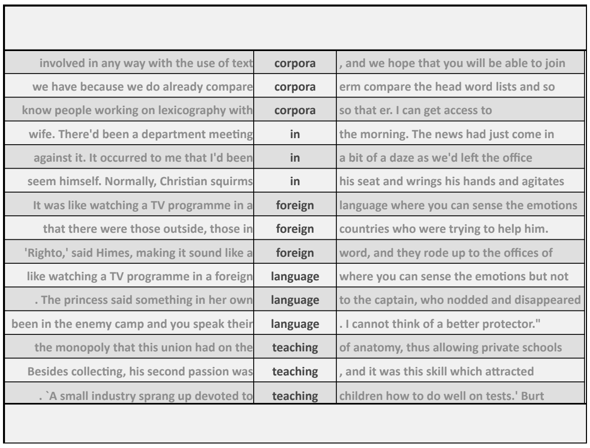


involved in any way with the use of text	corpora	, and we hope that you will be able to join
we have because we do already compare	corpora	erm compare the head word lists and so
know people working on lexicography with	corpora	so that er. I can get access to
wife. There'd been a department meeting	in	the morning. The news had just come in
against it. It occurred to me that I'd been	in	a bit of a daze as we'd left the office
seem himself. Normally, Christian squirms	in	his seat and wrings his hands and agitates
It was like watching a TV programme in a	foreign	language where you can sense the emotions
that there were those outside, those in	foreign	countries who were trying to help him.
'Righto,' said Himes, making it sound like a	foreign	word, and they rode up to the offices of
like watching a TV programme in a foreign	language	where you can sense the emotions but not
. The princess said something in her own	language	to the captain, who nodded and disappeared
been in the enemy camp and you speak their	language	. I cannot think of a better protector."
the monopoly that this union had on the	teaching	of anatomy, thus allowing private schools
Besides collecting, his second passion was	teaching	, and it was this skill which attracted
. `A small industry sprang up devoted to	teaching	children how to do well on tests.' Burt

Corpora in Foreign Language Teaching

- Vivienne Rogers
- School of Modern Languages, Newcastle University
- www.viviennerogers.info
- vivienne.rogers@education.ox.ac.uk

- Zöe Handley
- Department of Education, University of Oxford
- zoe.handley@education.ox.ac.uk
- http://humbox.ac.uk/profile/291

Linguistics Association of Great Britain Conference 2010

"Only when words are in their habitual environments, presented in their most frequent forms and their relational patterns and structures, can they be learnt effectively, interpreted properly and used appropriately"

(Wu, 1992: 32)

Plan

- What is a corpus?
- Basic corpus techniques
- Corpora in language learning
- Data-driven language learning (DDL)
- What the research says about DDL?
- Benefits of DDL
- Limitations of DDL
- Working within the limitations of DDL

What is a corpus?

"any collection of more than one text can be called a corpus: the term 'corpus' is simply the Latin for 'body', hence a corpus may be defined as any body of text"

(McEnery and Wilson, 2001: 29)

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

(Sinclair, 1996)

- Reference corpus
  - British National Corpus, Brown Corpus
  - Balanced sample, machine-readable form, annotated

Corpus linguistic techniques

- Concordancing
  - "using corpus software to find **every occurrence of a particular word or phrase**" (O'Keefe et al., 2001: 8)
- Word frequency counts and word lists
- Key word analysis
  - "**Key words** … are those whose frequency is unusually high in comparison with some norm" (O'Keefe et al., 2001: 12)
- Cluster analysis
  - **Cluster analysis** allows the user to generate a list of the most frequent 2-, 3-, 4-, 5-, or 6-word combinations (n-grams, word/lexical clusters/ bundles) from a corpus, i.e. collocations and colligations (O'Keefe et al., 2001)

Corpus linguistic techniques

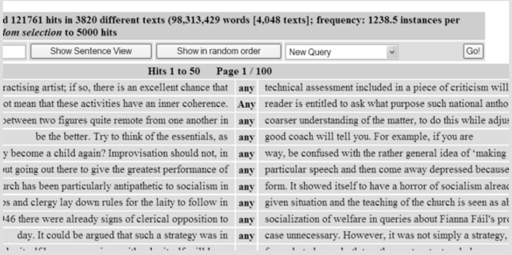
- Concgramming
  - "A **'concgram'** is all of the permutations of constituency variation and positional variation generated by the association of two or more words" (Greaves and Warren, 2007: 290)
- Lexico-grammatical profiles
  - Collocates
  - Chunks/idioms
  - Syntactic restrictions
  - Semantic restrictions
  - Semantic prosody

Concordancing

"using corpus software to find **every occurrence of a particular word or phrase**" … "The search word of phrase is often referred to as the 'node' and concordance lines are usually presented with the word/ phrase in the centre of the line with seven or eight words presented at either side. These are known as **Key-Word-In-Context displays (or KWIC concordances**"

(O'Keefe et al., 2001: 8)

Concordancing

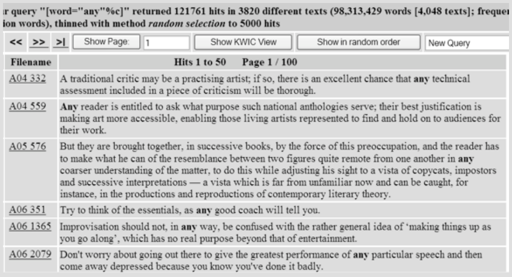


d 121761 hits in 3820 different texts (98,313,429 words [4,048 texts]; frequency: 1238.5 instances per
***lom selection* to 5000 hits**

```
                          Hits 1 to 50     Page 1 / 100
ractising artist; if so, there is an excellent chance that  any  technical assessment included in a piece of criticism will
ot mean that these activities have an inner coherence.      Any  reader is entitled to ask what purpose such national antho
between two figures quite remote from one another in        any  coarser understanding of the matter, to do this while adjus
        be the better. Try to think of the essentials, as   any  good coach will tell you. For example, if you are
y become a child again? Improvisation should not, in        any  way, be confused with the rather general idea of 'making
ut going out there to give the greatest performance of      any  particular speech and then come away depressed because
urch has been particularly antipathetic to socialism in     any  form. It showed itself to have a horror of socialism alreac
ps and clergy lay down rules for the laity to follow in     any  given situation and the teaching of the church is seen as al
46 there were already signs of clerical opposition to       any  socialization of welfare in queries about Fianna Fáil's pro
   day. It could be argued that such a strategy was in      any  case unnecessary. However, it was not simply a strategy,
```

KWIC Concordance: http://bncweb.lancs.ac.uk/

Concordancing



r query "[word="any"%c]" returned 121761 hits in 3820 different texts (98,313,429 words [4,048 texts]; freque
ion words), thinned with method *random selection* to 5000 hits

Filename	Hits 1 to 50 Page 1 / 100
A04 332	A traditional critic may be a practising artist; if so, there is an excellent chance that **any** technical assessment included in a piece of criticism will be thorough.
A04 559	**Any** reader is entitled to ask what purpose such national anthologies serve; their best justification is making art more accessible, enabling those living artists represented to find and hold on to audiences for their work.
A05 576	But they are brought together, in successive books, by the force of this preoccupation, and the reader has to make what he can of the resemblance between two figures quite remote from one another in **any** coarser understanding of the matter, to do this while adjusting his sight to a vista of copycats, impostors and successive interpretations — a vista which is far from unfamiliar now and can be caught, for instance, in the productions and reproductions of contemporary literary theory.
A06 351	Try to think of the essentials, as **any** good coach will tell you.
A06 1365	Improvisation should not, in **any** way, be confused with the rather general idea of 'making things up as you go along', which has no real purpose beyond that of entertainment.
A06 2079	Don't worry about going out there to give the greatest performance of **any** particular speech and then come away depressed because you know you've done it badly.

KWOC Concordance: http://bncweb.lancs.ac.uk/

Key Word Analysis

"**Key words** ... are those whose frequency is unusually high in comparison with some norm"

(O'Keefe et al., 2001: 12)

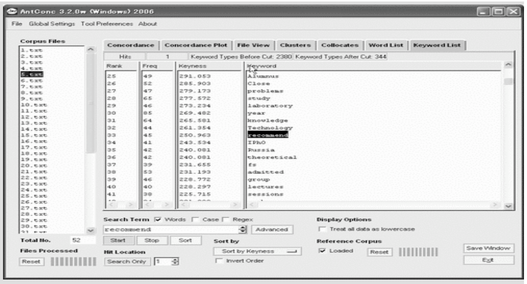
- *Wordsmith Tools* (Scott, 1999)
  - Compares the word list obtained from a small corpus with that obtained from a large reference corpus
  - Applications: genre analysis, forensic linguistics, stylistics, content analysis, text retrieval, and **Languages for Specific Purposes**

Key words from economics lecture relative to corpus of academic lectures

O'Keefe et al (2007:13)

1	Tax	8	poor	15	Higher	22	labour
2	Income	9	thousand	16	Percent	23	terms
3	System(s)	10	impact	17	Rates	24	Cost(s)
4	Average	11	equity	18	ordinary	25	characterised
5	basic	12	under	19	sixty	26	workers
6	rate	13	both	20	marginal	27	systems
7	supply	14	figures	21	scheme	28	negative

Key word Analysis


AntConc 3.2.0w (Windows) 2006
File Global Settings Tool Preferences About
Corpus Files
Concordance Concordance Plot File View Clusters Collocates Word List Keyword List
Search Term ☑ Words □ Case □ Regex
Advanced
Start Stop Sort
Sort by
Sort by Keyness
□ Invert Order
Display Options
□ Treat all data as lowercase
Reference Corpus
☑ Loaded
Reset
Save Window
Exit
Total No. 52
Files Processed
Reset
Hit Location
Search Only 1

1.txt
2.txt
3.txt
4.txt
5.txt
6.txt
7.txt
8.txt
9.txt
10.txt
11.txt
12.txt
13.txt
14.txt
15.txt
16.txt
17.txt
18.txt
19.txt
20.txt
21.txt
22.txt
23.txt
24.txt
25.txt
26.txt
27.txt
28.txt
29.txt
30.txt

Hits 1 Keyword Types Before Cut: 2380 Keyword Types After Cut: 344

Rank	Freq	Keyness	Keyword
25	49	291.053	Alumnus
26	52	285.903	Close
27	47	279.173	problems
28	65	277.572	study
29	46	273.234	laboratory
30	85	269.482	year
31	64	265.581	knowledge
32	44	261.354	Technology
33	45	250.963	recommend
34	41	243.534	IPhO
35	42	240.081	Russia
36	42	240.081	theoretical
37	39	231.655	fs
38	53	231.193	admitted
39	46	228.772	group
40	40	228.297	lectures
41	38	225.715	sessions

Search Term: recommend

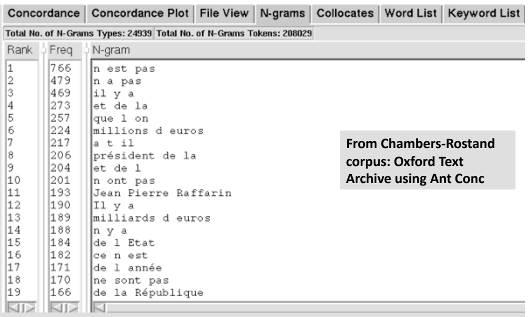
Cluster Analysis

Cluster analysis allows the user to generate a list of the most frequent 2-, 3-, 4-, 5-, or 6-word combinations (n-grams, word/lexical clusters/bundles) from a corpus, i.e. collocates.

(O'Keefe et al., 2001)

- Example application: Natural language processing (Part-of-Speech tagging), lexicography, **study of formulaic language**

Cluster Analysis



Total No. of N-Grams Types: 24939 Total No. of N-Grams Tokens: 208029

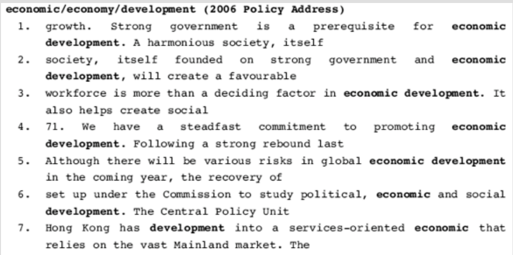
Rank	Freq	N-gram
1	766	n est pas
2	479	n a pas
3	469	il y a
4	273	et de la
5	257	que l on
6	224	millions d euros
7	217	a t il
8	206	président de la
9	204	et de l
10	201	n ont pas
11	193	Jean Pierre Raffarin
12	190	Il y a
13	189	milliards d euros
14	188	n y a
15	184	de l Etat
16	182	ce n est
17	171	de l année
18	170	ne sont pas
19	166	de la République

From Chambers-Rostand corpus: Oxford Text Archive using Ant Conc

Concgramming

- "A **'concgram'** is all of the permutations of constituency variation and positional variation generated by the association of two or more words" (Greaves and Warren, 2007: 290)
  - The words may be separated by a number of words
  - The words may appear in any order
- Permit the identification of meaningful word associations within a corpus, that is the 'aboutness' of a corpus or its **'phraseological profile'** (Greaves and Warren, 2007)

Concgramming



economic/economy/development (2006 Policy Address)

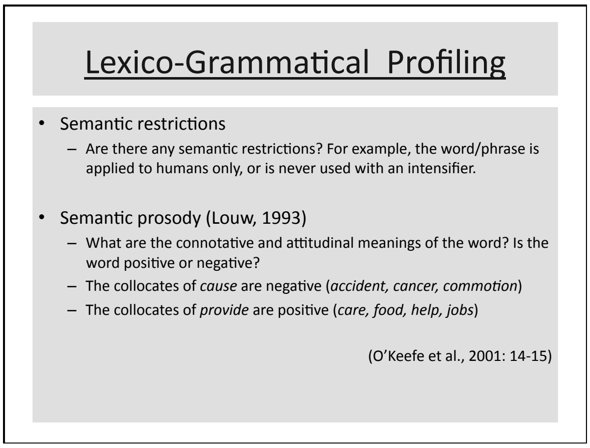
1. growth. Strong government is a prerequisite for **economic development**. A harmonious society, itself
2. society, itself founded on strong government and **economic development**, will create a favourable
3. workforce is more than a deciding factor in **economic development**. It also helps create social
4. 71. We have a steadfast commitment to promoting **economic development**. Following a strong rebound last
5. Although there will be various risks in global **economic development** in the coming year, the recovery of
6. set up under the Commission to study political, **economic** and social **development**. The Central Policy Unit
7. Hong Kong has **development** into a services-oriented **economic** that relies on the vast Mainland market. The

(Greaves and Warren, 2007: 299)

Lexico-Grammatical Profiling

- Collocates
  - Which word(s) occur most frequently and with statistical significance in the word's environment?
- Chunks/idioms
  - Does the word form part of any recurrent chunks? Is the word idiom-prone?
- Syntactic restrictions
  - Are there syntactic patterns which restrict the word? For example, are there prepositions that go with the word? What are its typical clause-positions (initial/medial/final)? Are there any tense/aspect restrictions?

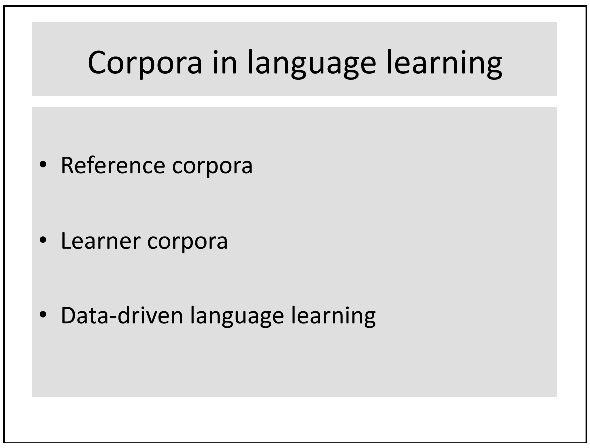
(O'Keefe et al., 2001: 14-15)



Lexico-Grammatical Profiling

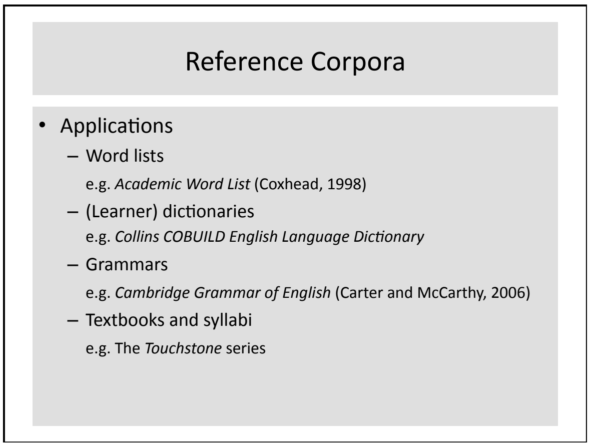
- Semantic restrictions
  - Are there any semantic restrictions? For example, the word/phrase is applied to humans only, or is never used with an intensifier.
- Semantic prosody (Louw, 1993)
  - What are the connotative and attitudinal meanings of the word? Is the word positive or negative?
  - The collocates of *cause* are negative (*accident, cancer, commotion*)
  - The collocates of *provide* are positive (*care, food, help, jobs*)

(O'Keefe et al., 2001: 14-15)



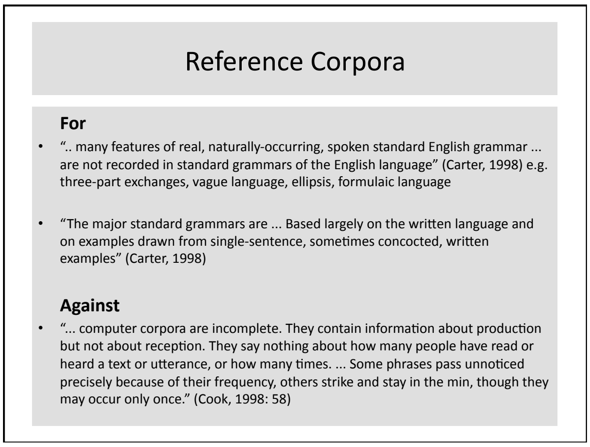
Corpora in language learning

- Reference corpora
- Learner corpora
- Data-driven language learning



Reference Corpora

- Applications
  - Word lists
    e.g. *Academic Word List* (Coxhead, 1998)
  - (Learner) dictionaries
    e.g. *Collins COBUILD English Language Dictionary*
  - Grammars
    e.g. *Cambridge Grammar of English* (Carter and McCarthy, 2006)
  - Textbooks and syllabi
    e.g. The *Touchstone* series



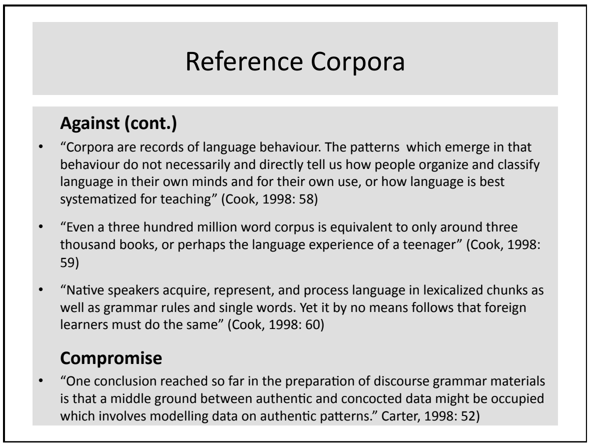
Reference Corpora

For

- ".. many features of real, naturally-occurring, spoken standard English grammar ... are not recorded in standard grammars of the English language" (Carter, 1998) e.g. three-part exchanges, vague language, ellipsis, formulaic language
- "The major standard grammars are ... Based largely on the written language and on examples drawn from single-sentence, sometimes concocted, written examples" (Carter, 1998)

Against

- "... computer corpora are incomplete. They contain information about production but not about reception. They say nothing about how many people have read or heard a text or utterance, or how many times. ... Some phrases pass unnoticed precisely because of their frequency, others strike and stay in the min, though they may occur only once." (Cook, 1998: 58)



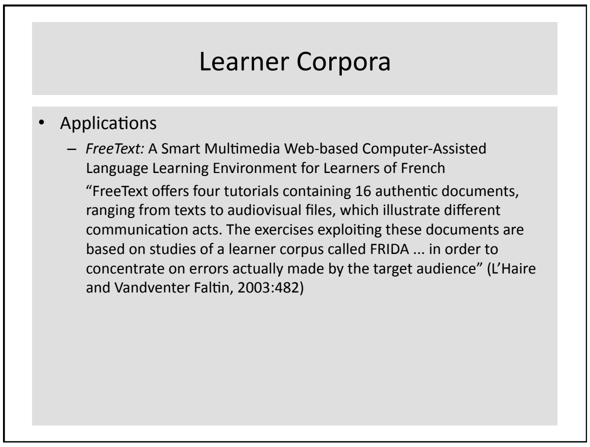
Reference Corpora

Against (cont.)

- "Corpora are records of language behaviour. The patterns which emerge in that behaviour do not necessarily and directly tell us how people organize and classify language in their own minds and for their own use, or how language is best systematized for teaching" (Cook, 1998: 58)
- "Even a three hundred million word corpus is equivalent to only around three thousand books, or perhaps the language experience of a teenager" (Cook, 1998: 59)
- "Native speakers acquire, represent, and process language in lexicalized chunks as well as grammar rules and single words. Yet it by no means follows that foreign learners must do the same" (Cook, 1998: 60)

Compromise

- "One conclusion reached so far in the preparation of discourse grammar materials is that a middle ground between authentic and concocted data might be occupied which involves modelling data on authentic patterns." Carter, 1998: 52)



Learner Corpora

- Applications
  - *FreeText:* A Smart Multimedia Web-based Computer-Assisted Language Learning Environment for Learners of French
    "FreeText offers four tutorials containing 16 authentic documents, ranging from texts to audiovisual files, which illustrate different communication acts. The exercises exploiting these documents are based on studies of a learner corpus called FRIDA ... in order to concentrate on errors actually made by the target audience" (L'Haire and Vandventer Faltin, 2003:482)

Learner Corpora

- Corpora
  - ICLE (Granger et al., 2002)
    - International Corpus of Learner English
    - Error tagged corpus of 2 million words of writing by learners of English from 19 different L1 backgrounds
  - FRIDA (Granger et al., 2001)
    - French Interlanguage Database
    - Error tagged corpus of 450, 000 words from essays written by French learners
  - Talkback project www.talkbank.org (MacWhinney 2007)
    - L2 French, Spanish, Danish, English, Welsh, Hebrew
    - Tagged spoken corpora with attached sound files.
    - Also contains speech data from patients with dementia and aphasia, as well as corpora coded for gesture etc.

Example activity with learner corpora: Developing more complex speech

- FLLOC (www.flloc.soton.ac.uk)
- Semi-elicited data with learners and native speakers
- Loch Ness story (LingDev, Newcastle corpora)
- Give class a selection of transcripts from different year groups (e.g. year 9-13, native speakers)
- Ask class to divide the transcripts according to proficiency.
- What clues did they use to categorize them?
- Lead into traditional exercises in use of discourse markers, connectives.
- Write their own story.

Data-Driven Language Learning

DDL, as described by Tim Johns, is intended "to confront the learner as directly as possible with the data, and to make the learner a linguistic researcher […] [someone who is able] to recognize and draw conclusions from clues in the data […]" (Johns, 2002: 108).

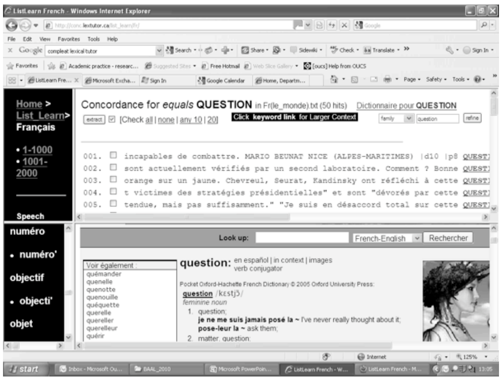
Tim Johns

Example Activities

- In vocabulary learning
  - Compleat lexical tutor (www.lextutor.ca)
- In reading/listening activities
  - Youth corpus (www.um.es/sacodeyl/)
- In teaching grammar
  - AntConc (http://www.antlab.sci.waseda.ac.jp/antconc_index.html)
  - Le Petit Prince (http://www.undlfoundation.org/lpp/sentences.txt)
- For other examples, please see Gavioli (1997), Dodd (1997), and Kaltenböck and Mehlmauser-Larcher (2005)

Vocabulary learning

- Compleat lexical tutor (www.lextutor.ca)
  - ListLearn – vocabulary lists for French and English divided into frequency bands of 1000 words. Links to audio, concordance and dictionary.
  - Hypertext – upload your own text. Links with concordance, audio and dictionary. Reading resources for words.
  - Concordances – English, French, German, Spanish (soon)
  - Cloze builder - upload your own document then decide what words to delete (e.g. every 5th word). Links to concordance.
  - N-gram- upload text and search for 3,4,5 word strings (useful for formulaic language)
  - Works best with Internet Explorer


ListLearn French - Windows Internet Explorer
http://conc.lextutor.ca/list_learn/fr/
File Edit View Favorites Tools Help
Google compleat lexical tutor
Search Share Sidewiki Check Translate Sign In
Favorites Academic practice - researc... Suggested Sites Free Hotmail Web Slice Gallery [oucs] Help from OUCS
ListLearn Fre... Microsoft Excha... Sign In Google Calendar Home, Departm...
Page Safety Tools
Home >
List_Learn>
Français
• 1-1000
• 1001-2000
Speech
numéro
• numéro'
objectif
• objecti'
objet
extract
Look up:
French-English
Rechercher
family
question
refine
Internet
125%
start Inbox - Microsoft Ou... BAAL_2010 Microsoft PowerPoin... ListLearn French - W... ListLearn French - M... 13:05

Concordance for *equals* **QUESTION** in Fr(le_monde).txt (50 hits) Dictionnaire pour **QUESTION**

☑ [Check all | none | any 10 | 20] **Click keyword link for Larger Context**

```
001. □ incapables de combattre. MARIO BEUNAT NICE (ALPES-MARITIMES) |d10 |p8 QUESTI
002. □ sont actuellement vérifiés par un second laboratoire. Comment ? Bonne QUESTI
003. □ orange sur un jaune. Chevreul, Seurat, Kandinsky ont réfléchi à cette QUESTI
004. □ t victimes des stratégies présidentielles" et sont "dévorés par cette QUESTI
005. □ tendue, mais pas suffisamment." "Je suis en désaccord total sur cette QUESTI
```

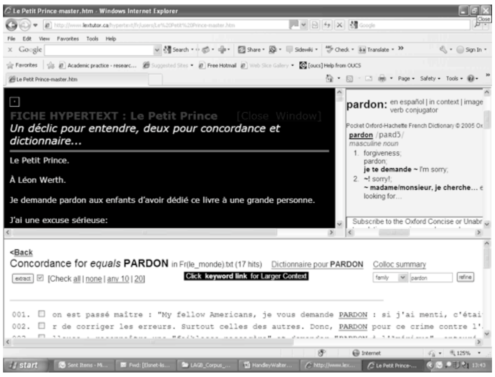
Voir également :
quémander
quenelle
quenotte
quenouille
quéquette
querelle
quereller
querelleur
quérir

question: en español | in context | images
verb conjugator

Pocket Oxford-Hachette French Dictionary © 2005 Oxford University Press:

question /kɛstjɔ̃/
feminine noun
1. question;
je ne me suis jamais posé la ~ I've never really thought about it;
pose-leur la ~ ask them;
2. matter, question;

List Learn



Le Petit Prince-master.htm - Windows Internet Explorer

http://www.lextutor.ca/hypertext/fr/users/Le%20Petit%20Prince-master.htm

FICHE HYPERTEXT : Le Petit Prince [Close Window]

Un déclic pour entendre, deux pour concordance et dictionnaire...

Le Petit Prince.

À Léon Werth.

Je demande pardon aux enfants d'avoir dédié ce livre à une grande personne.

J'ai une excuse sérieuse:

pardon: en español | in context | image verb conjugator

Pocket Oxford-Hachette French Dictionary © 2005 Ox

pardon /paʀdɔ̃/
masculine noun
1. forgiveness; pardon; **je te demande ~** I'm sorry;
2. ~! sorry!; **~ madame/monsieur, je cherche...** looking for...

Subscribe to the Oxford Concise or Unabr

<Back

Concordance for *equals* **PARDON** in Fr(le_monde).txt (17 hits) Dictionnaire pour **PARDON** Colloc summary

extract ☑ [Check all | none | any 10 | 20] **Click keyword link for Larger Context** family pardon refine

```
001. □ on est passé maître : "My fellow Americans, je vous demande PARDON : si j'ai menti, c'étai
002. □ r de corriger les erreurs. Surtout celles des autres. Donc, PARDON pour ce crime contre l'
```

Hypertext

Vocabulary/grammar

- Using AntConc
- (http://www.antlab.sci.waseda.ac.jp/software.html)
- Concordance lines (edited)
- Option 1: give list of sentences with unknown word – what does it mean?
- Option 2: replace keyword with blank
- Option 3: distribution of two L2 words with same L1 meaning, e.g. to know (savoir vs. connaître)
- Option 4: Idiomatic uses of word

Savoir versus connaître

à lui. Mais moi, malheureusement, je ne sais pas voir les moutons à travers les cais
...tats-Unis, le soleil, tout le monde le sait, se couche sur la France. Il suffirait
n éteint. Mais, comme il disait, On ne sait jamais! Il ramona donc également le vol
Mais il n'y a personne à juger! On ne sait pas, lui dit le roi. Je n'ai pas fait e
... . J'ai tellement de travail! je ne sais plus ... . Je suis sérieux, moi, je ne
i aperÁus il y a des années. Mais on ne sait jamais où les trouver. Le vent les prom
ssis auprès de moi. Quelle promesse? Tu sais .... une muselière pour mon mouton ....
it encore un effort: Ce sera gentil, tu sais. Moi aussi je regarderai les étoiles. T
st à dire .... pas tout à fait. Mais je sais bien qu'il est revenu à sa planète, car
n'est semblable si quelque part, on ne sait où, un mouton que nous ne connaissons p
uait au vent des cheveux tout dorés: Je connais une planète où il y a un monsieur cramo
ions d'un gros monsieur rouge? Et si je connais, moi, une fleur unique au monde, qui n'
sa chaise. Il voulut aider son ami: je connais un moyen de te reposer quand tu voudras
nes, là où il n'y en a qu'une seule. Je connais quelqu'un, dit le petit prince, qui ser
idée de notre planète à ceux qui ne la connaissent pas. Les hommes occupent très peu de pl
t, on ne sait où, un mouton que nous ne connaissons pas a, oui ou non, mangé une rose ....

Extracts from *Le Petit Prince* (http://www.undlfoundation.org/lpp/sentences.txt)

Reading/Listening

- Using Youth corpus (French, Spanish, German, Italian, Lithuanian, Romanian, English)
- http://www.um.es/sacodeyl/
- Based on videoed speech data
- Transcripts and resources available
- Searchable by topic, grammatical function etc.
- Students aged 11-18

'From textbook to data' or 'from data to textbook'?

"Th[e] principle of fidelity to the data is one which we ignore at our, and our students', peril. That danger is well illustrated by Groß, Müller and Wolff (1996), which uses concordance data to teach the old textbook rule for the use of *some* and *any* in English: *some* in positive statements, *any* in negative statements and in questions. Reference to any (!) KWIC concordance of *any* will show that generalisation to be false: the problem is that having decided on the generalisation in advance, it is all too easy to select only those citations that support it"

(Johns, 2002).

Web as Corpus

- **In the strictest sense of the term the Web is not a corpus** – it is not balanced in any way

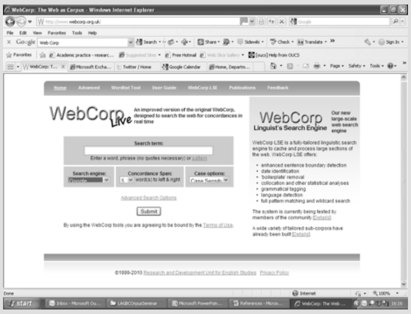
- **Advantages:**
  "constantly expanding, self-renewing machine-readable body of linguistic data, much richer in current language usage, infrequent expressions, text genres and domains than even the biggest standard reference corpus" (Krajka, 2009: 418) ...
  "freshness and spontaneity, completeness and scope, linguistic diversity, representativeness and free availability" (*ibid.*)

- **Disadvantages:**
  "huge rag bag of digital text" (Krajka, 2009: 418)
  - Unedited, non-native, etc.

  >> Teachers need to carefully select their corpus (Robb, 2003)

- **Example:** *Webcorp*: http://www.webcorp.org.uk/

Web as Corpus


WebCorp: The Web as Corpus - Windows Internet Explorer
http://www.webcorp.org.uk/
File Edit View Favorites Tools Help
Home Advanced Wordlist Tool User Guide WebCorp LSE Publications Feedback
Submit
Done Internet 100%
start

WebCorp Live An improved version of the original WebCorp, designed to search the web for concordances in real time

Search term:

Enter a word, phrase (no quotes necessary) or pattern

Search engine: Google | Concordance Span: 5 word(s) to left & right | Case options: Case Sensitive

Advanced Search Options

By using the WebCorp tools you are agreeing to be bound by the Terms of Use.

WebCorp Linguist's Search Engine Our new large-scale web search engine

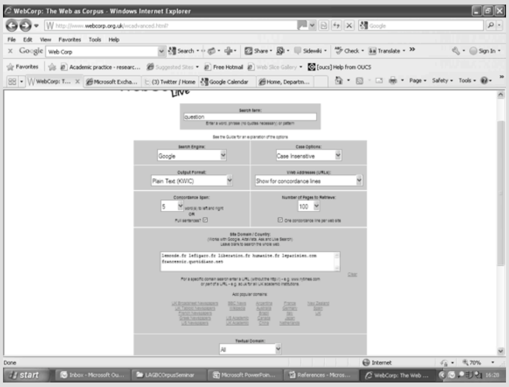
WebCorp LSE is a fully-tailored linguistic search engine to cache and process large sections of the web. WebCorp LSE offers:

- enhanced sentence boundary detection
- date identification
- 'boilerplate' removal
- collocation and other statistical analyses
- grammatical tagging
- language detection
- full pattern matching and wildcard search

The system is currently being tested by members of the community [Details].

A wide variety of tailored sub-corpora have already been built [Details].

©1999-2010 Research and Development Unit for English Studies Privacy Policy


WebCorp: The Web as Corpus - Windows Internet Explorer
http://www.webcorp.org.uk/wcadvanced.html?
File Edit View Favorites Tools Help
Done Internet 70%
start

Search term: question

Enter a word, phrase (no quotes necessary) or pattern

See the Guide for an explanation of the options

Search Engine: Google

Case Options: Case Insensitive

Output Format: Plain Text (KWIC)

Web Addresses (URLs): Show for concordance lines

Concordance Span: 5 word(s) to left and right OR Full sentences? ☑

Number of Pages to Retrieve: 100 ☑ One concordance line per web site

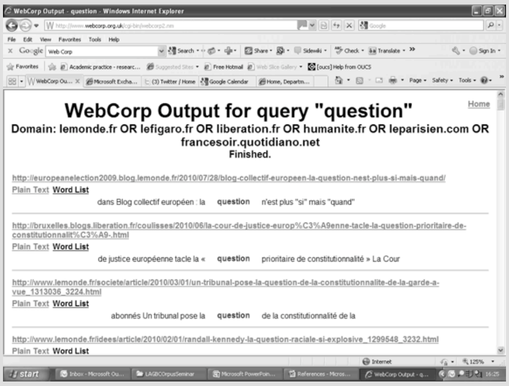
Site Domain / Country: (Works with Google, AltaVista, Ask and Live Search) Leave blank to search the whole web

lemonde.fr lefigaro.fr liberation.fr humanite.fr leparisien.com francesoir.quotidiano.net

For a specific domain search enter a URL (without the http://) - e.g. www.nytimes.com or part of a URL - e.g. ac.uk for all UK academic institutions

Add popular domains

Textual Domain: All


WebCorp Output - question - Windows Internet Explorer
http://www.webcorp.org.uk/cgi-bin/webcorp2.nm
File Edit View Favorites Tools Help
Home
Internet 125%
start

WebCorp Output for query "question"

Domain: lemonde.fr OR lefigaro.fr OR liberation.fr OR humanite.fr OR leparisien.com OR francesoir.quotidiano.net

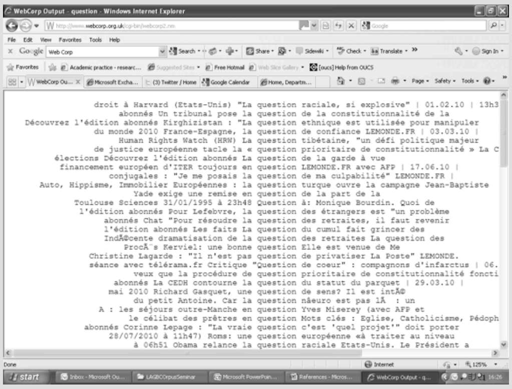
Finished.

http://europeanelection2009.blog.lemonde.fr/2010/07/28/blog-collectif-europeen-la-question-nest-plus-si-mais-quand/
Plain Text Word List
dans Blog collectif européen : la question n'est plus "si" mais "quand"

http://bruxelles.blogs.liberation.fr/coulisses/2010/06/la-cour-de-justice-europ%C3%A9enne-tacle-la-question-prioritaire-de-constitutionnalit%C3%A9-.html
Plain Text Word List
de justice européenne tacle la « question prioritaire de constitutionnalité » La Cour

http://www.lemonde.fr/societe/article/2010/03/01/un-tribunal-pose-la-question-de-la-constitutionnalite-de-la-garde-a-vue_1313036_3224.html
Plain Text Word List
abonnés Un tribunal pose la question de la constitutionnalité de la

http://www.lemonde.fr/idees/article/2010/02/01/randall-kennedy-la-question-raciale-si-explosive_1299548_3232.html
Plain Text Word List


WebCorp Output - question - Windows Internet Explorer
http://www.webcorp.org.uk/cgi-bin/webcorp2.nm
File Edit View Favorites Tools Help
Done Internet 125%
start

```
           droit à Harvard (Etats-Unis) "La question raciale, si explosive" | 01.02.10 | 13h3
                  abonnés Un tribunal pose la question de la constitutionnalité de la
  Découvrez l'édition abonnés Kirghizistan : "La question ethnique est utilisée pour manipuler
             du monde 2010 France-Espagne, la question de confiance LEMONDE.FR | 03.03.10 |
                  Human Rights Watch (HRW) La question tibétaine, "un défi politique majeur
             de justice européenne tacle la « question prioritaire de constitutionnalité » La C
      élections Découvrez l'édition abonnés La question de la garde à vue
       financement européen d'ITER toujours en question LEMONDE.FR avec AFP | 17.06.10 |
                conjugales : "Je me posais la question de ma culpabilité" LEMONDE.FR |
    Auto, Hippisme, Immobilier Européennes : la question turque ouvre la campagne Jean-Baptiste
                     Yade exige une remise en question de la part de la
           Toulouse Sciences 31/01/1995 à 23h48 Question à: Monique Bourdin. Quoi de
            l'édition abonnés Pour Lefebvre, la question des étrangers est "un problème
                 abonnés Chat "Pour résoudre la question des retraites, il faut revenir
                 l'édition abonnés Les faits La question du cumul fait grincer des
                 IndÃ©cente dramatisation de la question des retraites La question des
                     ProcÃ¨s Kerviel: une bonne question Elle est venue de Me
              Christine Lagarde : "Il n'est pas question de privatiser La Poste" LEMONDE.
              séance avec télérama.fr Critique "Question de coeur" : compagnons d'infarctus | 06.
                     veux que la procédure de question prioritaire de constitutionnalité foncti
                 abonnés La CEDH contourne la question du statut du parquet | 29.03.10 |
                mai 2010 Richard Gasquet, une question de sens? Il est intÃ©
                     du petit Antoine. Car la question nâeuro est pas lÃ  : un
              A : les séjours outre-Manche en question Yves Miserey (avec AFP et
                    le célibat des prêtres en question Mots clés : Eglise, Catholicisme, Pédoph
             abonnés Corinne Lepage : "La vraie question c'est 'quel projet'" doit porter
                28/07/2010 à 11h47) Roms: une question européenne «à traiter au niveau
                     à 06h51 Obama relance la question raciale Etats-Unis. Le Président a
```

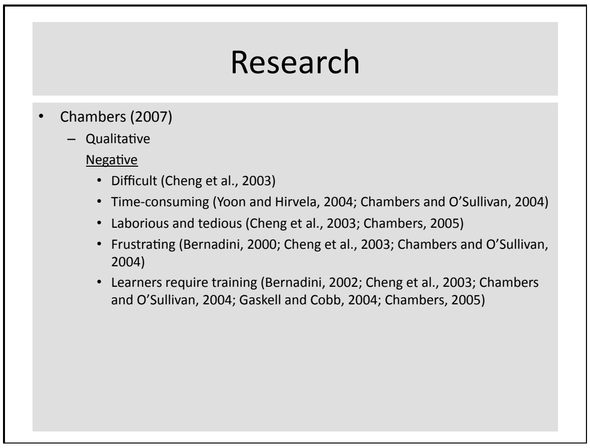
Research

- Boulton (2007)
  - Reviewed 39 empirical papers on DDL
  - In 34 studies English was the target language
  - **Only 2 studies focused on younger learners**
  - 36 studies were conducted in higher education institutes
    - 33 focused on language learning, 3 focused on linguistics
    - **Only 2 claim "low" levels and 2 "beginners"**
    - A variety of corpora were used (Bank of English, British National Corpus, ICE, MICASE, custom)
    - In most studies allowed directly accessed corpora using *WordSmithTools*
    - RQs: (1) Attitudes, (2) learners' practices, (3) learning outcomes
    - **Only 6 evaluate learning outcomes – these focus on lexicon/collocations**
  - Results: "learners attitudes are largely positive; in most cases they are remarkably capable of corpus techniques; corpora can be used as an effective reference tool, as well as for learning" (Boulton, 2007: 14)

Research

- Chambers (2007)
  - Quantitative
    - Stevens (1991): Concordance-based exercises on paper better than gap-filler exercises for vocabulary acquisition
    - Cobb (1997): On-screen concordance-based exercises are better than the use of other resources
  - Qualitative

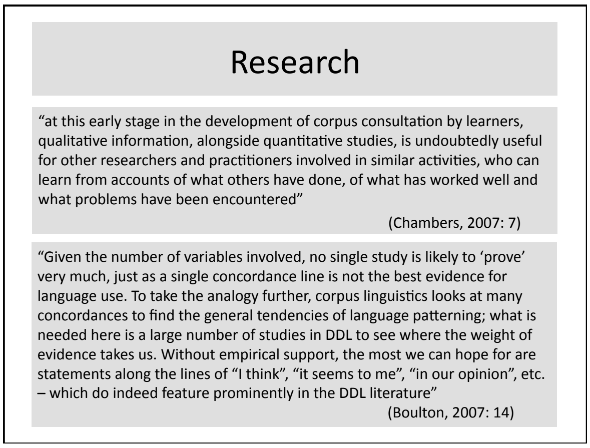
    Positive
    - "Appreciate the relevance of the corpus data" (Bernadini, 2002)
    - "Provide examples of language 'in context'" (Yoon and Hirvela, 2004; Chambers and O'Sulllivan, 2004)
    - "Appreciate the abundance of examples" in comparison with a dictionary (Cheng et al., 2003; Yoon and Hirvela, 2004; Chambers, 2005)
    - Appreciate the self-directed natural of DDL (Bernadini, 2002; Chambers, 2005)
    - Find the activity motivating (Chambers, 2005)



Research

- Chambers (2007)
  - Qualitative

    Negative
    - Difficult (Cheng et al., 2003)
    - Time-consuming (Yoon and Hirvela, 2004; Chambers and O'Sullivan, 2004)
    - Laborious and tedious (Cheng et al., 2003; Chambers, 2005)
    - Frustrating (Bernadini, 2000; Cheng et al., 2003; Chambers and O'Sullivan, 2004)
    - Learners require training (Bernadini, 2002; Cheng et al., 2003; Chambers and O'Sullivan, 2004; Gaskell and Cobb, 2004; Chambers, 2005)



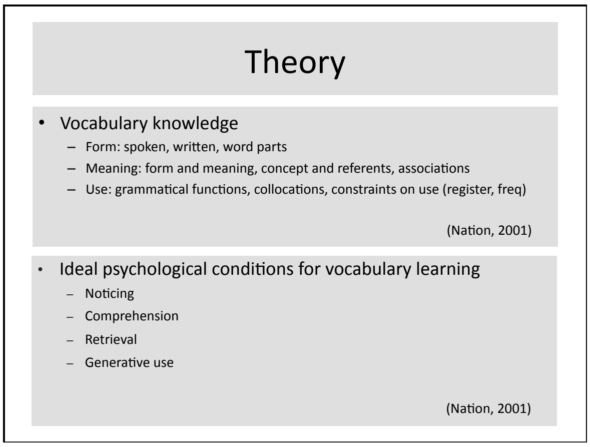
Research

"at this early stage in the development of corpus consultation by learners, qualitative information, alongside quantitative studies, is undoubtedly useful for other researchers and practitioners involved in similar activities, who can learn from accounts of what others have done, of what has worked well and what problems have been encountered"

(Chambers, 2007: 7)

"Given the number of variables involved, no single study is likely to 'prove' very much, just as a single concordance line is not the best evidence for language use. To take the analogy further, corpus linguistics looks at many concordances to find the general tendencies of language patterning; what is needed here is a large number of studies in DDL to see where the weight of evidence takes us. Without empirical support, the most we can hope for are statements along the lines of "I think", "it seems to me", "in our opinion", etc. – which do indeed feature prominently in the DDL literature"

(Boulton, 2007: 14)



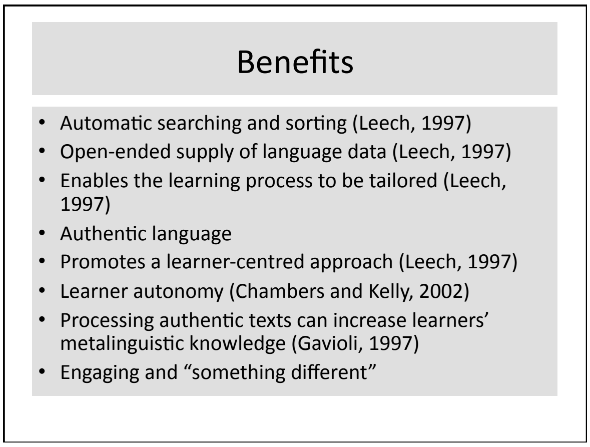
Theory

- Vocabulary knowledge
  - Form: spoken, written, word parts
  - Meaning: form and meaning, concept and referents, associations
  - Use: grammatical functions, collocations, constraints on use (register, freq)

(Nation, 2001)

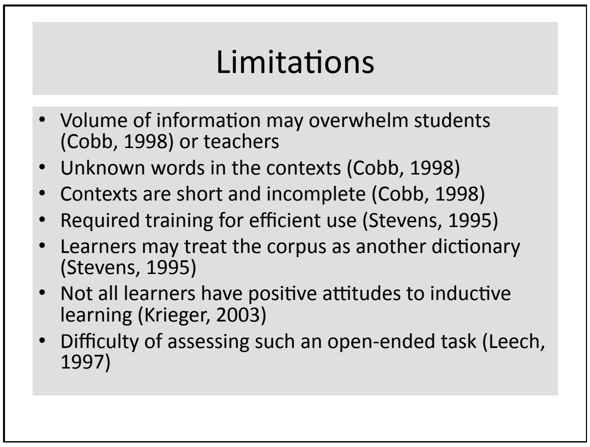
- Ideal psychological conditions for vocabulary learning
  - Noticing
  - Comprehension
  - Retrieval
  - Generative use

(Nation, 2001)



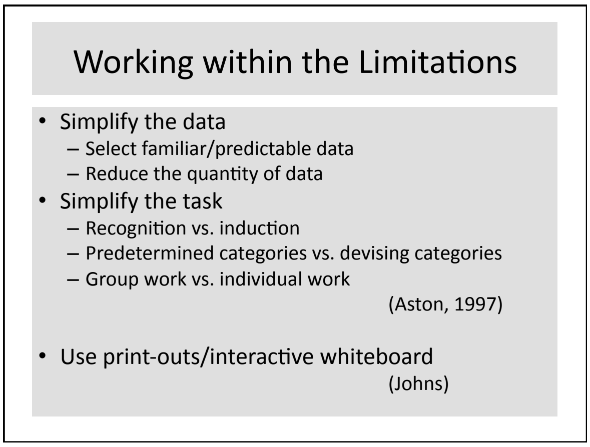
Benefits

- Automatic searching and sorting (Leech, 1997)
- Open-ended supply of language data (Leech, 1997)
- Enables the learning process to be tailored (Leech, 1997)
- Authentic language
- Promotes a learner-centred approach (Leech, 1997)
- Learner autonomy (Chambers and Kelly, 2002)
- Processing authentic texts can increase learners' metalinguistic knowledge (Gavioli, 1997)
- Engaging and "something different"



Limitations

- Volume of information may overwhelm students (Cobb, 1998) or teachers
- Unknown words in the contexts (Cobb, 1998)
- Contexts are short and incomplete (Cobb, 1998)
- Required training for efficient use (Stevens, 1995)
- Learners may treat the corpus as another dictionary (Stevens, 1995)
- Not all learners have positive attitudes to inductive learning (Krieger, 2003)
- Difficulty of assessing such an open-ended task (Leech, 1997)



Working within the Limitations

- Simplify the data
  - Select familiar/predictable data
  - Reduce the quantity of data
- Simplify the task
  - Recognition vs. induction
  - Predetermined categories vs. devising categories
  - Group work vs. individual work

(Aston, 1997)

- Use print-outs/interactive whiteboard

(Johns)

Recommended Reading

Tribble and Jones (1997). *Concordances in the classroom: A resource book for teachers.* Harlow: Longman [Example activities]

Boulton (2007). But where's the proof? The need for empirical evidence for data-driven learning. In Procs. *BAAL Conference 2007*. Edinburgh. (http://hal.archives-ouvertes.fr/docs/00/32/67/04/PDF/2007_boulton_BAAL_proof.pdf) [Review of the research evidence]

O'Keefe, McCarthy, and Carter (2007). *From corpus to classroom: Language use and language teaching.* Cambridge: Cambridge University Press. [General reference on corpora and concordancing, plus survey of uses in language learning and teaching]

Wichmann, Fligelstone, McEnery & Knowles (1997). *Teaching and Language Corpora.* London: Longman.

Johns (1991). You should be persuaded – Two samples of data-driven learning materials. In Johns and King (eds.), *Classroom concordancing. ELR Journal,* 4(1-16). (http://www.lexically.net/wordsmith/corpus_linguistics_links/Tim%20Johns%20and%20DDL.pdf) [Data-driven language learning in the words of Tim Johns]

Diniz (2005). Comparative review: TextSTAT 2.5, AntConc 3.0, and Compleat Lexical Tutor 4.0. *Language Learning & Technology*, Vol. 9, No. 3, pp. 22-27. (http://llt.msu.edu/vol9num3/pdf/review2.pdf) [Review of free concordancing tools for language learning]

Other References

McEnery and Wilson (2001). *Corpus Linguistics.* Edinburgh: Edinburgh University Press.

Coxhead (1998). *An Academic Word List.* English Language Institute Occasional Paper Number 18, Wellington: Victoria University of Wellington.

Krajka (2009). Concordancing 2.0: On Custom-Made Corpora in the Classroom. In *The Handbook of Research on Web 2.0 and Second Language Acquisition* edited by M. Thomas. IGI Publishing: Hershey, PA

Greaves and Warren (2007). Concgramming: A computer-driven approach to learning the phraseology of English. ReCALL. 19(3): 287-306

Carter (1998). Orders of reality: CANCODE, communication, and culture. ELT Journal. 52(1): 43-56

Cook (1998). The uses of reality: a reply to Ronald Carter. ELT Journal. 52(1): 57-63

Materials

Thornbury (2004). *Natural Grammar.* Oxford: Oxford University Press

McCarthy and O'Dell (1999). *English Vocabulary in Use.* Cambridge: Cambridge University Press

McCarthy and O'Dell (2001). *Basic Vocabulary in Use.* Cambridge: Cambridge University Press

McCarthy and O'Dell (2002). *English Idioms in Use.* Cambridge: Cambridge University Press

McCarthy and O'Dell (2004). *English Phrasal Verbs in Use.* Cambridge: Cambridge University Press

McCarthy and O'Dell (2005). *English Collocations in Use.* Cambridge: Cambridge University Press

McCarthy and O'Dell (2008). *Academic English in Use.* Cambridge: Cambridge University Press

Thurston and Candlin (1997). *Exploring Academic English: A Workbook for Student Essay Writing.* Sydney: NCELTR

Barlow and Burdine (2006). *Phrasal Verbs in American English.* Houston, Texas: Athelstan.

Resources and Tools

Webcorp: http://www.webcorp.org.uk/
Output: KWIC for web pages from selected domains / so multilingual
Options

- Search engine
- Case sensitivity
- Output format (plain/HTML)
- Web addresses for corpus lines
- Concordance span (no. words left and right) / whole sentences
- Number of pages to retrieve
- Site domain (e.g. .ac.uk)
- Textual domain (topic)
- Word filter (extra words which must (not) appear in the concordance lines)

Youth Corpus: http://www.um.es/sacodeyl/

- Available in English, French, Spanish, Italian, Lithuanian and Romanian
- Video and transcripts of interviews with 11-18 year olds.
- Fixed range of topics/categories including family, free time, elections etc.
- Search engine (by word, topic or grammatical feature)
- Teacher/learner resources available for some topics
- Requires RealPlayer for playback of video clips.

AntConc http://www.antlab.sci.waseda.ac.jp/software.html
Description: http://www.antlab.sci.waseda.ac.jp/research/iwlel_2004_anthony_antconc.pdf
Options/features:

- Use own corpus
- Concordancer
- Frequency lists
- Keyword generator
- Cluster and lexical bundle analysis

Compleat lexical tutor http://www.lextutor.ca
French, English Spanish
Options/Features

- Frequency > Word profile
- Range > Word profile + (Text_lex_compare) recycling index
- Concordancer (French, German, English a range of corpora + Custom for French and English)
- Story concordancer – links words to concordances from the story
- N-Gram Phrase Extracter (2 – 5 words)
- VocabProfile (Word profile)
- TextLexCompare (Compare word profile with word lists)
- Links reading text to concordancer and WordNet (English only)
- Keywords Extractor (English only)
- Multi_Conc (Multiple choice concordancer – English only)
- ID Word Identification Quiz (Guess the word that fills the concordance – English only)
- Cloze Builder (English and French)
- HyperText (Put in own text and link to concordances to help with guessing for further contexts)

ConcApp (http://www.edict.com.hk/PUB/concapp/)
Options/Features:

- Use own corpus / custom corpus (untagged/annotated)

- Phrase, word, prefix, suffix searches
- Concordancer
- Collocations
- Concgramming
- WordProfile
- Compare word profile with word lists (2 lists provided English)
- Link to Net Dictionary

TextSTAT
Download: http://neon.niederlandistik.fu-berlin.de/textstat/
User guide: http://sites.google.com/site/genabennett2/TextSTATusersguide.pdf?attredirects=0
Options/Features:

- User compiled corpus creation
- Frequency lists for the corpus
- Keyword searches
- Wild card searches
- Searches for two expressions with a number of words between them

Simple Concordance Program (http://www.textworld.com)
Options/features:

- Use own corpus
- Keyword search
- Prefix/suffix/anywhere search
- Case sensitivity
- Frequency lists
- Word profile

Corpora

French

- Chambers-Rostand Corpus of Journalistic French: http://www.ota.ahds.ac.auk/%20texts/2491.html
- Lexicometrie: Corpus of classic French literary texts
- http://ota.ahds.ac.uk/headers/2466.xml
- Le Corpus BAF (English-French Parallel): http://rali.iro.umontreal.ca/

Italian

- Banca dai dell'italiano parlato (BADIP): http://languageserver.uni-graz.at/badip/
- Corpus di Italiano Scritto (CORIS): http://corpus.cilta.unibo.it:8080/CORISCorpQuery.html

Spanish

- Corpus Oral de Referencia del Espanol Contempoaneo (COREC): http://www.lllf.uam.es/corpus/corpus_oral.html (Sample: http://www.llf.uam.es/corpus/corpus_lee.html#B4)
- The CREA Corpus of Spanish: http://www/rea.es/ AND http://corpus.rae.es/creanet.html

Multilingual

- TRACTOR archive: http://www.corpus.bham.ac.uk/ccl/services.htm#tractor

Corpus-based dictionaries

French

- Beachesne (2001). Dictionnaire des cooccurrences. Montreal: Gurein.
- Binon, Verlinde, Van Dyck, and Bertels (2000). Dictionnaire d'apprentissage du francais des affaires. Paris: Didier (www.projetdafa.net)
- Gonzalez Rodriguez (2004). Dictionnaire des collocations. (www.tonitraduction.net)
- Zingle and Brobeck-Single (2004). Dictionnaire combinatoire du francais. Expressions, locutions, et constructions. Paris: La Maison du Dictionnaire

German

- Elexiko: www.elexiko.de
- Neuman et al. (DATE). A Corpus-based lexical resource of German idioms. Produced as part of the "Collocations in the German Language" project at Berlin-Brandenburg Academy of Sciences.

Multilingual

- Worschatz Universitat Leipzig: http://corpora.informatik.uni-leipzig.de/?dict=es
- >> Search 59 Corpus-based monolingual dictionaries

Corpus-based Curriculae

- McCarthy, McCarten and Sandiford (2005/6). *Touchstone. Student's Book [1-4].* Cambridge: Cambridge University Press.