Corpus Linguistics for Language Learning and Teaching

Martin Wynne martin.wynne@bodleian.ox.ac.uk

Bodleian Libraries
Faculty of Linguistics, Philology and Phonetics
Subject: Les Francais des Corpus – Aftermath

Dear colleagues,

First, many thanks for presenting at /attending the Francais des Corpus Workshop and for making it such a success. I promised I would keep you in touch with one another and hope that the full list of your e-mail addresses above makes that possible.

...
BNCWeb:
There are 1486 different types in your collocation database for the query "[word="aftermath"] [word="of"]".
(Your query "aftermath of" returned 544 hits in 337 different texts)
The selected range was 1 to 4.
Corpus basis for calculation: the whole BNC.
Type of calculation: Log-likelihood
Tag restriction: any noun
Collocates occur at least 5 times in the whole BNC.
Words collocate at least 5 times.
What is a corpus?

“…a collection of pieces of language, selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language.”

(Sinclair 1996)
What is Corpus Linguistics?

(1) Focus on linguistic performance, rather than competence
(2) Focus on linguistic description, rather than linguistic universals
(3) Focus on quantitative, as well as qualitative models of language
(4) Focus on a more empiricist, rather than rationalist view of scientific inquiry.

(Leech 1992)
How do you know things about language?
What does your knowledge and experience tell you about the use of *try to* & *try and*?
Fill in the blanks

1. Did you **try** ... talk her out of swimming?
2. Mr. Kissinger, **try** ... explain to us what might happen
3. He did it to **try** ... score points
4. They both wanted to **try** ... have a family
5. They **try** ... treat you like machines
6. Sometimes, people **try** ... make fun of you by imitating you.
7. Now the government will **try** ... sell all of this.
8. Did you **try** ... get out of it?
9. I will **try** ... understand this.
Fill in the blanks

1. Did you **try and** talk her out of swimming?
2. Mr. Kissinger, **try and** explain to us what might happen
3. He did it to **try and** score points
4. They both wanted to **try and** have a family
5. They **try to** treat you like machines
6. Sometimes, people **try to** make fun of you by imitating you.
7. Now the government will **try to** sell all of this.
8. Did you **try to** get out of it?
9. I will **try** understand this.
• “Try and do something is incorrect for try to do...” [Partridge and Greet 1947]

• “Try and is well established in conversational use ..Try to is to be preferred in serious writing” [Plain Words 1986]

• “… try and has been socially acceptable for these two centuries … is not used in an elevated style” [Webster’s Dictionary 1989]
What are the factors governing the choice and distribution of *try to* vs. *try and*?

How would you investigate this question?
Try to or try and? Verb complementation in British and American English

Hommerberg & Tottie (2007)
ICAME Journal 31:45-64
http://icame.uib.no/ij31/ij31-page45-64.pdf
Corpus linguistics in language pedagogy

- Primary data for developing dictionaries and grammars
- Informing textbooks and other teaching materials
- Preparing materials for classes (e.g. as a source of examples)
- Data-driven learning in the classroom
- Studying learner language
- Developing new theories (e.g. 'English as a Lingua Franca')
Why not just Google it?
Problems with language on the web

- Biased distribution of text-types and genres
- Repeated and reused text
- Unknown provenance (who wrote this, when, why?)
- Native and non-native producers of language
- Mixture of varieties
- Unclear separation of linguistic elements of the webpage
- Accessing the hidden web
- Accessing language embedded in audio and video streams
- Repeated and reused text
- Lack of persistence of source data and location identifiers
- Unknown (or undesirable) sampling and ranking strategies of search engines
Problems with language in the corpus

- Limited size
- Expensive, time-consuming and slow to make
- Usually limited to out of copyright texts
- Not up to date
- Design decisions were made by someone else
- Not comparable to other corpora
- Access restrictions
- Difficult to deploy in the classroom
- Often, only restricted online access with limited functionality, processing, and not connected to other resources
“I'm just going out to commit certain deeds”
Links for Practical Work

- [http://bncweb.lancs.ac.uk/](http://bncweb.lancs.ac.uk/) (register using ac.uk email address)
- [http://corpus.byu.edu](http://corpus.byu.edu) (via OxLip+)
- All links can be found via: [https://ota.ox.ac.uk/oxonly/oxford.xml](https://ota.ox.ac.uk/oxonly/oxford.xml)
Data-driven language learning in the classroom

- Do corpora reveal 'real language'?
- Do we want to teach real English?
- Can teachers deal with unrestricted texts in the classroom?
- Should teachers control the rate and order of exposure to linguistic features?
- Reading concordances and collocation lists, and interpreting quantitative data require skills
- There are difficult technical issues in effective deployment of corpora in the classroom
Antconc: a stand-alone concordancer

- http://www.antlab.sci.waseda.ac.jp/software.html

- Use with any 'plain' text (txt, html, xml)

- Multilingual capabilities

- Does not interpret mark-up or metadata
Finding resources

https://ota.ox.ac.uk/oxonly/oxford.xml
References

**References**


**Online resources**

- Corpora for users in the University of Oxford https://ota.ox.ac.uk/oxonly/oxford.xml
- Brigham Young Corpora http://corpora.byu.edu/ (also via Oxf
- British National Corpus http://ota.oerc.ox.ac.uk/bncweb/cgi/BNCweb.pl/
- VOICE. 2013. The Vienna-Oxford International Corpus of English (version 2.0 Online) http://voice.univie.ac.at, also available for download from the Oxford Text Archive (http://www.ota.ox.ac.uk).