Raising teachers’ awareness to corpora

TaLC 7- Paris
Ana Frankenberg-Garcia
ISLA, Lisboa
From TaLC 1994 (Lancaster) To TaLC 7 (Paris)

Corpus availability

Corpora in the classroom fans
But do language teachers actually use corpora?

Two ways of using corpora in language teaching

**Indirectly**
Teachers (and learners) use corpus-based materials mediated by experts

- e.g. dictionaries, texts, books, grammars

**Directly**
Teachers (and learners) use corpora and concordances hands-on

i.e. data-driven learning
Do language teachers use corpora indirectly?

Yes

At least in the EFL context (other languages?)
Indirect use of corpora

A few EFL examples

Dictionaries
COBUILD (1987), Oxford Collocations (2002) and many others...

Grammars
COBUILD (1990), Longman (1999)...

Text books
COBUILD English course (1989), Touchstone series (2004)...

No need to understand corpora
Many users don’t even know what a corpus is (Mukherjee 2004)
Do language teachers use corpora directly?  

**no**

Email survey (Tribble 2001)

52.8% of respondents used corpora in teaching

But the survey was circulated on Corpora and Linguist lists and its readers are:
- an unrepresentative minority
- far more likely to know about corpora than the average language teacher!
Do language teachers use corpora directly? 
again, no

Use of corpora in German secondary schools (Mukherjee 2004)

248 qualified English language teachers
10.9% familiar with corpus linguistics (but do they use it?)
9.7% not familiar but had heard of it
79.4% didn’t know anything about it
Why don’t teachers use corpora directly in the classroom?

Main reasons (Tribble 2001)

- 29.2% No access to software
- 23.6% Not enough knowledge about the potential of corpora
- 20.2% No time to prepare corpus materials
- 12.4% Not confident about using computers to analyse language

50.6% Did not (or could not?) answer why
A growing area of concern

TaLC 2006

Yvonne Breyer  *How to teach with corpora: Integrating corpus linguistics into initial teacher training*

Ute Römer  *Corpus research and practice: What help do teachers need and what can we offer?*

Alex Boulton  *Bringing corpora to the masses – Free and easy tools for language teaching and learning*

Fanny Meunier and Cédrick Fairon  *Empowering teachers and learners corpus literacy Using the RSS technology to automate tailor-made corpus collection*

Francesca Bianchi and Elena Manca  *Discovering language through corpora Needed abilities and student difficulties in corpus analysis*
There seems to be a clear need to:

- Train learners to use corpora
- Train teachers to use corpora
- Improve the usability of corpus resources
Where can teachers learn about corpora?

- General introductions to corpora
- Corpus-specific tutorials
- Books and articles about using corpora in language teaching
General introductions to corpora
http://bowland-files.lancs.ac.uk/monkey/ihe/linguistics/contents.htm

Corpus Linguistics
By Tony McEnery and Andrew Wilson

<table>
<thead>
<tr>
<th></th>
<th>Section 1</th>
<th>Section 2</th>
<th>Section 3</th>
<th>Section 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Early</td>
<td>Early Corpus</td>
<td>What is a Corpus</td>
<td>Quantitative Data</td>
<td>The Use of Corpora in Language Studies.</td>
</tr>
<tr>
<td>Corpus</td>
<td>Linguistics and the</td>
<td>and What is it?</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Chomskyan Revolution.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

What you won't find at this website.

Web pages to be used to supplement the book "Corpus Linguistics" published by Edinburgh University Press
ISBN: 0-7486-0508-7 (cased) and 0-7486-0482-0 (paperback)
written by: Tony McEnery and Andrew Wilson.
General introductions to corpora

http://www.georgetown.edu/faculty/ballc/corpora/tutorial.html

Tutorial:
Concordances and Corpora

Catherine N. Ball
Department of Linguistics
Georgetown University
Washington DC 20057
cball@gwvax.georgetown.edu

This three-hour tutorial is derived from my Corpus Linguistics course (LING-468), which is offered every other Spring (current session: Spring 1997).

Enter your e-mail address to receive e-mail from the URL-Minder when this page is updated.

Your Internet e-mail address:

[Submit button]

Table of Contents | Figures and Tables | Copyright notice

Table of Contents

- 1 Tutorial Outline
- 2 Electronic text: Texts, collections and corpora
  - 2.1 Corpora, text collections, and texts
  - 2.2 Finding electronic texts
  - 2.3 Constructing a corpus
  - 2.4 Applications: corpus-based research and teaching
- 3 Concordancers
  - 3.1 Overview
  - 3.2 Features
General introductions to corpora


ICT4LT Module 2.4
Using concordance programs in the Modern Foreign Languages classroom

This is a very large module, around 50 pages. Please wait for it to load completely; it may take several minutes.

This Web page is designed to be read from the printed page. Use File / Print in your browser to produce a printed copy. After you have digested the contents of the printed copy, come back to the onscreen version to follow up the hyperlinks.

Contents

- Aims
- Authors of this module
- Introduction by Graham Davies
- 1. What are concordances and how can they help language teachers?
- 2. Concordance software and corpora
- 3. How does the use of concordances fit in with language teaching methodologies?
- 4. Some practical illustrations
- 5. Preparing yourself and your students for working with concordancers
- 6. The future of concordancing in the classroom
- Bibliography and references: Printed publications and websites
- Feedback

Aims
Using authentic 'real language' in teaching and learning: an online corpus tutorial

This on-line tutorial guides you step-by-step through the basics of collecting and analyzing 'real language.' If you are a language teacher working in school, college or university, then this website is for you. In straightforward language, it takes you on a tour of this exciting area. You will have the opportunity to share the results of worldwide corpus projects, and to carry out guided corpus projects of your own.

In order to use this tutorial, all you need is access to the Internet, a computer that uses Microsoft Word (details of versions that will support wordsmith), and basic computer skills.

To get the most out of the tutorial it is best to do the units in the order that they are presented, and to spend a minimum of 1.5 hours on each one.

Continue to the Corpus Tutorial Overview for descriptions of the tutorial chapters.
Corpus-specific tutorials
http://www.natcorp.ox.ac.uk/using/index.xml

What can I do with the BNC?
The BNC is a corpus - a collection of texts. It is presented in a way that makes possible almost any kind of computer-based research on the nature of the language. Obvious application areas include lexicography, natural language understanding (NLP) systems, and all branches of applied and theoretical linguistics.

The BNC material is made available under certain conditions, summarized in the BNC End User Licence (also available in pdf format). All rights in the texts are reserved. For further information, see the BNC copyright page.

Uses of the BNC
What's the plural of corpus? In what social situations is wicked a term of approval? Why does it "sound wrong" to say The good weather set in on Thursday although The bad weather set in on Thursday is perfectly acceptable? If I can say I live a stone's throw away from here, can I also say I'm going a stone's throw away from here?

Large language corpora can help provide answers for these kinds of questions -- if only because they encourage linguists, lexicographers, and all who work with language to ask them. The purpose of a language corpus is to provide language workers with evidence of how language is really used, evidence that can then be
A Ten-step Introduction to Concordancing through the Collins Cobuild Corpus Concordance Sampler

Table of Contents
1. Introduction
2. Getting started: find a word
3. Searching for lemmas and word families
4. Part of Speech Searches
5. More than one word
6. Searching for structures
7. Language varieties
8. Collocation
9. Searching for Answers
10. Further study

1. Introduction

If you are a student of English, a teacher, a translator or if you are writing in English, analysing English, or have any questions about how English works, concordancing can be of great benefit to you.

This mini-course is designed to introduce concordancing to students and teachers, native and non-native speakers of English and people with a general interest in language and learner autonomy. It is therefore catering its net over a wide range of people. It does not assume a great deal of linguistic knowledge: all required terminology, both in computer use and linguistics, is explained as it is introduced. The main purpose of these sessions is to introduce the techniques involved in searching for answers to questions. Just what can be asked will be revealed at every step, as we see how searches can be formed and refined.

This first Session is a short introduction. In Session 2 we will dive directly into concordancing using the Collins Cobuild Corpus Concordance Sampler. This has been chosen because it is readily accessible through the internet and because of its rich variety of functions that demonstrate many features of full concordancers. Throughout these sessions it will be shortened to CCS.

If you are a non-native speaker of English, it is likely that you will want your English to be as close as possible to the norms of English. You might even think that English is spoken and written with so much variation that the norms are too unstable to be grasped. There is, of course, a core language which represents the vast majority of English, without which it could not be called one language. It is not always easy to gain access to these norms, i.e. the most likely way a native speaker would express something, and grammar books are not always able to answer questions, especially about word use: modern dictionaries, based on corpus research, provide more reliable information. Corpus-based grammar books now exist also, but as far as they are of more use to the language professional than the general user. One of these is the Longman Grammar of Written and Spoken English. Read a review of it in TESL-EJ.
Corpus searching with AC/DC

A concordance - concordância - finds all the sentences containing the search word or phrase, and lists them in a standard format - in this case, in order of occurrence in the corpus, with the search phrase emboldened. (Other programs allow more control over the layout of the output)

Extract from concordance of SEMPRE (14.1.2002)

É uma das mais antigas discotecas do Algarve, situada em Albufeira, que continua a manter os traços decorativos e as clientelas de sempre, e continua a manter os traços decorativos e as clientelas de sempre. É um pouco a versão de uma espécie de «outro lado» da n

E razão desta escolha é, obviamente, a progressão demente da Frente Nacional, que prospera sempre a apontar o imigrante como bode expiatório e simultaneamente como a fonte de todos males do povo francês.

Carrington fez sempre questão de salientar que as hipóteses de sucesso do cessar-fogo dependem sobretudo dos beligerantes -

Com argumentos economicistas e de operacionalidade, o Executivo de Cavaco Silva sempre se escusou a concretizar o SIED, cujas competências foram, entretanto, transferidas para o SIM (Serviços de Informações Militares), por via de um polêmico acto administrativo do Governo, que assim chamava a si matérias da exclusiva competência da AR.

The current version of AC/DC displays the whole sentence containing the word(s) requested. This has advantages and disadvantages over the previous method of displaying a fixed number of characters on either side, as can be seen from an earlier version of the first example, where que is truncated to an ambiguous e, but part of the following sentence is included:

e continua a manter os traços decorativos e as clientelas de sempre, e continua a manter os traços decorativos e as clientelas de sempre. É um pouco a versão de uma espécie de «outro lado» da n
Corpus-specific tutorials
http://www.linguateca.pt/COMPARA/Tutorial.doc

1. QUICK START TO THE CORPUS

Choosing your working language .................................................. 2
Start using COMPARA ................................................................. 2
Searches that work .................................................................. 3
  Query syntax ........................................................................ 3
  Realistic queries ................................................................... 4
Where are the results from? ....................................................... 5
Books and articles about using corpora in language teaching


Where can teachers learn about corpora?

- Corpus specific tutorials
- General introductions to corpora
- Books and articles about using corpora in language teaching

Are they not enough?
What else can we do?

Few teachers use corpora
no studies yet of how they use them

Some studies of how novice users behave
and most teachers are novice users

Starting point

novice-user behaviour
Novice-user behaviour

Bernardini (2000)
Translation students using the BNC

Kennedy & Miceli (2001)
Intermediate students using the Contemporary Written Italian Corpus

Chambers (2004)
Undergraduate language students using corpora to write essays

Frankenberg-Garcia (2005)
Translation students combining the use of corpora, termbanks, the Web and paper references

Santos & Frankenberg-Garcia (submitted 2005)
Anonymous user logs of the COMPARA corpus

Help messages to COMPARA
4th year undergraduates using corpora in applied translation

Corpus skills that come as second nature to experts are not obvious to everyone
Novice-user behaviour

Corpus-specific problems
different search interfaces and CQLs

Need to improve human-computer interaction

A number of very basic problems,
no matter which corpus is used
Novice-user behaviour

Choosing between different types of corpora

Using a general language corpus to look up technical terms
 e.g. choosing the BNC to look up electrostatic precipitator

Using a corpus from the early nineties to look up new words in the language
 e.g. choosing the BNC to look up bluetooth

Harald Bluetooth
Novice-user behaviour

Choosing between different types of corpora

Using a parallel corpus of fiction to look up words unlikely to turn up in it

e.g. choosing COMPARA to look up the translation of
Special Tax Indemnity
cupuaçu
Novice-user behaviour

Using sub-corpora

Not using them at all
- using the whole BNC all the time
- not separating written from spoken language in Collins Worbanks Online (COBUILD)
- not separating translated from untranslated language in COMPARA

Using them too restrictively
- using only the Brazilian translations in COMPARA for general queries that needn’t be restricted to translated Brazilian Portuguese
Novice-user behaviour

Formulating corpus queries

Too general
What does DC (in a Colin Dexter novel) mean?
NU look up in the BNC: DC

Too restrictive
Can you perform a contract?
NU look up in COMPARA: perform a contract

No follow-up queries
Can’t find out what DC means.
Can’t perform a contract
Novice-user behaviour

Formulating corpus queries

Dictionary strategies - uninflected forms

COMPARA log files

“coxear” : hobble, hobbled
Lemma “coxear” : hobble, hobbeld, limps, limping, creeps

“cutucar” : NO HITS
Lemma “cutucar”: poking, nudges, shaken
Novice-user behaviour

Formulating corpus queries

Search-engine strategies: leaving out stop words

COMPARA log files

“congratulations” “World” “Cup” : NO HITS

“Virgem” “lábios” “mel”
“a” “virgem” “dos” “lábios” “de” “mel” : NO HITS

: the maiden with lips of honey
the virgin with the honey lips
the maiden of the honied lips
Novice-user behaviour

Formulating corpus queries

Search engine strategies: case insensitive

COMPARA log files

CONTABILIDADE : NO HITS
contabilidade : accounting, accountant’s, account, books, doing the books, book-keeping

“i’d” “love” “to” : NO HITS
“I’d” “love” “to” : adorava, adoraria
gostaria muito, bem gostava
quem me dera
Novice-user behaviour

Formulating corpus queries

Search engine strategies: no accents

COMPARA log files

cunteudo : NO HITS

cunteúdo : contents, content, upshot, inside, load
Novice-user behaviour

Formulating corpus queries

Misconceptions about the kind of information that can be retrieved from a corpus

COMPARA log files
Na sequência de conversa com o Dr. Magalhães Ramalho e tendo existido algumas dúvidas quanto ao valor atribuído ao imóvel, venho por este meio clarificar o seguinte

d this still did not give me the happiness I thought it would or for which I sought
Novice-user behaviour

Formulating corpus queries

Misconceptions about the way chunks of words behave

COMPARA log files
water shining
bill quantities
calling with the palm
mad honey
like a manor
Novice-user behaviour

Interpreting corpus data

Not taking corpus size into account
2 hits/20 K words = 2 hits/20 M words!

Not taking corpus composition into account
Not in the BNC, therefore not English!

No experience of dealing with undedited data
Found it in the BNC, therefore it’s English!

Making a summary analysis of results
Found it, never mind near what! (not checking the co-text)
Found it, never mind where! (not checking the context)

Being lured by misleading near matches
Looks like it, yeah, yeah... That’s it!
Need to develop corpus awareness

Language teachers are familiar with dictionaries, grammar books, texts books (and the Web).

Difficult to grasp that corpora do not work in the same way.
Need to develop corpus awareness

Corpus size

Pocket dictionary

100 K words

OED

100 M words
Need to develop corpus awareness

Corpus composition

Learner dictionary  Bilingual dictionary  Thesaurus  Encyclopaedia

General language corpus  Newspaper corpus  Multilingual corpus  Spoken language corpus
Need to develop corpus awareness

Formulating corpus queries

Dictionary strategies
uninflected forms

Too limited!
Need to develop corpus awareness

Formulating corpus queries

Web-browsing strategies
No stop words, no accents, case-insensitive anything (even spelling mistakes and the most outrageous things)

 Doesn’t work!
Need to develop corpus awareness

Interpreting corpus data

Dictionaries, grammars, text books, etc.
Written by experts, carefully edited, revised, explained...

Mistakes, idiosyncrasies...
Too many or not enough hits...
Relative frequencies...
Unexpected things...
My own conclusions???

CORPORA

!?
Where can teachers learn about corpora?

- General introductions to corpora
- Corpus-specific tutorials
- Books and articles about using corpora in language teaching
- Basic corpus skills
Raising teachers’ awareness to the basics of corpora

Examples of hands-on, task-based consciousness-raising exercises

To help teachers understand
1. Different types of corpora
2. How to retrieve information from a corpus
3. How to evaluate that information
Raising teachers’ awareness to the basics of corpora

To begin with, teachers don’t have to make their own corpus

Easy, online access
Different sizes
Different text types

A few EN examples
The BNC (simple search)

http://www.natcorp.ox.ac.uk/using/index.xml.ID=simple

About the BNC

The British National Corpus (BNC) is a 100 million word collection of samples of written and spoken language from a wide range of sources, designed to represent a wide cross-section of current British English, both spoken and written. [more]

Search the Corpus

Look up: [Enter text]

You can search for a single word or a phrase, restrict searches by part of speech, search in parts of the corpus only and much more.

The search result will show the total frequency in the corpus and up to 50 examples. [more information]

News from the BNC

- New BNC website
- BNC goes XML
- New version of Xaira
Corpus Concordance Sampler

The Collins WordbanksOnline English corpus is composed of 56 million words of contemporary written and spoken text. To get a flavour of the type of linguistic data that a corpus like this can provide, you can type in some simple queries here and get a display of concordance lines from the corpus. The query syntax allows you to specify word combinations, wildcards, part-of-speech tags, and so on.

Type in your query: impediment

Which sub-corpora should be searched?

- British books, ephemera, radio, newspapers, magazines (36m words)
- American books, ephemera and radio (10m words)
- British transcribed speech (10m words)

To get sample concordances, press this button:

To set concordance width (in characters), make a selection:

Note that output from this demo facility will be restricted to 40 lines of concordance, each with a maximum width of 250 characters. The lines to be displayed will be selected at random.

Collocation Sampler

Type in your word:

Select a significance score to be calculated:
OPUS - Corpus query (CWB)

CQP query (CWB)
A CQP query consists of a regular expression over attribute expressions.

Example queries

Select

Show max hits and skip non-aligned segments

Horizontal
(advanced search)
COMPARA Simple Search

A simple search enables you to search the whole of COMPARA either in the Portuguese-English or in the English-Portuguese direction.

From Portuguese to English

Enter a word or a sequence of words in Portuguese. Put quotation marks around each separate word (e.g. "Até" "logo")

Search (from Portuguese to English)  Clear

Case insensitive  Diacritic insensitive

Help

From English to Portuguese

Enter a word or a sequence of words in English. Put quotation marks around each separate word (e.g. "See" "you")

Search (from English to Portuguese)  Clear

Case insensitive  Diacritic insensitive

Help
Business Letter Corpus
http://ysomeya.hp.infoseek.co.jp/

Business Letter Corpus
Online KWIC Concordancer

Search String
Search Type
Equal to
Line Width
40
Search Corpus
Business Letter Corpus (BLC 2000)
Sort Type
Right

Click [Submit] or [Reset]

Your search can be rejected when the server is busy (or when you overload the server by submitting a search string whose frequency is extremely high). If it happens, wait a few minutes before you try again. In the meantime, you may want to try your search with a smaller corpus, which is less likely to be rejected.

Bigram plus
A search program for any two-word combination within intervening words.

Instructions for first-time users (Last updated: Mar 22, 01)
Sample Exercises for Data-driven Learning (Japanese Only; Last updated: Mar 13, 01)
Regular Expressions for Beginners
Guide to Letter Format & Conventions (Japanese Only)
Bilingual Model Business Letters (not ready yet)
BLC Selected Parts
Raising teachers’ awareness to the basics of corpora

<table>
<thead>
<tr>
<th>(Sub-)Corpus</th>
<th>Size</th>
<th>Type of English</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>BNC</td>
<td>100 M</td>
<td>General British</td>
<td>early to mid 1990s</td>
</tr>
<tr>
<td>Collins speech</td>
<td>10 M</td>
<td>Spoken British</td>
<td>contemporary?</td>
</tr>
<tr>
<td>COMPARA - EN</td>
<td>1.5 M</td>
<td>Original and translated fiction</td>
<td>1837- 2002</td>
</tr>
<tr>
<td>Business Letter Corpus</td>
<td>1 M</td>
<td>US and UK business letter samples</td>
<td>since 2000</td>
</tr>
</tbody>
</table>

But what does this mean?
Raising teachers´ awareness to the basics of corpora

1. understanding different corpora
Understanding different corpora

different corpora exercise

Something old
Something new
Something common
Something rare
Something oral
Something written
Something technical
Something regional
Something sentimental
Something religious
Something political
Something foreign

counterpane
MP3
with
epicure
d’you
amiable
pelagic
lass
darling
rosary
coalition
rapporteur
Understanding different corpora

<table>
<thead>
<tr>
<th>different corpora exercise</th>
</tr>
</thead>
<tbody>
<tr>
<td>old</td>
</tr>
<tr>
<td>BNC</td>
</tr>
<tr>
<td>-----</td>
</tr>
<tr>
<td>counterpane</td>
</tr>
<tr>
<td>MP3</td>
</tr>
<tr>
<td>d'you</td>
</tr>
<tr>
<td>amiable</td>
</tr>
<tr>
<td>pelagic</td>
</tr>
<tr>
<td>lass</td>
</tr>
<tr>
<td>Darling</td>
</tr>
<tr>
<td>rosary</td>
</tr>
<tr>
<td>coalition</td>
</tr>
<tr>
<td>rapporteur</td>
</tr>
</tbody>
</table>

Different corpora will give you different results.
Understanding different corpora

going to know a specific corpus exercise

• Choose a corpus
• Read the information about it
• Based on this info, try to predict:
  – Frequent words and expressions
  – Words and expressions you won’t find in the corpus
• Test your predictions
Understanding different corpora

getting to know a specific corpus exercise

Business Letter Corpus

Frequent
Yours sincerely 1462
looking forward to 159
Thank you for 1312
I am pleased to 78
We regret 79

Unlikely
Who’s there? 0
I love you 0
very funny 0
Cheerio 0
soup 3

At least we can provide a bowl of soup and a safe place to sleep. the IRS can be as frustrating as eating soup with a fork. relayed to him how much you enjoyed the soup.
Understanding different corpora

corpus size exercise

old new common rare oral written technical regional sentimental religious political foreign

counterpane MP3 with epicure d’y amiable pelagic lass darling rosary coalition rapporteur

BNC BNC sampler 2 M (1/50)

41 0

When size matters...

2K 2 2 K 41 0 29 29
Raising teachers’ awareness to the basics of corpora

2. retrieving information from a corpus
Retrieving information from a corpus
Corpora are not like dictionaries exercise
Carry out a search for look in Collins Wordbanks online
Retrieving information from a corpus

Corpora are not like dictionaries exercise

Now do a search for looks

supporters. The Office programme for the day looks like this [p] 1. 2000 "Weather Alert" [h] [p] is a quiet and powerful workhorse that looks like a thoroughbred -- particularly:

Mon-Sat 10-5. Fascinating exhibition that looks at working-class decoration. Include:

with HOLD, [p] 3.45am, ITV: AUSTIN CITY LIMITS looks at Progressive Country Music For f's plaster pillars lend grandeur. Below: Cream looks as welcoming outside as in. Sunflower:

other hand, with his blonde hair and sunglassess, looks like he should be giving surfing les:

A MESS OF IT AS LEEDS DID LAST YEAR. 90 MINUTES looks WHAT WENT WRONG AND WHAT HOWARD W.

PHOTOS WITH CAPTIONS [/c] [h] Inspirational New looks [/h] History and fantasy set the scene:

Box 29580 [p] SHORTS/LEGS INTEREST. Slim 33, looks better. MACPHEERSON: So, wild bore on the menu looks musical talent could be substituted for good looks:

of the talks. Malcolm Harlott of the BCC looks comfortable doing business with a woman who looks can trace his lineage back to the Egyptians and looks stood out in the crowd with his striking good looks: You couldn't forget him in a hurry!

can forgive much in an independent film that looks at itself in a crazy mirror never chiff safety and audio equipment. Simon Price, who looks, although she was "just a housewife"

be placed? The danger of design by committee looks, ominuous. [p] The Sports Council is:

dead men were "martyrs" [p] [h] New instruction looks wide open to trainers' excuses: Racing

world's leading rock acts when touring Asia, looks set to be remembered only as a white:

and she's a marvellous woman [p] Marianne now looks after the couple's three children -

swamp the High Street next summer. [p] Other looks to emerge paid lip-service to the Se:

famous Western lawman. [p] He enjoys tender looks for Jaguar. So how do companies decide just who looks been very proud. [p] [h] On paper this trip looks dull, People Today [/h] [b] Christopher

old. And your picture was so unflattering. He looks stunning on his new video. [p] Jenny
Retrieving information from a corpus

Corpora are not like dictionaries exercise

Now try a search for looked

might say, fruit cakes. [p] [p] I looked hopefully through this year's offers. Uxbridge, you’ve lost your leg Lord Uxbridge turns. There is a half-smile on her face. She looked down and said, 'By God, Sir, so I have.'

But, speaking through an interpreter, he came coming home with me for a drink?' [p] Alistair looked back as well as forward. I have come to his watch. 'I'm not sure that I can' looked as if the words hurt. 'She told us all looked straight ahead, absorbed, still from looked from the snowperson to Brian. 'I think looked as though two of them had fired Mann looked into the closet, found both her big looked up at Pete in alarm. Bob and Jupiter looked at each other, shrugged, and took the looked on in amazement. They could not under looked at, so they wanted to close it quick looked left, looked right, then continued s looked the more dangerous side, but Beardsley looked an improved horse until caught out by looked more successful. On my second visit, looked like a broken man this season. With looked after by their grandmother. [p] And looked at her privates. [p] WYRE: So she d looked like being the south London side's he looked good. [p] And when that free kick f looked frightened. [p] Come on, " the Big No looked in the midst of its gold rush days. looked startled. [p] Ted, do you see that a looked thoughtful, and for a moment she ima
Retrieving information from a corpus

Corpora are not like dictionaries exercise

Do the same for looking

Uninflected forms not always good idea!
Retrieving information from a corpus

Corpora are not like dictionaries exercise

Read the information on the CQL and try and find out how to obtain results for look, looks, looked and looking all in one go.

international community to realise this and to fashion trends, but also need their clothes to fabrics and trimmings to create a truly special two kids. Now only 5 per cent of our families counterparts. [p] [F] This book takes a broad view of how to stand indicts, and eras whilst devoted in Pride And Prejudice [p] Though she could not 16 months and on last night's form Palace don't new model for our activities and must therefore in marriage. He may even first dare to

It's a kineticcally shot, abruptly edited. Your Ears [p] With the beautiful, wood-grained history books? [FOX] I don't know. Have a white quartz there. But come and have a [2F0] married her master but she didn't really practical emotional to develop skills to when you saw it [F02] Yeah [M02] because of the

14 a wee, she scored on his debut and never nearest neighbour. Where else was there? She did. The first Orphan knew of it was when she somebody who had had too much to drink. He its snout pointing upward. [p] The Wolf's hair was thick and shiny. Was it as soft as it then Xavier appeared. Although they hadn't [2F1] I [2F0] I got into the cafe and then I rather than spike lee, even to the point of No 77606. Tel NORTH-WEST professional yowl

looking outside Hollywood if Hollywood itse

looking lesbian. 37. honest, sensitive, car-
Retrieving information from a corpus

Corpora are not like dictionaries exercise

Go back to your results for look. Is it always a verb?

Corpora are not like dictionaries exercise.

Go back to your results for look. Is it always a verb?
Retrieving information from a corpus

Corpora are not like dictionaries exercise

Read the information on the CQL and try and find out how to obtain results only for noun forms of the word look

begging or sleeping rough on the streets. One look in their eyes and you can tell that they are not MFI. (p. 103) with captions this was not enough for extra interest. Is an area of the basement, we stood in the greying beard stall, I panicked and beside the door illustrate the reasons why I didn’t say “what did he say?” Many of the artists in here were hidden in the range has covered streets, allowing a job for a better same old anonymous. Die-hard romantics can achieve a similar look at home as well - far more enticing than you walk through a cosmetics department, take a look around. The floor is filled with people. It’s just too rich.” (p. 46) From the baffled an’ the guy maybe leaves huh f’ another girl an’ Her short hair. Her small body. And that same spacious, but there is a lived-in, cluttered strong tapered tail. It has a tousled kind of teens and teen-agers. They stop for a brief look and then return to their afternoon play.
Retrieving information from a corpus

Corpora are not like web browsers exercise

Look up the English for Protocole sur les privilèges et immunités in the EUROPARL corpus

First try (without stop words)
Retrieving information from a corpus

Corpora are not like web browsers exercise

OPUS - Corpus query (CWB)

Protocole sur les privilèges et immunités
Protocol on the privileges and immunities
Protocol of the privileges and immunities
Protocol on privileges and immunities
Retrieving information from a corpus

Corpora are not like web browsers exercise

Third try (stop words + case insensitive)
Retrieving information from a corpus

Corpora are not like web browsers exercise

Fourth try (case-insensitive + wildcards instead of stop words)
Retrieving information from a corpus

Corpora are not like web browsers exercise

OPUS - Corpus query (CWB)

corpus | languages | show attributes | alignments
---|---|---|---
EUROPARL | da de el en es fi fr it nl | | |

CQP query (CWB) A CQP query consists of a regular expression over attribute expressions. Introduction of the query syntax

protocole sur les privilèges et immunités (16)
protocole sur les privilèges et les immunités (6)
protocole sur les immunités (3)
protocole des privilèges et immunités (1)
protocole des immunités (1)
protocole sur les prérogatives et les immunités (1)
protocole relatif aux immunités (1)
different English equivalents
Retrieving information from a corpus

Corpora are not like web browsers exercise

**OPUS - Corpus query (CWB)**

**CQP query (CWB):**
A CQP query consists of a regular expression over attribute expressions.

**Introduction of the query syntax**

**Example queries**

- "protocole" [] [1,5] "immunités"

**alignments**

- positional annotation
- word, id, len, pos
- skip non-aligned segments
- vertical, KWIC
- horizontal

**Query string:** ""protocole" [] [1,5] "immunités"

- 0 hits found

**Sixth try**

case-insensitive
any 1 to 5 words between **protocole** and **immunités**
no accents
Retrieving information from a corpus

Protocole sur les privilèges et immunités

It was okay as far as I could see

EUROPARL

COMPARA
It was okay as far as I could see 0
was okay as far as I could see 0
okay as far as I could see 0
as far as I could see  3
far as I could see   3
as I could see   4
I could see 117
could see 249
see 2214
It 16005
It was 3268
It was okay 2
It was okay as 0
Retrieving information from a corpus

It’s English, but it’s not in the BNC

“As a rule of thumb you need a litre of paint to every 12 square metres of wall”
As a rule of thumb, you need a litre of paint to every 12 square metres of wall.

290

Which ones are likely to turn up?
Which ones won’t turn up?
Which one will be the most frequent one?
Raising teachers’ awareness to the basics of corpora

3. evaluating corpus data
Unlike dictionaries, the language of corpora is not revised (so corpora can include mistakes) But correct things tend to be a lot more frequent

*Reckonize
Recognize
*Payed
Paid
*Accomodation
Accommodation
Evaluating corpus data

CETEMPúblico
Portuguese National Newspaper
180 M words

*caiem : caem
44 : 896

DIACLAV
4 Portuguese Regional Newspapers
6 M words

6 : 11

Frequencies are relative...
Evaluating corpus data

co-text exercise

Look up congratulations + PREPOSITION in Collins Online

http://www.collins.co.uk - CorpusPopup - Microsoft Internet Explorer

and venues are as follows [p] LIST [h] Congratulations to [/h] John Onslow for his promoting of the two peoples he is in the United States, having unified the two people who have support. Although I was only at the post-show celebration for the 2008 Eurovision Song Contest, I did not check through their recent and controversial programme. Congratulations to the winners of our Children's Competition [h] by the club's Gallery is a company who've supported many from poor to rich and to the 11 other months. Congratulations on the birth of Thomas had the Chief Executive, Sir, congratulations to The Times for supporting the programme and the programme.

Don Luigi Stark, a better voyeur [p] than, when hundreds of people visited [h] congratulations on her marriage. [p] I was sitting at home and it would make me smile back on faces [h] congratulations on your bronze medal, you're fine. [p] [h] Rich B; Today Competition [h] p[ ] CONGRATULATIONS to Edward Futer, of Wallington, me as well [p] Mason received equally warm congratulations from Olazabal and third placed C. Slam winners; Today Competition [h] p[ ] CONGRATULATIONS to Peter Jonas, the latest of our Famous International Golf Club [h] p[ ] MANY congratulations to [p] P Jones and A Critchley, she received it in less than 48 hours. Congratulations to the Post Office. If Michael
Evaluating corpus data

co-text exercise

Look up Congratulations + (on|from|to) : what comes next?

Co-text is important:

Congratulations on
Congratulations to
Congratulations from
used for different purposes
Evaluating corpus data

context and medium exercise

Lookup *whatsit* in different sub-corpora of Collins Wordbanks online

**Corpus Concordance Sampler**

The Collins Wordbanks*Online* English corpus is composed of 56 million words of contemporary written and spoken text. To get a flavour of the type of linguistic data that a corpus like this can provide, you can type in some simple queries here and get a display of concordance lines from the corpus. The [query syntax](#) allows you to specify word combinations, wildcards, part-of-speech tags, and so on.

*Type in your query:* [whatsit]

*Which sub-corpora should be used?*

- British books, ephemera
- American books
- British travel

To get sample concordance lines, click button.
To set concordance lines, select concordance line and make a selection.

*Note that output from this tool is not filtered and may contain data from the 20th century with a maximum width of 250 characters. The lines to be displayed can be selected.**

**Collocation Sampler**

*Type in your word:*

*Select a significance score to be calculated:*
To summarize

Novice-user behaviour suggests that:

Teachers need help to understand
1. Different types of corpora
2. How to retrieve information from a corpus
3. How to evaluate that information

A few simple, hands-on, task-based consciousness-raising exercises

Too obvious for experts, but not self-evident for novice users

Many more are possible!
In conclusion

• Recognize that corpus skills are not obvious
• Important to:
  – raise teachers’ awareness to different types of corpora
  – train teachers in basic corpus skills

General introductions to corpora

Corpus-specific tutorials

Books and articles about using corpora in language teaching