Finding the Words Which Are Most X

Adam Kilgarriff
Lexical Computing Ltd

Pavel Rychlý
Masaryk University

Which English words are most distinctive of American English? Which Spanish verbs have a strong tendency to occur in the gerund? Which English nouns are most often used in the plural? All these questions can be answered in quite a straightforward with a suitable corpus with appropriate markup. The task usually takes a moderate amount of programming. We present a tool which means that it is easy to produce lists of this kind-and many others-which needs no further programming. The work takes place in the framework of a leading corpus query tool.

Introduction

Which words are most distinctive of Business English? Which English nouns are most often in the plural? Which Spanish verbs tend to be used in the gerund?

Given a suitable corpus, with appropriate markup, these questions are not difficult to answer. For each word, we count the frequency of the word in the condition and compare it with the frequency of the word where the condition does not hold. We then sort the words according to the ratio and the words at the top of the list are the ones that answer our question. For the opening questions, we have done this and provide some answers in Table 1.

A lexicographer would often like to ask such questions. Should this word be marked as "business"? Yes, if it is strongly associated with business contexts, but how can I check if it is? Should this English noun be marked as usually plural, or should that Spanish verb have its gerund highlighted? The lexicographer is rarely in a position to check. Even if the right corpus, with the right markup, is available, it is still a programming task to do the counting, compute the statistics, sort the list, and make the results accessible to the lexicographer. It does not usually happen.

Within our corpus tool, the Sketch Engine (Kilgarriff et al. 2004), we now make it easy to produce such lists. We have well-marked-up, large corpora for a number of languages. The corpus data is stored in ways that support efficient and sophisticated searching and counting. We have often encountered queries of the form "which words are most X?". We have now developed a general response: users can specify the X, and the software will generate the list.

The Sketch Engine

The Sketch Engine is a corpus query tool, designed particularly for lexicography. It is available at *http://www.sketchengine.co.uk* (with self-registration for free trial accounts) where it is already loaded with corpora for nine major world languages (Chinese, English, French, German, Italian, Japanese, Portuguese, Russian, Spanish) as well as several smaller ones. Its functions include word sketches—one page summaries of a word's grammatical and collocational behaviour—and a distributional thesaurus. It is in use at Oxford University Press, Collins, Chambers Harrap, Macmillan (all UK), Le Robert (France), the FrameNet Project (USA), Institute of Dutch Lexicology, Institute for the Czech National Corpus, and Patakis Publishers (Greece) amongst others.

Business English keywords		Highly plural English nouns		Spanish verbs often found in gerund	
shareholders	discount	grandparent	follower	intercalar	estudiar
Companies	employee	bacterium	bound	retomar	obviar
investors	earnings	demonstrator	miner	prescindir	tomar
Stock	loans	african	shopper	promediar	ampliar
equity	turnover	immigrant	supporter	pulsar	recurrir
investments	profits	settler	russian	partir	preparar
buyer	corporate	indian	muslim	depender	alargar
transactions	profitable	tooth	allegation	incluir	funcionar
buyers	firms	contemporary	resource	acortar	parafrasear
accounting	customer	palestinian	inmate	redefinir	multiplicar
assets	competitors	tear	theorist	sumar	manipular
Exchange	markets	voter	academic	procurar	sustituir
subsidiary	stocks	liberal	jew	hablar	comparar
banking	taxation	antique	unionist	usar	eliminar
Financial	liability	bureaucrat	delegate	utilizar	excluir
premium	Business	fluctuation	american	minimizar	seguir
suppliers	marketing	symptom	kilometre	esperar	generar
retail	Market	socialist	grandchild	forjar	retransmitir
asset	shares	environmentalist	sock	aumentar	especificar

Table 1. Lists of words which are "most X"

Notes: Corpora were BNC for English and a similar-sized web corpus for Spanish. Results for 2nd and 3rd lists use POS-tagger output: CLAWS for English, TreeTagger for Spanish. Only words with frequency over 100 included. Nouns occurring exclusively in plural excluded from "plurals" results.

The specification language

The Sketch Engine offers a number of ways for users to make queries (see the User Guide at *http://trac.sketchengine.co.uk/wiki/SkE/DocsIndex*). Internally, all queries are interpreted in CQL, the Corpus Query Language, as first developed at the University of Stuttgart (Christ and Schulze 1994). The language is fully described in the Sketch Engine documentation.

The corpora in the Sketch Engine are, wherever possible, lemmatised and part-of-speech tagged, so, for each word, we know the lemma and the part of speech tag. Here we assume a corpus which has been processed in this way.

The list-making functions are to be found under the "word list" button. On clicking the button the user sees a form with three parts: frequency list, keyword list, advanced and saved.

Frequency list: this allows the user to generate a frequency list for all words (or all lemmas, or all part-of-speech tags) for the corpus. Lists may either be based on simple frequency or on ARF (Average Reduced Frequency). ARF addresses the problem of words sometimes occurring many times in one or two documents and little elsewhere (Hlaváčová 2006). They then have a high frequency in the corpus but this is not a good guide to the word's frequency in the type of language the corpus represents. ARF discounts the frequency of such words.

Keyword list: this function, like the WordSmith[1] Tools keyword function, identifies the words which are distinctive of one corpus (or subcorpus) in relation to another. It was used to find the business English words in Table 1, by comparing the BNC business English subcorpus (formed by taking all BNC documents marked "business" or "commerce") with BNC frequencies overall. In the Sketch Engine it is simple to define subcorpora according to information in document headers and then find the distinctive words of that subcorpus. Either ARF or simple frequencies may be used, and the user is given a choice of statistics for making the comparison.

Advanced: "Most plural" and "most gerund" (the second and third lists in Table 1) use the advanced list-making function. Here, the user specifies

the kinds of objects they want a list of (typically, words or lemmas);

the condition they are interested in (e.g., that a noun is plural; Q1);

the condition for comparison (e.g., that a noun is singular; Q2);

the conditions can be any CQL query, or can make use of data collected for word sketches.

Additional constraints may include:

regular expressions limiting the words or lemmas to be considered

a minimum frequency, below which words are not to be listed.

here is the plurals example:

```
=sing_plur
Q1 [lemma="%s" & tag="NN2"]
Q2 [lemma="%s" & tag="NN1"]
```

The first line gives a name to the list. Q1 states what we are counting – items with a specific value for the lemma and with part-of-speech tag NN2 (plural noun). Q2 states what we are comparing it with: items with the same lemma and with part-of-speech tag NN1 (singular noun). Once the counts are established, the ratio between the Q1 and Q2 values will be established for all nouns. The second column of Table 1 shows the words with frequency over 100 at the top of the list.

A full description of the formalism, with further examples, is available in the documentation at *http://trac.sketchengine.co.uk/wiki/SkE/DocsIndex.*[2]

A list specification may be entered into the "word list" form or may be prepared in a file, and uploaded.

Lists can be computationally intensive to generate, and may take minutes or hours to produce, so, once generated, they are stored, and can be retrieved if they are wanted again, rather than recomputed.

Conclusion

We present a fast, flexible, general-purpose mechanism for preparing lists of words meeting some criterion of interest to the user. We believe this will prove a useful addition to the corpus lexicographer's toolbox.

---

[1] *http://www.lexically.net/wordsmith/.*

[2] As we receive more feedback from users, we may modify the formalism. Users should check the website rather than working from the example presented here.

References

Christ, O.; Schulze, M. (1994). *The IMS Corpus Workbench: Corpus Query Processor (CQP) User's Manual.* Stuttgart: University of Stuttgart. *http://www.ims.uni-stuttgart.de/projekte/ CorpusWorkbench/.*

Hlaváčová, J. (2006). "New Approach to Frequency Dictionaries". Proc. LREC, Genoa, Italy.

Kilgarriff, A.; Smrz, P.; Rychlý, P.; Tugwell, D. (2004). "The Sketch Engine". Proc. Euralex, Lorient, France.