

# 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

+ computers & Internet  
+ free online texts & corpora

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

	<p>^what a _bout a cigarle: *((4 sylls))* *[^wlon't have one th/a ^aren't you . going to si ^[/Am]# - ^have my _coffee in p= ^quite a nice . room to (a ctally))# *^lisn't it# *^y/les#*---</p>	<table border="1"><thead><tr><th>Verb</th><th>A</th><th>B</th><th>C</th><th>D</th></tr></thead><tbody><tr><td>can</td><td>210</td><td>148</td><td>59</td><td>89</td></tr><tr><td>could</td><td>120</td><td>49</td><td>36</td><td>23</td></tr><tr><td>may</td><td>160</td><td>86</td><td>15</td><td>46</td></tr><tr><td>might</td><td>24</td><td>29</td><td>13</td><td>4</td></tr><tr><td>must</td><td>43</td><td>34</td><td>12</td><td>28</td></tr><tr><td>ought</td><td>3</td><td>4</td><td>0</td><td>1</td></tr><tr><td>shall</td><td>12</td><td>4</td><td>0</td><td>10</td></tr></tbody></table>	Verb	A	B	C	D	can	210	148	59	89	could	120	49	36	23	may	160	86	15	46	might	24	29	13	4	must	43	34	12	28	ought	3	4	0	1	shall	12	4	0	10	
Verb	A	B	C	D																																							
can	210	148	59	89																																							
could	120	49	36	23																																							
may	160	86	15	46																																							
might	24	29	13	4																																							
must	43	34	12	28																																							
ought	3	4	0	1																																							
shall	12	4	0	10																																							
Section 1	Section 2	Section 3	Section 4																																								
<a href="#">Early Corpus Linguistics and the Chomskyan Revolution.</a>	<a href="#">What is a Corpus and What is in it?</a>	<a href="#">Quantitative Data.</a>	<a href="#">The Use of Corpora in Language Studies.</a>																																								

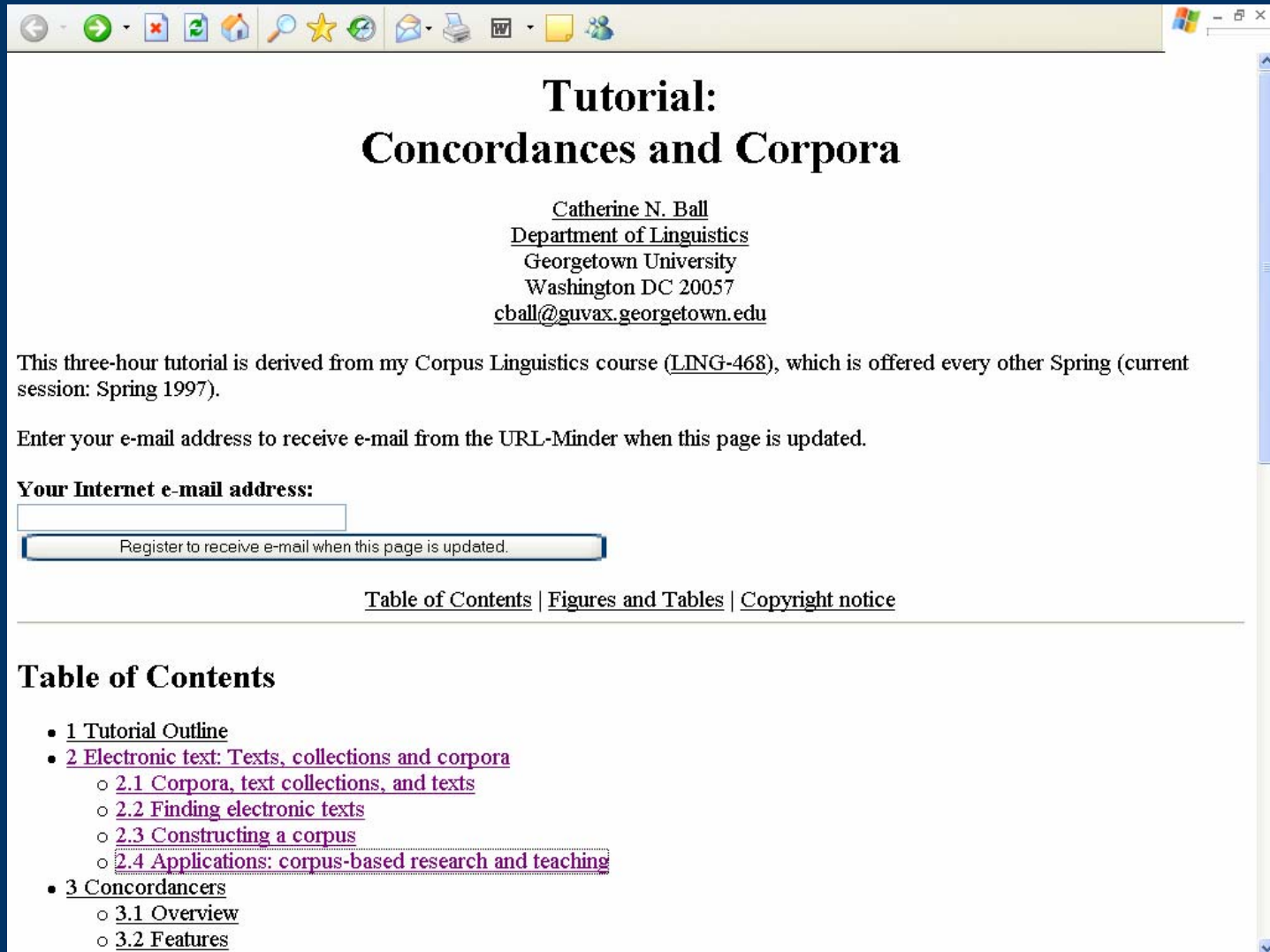
[What you won't find at this website.](#)



Web pages to be used to supplement the book "Corpus Linguistics" published by [Edinburgh University Press](#)  
ISBN: 0-7486-0808-7 (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@guvax.georgetown.edu](mailto:cball@guvax.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:**

  
 Register to receive e-mail when this page is updated.

[Table of Contents](#) | [Figures and Tables](#) | [Copyright notice](#)

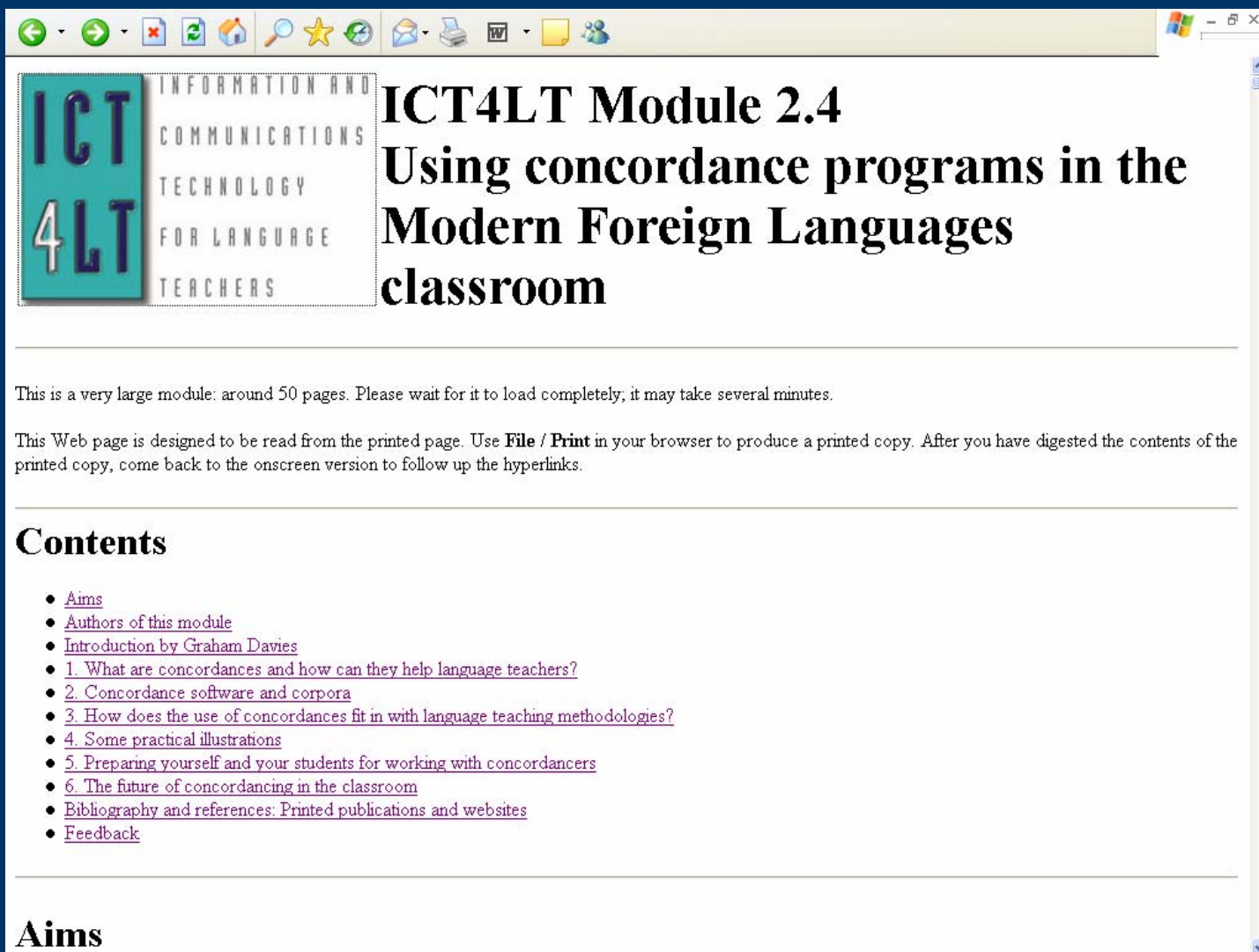
---

## 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

[http://www.ict4lt.org/en/en\\_mod2-4.htm](http://www.ict4lt.org/en/en_mod2-4.htm)



The screenshot shows a web browser window with a standard toolbar at the top. The main content area features a logo on the left with the text 'ICT 4LT' in large, stylized letters, and 'INFORMATION AND COMMUNICATIONS TECHNOLOGY FOR LANGUAGE TEACHERS' in smaller text to its right. The main heading is 'ICT4LT Module 2.4 Using concordance programs in the Modern Foreign Languages classroom'. Below the heading, there is a paragraph of text, followed by another paragraph, and then a 'Contents' section with a list of links. At the bottom left, the word 'Aims' is visible.

## 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.

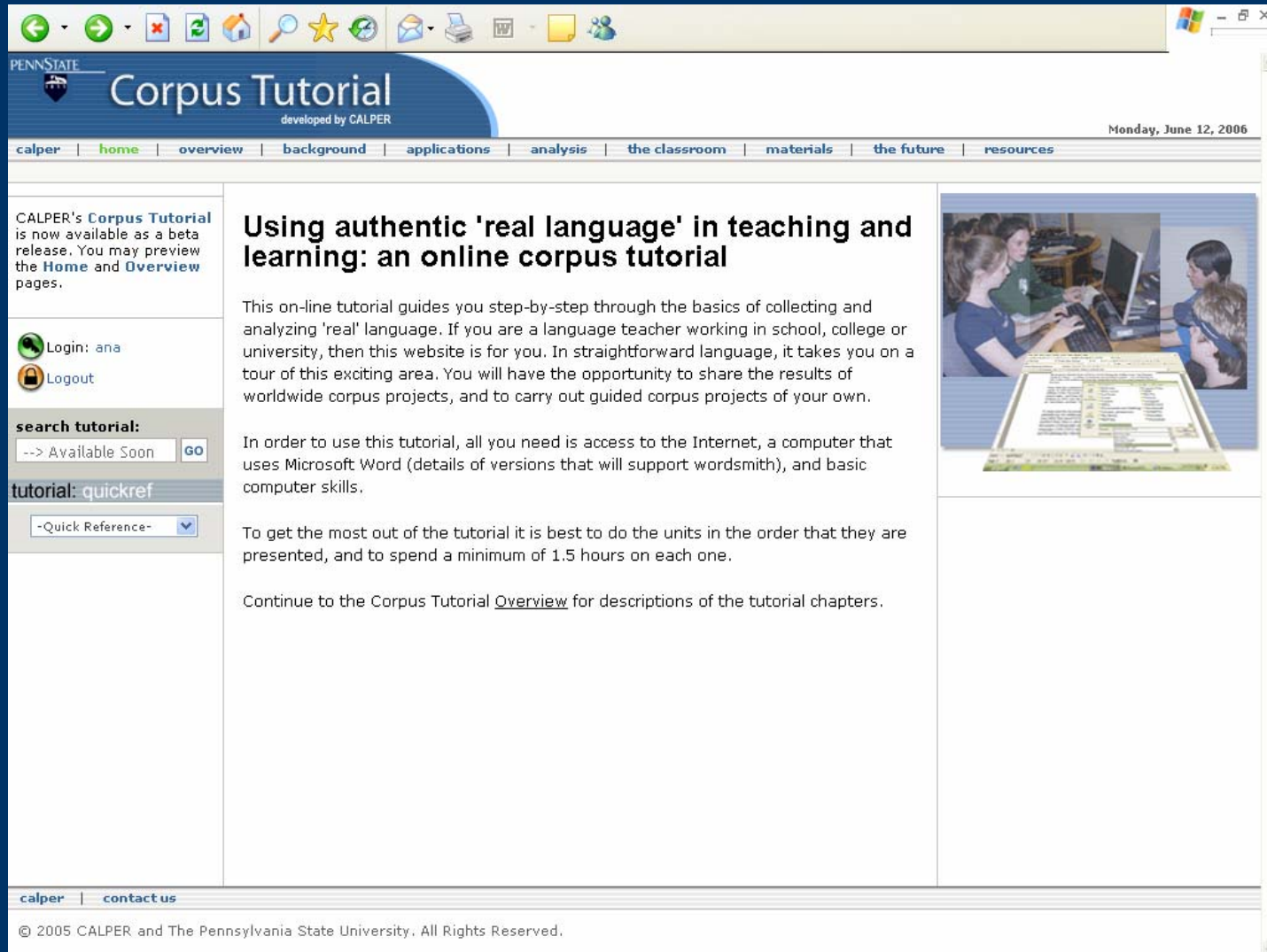
### Contents

- [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

# General introductions to corpora

<http://calper.la.psu.edu/corpus/tutorial/index.php>



The screenshot shows a web browser window displaying the Penn State Corpus Tutorial website. The browser's address bar shows the URL <http://calper.la.psu.edu/corpus/tutorial/index.php>. The website header includes the Penn State logo and the title "Corpus Tutorial" with the subtitle "developed by CALPER". A navigation menu at the top lists: [calper](#) | [home](#) | [overview](#) | [background](#) | [applications](#) | [analysis](#) | [the classroom](#) | [materials](#) | [the future](#) | [resources](#). The date "Monday, June 12, 2006" is displayed in the top right corner.

**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.

On the right side of the page, there is an image showing three people (two women and one man) sitting around a table, looking at a laptop screen. In front of them is a large, colorful diagram or chart with text and arrows, likely representing a corpus analysis or a tutorial interface.

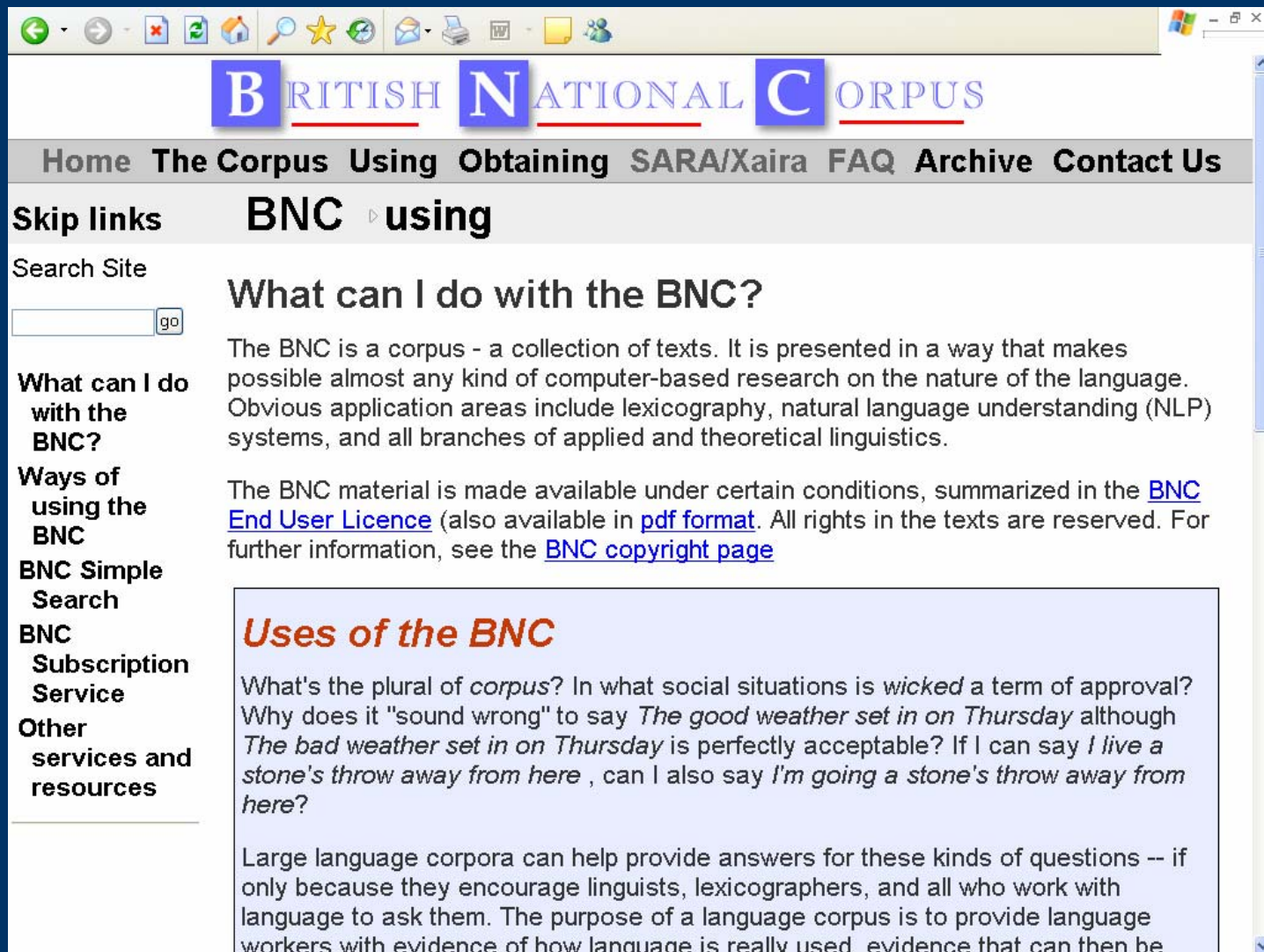
The left sidebar contains a login section with "Login: ana" and a "Logout" button. Below that is a "search tutorial:" section with a search box containing "--> Available Soon" and a "GO" button. At the bottom of the sidebar, there is a "tutorial: quickref" section with a dropdown menu set to "-Quick Reference-".

At the bottom of the page, there is a footer with the text "© 2005 CALPER and The Pennsylvania State University. All Rights Reserved." and a "contact us" link.



# Corpus-specific tutorials

<http://www.natcorp.ox.ac.uk/using/index.xml>



The screenshot shows a web browser window displaying the British National Corpus (BNC) website. The browser's address bar is empty, and the page title is "BRITISH NATIONAL CORPUS". The navigation menu includes "Home", "The Corpus", "Using", "Obtaining", "SARA/Xaira", "FAQ", "Archive", and "Contact Us". The main content area is titled "BNC > using" and features a search box with a "go" button. The page content includes a section titled "What can I do with the BNC?" which explains that the BNC is a corpus of texts used for computer-based research in linguistics. It also mentions that the BNC material is available under certain conditions, summarized in the "BNC End User Licence" (available in pdf format) and that all rights in the texts are reserved. A sidebar on the left contains links for "Skip links", "Search Site", "What can I do with the BNC?", "Ways of using the BNC", "BNC Simple Search", "BNC Subscription Service", and "Other services and resources". A highlighted box titled "Uses of the BNC" contains a paragraph of text discussing the plural of "corpus" and the use of "wicked" as a term of approval, along with a question about the plural of "stone's throw".

**BRITISH NATIONAL CORPUS**

Home The Corpus Using Obtaining SARA/Xaira FAQ Archive Contact Us

Skip links **BNC > using**

Search Site

**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*?

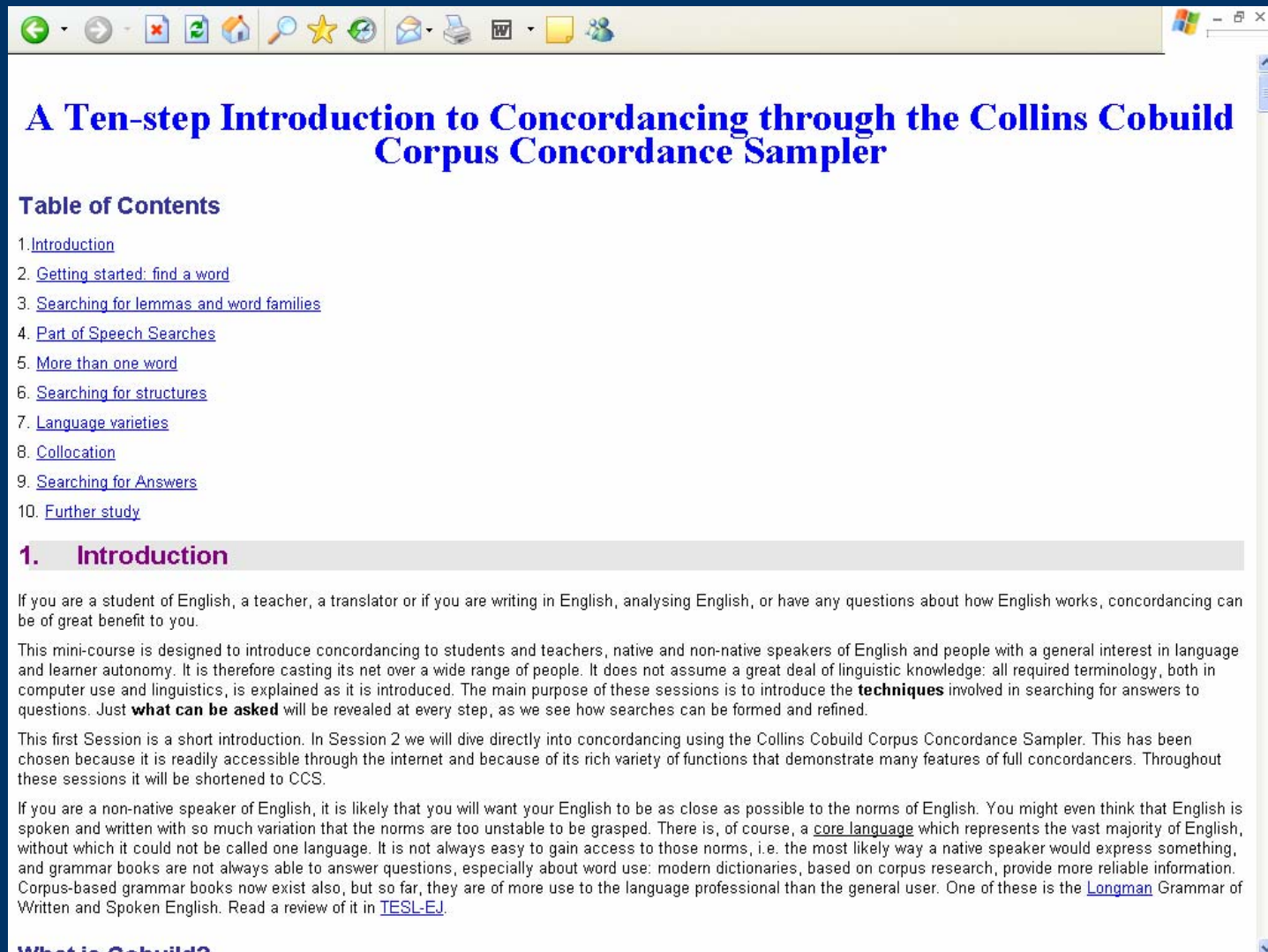
Large language corpora can help provide answers for these kinds of questions -- if only because they encourage linguists, lexicographers, and all who work with language to ask them. The purpose of a language corpus is to provide language workers with evidence of how language is really used, evidence that can then be



# Corpus-specific tutorials

<http://web.quick.cz/jaedth/Introduction%20to%20CCS.htm>

By James Thomas, Masaryk University, Czech Republic



**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 casting 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.

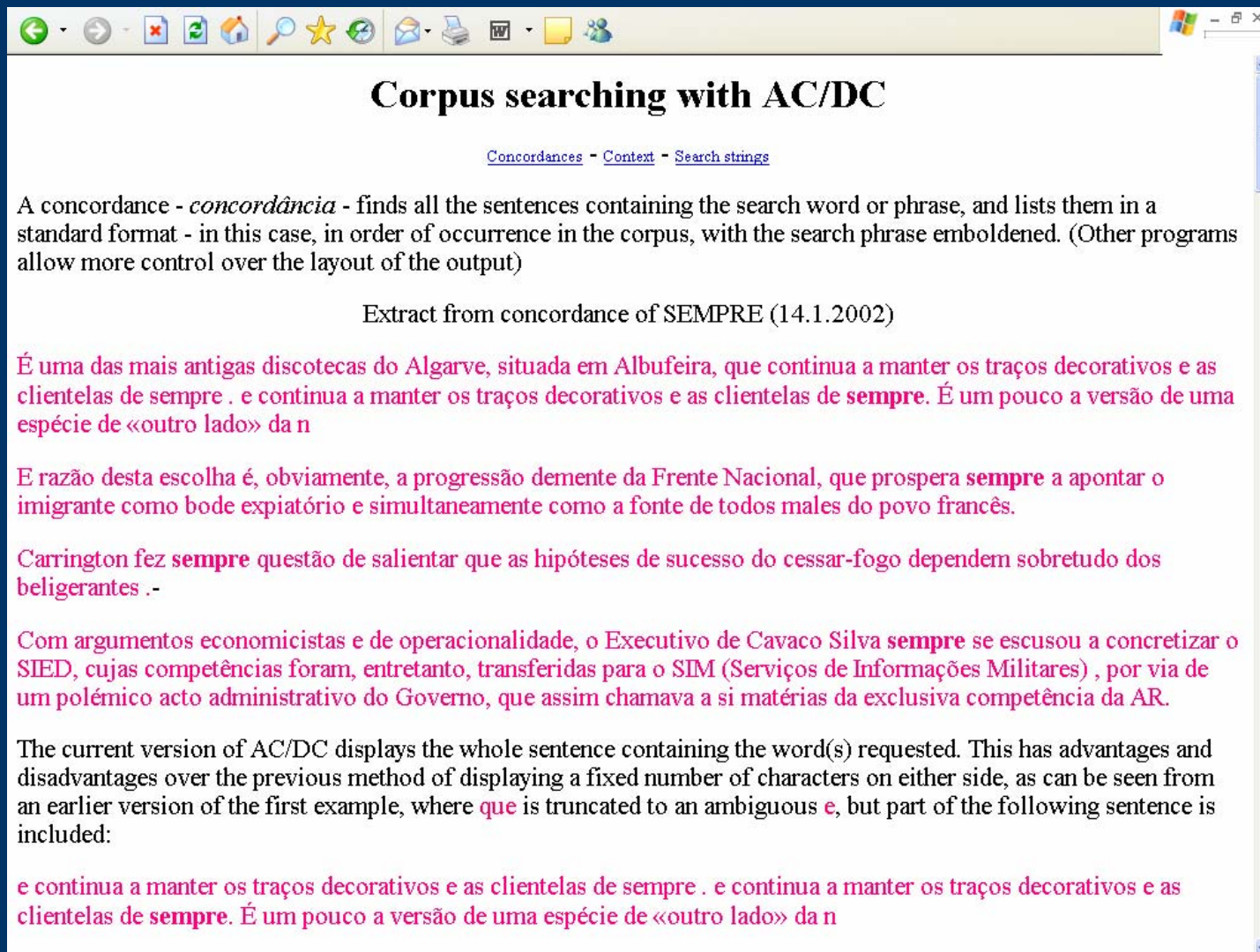
If you are a non-native speaker of English, it is likely that you will want your English to be as close as possible to the norms of English. You might even think that English is spoken and written with so much variation that the norms are too unstable to be grasped. There is, of course, a core language 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 those 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 so far, 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.

**What is Cobuild?**

# Corpus-specific tutorials

<http://users.ox.ac.uk/~srp/corpussearching.html>

By Stephen Parkinson, Oxford University



The screenshot shows a web browser window with a yellow title bar and a toolbar containing various icons. The main content area has a white background and a blue border. The title 'Corpus searching with AC/DC' is centered at the top in a bold, black font. Below the title are three links: 'Concordances', 'Context', and 'Search strings'. The text explains the function of a concordance and provides an example from a corpus. The example text is in pink and contains the word 'sempre' in bold. The text is followed by a paragraph explaining the current version of the AC/DC software and its advantages over a previous version.

## Corpus searching with AC/DC

[Concordances](#) - [Context](#) - [Search strings](#)

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 polémico 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>

**COMPARA English Tutorial**  
Ana Frankenberg-Garcia 28/09/2004

## 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

- Aston, G. (ed.) (2001) *Learning with corpora*. Houston: Athelstan.
- Johns, T. & P. King (eds.). (1991) *Classroom Concordancing*. Birmingham: The University of Birmingham Centre for English Language Studies.
- Sinclair, J. (ed.) (2004) *How to Use Corpora in Language Teaching*. Amsterdam: John Benjamins.
- Tribble, C. & G. Jones. (1997) *Concordancing in the classroom: a resource guide for teachers*. Houston: Athelstan.
- TaLC Proceedings 1994, 1996, 1998, 2000, 2002, 2004

and many more....

# Where can teachers learn about corpora?

General introductions to corpora

Corpus specific tutorials

**Are they not enough?**

Books and articles about using corpora in language teaching

# 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 Workbanks 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, hobbled, 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” : **NO HITS**

“a” “virgem” “dos” “lábios” “de” “mel” : 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**

conteudo : **NO HITS**

conteúdo : 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

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**

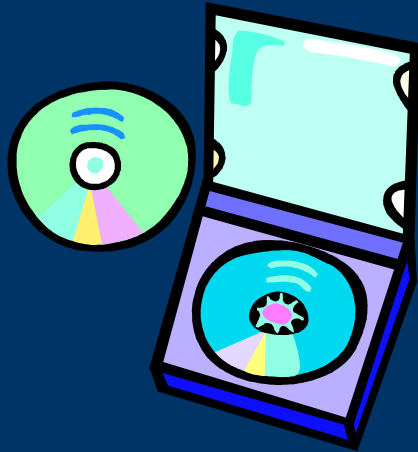
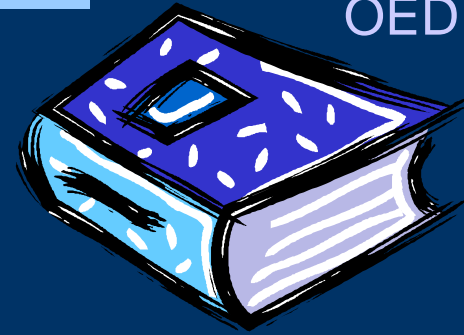
# Need to develop corpus awareness

## Corpus size

Pocket dictionary



OED



100 K words



100 M words

# Need to develop corpus awareness

## Corpus composition



Learner dictionary



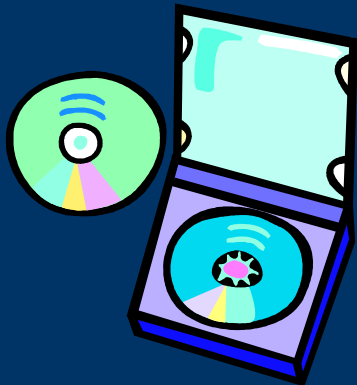
Bilingual dictionary



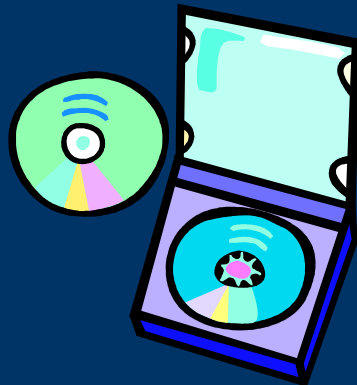
Thesaurus



Encyclopaedia



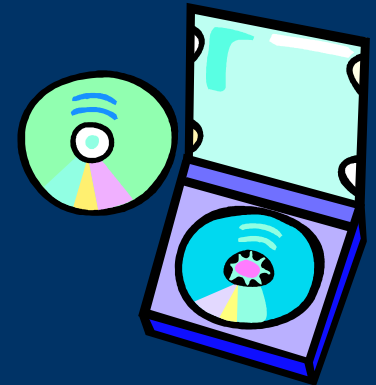
General language corpus



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!

CORPORA



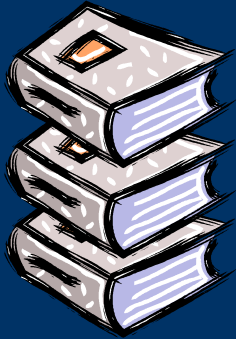


# Need to develop corpus awareness

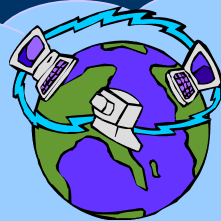
## Interpreting corpus data

**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?

**basic corpus skills**



General  
introductions  
to corpora

Corpus-specific  
tutorials

Books and articles  
about using corpora  
in language teaching

# 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

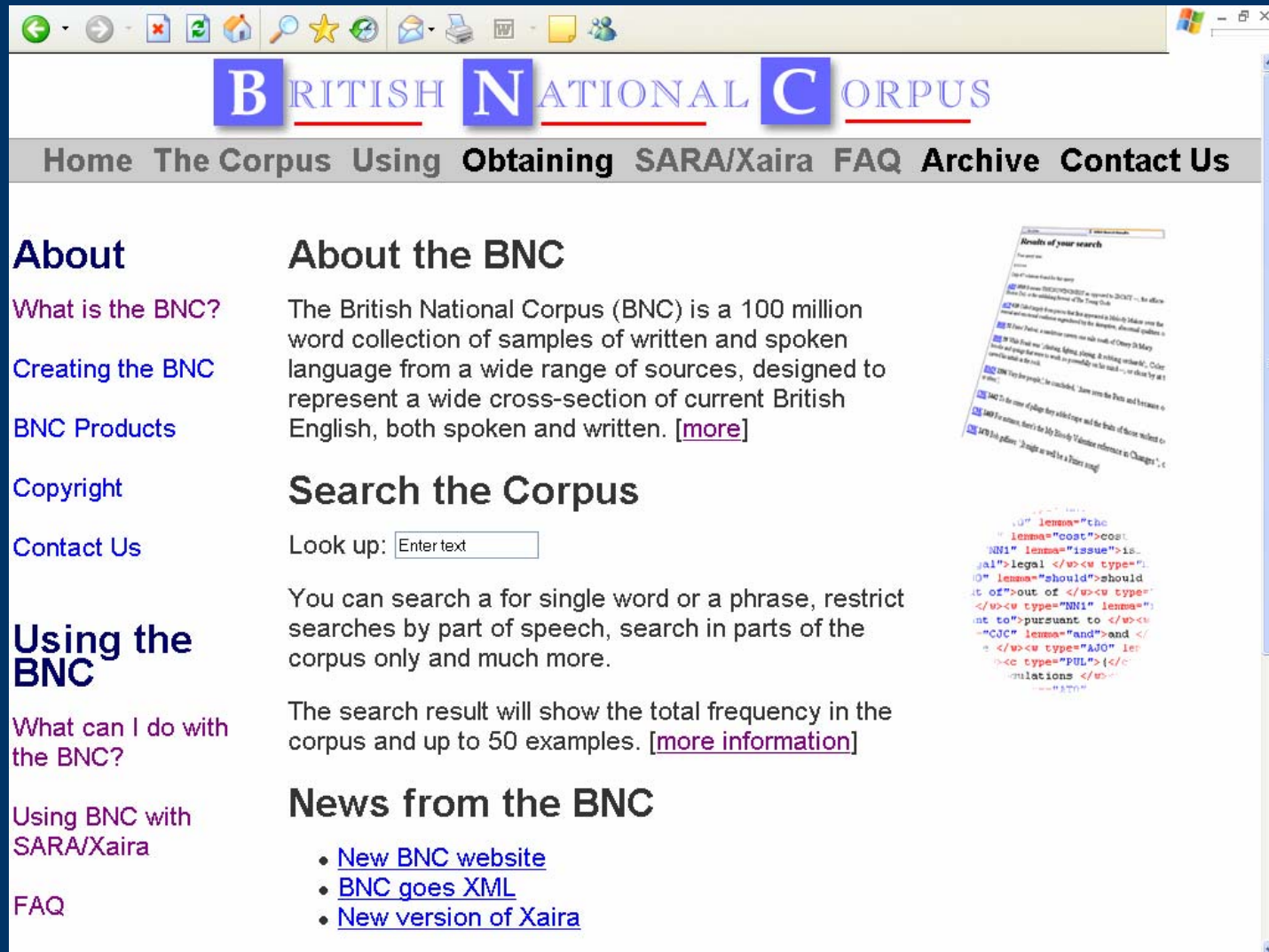
To begin with, teachers don't have to make their own corpus



A few EN examples

# The BNC (simple search)

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



**BRITISH NATIONAL CORPUS**

Home The Corpus Using Obtaining SARA/Xaira FAQ Archive Contact Us

## About

[What is the BNC?](#)

[Creating the BNC](#)

[BNC Products](#)

[Copyright](#)

[Contact Us](#)

## 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:

You can search a for 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. [\[more information\]](#)

## Using the BNC

[What can I do with the BNC?](#)

[Using BNC with SARA/Xaira](#)

[FAQ](#)

## News from the BNC

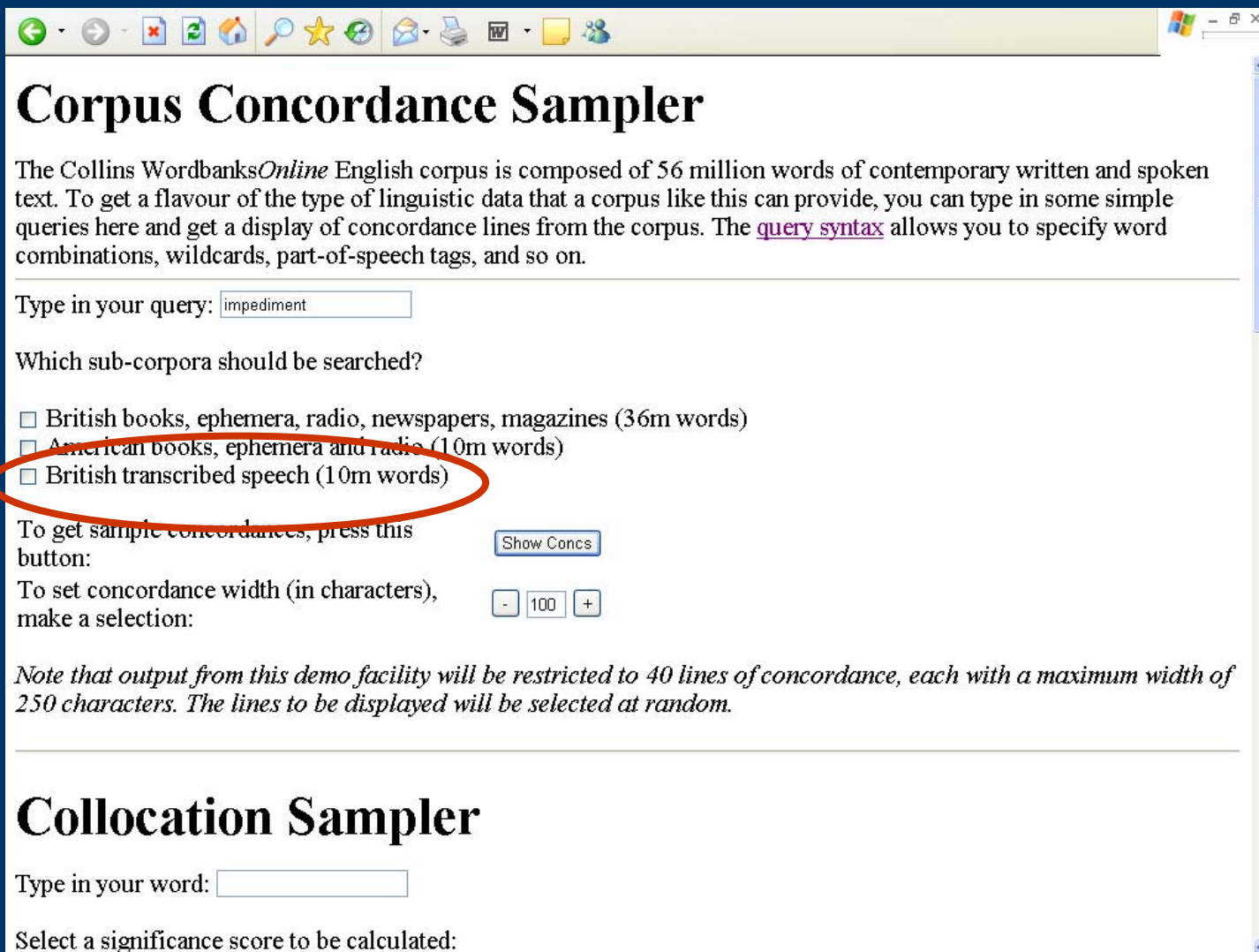
- [New BNC website](#)
- [BNC goes XML](#)
- [New version of Xaira](#)

**Results of your search**

```
... lemma="the"
... lemma="cost">cost
NN1" lemma="issue">is_
gal">legal </w><v type="
O" lemma="should">should
it of">out of </w><v type="
</w><v type="NN1" lemma="
nt to">pursuant to </w><v
="CJC" lemma="and">and </
</w><v type="AJ0" les
><c type="PUL">{</c
...ulations </w><v
... "STY"
```

# Collins Wordbanks Online Demo

<http://www.collins.co.uk/corpus/CorpusSearch.aspx>



**Corpus Concordance Sampler**

The Collins Wordbanks<sup>Online</sup> 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:

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:

# EUROPARL

<http://logos.uio.no/cgi-bin/opus/opuscqp.pl?corpus=EUROPARL;lang=en>

**OPUS - Corpus query (CWB)**

corpus	languages
EUROPARL	da de el es fi fr it pt sv

[EUconst](#)  
[KDE](#)  
[KDEdoc](#)  
[OpenOffice.org](#)  
[PHP](#)

**CQP query (CWB)**      show attributes      alignments

A CQP query consists of a regular expression over *attribute expressions*.

[Introduction of the query syntax](#)      positional annotation

[Example queries](#)

[word="a.\*"]

select    show max 20 hits and     skip non-aligned segments     vertical     KWIC

horizontal

(advanced search)

da     de     el  
 es     fi     fr  
 it     nl     pt  
 sv



# COMPARA

<http://www.linguateca.pt/COMPARA/>

[[Esta página em português](#)]

## 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")

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")

Case insensitive  
 Diacritic insensitive

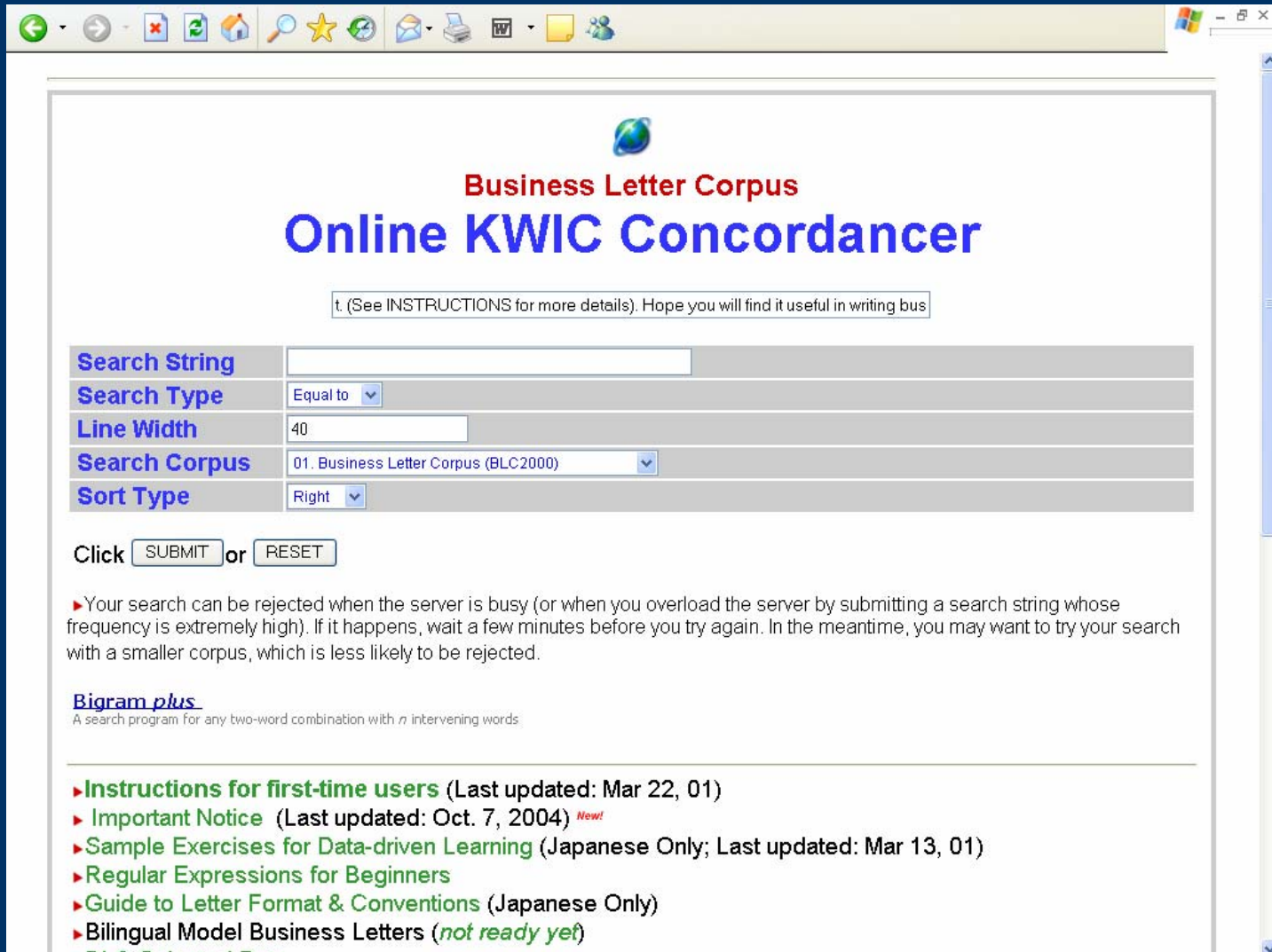
[Help](#)

<a href="#">SIMPLE SEARCH</a> <a href="#">COMPLEX SEARCH</a> Contents: corpus texts and quantitative snapshot	<a href="#">A brief description of COMPARA</a> <a href="#">Acknowledgements</a> <a href="#">Building COMPARA</a> <a href="#">How to contribute</a>	<a href="#">Publications</a> <a href="#">Questions from users</a> <a href="#">Search Help</a> <a href="#">The DISPARA system</a>
---	---	---




# Business Letter Corpus

<http://ysomeya.hp.infoseek.co.jp/>



The screenshot shows a web browser window with a yellow title bar and a standard Windows XP-style taskbar. The browser's address bar is empty. The main content area of the browser displays the following:

  
**Business Letter Corpus**  
**Online KWIC Concordancer**

t (See INSTRUCTIONS for more details). Hope you will find it useful in writing bus

<b>Search String</b>	<input type="text"/>
<b>Search Type</b>	Equal to <input type="button" value="v"/>
<b>Line Width</b>	40 <input type="text"/>
<b>Search Corpus</b>	01. Business Letter Corpus (BLC2000) <input type="button" value="v"/>
<b>Sort Type</b>	Right <input type="button" value="v"/>

Click  or

▶ 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 with *n* intervening words

---

- ▶ **Instructions for first-time users** (Last updated: Mar 22, 01)
- ▶ **Important Notice** (Last updated: Oct. 7, 2004) *New!*
- ▶ **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 Data

# Raising teachers' awareness to the basics of corpora

(Sub-)Corpus	Size	Type of English	Time
<b>BNC</b>	100 M	General British	early to mid 1990s
<b>Collins speech</b>	10 M	London British	contemporary?
<b>EUROPARL - EN</b>	28 M (?)	French, international & translated (EP debates)	1998 – 2003
<b>COMPARA - EN</b>	1.5 M	Original and translated fiction	1837- 2002
<b>Business Letter Corpus</b>	1 M	US and UK business letter samples	since 2000

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	counterpane
Something new	MP3
Something common	with
Something rare	epicure
Something oral	d'you
Something written	amiable
Something technical	pelagic
Something regional	lass
Something sentimental	darling
Something religious	rosary
Something political	coalition
Something foreign	rapporteur

# Understanding different corpora

## different corpora exercise

		<u>BNC</u>	<u>COs</u>	<u>EUR</u>	<u>COM</u>	<u>BLC</u>
old	counterpane	41	0	0	0	0
new	MPs				0	0
common					K	7K
rare						0
oral						0
written					16	2
technical	paralyse		0	25	0	0
regional	lass	414	27	0	0	0
sentimental	darling	2K	+40		38	0
religious	rosary	85	1	1	31	0
political	coalition	2K	12	413	1	3
foreign	rapporteur	29	0	16K	0	0

Different corpora  
will give you  
different results

# Understanding different corpora

getting 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

		BNC	BNC sampler	2 M (1/50)
old	counterpane	41	0	
new	MP2			
common				
rare				
oral				
written				
technical				
regional				
sentimental	da			
religious	rosary		2	
political	coalition	2K	41	
foreign	rapporteur	29	0	

When  
size matters...



# 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

onal community to realise this and to look again at their development, energy and  
appreciated and put to good use. [p] I look forward to hearing from you. And on behalf of  
while return on your money. So you can look forward to collecting a considerable cash lump  
We can advise people how to create a look that reflects their personality, and it's more  
because he said the standard hub caps look to flashy, but as he well knows black wheels  
emed so stylish; they made womens feet look smaller. And suddenly Reebok was a bandwagon.  
bags so heavy under his eyes that they look like make-up, is cool yet frantic, hurling  
l cows, and launch themselves as a new-look party with a different, modern image. [h]  
o the guys you're gonna see in here, I look like a goddamned preppie. Do me a favour,  
development economics to take a starker look at itself. How Marxism and developmentalism  
lf sit on the window ledge and turn to look out and down. Even though his shirt stuck to  
b up the trail over the mountains and look down at the village. What were you there? A  
in the eyes of your offspring. [p] I look round our own house. Good grief, it's like  
curlers. They really do make your eyes look more open, alive and livelier than they are.  
f the best lines in the book, when you look for the crumple zone in a Mini, you find it's  
[p] In so doing, the Court had to look at the interference complained of in the light  
lay before a search party was sent to look for the men of the Antares. [p] Since the  
sh LADIES and gentlemen. If you take a look to the right of the aircraft we have a  
Janet with the two runners-up in the Look of '95 competition Natalie Lowe, 16, left, and  
or so, and when I get back he doesn't look as if he's moved at all. [p] I do the  
that it keeps you from being lonely, look closer. Is it really providing the comfort and  
lers of a younger man. 'By the Jesus!' Look at him, still a doozy of a boy. 'Walker Owen!  
him on the lips. 'You want coffee? You look like you need some.' I'm nah stayin' tha"  
ey look at the person who is talking, look away at the beginning of their own speaking  
to work for, yet one has to be on the look-out all the time. Life has to go on, and one  
partments, giving them the first hard look at the economic impact of the Persian Gulf  
" a fascinating work would do well to look further in the oratorios--the English

# Retrieving information from a corpus

## Corpora are not like dictionaries exercise

Now do a search for looks

Close Window

supporters. The Office programme for the day  
[/h] [p] is a quiet and powerful workhorse that  
Mon-Sat 10-5. Fascinating exhibition that  
with HOLE. [p] 3.45am, ITV: AUSTIN CITY LIMITS  
plaster pillars lend grandeur. Below: Cream  
other hand, with his blond hair and sunglasses,  
A MESS OF IT AS LEEDS DID LAST YEAR. 90 MINUTES  
PHOTOS WITH CAPTIONS [/c] [h] Inspirational New  
Box 29590 [p] SHORTS/LEGS INTEREST. Slim 33,  
better. MACPHERSON : So, wild boar on the menu  
musical talent could be substituted for good  
of the talks. Malcolm Haslett of the BBC  
comfortable doing business with a woman who  
can trace his lineage back to the Egyptians and  
stood out in the crowd with his striking good  
can forgive much in an independent film that  
safety and audio equipment. Simon Price, who  
servant" to Jeanne. The latter had film goddess  
be placed? The danger of design by committee  
dead men were `martyrs [p] [h] New instruction  
world's leading rock acts when touring Asia,  
and she's a marvellous woman [p] Marianne now  
swamp the High Street next summer. [p] Other  
famous Western lawman. [p] He enjoys tender  
for Jaguar. So how do companies decide just who  
been very proud. [p] [h] On paper this trip  
old. And your picture was so unflattering. He

looks like this [p] 1. 2000 `Weather Alert'  
looks like a thoroughbred -- particularly :  
looks at working-class decoration. Include:  
looks at Progressive Country Music For f s  
looks as welcoming outside as in. Sunflowe:  
looks like he should be giving surfing les:  
LOOKS AT WHAT WENT WRONG AND WHAT HOWARD W:  
Looks [/h] History and fantasy set the scen:  
looks younger 5'9 craves reciprocal leg ti:  
looks good in a restaurant? DAUNCEY : Oh ye:  
looks, and a nice dance routine which sold  
looks at the background to the dispute: Of:  
looks like a teeny-bopper. We need to put :  
looks exactly like an ancestor of 2000 year:  
looks. You couldn't forget him in a hurry!  
looks at itself in a crazy mirror never che:  
looks after the car cleaning side, has spe:  
looks, although she was `just a housewife"  
looks ominous. [p] The Sports Council is :  
looks wide open to trainers' excuses;Racing  
looks set to be remembered only as a white:  
looks after the couple's three children - :  
looks to emerge paid lip-service to the Ser:  
looks and kisses with soon-to-die wife Anna:  
looks right with their products [p] AN act:  
looks dull;People Today [/h] [b] Christophe:  
looks stunning on his new video. [p] Jenny

Done Internet

start Eudora - [In] Microsoft PowerPoint ... 4 Internet Explorer PT 8:08

# Retrieving information from a corpus

## Corpora are not like dictionaries exercise

Now try a search for looked

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

Close Window

might say, fruit cakes. [p] [p] I looked hopefully through this year's offerin

Uxbridge, you've lost your leg Lord Uxbridge looked down and said, `By God, Sir, so I ha

turns. There is a half-smile on her face. She looked radiant I always feel like I'm just l

But, speaking through an interpreter, he looked back as well as forward. I have come

coming home with me for a drink?" [p] Alistair looked at his watch. `I'm not sure that I on

[p] She was here," Friedman finally said. He looked as if the words hurt. `She told us al

It is right that he should hear me. [p] Franzi looked straight ahead, absorbed, still frow

stepped back to admire their handiwork. Autumn looked from the snowperson to Brian. `I thi

He was looking at their rifles not at them. It looked as though two of them had fired Mann

edge against the palm of his left hand as he looked into the closet, found both her big

The teenagers turned from the gates and looked up at Pete in alarm. Bob and Jupiter

wider at some safer place, but Rhodry and Enj looked at each other, shrugged, and took the

got stuck, then another. Local villagers looked on in amazement. They could not unde

And Bronwen was the only girl Emyr had ever looked at, so they wanted to close it quick

two roads. The man wiped his sweating face, looked left, looked right, then continued s

beginning to fade, and Newcastle increasingly looked the more dangerous side, but Beardsl

but preference is for Smith's Band, who looked an improved horse until caught out by

dealers. But, as the evening wore on, they looked more successful. On my second visit,

goals in a six-year NFL career, has at times looked like a broken man this season. With

year-old sister Victoria were yesterday being looked after by their grandmother. [p] And

did you do [p] BLACK: I took her home and I looked at her privates. [p] WYRE: So she d

can come back and finish the job [p] Armstrong looked like being the south London side's he

Thomson pulled off a few saves, the omens looked good. [p] And when that free kick f

Leon stood stiffly in front of the truck. She looked frightened. [p] Come on," the Big Na

City, for example, shows how the town may have looked in the midst of its gold rush days. f

with a shrug of his shoulders. The supervisor looked startled. [p] Ted, do you see that a

are. I could help, like Liddie does." [p] He looked thoughtful, and for a moment she ima

Done Internet

start Eudora - [In] Microsoft PowerPoint ... 4 Internet Explorer PT 8:09

# Retrieving information from a corpus

Corpora are not like dictionaries exercise

Do the same for looking

Close Window

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

Some are away from home for the first time looking for work or starting a new job. Other  
he hears that his ex-wife (Sally Field) is looking for a new partner, he does the only  
not homosexuals. Often the best way to cause someone  
similar timetable to... to enhance his  
in their rush to please... to the job. I  
102222. [p] FREE... and loving,  
Dunningford... of initiative  
that something... return to  
the Croats of... Evgen Pavlov  
marble pillars... tunnel. It  
Rich Fisher... today it may  
that colon... representatives of the  
before thoughts... had gone beyond his  
cities as Bradford... on capital as a potent  
even put a puppet over... first for a known goalscorer, but  
most people think are sensible. It is... at making it easier for leaseholder  
that being photographed in the company of good-looking women, as well as other heads of state  
[p] Taylor admitted: "I think John will be looking at the situation and wondering. In  
say when playing the part of a rape victim [p] Looking furious, Gillian... No, Mr C  
small waterfalls. [p] It sits on a low hill, looking out on Liberation... [p] [sh]  
with a majority of 4,888, but is thought to be looking for a more secure constituency. [p]  
They were in Leo's desk, where he had been looking for stamps. Leo, after all, was ent  
When he finally ambled forward, he avoided looking into the faces of the passers-by, fe  
t we get along?" Mr. Trancas frowned slightly, looking puzzled. Mr. Elliot held up a hand  
t be silly," walking away from him, not even looking at him, thinking of something else  
by her daily work, had grown tired of looking after Gregor as she did formerly, th  
I do, which is to hit the road for Hoopdance, looking for a better time. I cruise until I

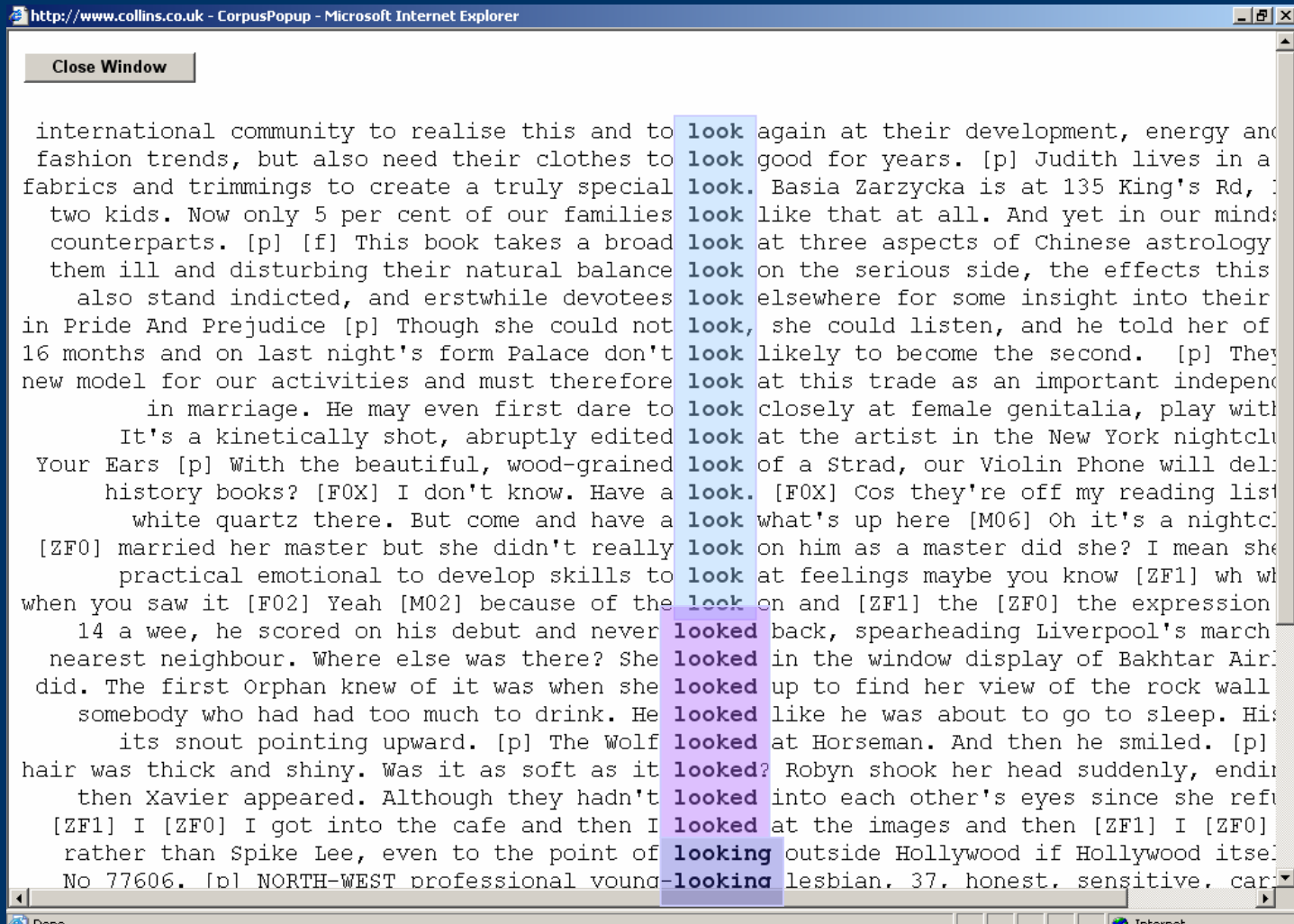
Done Internet

start Eudora - [In] Microsoft PowerPoint ... 4 Internet Explorer PT 8:10

# 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.



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

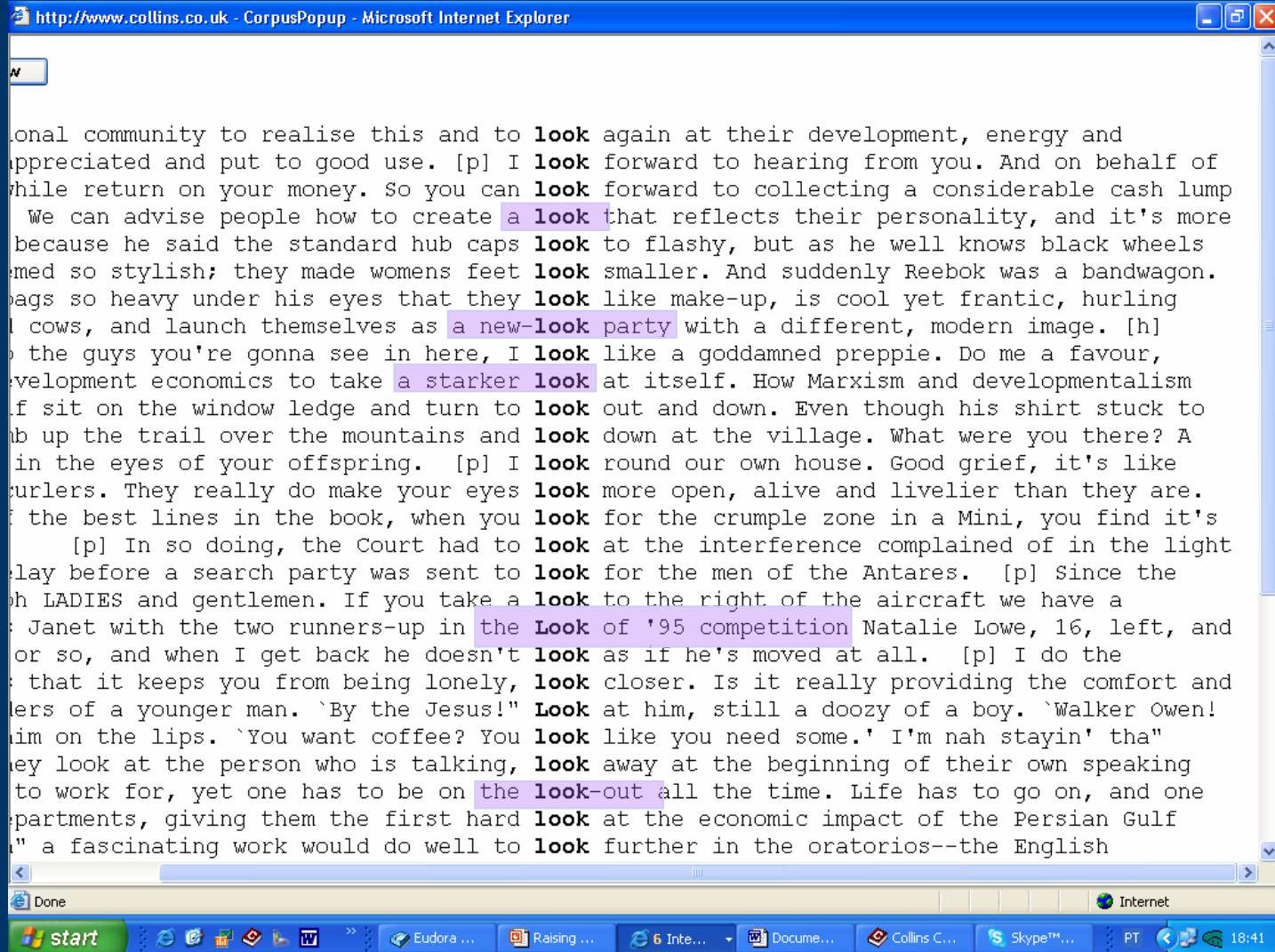
Close Window

international community to realise this and to look again at their development, energy and fashion trends, but also need their clothes to look good for years. [p] Judith lives in a fabrics and trimmings to create a truly special look. Basia Zarzycka is at 135 King's Rd, two kids. Now only 5 per cent of our families look like that at all. And yet in our minds counterparts. [p] [f] This book takes a broad look at three aspects of Chinese astrology them ill and disturbing their natural balance look on the serious side, the effects this also stand indicted, and erstwhile devotees look elsewhere for some insight into their in Pride And Prejudice [p] Though she could not look, she could listen, and he told her of 16 months and on last night's form Palace don't look likely to become the second. [p] They new model for our activities and must therefore look at this trade as an important independ in marriage. He may even first dare to look closely at female genitalia, play with It's a kinetically shot, abruptly edited look at the artist in the New York nightclu Your Ears [p] With the beautiful, wood-grained look of a Strad, our Violin Phone will del history books? [FOX] I don't know. Have a look. [FOX] Cos they're off my reading list white quartz there. But come and have a look what's up here [M06] Oh it's a nightcl [ZF0] married her master but she didn't really look on him as a master did she? I mean she practical emotional to develop skills to look at feelings maybe you know [ZF1] wh wh when you saw it [F02] Yeah [M02] because of the look on and [ZF1] the [ZF0] the expression 14 a wee, he scored on his debut and never looked back, spearheading Liverpool's march nearest neighbour. Where else was there? She looked in the window display of Bakhtar Air did. The first Orphan knew of it was when she looked up to find her view of the rock wall somebody who had had too much to drink. He looked like he was about to go to sleep. His its snout pointing upward. [p] The Wolf looked at Horseman. And then he smiled. [p] hair was thick and shiny. Was it as soft as it looked? Robyn shook her head suddenly, endi then Xavier appeared. Although they hadn't looked into each other's eyes since she refu [ZF1] I [ZF0] I got into the cafe and then I looked at the images and then [ZF1] I [ZF0] rather than Spike Lee, even to the point of looking outside Hollywood if Hollywood itse No 77606. [p] NORTH-WEST professional vound-lookinga 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?



The screenshot shows a Microsoft Internet Explorer browser window with the address bar displaying "http://www.collins.co.uk - CorpusPopup - Microsoft Internet Explorer". The main content area shows a list of text excerpts from a corpus, with the word "look" highlighted in purple in several instances. The excerpts are:

- ...onal community to realise this and to **look** again at their development, energy and appreciated and put to good use. [p] I **look** forward to hearing from you. And on behalf of while return on your money. So you can **look** forward to collecting a considerable cash lump
- We can advise people how to create **a look** that reflects their personality, and it's more because he said the standard hub caps **look** to flashy, but as he well knows black wheels emed so stylish; they made womens feet **look** smaller. And suddenly Reebok was a bandwagon. bags so heavy under his eyes that they **look** like make-up, is cool yet frantic, hurling l cows, and launch themselves as **a new-look** party with a different, modern image. [h] o the guys you're gonna see in here, I **look** like a goddamned preppie. Do me a favour, evelopment economics to take **a starker look** at itself. How Marxism and developmentalism f sit on the window ledge and turn to **look** out and down. Even though his shirt stuck to b up the trail over the mountains and **look** down at the village. What were you there? A in the eyes of your offspring. [p] I **look** round our own house. Good grief, it's like irlers. They really do make your eyes **look** more open, alive and livelier than they are. f the best lines in the book, when you **look** for the crumple zone in a Mini, you find it's
- [p] In so doing, the Court had to **look** at the interference complained of in the light lay before a search party was sent to **look** for the men of the Antares. [p] Since the oh LADIES and gentlemen. If you take a **look** to the right of the aircraft we have a e Janet with the two runners-up in **the Look** of '95 competition Natalie Lowe, 16, left, and or so, and when I get back he doesn't **look** as if he's moved at all. [p] I do the e that it keeps you from being lonely, **look** closer. Is it really providing the comfort and lers of a younger man. `By the Jesus!" **Look** at him, still a doozy of a boy. `Walker Owen! im on the lips. `You want coffee? You **look** like you need some.' I'm nah stayin' tha" ey look at the person who is talking, **look** away at the beginning of their own speaking to work for, yet one has to be on **the look-out** all the time. Life has to go on, and one epartments, giving them the first hard **look** at the economic impact of the Persian Gulf " a fascinating work would do well to **look** further in the oratorios--the English

The browser window also shows a taskbar at the bottom with various application icons and a system tray on the right displaying the time as 18:41.



# 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

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

Close Window

begging or sleeping rough on the streets. One **look** in their eyes and you can tell that th  
at MFI. [p]  
with captions  
sweaters this w  
for extra inter  
is ar  
bow, the  
basement, we ste  
greying beard  
stall, I par  
and beside the  
illustrate t  
the reasons why  
did he say?" M  
artists in her  
were hidden in t  
range has cover  
Streets, allow  
job for a be  
same old anyo  
them. Die-hard romantics can achieve a similar **look** at home as well - far more enticing th  
you walk through a cosmetics department, take a **look** around. The floor is filled with peop  
It's just too rich." [p] From the baffled **look** on Ted's face I knew I had to interpre  
an' the guy maybe leaves huh f' another girl an **Look**, Miguel, if you really quit this shit,  
Her short hair. Her small body. And that same **look** on her face. She has the back of her h  
spacious, but there is a lived-in, cluttered **look** about it. There are four different ent  
strong tapered tail. It has a tousled kind of **look** to it with no feature exaggerated. Chac  
teens and teen-agers. They stop for a brief **look** and then return to their afternoon pla

**POS tags**  
**look/NOUN**

**No tags**  
**a+2look, the+2look**



# 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

**OPUS - Corpus query (CWB)**

corpus	languages
EUROPARL	<a href="#">da</a> <a href="#">de</a> <a href="#">el</a> <a href="#">en</a> <a href="#">es</a> <a href="#">fi</a> <a href="#">fr</a> <a href="#">it</a> <a href="#">nl</a> <a href="#">pt</a> <a href="#">sv</a>

[EUconst](#)  
[KDE](#)  
[KDEdoc](#)  
[OpenOffice.org](#)  
[PHP](#)

**CQP query (CWB)**  
A CQP query consists of a regular expression over *attribute expressions*.  
[Introduction of the query syntax](#)  
[Example queries](#)

show attributes      alignments

positional annotation

word  id  lem  pos

skip non-aligned segments  vertical  KWIC

horizontal

([advanced](#) search)

Query string: ""Protocole" "privilèges" "immunités""  
0 hits found

fr en

First try (without stop words)

# Retrieving information from a corpus

Corpora are not like web browsers exercise

The screenshot shows the OPUS - Corpus query (CWB) interface. It includes a search bar with the query string: "Protocole" "sur" "les" "privilèges" "et" "immunités". The interface has several sections: a corpus list on the left, a search configuration area with options for positional annotation and alignment, and a results section at the bottom. A purple oval highlights the search string and the 'show' button. A blue box highlights the first four search results.

corpus	languages
EUROPARL	da de el en es fi fr it nl pt sv
<a href="#">EUconst</a>	
<a href="#">KDE</a>	
<a href="#">KDEdoc</a>	
<a href="#">OpenOffice.org</a>	
<a href="#">PHP</a>	

**CQP query (CWB)**  
A CQP query consists of a regular expression over *attribute expressions*.  
[Introduction of the query syntax](#)  
[Example queries](#)

show attributes: positional annotation  
alignments:  da  de  el  en  es  fi  it  nl  pt  sv

select show max 20 hits and  skip non-aligned segments  vertical  KWIC  
 horizontal  
([advanced search](#))

Query string: ""Protocole" "sur" "les" "privilèges" "et" "immunités""

6 hit

371	er
156	g to
262	to eges

**Protocole sur les privilèges et immunités**  
**Protocol on the privileges and immunities**  
**Protocol of the privileges and immunities**  
**Protocol on privileges and immunities**

datant de 1965 , et à l' Acte de 1976 relatif à l' élection des représentants au Parlement européen , et à la Constitution ou à l' ordonnance juridique de chacun des États membres où nous trouvons des situations tellement différentes , par exemple le fait qu' au Royaume-Uni , l' immunité and Immunities of the European Communities , the 1976 Act concerning the Election of Representatives of the European Parliament , and the Constitution or legislation of each Member State . The situations in the Member States are very diverse : no parliamentary immunity in the

# Retrieving information from a corpus

## Corpora are not like web browsers exercise

**OPUS - Corpus query (CWB)**

corpus	languages
EUROPARL	<a href="#">da</a> <a href="#">de</a> <a href="#">el</a> <a href="#">en</a> <a href="#">es</a> <a href="#">fi</a> <a href="#">fr</a> <a href="#">it</a> <a href="#">nl</a> <a href="#">pt</a> <a href="#">sv</a>

[EUconst](#)  
[KDE](#)  
[KDEdoc](#)  
[OpenOffice.org](#)  
[PHP](#)

**CQP query (CWB)**  
A CQP query consists of a regular expression over *attribute expressions*.  
[Introduction of the query syntax](#)  
[Example queries](#)

show attributes      alignments

positional annotation

word  id  lem  pos

da  de  el  
 en  es  fi  
 it  nl  pt  
 sv

select show max 20 hits and  skip non-aligned segments  vertical  KWIC  
 horizontal  
([advanced](#) search)

Query string: `""protocole"%c "sur"%c "les"%c "privilèges"%c "et"%c "immunités"%c"`  
16 hits found

	fr	en
1877788	C'est la raison pour laquelle nous avons, dans le cadre de la négociation avec la Commission et le Conseil, soumis tous les passages sensibles du règlement sur l'OLAF aux réserves du <b>protocole sur les privilèges et immunités</b> .	That was why, during negotiations with the Commission and the Council, we made all the sensitive passages in the OLAF regulation subject to the Protocol on Privileges and Immunities.
3717295		der munities
7145699	II. modifiant le règlement (Euratom, CECA, CEE) 549/69 déterminant les catégories de fonctionnaires et agents des Communautés auxquels s'appliquent les dispositions de l'article 12, de l'article 13 deuxième alinéa et de l'article 14 du <b>protocole sur les privilèges et immunités</b> des Communautés (COM (2001) 50 - C5-0058 / 2001 - 2001 / 0028 (CNS)).	The next item is the report (A5-0194 / 2001) by Mr Miller, on behalf of the Committee on Legal Affairs and the Internal Market, on the proposals for Council regulation introducing special measures to terminate the service of officials of the Commission of the European Communities as part of the reform of the Commission [COM (2001) 50 - C5-0057 / 01 - 2001 / 0027 (CNS)] and on the proposal for a Council regulation amending

**Third try (stop words + case insensitive)**



# Retrieving information from a corpus

Corpora are not like web browsers exercise

The screenshot shows the OPUS Corpus query (CWB) interface. The main heading is "OPUS <sup>Stop</sup> Corpus query (CWB)". Below this, there are several sections:

- corpuses:** EUOPARL, EUconst, KDE, KDEdoc, OpenOffice.org, PHP.
- languages:** da, de, el, en, es, fi, fr, it, nl, pt, sv.
- CQP query (CWB):** A CQP query consists of a regular expression over *attribute expressions*. It includes links for "Introduction of the query syntax" and "Example queries". A search box contains the query: `"protocole"%c ".*" ".*" "privilèges"%c`. Below the search box are options for "select", "show max" (set to 20), "hits and", "skip non-aligned segments", "vertical" (selected), "KWIC", "horizontal", and "(advanced search)".
- show attributes:** positional annotation, word (checked), id, lem, pos.
- alignments:** A grid of checkboxes for language pairs: da, de, el, en (checked), es, fi, it, nl, pt, sv.

The query string is: `""protocole"%c ".*" ".*" "privilèges"%c ".*" "immunités"%c"`. 16 hits found.

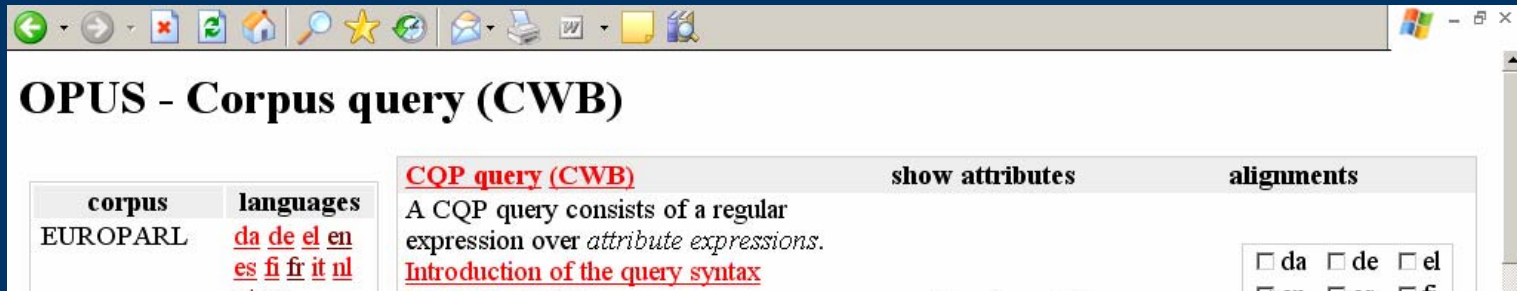
The results are displayed in a table with two columns: **fr** and **en**.

	fr	en
1877788	C ' est la raison pour laquelle nous avons , dans le cadre de la négociation avec la Commission et le Conseil , soumis tous les passages sensibles du règlement sur l' OLAF aux réserves du <b>protocole sur les privilèges et immunités</b> .	That was why , during negotiations with the Commission and the Council , we made all the sensitive passages in the OLAF regulation subject to the Protocol on Privileges and Immunities .
3717205	L 'immunité des députés de ce Parlement est prévue par l'	The immunities of Members of the House arise under
7145699	II . modifiant le règlement ( Euratom , CECA , CEE ) 549 / 69 déterminant les catégories de fonctionnaires et agents des Communautés auxquels s ' appliquent les dispositions de l' article 12 , de l' article 13 deuxième alinéa et de l' article 14 du <b>protocole sur les privilèges et immunités</b> des Communautés ( COM ( 2001 ) 50 - C5-0058 / 2001 - 2001 / 0028 ( CNS ) ) .	The next item is the report ( A5-0194 / 2001 ) by Mr Miller , on behalf of the Committee on Legal Affairs and the Internal Market , on the proposals for Council regulation introducing special measures to terminate the service of officials of the Commission of the European Communities as part of the reform of the Commission [ COM ( 2001 ) 50 - C5-0057 / 01 - 2001 / 0027 ( CNS ) ]

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	CQP query (CWB)	show attributes	alignments
EUROPARL	<a href="#">da</a> <a href="#">de</a> <a href="#">el</a> <a href="#">en</a> <a href="#">es</a> <a href="#">fi</a> <a href="#">fr</a> <a href="#">it</a> <a href="#">nl</a>	A CQP query consists of a regular expression over <i>attribute expressions</i> . <a href="#">Introduction of the query syntax</a>		<input type="checkbox"/> da <input type="checkbox"/> de <input type="checkbox"/> el <input type="checkbox"/> es <input type="checkbox"/> fi <input type="checkbox"/> fr <input type="checkbox"/> it <input type="checkbox"/> nl

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)**

corpus	languages
EUROPARL	<a href="#">da</a> <a href="#">de</a> <a href="#">el</a> <a href="#">en</a> <a href="#">es</a> <a href="#">fi</a> <a href="#">fr</a> <a href="#">it</a> <a href="#">nl</a> <a href="#">pt</a> <a href="#">sv</a>
<a href="#">EUconst</a>	
<a href="#">KDE</a>	
<a href="#">KDEdoc</a>	
<a href="#">OpenOffice.org</a>	
<a href="#">PHP</a>	

**CQP query (CWB)**  
A CQP query consists of a regular expression over *attribute expressions*.  
[Introduction of the query syntax](#)  
[Example queries](#)

show attributes:  positional annotation  
 word  id  lem  pos

alignments:  da  de  el  
 en  es  fi  
 it  nl  pt  
 sv

select show max 20 hits and  skip non-aligned segments  vertical  KWIC  
 horizontal  
(advanced search)

Query string: ""protocole" [] {1,5} "immunités"  
0 hits found

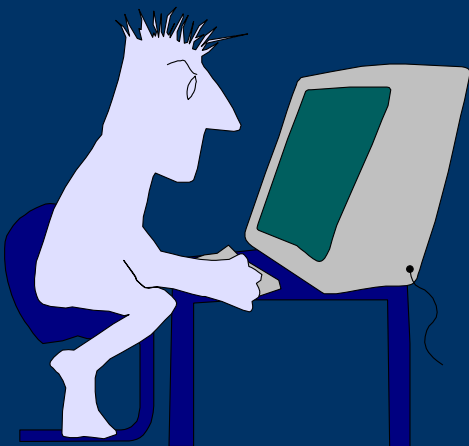
fr en

**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

EUROPARL



It was okay  
as far as  
I could see

COMPARA



# Retrieving information from a corpus

chunks of language exercise 1: reduction and expansion

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

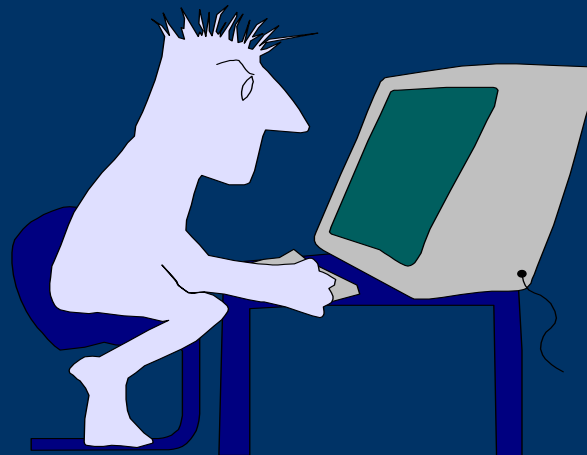
see 2214

It 16005

It was 3268

It was okay 2

It was okay as 0



COMPARA



# 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"



# Retrieving information from a corpus

## Chunks of language exercise 2: tri-gram

As a rule 290

a rule

ru

Which ones are likely to turn up?

Which ones won't turn up?

Which one will be the most frequent one?

of paint to 0

paint to every 6

to every 12 0

every 12 square 0

12 square metres 0

square metres of 30

metres of wall 0

of 39

litre of p



# Raising teachers' awareness to the basics of corpora

## 3. evaluating corpus data



# Evaluating corpus data

unedited data exercise

Dictionaries BNC

\*Reckless  
Reckless

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

\*Accomodation  
Accommodation

0 46  
1 4361

# Evaluating corpus data

count carefully exercise

## 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

The screenshot shows a Microsoft Internet Explorer browser window with the address bar displaying <http://www.collins.co.uk>. The page content is a corpus search for the word "congratulations". A large blue cloud-shaped overlay is positioned in the center of the page, containing the text "congratulations to", "congratulations on", and "congratulations from". The background text is a list of search results, each consisting of a snippet of text followed by the word "congratulations" and a preposition. The results include:

- and venues are as follows [p] LIST [h] **Congratulations** to [h] John Onslaw for his pro...
- readers quoted here to write to me. [p] **Congratulations** on such an adept merger. Just a...
- Croatia. [h] Prize Draw WINNERS! [/h] [p] **Congratulations** to the winners of our Children...
- hope that those who have sent messages of **congratulations** including the Director of...
- Although I was only six when ... 198 Eurovision Song Co...
- the two peoples he s... its smooth tran...
- post-show she... lick through the...
- COMPETITION [/... our recent and...
- by the club's f... the 11 other mo...
- Gallery is... Company who've...
- [/h] The... ving unificat...
- have support... a number of...
- he was struc... many from poor...
- controversie... sense of timing"
- the program... Tuesday morning we we...
- Thomas had t... mmons on the birth of...
- Don Luigi Ste... me. Giuditta took her wed...
- a better voyeur [p] Ta... ons to rania Glyde, whose provocati...
- when hundred... le visited to C... ations on her marriage. [p] I was sit...
- Chief Exec... r of Arco [p] Sir, **congratulations** to The Times for supporting the...
- Short was last night sent a message of **congratulations** from John Major after beating A...
- it would... me smiles back on faces [h] **Congratulations** on your bronze medal, you're fi...
- [p] [h] Rich 8; 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 o...
- Fame; International Golf Club [/h] [p] **congratulations** to [p] P JONES and A CRITCHLEY,
- She received it in less than 48 hours. **Congratulations** to the Post Office. If Michael I...

The browser's taskbar at the bottom shows the Windows Start button, several application icons, and the system tray with the time 20:27.

# Evaluating corpus data

## co-text exercise

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

Close Window

and venues are an... [h] law for his pro  
readers quote... rger. Just a  
Croatia. [h] Children  
hope that th... rector of  
Although... Song Co  
the two peop... both tran  
post... rough the  
COM... cent and  
by... her mo  
Gal... who've  
[h]... significa  
have sup... ber of  
he was... from poor  
contro... of timing"  
the p... ay morning we we  
Thoma... on the birth of  
Don... uditta took her we  
a better voyeu... yde, whose provocati  
when hundreds of... s... marriage. [p] I was sit  
Chief Executive O... ns to The Times for supporting the  
Short was last night sent a messa... tions from Job... ter beating A  
it would put some smiles back on faces [h]... tulations on your... you're fi  
[p] [h] Rich 8;Today Competition [/h] [p] **CONGRATULATIONS** to Edward Wallington,  
me as well [p] Mason received equally warm **congratulations** from Olazas... and third placed  
Slam winners;Today Competition [/h] [p] **CONGRATULATIONS** to Peter Jonas, t... t of on  
Fame;International Golf Club [/h] [p] **congratulations** to [p] P JONES and... CHLEY,  
She received it in less than 48 hours. **Congratulations** to the Post Office. If Michael

Done Internet

start Eudora... Collins C... Microsof... Docume... 5 Inter... 2 Wind... PT 20:27

# Evaluating corpus data

context and medium exercise

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

**Corpus Concordance Sampler**

The Collins Wordbanks<sup>Online</sup> 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:

Which sub-corpora should be searched?

- British books, ephemera
- American books, ephemera
- British transcripts

To get sample concordance lines, click the **Search** button:  
To set concordance options, click the **Options** button:  
To make a selection of concordance lines, click the **Select** button.

*Note that output from this sampler is limited to 250 characters. The lines to be displayed will be sorted in order of significance with a maximum width of 250 characters.*

---

**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

Books and articles  
about using corpora  
in language teaching

Corpus-specific  
tutorials