2.3.4 Employing corpora for language learning

Transcript

Slide 1

Hello! Welcome to the subsection on Using Corpora for language learning. of the ENRICH course. USING CORPORA is part of the section on methods and approaches, since most recently CORPORA, once mostly object of language research studies, have been more and more used in second language teaching and are central in Data Driven Learning, a language teaching and learning approach that has recently been re-valued.

Slide 2

Any idea of what language corpora are? If not, can you try and ***guess*** what they are and what their function might be in language teaching/learning?

- Look at the picture below: can it help you understand what Language Corpora might be?
- There are people walking into a PC, as if they could all be contained inside. Real people whose thoughts, voices and ideas are being swallowed inside, how comes? How comes that a tool as small as a PC can contain an apparently huge number of people's voices?
- In a way that is what corpora are, the largest possible collections of spoken and written language, so it is those people's voices, what they wrote, their language that can be saved into the PC and can be retrieved as we wish in order to observe language in use and understand it.

Now stop this video and take 5 minutes to watch the short YouTube video "Corpus linguistics: The basics" that provides an overview of language corpora. Click the link https://youtu.be/32RjJ-lA-8Q

- Check your guesses about corpora with what is said in the video.
- What have you learnt about Corpora? Does it match your intuitions?

Click the link under the video to share your views on the Forum

Slide 3

The aims of this subsection are to:

- **introduce** you to Language Corpora and to its use in ELT
- learn how to **consult and use corpora** in ELT

- encourage your use of corpora as an approach to teaching, particularly in an ELF aware perspective in multilingual classrooms
- **observe uses of non-standard English in standard language corpora and to explore ELF in ELF corpora**
- introduce you to the **Data Driven Learning** approach

Slide 4

So, let's try and define what a Corpus is. A Corpus is

- A collection of a large amount of **authentic texts**, **written**, **spoken** or **multimedia**, stored in a computer
- It is a **principled** collection of texts – written or spoken – that can be analysed with specially designed software
- A corpus is **principled** because texts are **selected** for inclusion according to pre-defined research purposes
- But what is most important is that a corpus is **not a dictionary**

Now stop this video and click the link to the video on Corpora and Spoken Language. Take your time to watch the video where Professor Michael McCarthy reflects upon corpora and spoken language. Michael McCarthy talks about corpora, describes their use in ELT, and illustrates how they can be a powerful tool for learners to understand spoken language. This video is particularly useful to understand corpora, but mostly to highlight their relevance for learning more about spoken language, too often disregarded in language classrooms and very rarely presented in terms of authentic language.

What have you learnt on corpora and on spoken language from Professor McCarthy's reflections?

Slide 5

Let's now learn more about a research area closely connected to corpora: **Corpus Linguistics**. What is it?

By Corpus Linguistics we refer to a research approach for describing authentic language in use, but also a **collection of methods** for studying language by using software packages (**concordances**).
It is a way to build a corpus using data well matched to a specific **research question**

Slide 6

So far, we have gradually learnt about corpora, about corpus linguistics and we have begun to see the implications of using corpora for language teaching. **Why are corpora useful in language teaching? Because...**

- **Large amounts of data** unveil tendencies and what's normal or typical in **real-life language use**
- Corpora show us what grammar books do not: **anomalies** in written & spoken language that seem to violate rules but **are acceptable and authentic**
- Corpora **can reveal instances of very rare cases, that we wouldn't get from just looking at single written or spoken texts**

Slide 7

But also, because…

- Corpora are a **valuable resource** in the language teaching and learning **process**
- They provide **significant insights** into **authentic language use** and strengthen **learners' autonomy** because learners can observe language and draw their own conclusions.

We can ask learners to use corpora for:

- Extracting information from texts
- Comparing different texts from different languages
- Identifying most frequent words in a language
- Understand collocations
- Unveiling unusual language occurrences
- Observing spoken language features
- Creating their own corpora for language projects
- Identify occurrences of non-standard language commonly used in authentic conversations

Slide 8

Let's now try and understand what corpora may reveal about English, the language we teach. For example,

- **do you know which the most frequent words in English are? And in your language?** Take a note and get back to it later on.

Now, stop the video, and let's do the activity, make sure you've got a copy of the table of the following slide

Slide 9

The table is subdivided into columns with **the 50 most frequent words in written (W) and in spoken (S) English**

Look carefully at the 2 columns

- Think of what you had first answered when you were asked whether you knew the most frequent words in English and in your language

Is there anything in the table that matches any of your original guesses?

- What do you notice that you did not expect to find?
- Would you use this table and this task with your students?

This table reveals several features of English, for example:

- The most frequent words are almost all grammar words and not content words (contrary to what we might have expected);
- Short forms are much more frequent than expected, particularly in spoken language;
- Colloquial forms are very frequent;
- Discourse markers (too often not taught when teaching spoken language)

This 'noticing' activity is a way of asking learners to 'observe' features of language that they would never normally be asked to do.
This is possible because of corpora that can reveal what is not possible to be shown in a coursebook or in a grammar book

When finished, go to the FORUM and share your findings but remember to

Click the link under the video to share your views on the Forum

Slide 10

Learning about language corpora implies getting familiar with the notions and terms most frequently encountered in this field and being able to refer to them. For example, COLLOCATION and COLLIGATION are central in working with corpora, they are

- concepts associated with the distributional properties of linguistic items in actual language use
- They refer to the ***likelihood of occurrence*** of:
    - (two or more) lexical items = ***collocation***
    - grammatical categories = ***colligation***

COLLOCATION refers to words that **usually tend to occur together** showing a **frequency** higher than what would be expected by chance, for example when we say "heavy rain"**.**

COLLIGATION is the **significant occurrence of a word with grammatical words** or with grammatical categories, for example when using 'as', it usually occurs followed by an adverb or by a noun
In "as much as you can"

Slide 11

Other two notions often associated with corpora are those referred to **tools** used when consulting corpora, as the **FREQUENCY LIST** of a corpus that includes its most frequent words with their numbers of occurrences **(e.g., most frequent**

words in English) as the one you saw in the previous activity on most frequent written and spoken terms in English.

or

the notion of **CONCORDANCES** that show **the shades of meaning of a word in real contexts and the syntactic and grammatical contexts where the words are used**

As an example, have a look at the table below where the **Concordances of the word 'problem' – our 'node' word - are observable in different strings. What do you notice? For sure this is a new way of looking at language by observing the numerous occurrences a word.**

Slide 12

Let's now have a closer look at how corpora can actually provide a way **for teachers to teach** and **for learners to learn** using a non-traditional approach. Think of some possibilities and jot them down, let's now look at an example of activity on language use where learners – and you - are asked to understand ways of using the word JOB.

- **What and How would you teach your learners to understand and use the word JOB? Think of situations where you did it and think of how you did it.**

Now, briefly stop the video, look at the table in this slide and

- **Observe the occurrences of JOB in the following concordance lines from the BNC Corpus**
- **What precedes and what follows JOB? Does the meaning change in each line?**
- **Would you use this type of activity in your lessons? Yes or No? Why?**
- **Do you see any advantage in using corpora to show your learners different uses if a word? If yes, which one/s?**

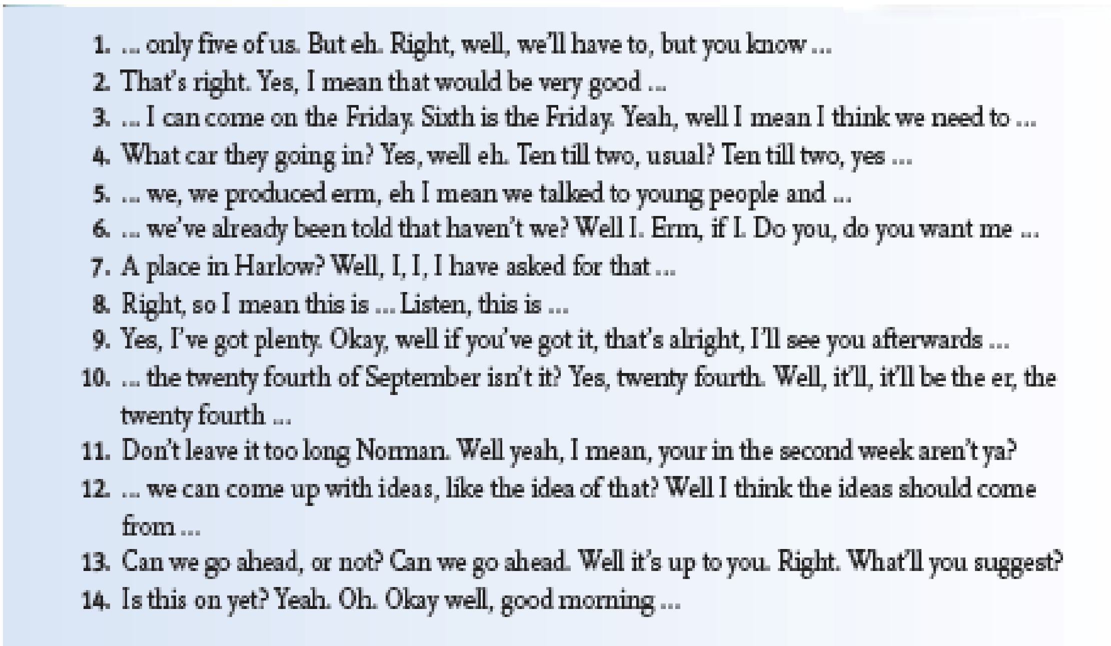
When finished, go to the FORUM and share your findings but remember to Click the link under the video to share your views on the Forum **

Slide 13

Let's now practice with an activity based upon the observation of concordance lines of authentic extracts of spoken language that could be used in a lesson to elicit learners noticing the main features of spoken language as well as on non-standard use of English.

Make sure you have **a copy of the handout for Activity 3**, **stop this video** and answer the following questions:

- **What do you notice? What can you ask your learners to notice?**



1. ... only five of us. But eh. Right, well, we'll have to, but you know ...
2. That's right. Yes, I mean that would be very good ...
3. ... I can come on the Friday. Sixth is the Friday. Yeah, well I mean I think we need to ...
4. What car they going in? Yes, well eh. Ten till two, usual? Ten till two, yes ...
5. ... we, we produced erm, eh I mean we talked to young people and ...
6. ... we've already been told that haven't we? Well I. Erm, if I. Do you, do you want me ...
7. A place in Harlow? Well, I, I, I have asked for that ...
8. Right, so I mean this is ... Listen, this is ...
9. Yes, I've got plenty. Okay, well if you've got it, that's alright, I'll see you afterwards ...
10. ... the twenty fourth of September isn't it? Yes, twenty fourth. Well, it'll, it'll be the er, the twenty fourth ...
11. Don't leave it too long Norman. Well yeah, I mean, your in the second week aren't ya?
12. ... we can come up with ideas, like the idea of that? Well I think the ideas should come from ...
13. Can we go ahead, or not? Can we go ahead. Well it's up to you. Right. What'll you suggest?
14. Is this on yet? Yeah. Oh. Okay well, good morning ...

For example you could focus your learners' attention on the use of **fillers, like 'erm'; or short forms, as '*cos*'; or repetitions, as I..I.., Ten till two....; slang forms, ; discourse markers as *Well, yeah, You know, I mean....*; as well as non-standard forms of English, as *What car they going in*?)**

When finished, go to the FORUM and share your findings but remember to Click the link under the video to share your views on the Forum **

Slide 14

In our brief introduction to corpora, to the main notions, to their use in ELT, it is important to know that there are **numerous and diverse corpora** for almost every language, most of them created for research purposes but that can be used also for teaching purposes. For example,

1. **General English corpora**—very large! e.g.,
   - The British National Corpus (BNC) (100 million words of spoken and written British English)
   - The Collins Co-build is an analytical database of English (over 4.5 billion words)
2. **Specialised corpus—e.g.,**
   - The Michigan Corpus of Spoken English (MICASE)
3. **Learner corpus—language use created by people learning a particular language. e.g.,**

- The International Corpus of Learner English.

4. **Comparable corpora—a corpus formed by 2 languages, e.g.,** English and Spanish—exactly the same texts translated
5. **Parallel Corpora—two or more collections of texts in different languages**
6. **There are also Corpora of English as a Lingua Franca:** VOICE, ELFA and The Asian Corpus of English (ACE)

Most of these corpora can be easily consulted by teachers and learners either directly or through specific platforms.

Slide 15

The main corpora of ELF are 3: VOICE, ALFA and ACE.

The Vienna-Oxford International Corpus of English (VOICE), compiled at the Department of English at the University of Vienna, is a structured collection of language data capturing spoken ELF interactions mostly among university students;

VOICE comprises transcripts of ***naturally occurring, non-scripted*** **face-to-face interactions in English as a lingua franca** (ELF).
The speakers recorded in VOICE are experienced ***ELF speakers*** from a wide range of first language backgrounds. So far, VOICE includes approximately **1250** ELF speakers with approximately ***50 different first languages***

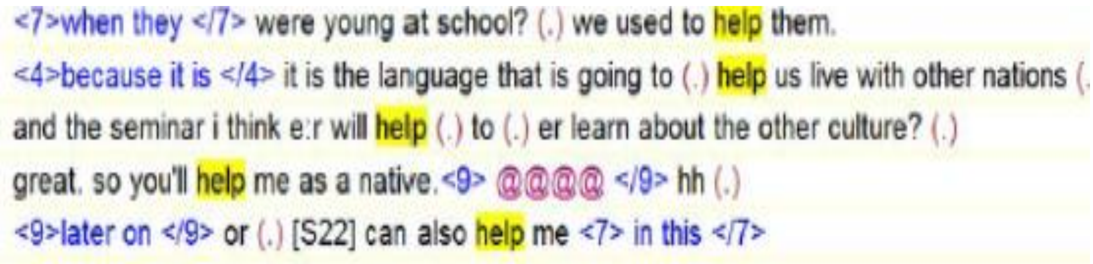
It aims to provide a general basis for analyses of English as a lingua franca (ELF) talk on all linguistic levels.

VOICE provides users with a computer-readable corpus of English as it is spoken by this non-native speaking majority of users in different contexts. These speakers use English successfully on a daily basis all over the world, in their personal, professional or academic lives.
Therefore, they are not seen as language learners but as **language users in their own right.**
It is therefore clearly worth finding out just **how they use the language**. This is exactly what VOICE seeks to make possible.
VOICE researchers seek to gain access to interactions where people of different first language backgrounds meet and use English as their preferred language, i.e. as their lingua franca of choice.

The VOICE website provides all the help you may need to consult the corpus.

Slide 16

Let's have a look at an extract from VOICE, in the extract you will notice how ELF speakers use the verb HELP. Is there anything in particular that strikes you?
For example, look at the way S1 asks for help; she says, "Help me out" and the others respond. Most interesting is also the emergence of speakers' awareness of cultural differences as in the last few lines.



<7>when they </7> were young at school? (.) we used to help them.
<4>because it is </4> it is the language that is going to (.) help us live with other nations (.
and the seminar i think e:r will help (.) to (.) er learn about the other culture? (.)
great. so you'll help me as a native.<9> @@@@ </9> hh (.)
<9>later on </9> or (.) [S22] can also help me <7> in this </7>

Slide 17

Another Corpus of English as a Lingua Franca is ELFA developed in the University of Helsinki, Finland.

ELFA (English as a Lingua Franca in the Academia) investigates English as it is currently used by speakers with different mother tongues. It explores the practices, regularities, and patterns of English as a lingua franca (ELF) as it appears in natural use by 'expert users' in university settings. The project has compiled two large, electronically stored databases: ELFA of academic speech, and WrELFA of academic writing, freely available to the research community.

- The ELFA corpus contains 1 million words of transcribed spoken academic ELF.
- It includes approx. 650 speakers representing 51 first languages.
- The percentage of speech by native English speakers is 5%.

You can explore ELFA by using the link provided.

Slide 18

The third most recent corpus of ELF is the Asian Corpus of English.

It contains 1 million words
Data nature: naturally occurring, spoken, interactive ELF in Asia
Speech events are interviews; press conferences; service encounters; seminar discussions; working group discussion; workshop discussions; meetings; panels; question-and-answer sessions.

Slide 19

Let's look at two platforms that allow us to use corpora.
Why are these types of platforms useful for us language teachers? Let's remember that corpora have many different uses, including:

- finding out how native speakers actually speak and write,
- finding the frequency of words, phrases collocates,
- looking at language variation and change; e.g. historical, dialects, and genres
- gaining insight into culture; for example, what is said about different concepts over time and in different countries,
- designing authentic language teaching materials and resources.

The first one is the one at **BYU site** that was created by Mark Davies, Professor of Linguistics at Brigham Young University. These are probably the most widely-used corpora currently available.

In addition to the ten corpora (and the Google Books (Advanced) interface), there are the TV corpus and the Movie Corpus in the BYU platform, but there are also many corpus-based resources. These allow you to:

- See detailed entries for the top 60,000 words in English (definitions, genre variation, collocates, concordance lines, synonyms) -- all on one page
- Enter and analyze your own text, find keywords from your text, compare phrases to COCA, and see detailed information (see above) for each word
- Get detailed information from the Academic Vocabulary List (including detailed information on each word, and analyzing your own academic texts)
- Download large amounts of corpus-based data, including word frequency, collocates, and n-grams
- Download the entire corpus for offline use (COCA, COHA, GloWbE, NOW, NOW monthly updates, Wikipedia, Spanish)

Slide 20

The other most recent and extremely powerful platfom is **Sketch Engine** contains 500 ready-to-use corpora in 90+ languages, each having a size of up to 30 billion words to provide a truly representative sample of language
Sketch Engine is a tool to explore how language works. Its algorithms analyze authentic texts of billions of words (text corpora) to identify instantly what is typical in language and what is rare, unusual or emerging usage. It is also designed for text analysis or text mining applications.
Sketch Engine is used by linguists, lexicographers, translators, students and teachers. It processes texts of billions of words and, within seconds, finds

instances of the word, phrase or phenomenon and presents the results in the form of Word Sketches, concordances or word lists.

Watch the introduction to Sketch Engine:
https://www.youtube.com/watch?v=_MtcVMK7AGU

If you wish to learn more about how Sketch Engine works, watch the presentation indicated in the references that will guide you through its main functions : https://www.youtube.com/watch?v=PL5wOxwQ1S4

Slide 21

To conclude this preliminary introduction to Corpora in language teaching within an ELF aware perspective, let's briefly explore **Data Driven Learning (DDL),** an approach to language learning that

- promotes the direct use of corpora for learning purposes;
- values corpora as tools with which learners can engage directly for guided or autonomous learning and
- where real language data are investigated by learners, and learner-centered activities focus on language discovery

DDL is an approach where "the language-learner is also, essentially, a research worker whose learning needs to be driven by access to linguistic data – hence the term 'data-driven learning' (DDL) to describe the approach" (Johns, 1991: 2).

Slide 22

Let's now try and explore features of non-standard English – ELF – use in an extract from the Asian Corpus of English.

Make sure you have **a copy of the handout for Activity 4**. S**top this video, go through all the lines of the extract** and answer the following questions:

What features of non-standard English (ELF) use do you notice?

For example, you may observe line 1, 5 or line 8, 13 and 14 in the interactions. Communication takes place and interactants understand each other using repetitions at times and adjusting to the interlocutor.

Slide 23

To conclude Activity 4, It would be important for you to reflect upon what has been presented so far about Language Corpora. I think you would agree that **Corpora for language learning can be a very useful tool for exploring**

authentic uses of English and contribute to teaching English in an ELF aware perspective

But, using corpora might imply a shift in perspective in language teaching and the use of materials and tools that most of you have never used before. This might sound slightly upsetting, and imply the adoption of a new approach, so what is your personal reaction?

Please respond to the following questions bearing in mind your own experience and the context you work in since you are the one who knows the context. Jot down your answers and keep them for future discussions.

- To what extent can language corpora help EL teachers to better scaffold their students' learning?
- How can language corpora enhance an ELF-aware approach in ELT?
- What might be the pedagogical advantages of using Data Driven Learning?
- Make a list of what you foresee as pros and cons of using corpora in the ELT class, what is their balance?

When you have finished, go to the FORUM and share your findings but remember to Click the link under the video to share your views on the Forum **

Slide 24

To conclude, bearing in mind the aims of this subsection initially stated in slide 3, let's look at what we have learnt so far about Corpora and about their use in language learning.

We understood

- What a Corpus and What Corpus Linguistics is.
- Why and how to use Corpora in ELT.
- Which Corpora may better unveil features of standard and non-standard English and we learnt about the 3 main Corpora of English as a Lingua Franca.
- (We understood) the meaning and use of different terms, notions and tools related to Language Corpora.
- (We understood) Which platforms can help us and our learners better consult corpora, in particular BYU & Sketch Engine.
- We learnt about DDL: a new approach for language learning capable of enhancing learners' autonomy and ELF Awareness

In view of your future language teaching, I do hope that you have been enriched by discovering Corpora for learning

So, GOODBYE AS LONG AS IT GOES!