillmore (1992)).

11 Key Terms

- Collocation;
- Sublanguage;
- (Word) Token vs. Type; Tokenization;
  (8) The dog took the bone into the garden.
- Lemma vs. word form: *walk walks*, *walked*, *walking*; Lemmatization. (vs word family);
- Tag, Tagging, POS Tag, Mark-up;
- Parsed Corpus, Treebank

12 Reading

You will find some useful discussion of corpus linguistics in Meyer (2002, Ch1), and Biber et al. (1998, Ch1).

13 References

Biber, Douglas, Conrad, Susan and Reppen, Randi. 1998. *Corpus Linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.

Fillmore, Charles. 1992. Corpus Linguistics or Computer-Aided Armchair Linguistics. In J Svartvik (ed.), *Directions in Corpus Linguistics*, pages 35–60, Mouton de Gruyter.

Meyer, Charles F. 2002. *English Corpus Linguistics: an introduction*. Cambridge: Cambridge University Press.