ENRICH

English as a Lingua Franca practices
for inclusive multilingual classrooms

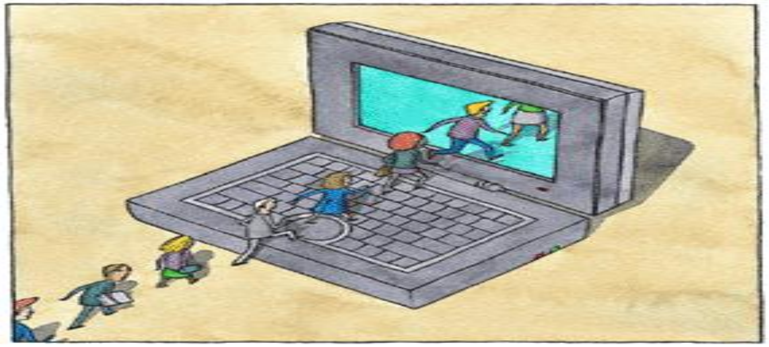
Corpora for language learning

Lucilla Lopriore
Valeria Fiasco
Roma Tre University

Activity 1—What are corpora?

- Do you ***know*** what language corpora are? If not, can you ***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?
- Now watch the short video and check your guesses

Click the link under the video to share your views



What is the aim of this sub-section?

- To **introduce** you to Language Corpora in ELT
- To learn how to **consult and use corpora** in ELT
- To encourage the use of corpora in language teaching and learning as an approach to teaching
- To **explore ELF in language corpora**
- To introduce you to **Data Driven Learning**

What is a Corpus?

- A collection of a large amount of **authentic texts**, **written**, **spoken** or **multimedia**, stored in a computer
- A **principled** collection of texts available for analysis with specially designed software
- A corpus is **principled** because texts are **selected** for inclusion according to pre-defined research purposes
- A corpus is **not a dictionary**

Task: Watch the video in the link* and learn about Corpora and spoken language

What is Corpus Linguistics?

- A **research approach** for describing authentic language in use
- It is a **collection of methods** for studying language
- Software packages (**concordancers**) are used in order to study them
- A corpus is built using data well matched to a **research question** it is built to investigate

Why use corpora 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

Why use corpora in language teaching? Because…

- Corpora represent a **valuable resource** in the language teaching and learning process
- They provide **significant insights** into **authentic language use** and strengthen **learners' autonomy**

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
- Learning about collocations
- Unveiling unusual language occurrences
- Observing spoken language features
- Creating their own corpora for language projects

What do you expect corpora may reveal about English?

ACTIVITY 2

a. What do you expect corpora might reveal about English?

b. Can you guess which the most frequent words in English are? And in your own language?

c. In the next slide there is a table with **2 columns** with **the 50 most frequent words in written (W) and in spoken (S) English**

d. Look carefully at the 2 columns

- Think of what you answered in **b**, is there anything that matches your guesses?
- What do you notice that you did not expect to find?
- Would you use this table and this task with your students?

When finished, go to the FORUM and share your findings

- **What do you notice? What did you not expect to find?**

✍ W	💬 S	✍ W	💬 S	✍ W	💬 S	✍ W	💬 S	✍ W	💬 S
1. THE	1. THE	11. IT	11. IN	21. BE	21. THEY	31. ARE	31. FOR	41. THEIR	41. DONXT
2. TO	2. I	12. ON	12. WAS	22. MY	22. WELL	32. AN	32. THIS	42. SHE	42. SHE
3. OF	3. YOU	13. HE	13. IS	23. HAVE	23. WHAT	33. THIS	33. JUST	43. WHO	43. THINK
4. A	4. AND	14. IS	14. ITXS	24. FROM	24. YES	34. HAS	34. ALL	44. IF	44. IF
5. AND	5. TO	15. WITH	15. KNOW	25. HAD	25. HAVE	35. BEEN	35. THERE	45. HIM	45. WITH
6. IN	6. IT	16. YOU	16. NO	26. BY	26. WE	36. UP	36. LIKE	46. WE	46. THEN
7. I	7. A	17. BUT	17. OH	27. ME	27. HE	37. WERE	37. ONE	47. ABOUT	47. AT
8. WAS	8. YEAH	18. AT	18. SO	28. HER	28. DO	38. OUT	38. BE	48. WILL	48. ABOUT
9. FOR	9. THAT	19. HIS	19. BUT	29. THEY	29. GOT	39. WHEN	39. RIGHT	49. ALL	49. ARE
10. THAT	10. OF	20. AS	20. ON	30. NOT	30. THATXS	40. ONE	40. NOT	50. WOULD	50. AS

FREQUENCY in written (✍ W) and spoken (💬 S) language

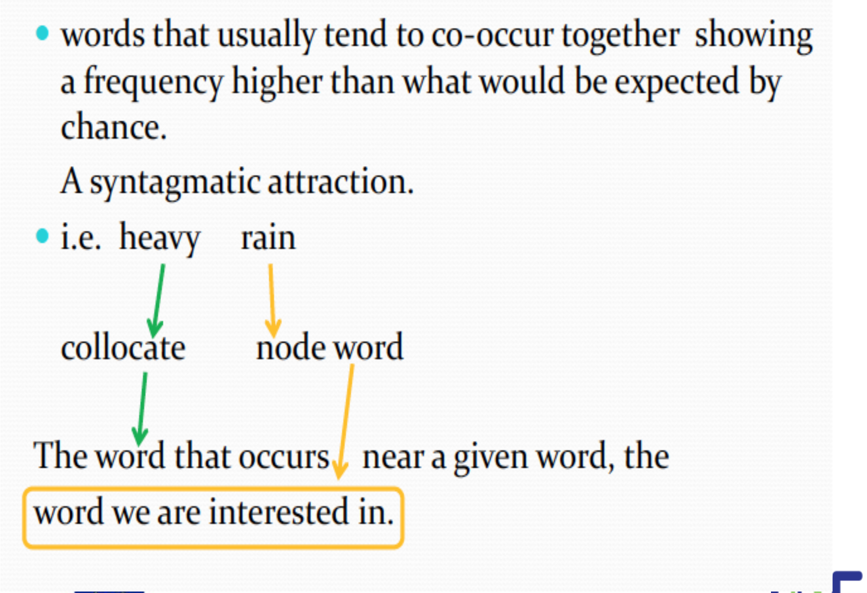
Adapted from Carter et al. 1999

Corpus termimology

Collocation and colligation

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

COLLOCATION



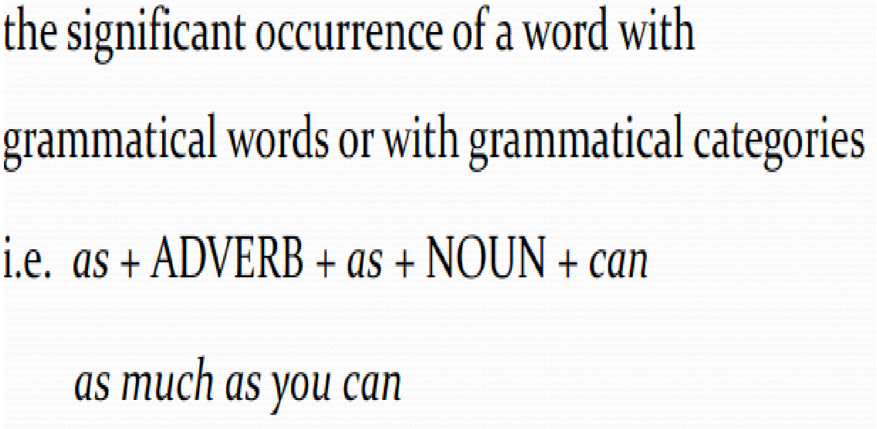
- words that usually tend to co-occur together showing a frequency higher than what would be expected by chance.

  A syntagmatic attraction.
- i.e. heavy rain

collocate node word

The word that occurs near a given word, the word we are interested in.

COLLIGATION



the significant occurrence of a word with grammatical words or with grammatical categories

i.e. *as* + ADVERB + *as* + NOUN + *can*

as much as you can

Corpora tools

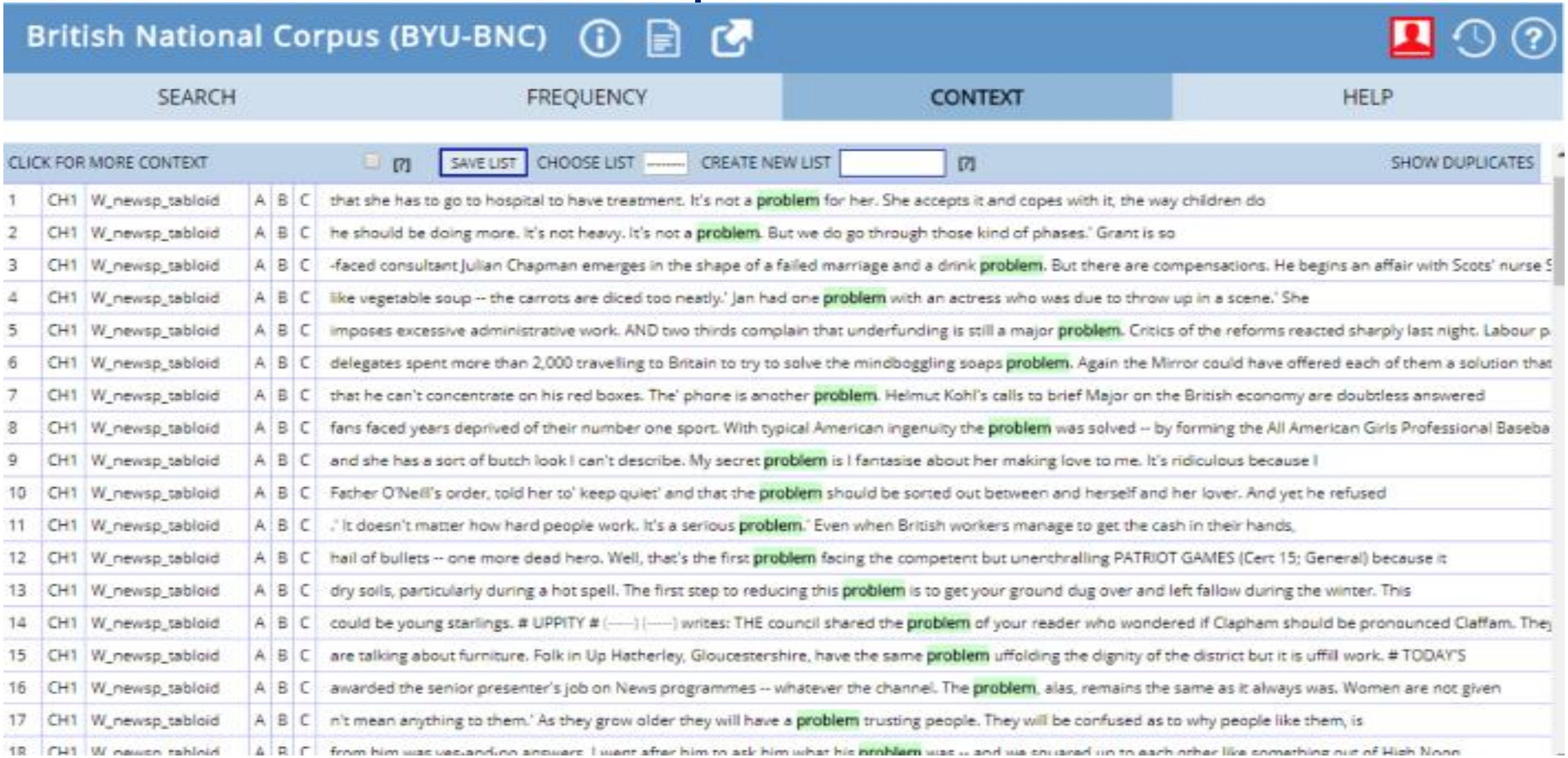
FREQUENCY LISTS

The frequency list of a corpus includes its most frequent words with their numbers of occurrences (eg most frequent words in English)

CONCORDANCES

They show the shades of meaning of a word in real contexts and the syntactic and grammatical contexts where the words are used

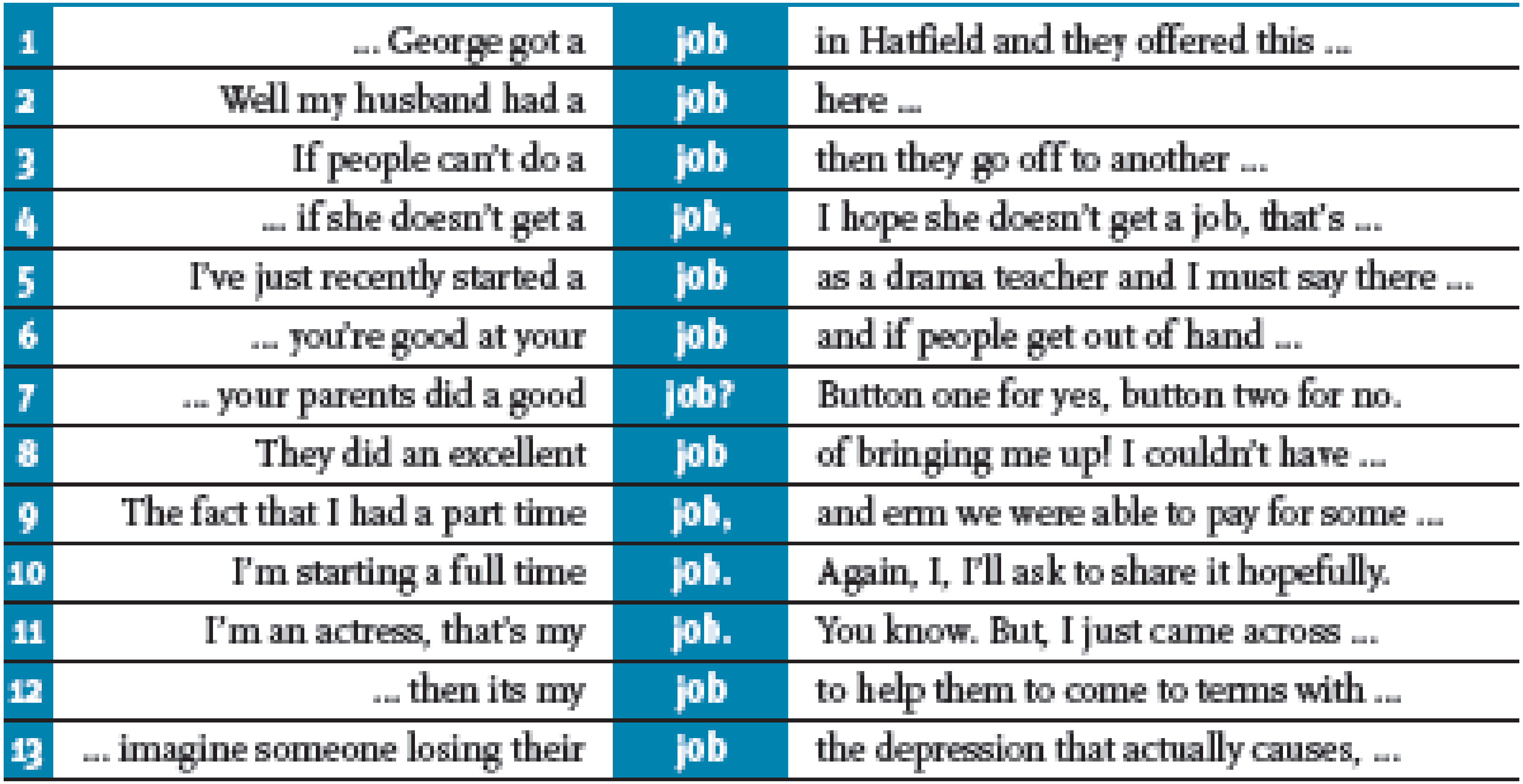
See below **Concordances of the word ‘problem’**:



1	CH1	W_newsp_tabloid	A	B	C	that she has to go to hospital to have treatment. It's not a **problem** for her. She accepts it and copes with it, the way children do
2	CH1	W_newsp_tabloid	A	B	C	he should be doing more. It's not heavy. It's not a **problem**. But we do go through those kind of phases.' Grant is so
3	CH1	W_newsp_tabloid	A	B	C	-faced consultant Julian Chapman emerges in the shape of a failed marriage and a drink **problem**. But there are compensations. He begins an affair with Scots' nurse S
4	CH1	W_newsp_tabloid	A	B	C	like vegetable soup -- the carrots are diced too neatly.' Jan had one **problem** with an actress who was due to throw up in a scene.' She
5	CH1	W_newsp_tabloid	A	B	C	imposes excessive administrative work. AND two thirds complain that underfunding is still a major **problem**. Critics of the reforms reacted sharply last night. Labour p
6	CH1	W_newsp_tabloid	A	B	C	delegates spent more than 2,000 travelling to Britain to try to solve the mindboggling soaps **problem**. Again the Mirror could have offered each of them a solution that
7	CH1	W_newsp_tabloid	A	B	C	that he can't concentrate on his red boxes. The' phone is another **problem**. Helmut Kohl's calls to brief Major on the British economy are doubtless answered
8	CH1	W_newsp_tabloid	A	B	C	fans faced years deprived of their number one sport. With typical American ingenuity the **problem** was solved -- by forming the All American Girls Professional Baseba
9	CH1	W_newsp_tabloid	A	B	C	and she has a sort of butch look I can't describe. My secret **problem** is I fantasise about her making love to me. It's ridiculous because I
10	CH1	W_newsp_tabloid	A	B	C	Father O'Neill's order, told her to' keep quiet' and that the **problem** should be sorted out between and herself and her lover. And yet he refused
11	CH1	W_newsp_tabloid	A	B	C	.' It doesn't matter how hard people work. It's a serious **problem**.' Even when British workers manage to get the cash in their hands,
12	CH1	W_newsp_tabloid	A	B	C	hail of bullets -- one more dead hero. Well, that's the first **problem** facing the competent but unenthralling PATRIOT GAMES (Cert 15; General) because it
13	CH1	W_newsp_tabloid	A	B	C	dry soils, particularly during a hot spell. The first step to reducing this **problem** is to get your ground dug over and left fallow during the winter. This
14	CH1	W_newsp_tabloid	A	B	C	could be young starlings. # UPPITY # (----) (----) writes: THE council shared the **problem** of your reader who wondered if Clapham should be pronounced Claffam. They
15	CH1	W_newsp_tabloid	A	B	C	are talking about furniture. Folk in Up Hatherley, Gloucestershire, have the same **problem** uffolding the dignity of the district but it is uffill work. # TODAY'S
16	CH1	W_newsp_tabloid	A	B	C	awarded the senior presenter's job on News programmes -- whatever the channel. The **problem**, alas, remains the same as it always was. Women are not given
17	CH1	W_newsp_tabloid	A	B	C	n't mean anything to them.' As they grow older they will have a **problem** trusting people. They will be confused as to why people like them, is
18	CH1	W_newsp_tabloid	A	B	C	from him was yes-and-no answers. I went after him to ask him what his **problem** was -- and we squared up to each other like something out of High Noon

Corpora for Language Learning

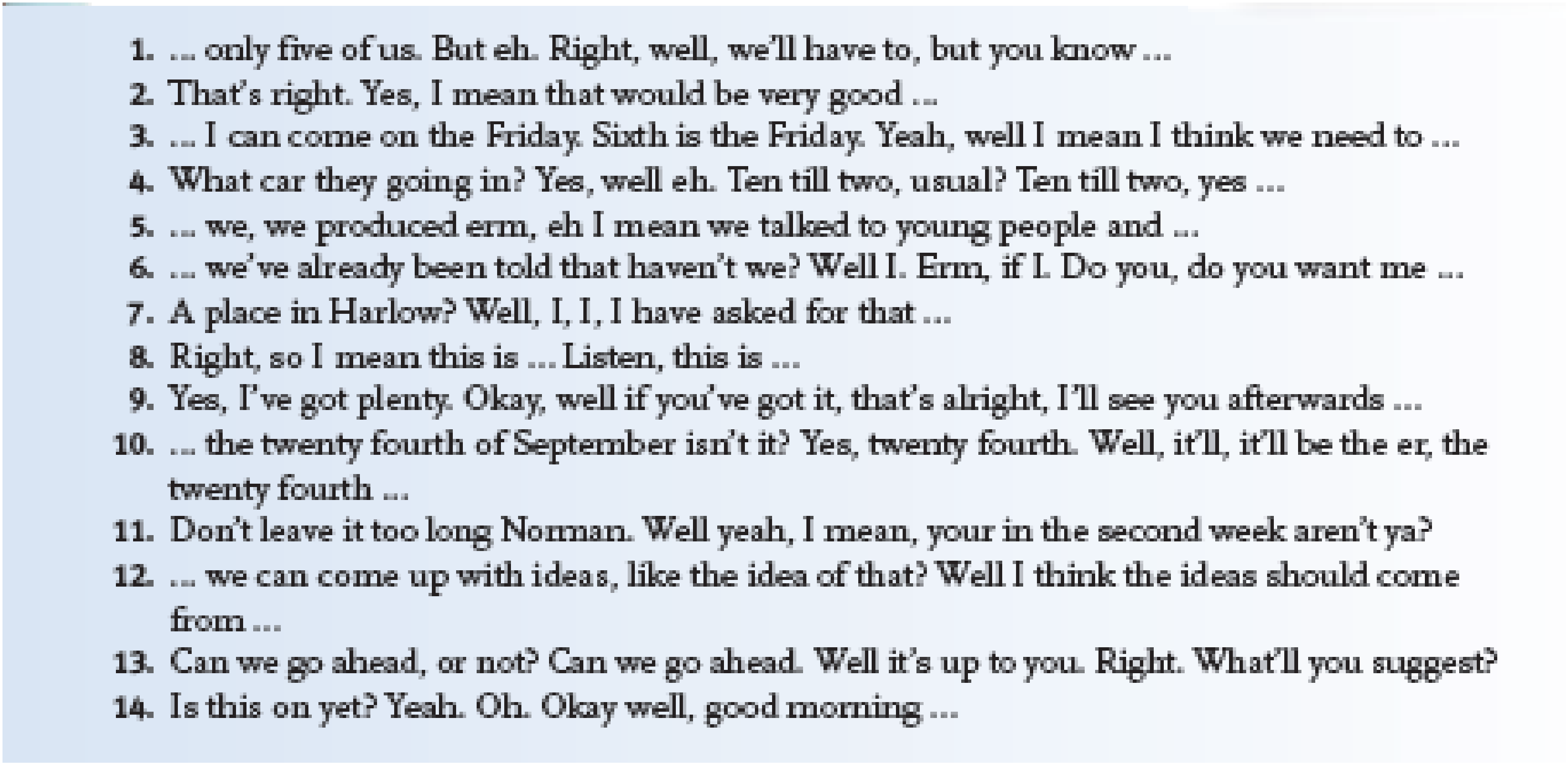
- What and How would you teach your learners to understand and use the word JOB?
- Observe the occurrences of JOB in the following concordance lines from the BNC Corpus
- What precedes and what follows JOB? Does the meaning change in each line?



1	... George got a	job	in Hatfield and they offered this ...
2	Well my husband had a	job	here ...
3	If people can't do a	job	then they go off to another ...
4	... if she doesn't get a	job,	I hope she doesn't get a job, that's ...
5	I've just recently started a	job	as a drama teacher and I must say there ...
6	... you're good at your	job	and if people get out of hand ...
7	... your parents did a good	job?	Button one for yes, button two for no.
8	They did an excellent	job	of bringing me up! I couldn't have ...
9	The fact that I had a part time	job,	and erm we were able to pay for some ...
10	I'm starting a full time	job.	Again, I, I'll ask to share it hopefully.
11	I'm an actress, that's my	job.	You know. But, I just came across ...
12	... then its my	job	to help them to come to terms with ...
13	... imagine someone losing their	job	the depression that actually causes, ...

Corpora for Language Learning
ACTIVITY 3

- The following are extracts from spoken English corpora
- They can be used by learners to “observe” spoken language
- **What do you notice? What can you ask your learners to notice?** (e.g. use of fillers, like ‘erm’; short forms, as ‘*cos*’; repetitions; slang forms; discourse markers as *Well, You know, I mean*….)



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 …

Types of English corpora

1. **General English corpora**—very large! e.g.,

- The **British National Corpus (BNC)** (100 million words of spoken and written British English)
- The **Collins Cobuild** 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 **Corpora of English as a Lingua Franca**: **VOICE, ELFA and The Asian Corpus of Engish**

Corpora of English as a Lingua Franca

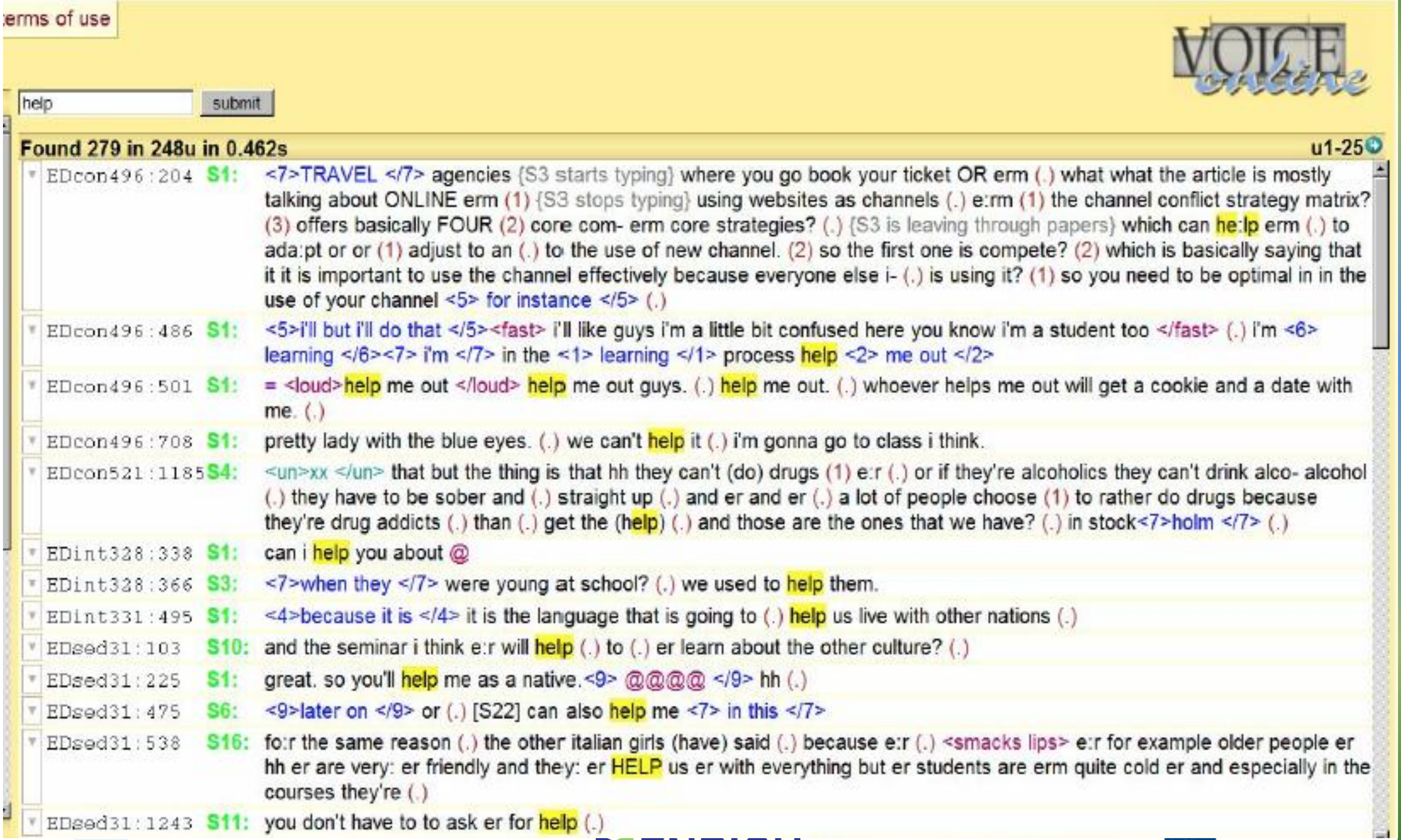
VOICE **https://www.univie.ac.at/voice/**

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;

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

ELF: an extract from VOICE

What do you notice?



Found 279 in 248u in 0.462s u1-25

EDcon496:204	S1:	<7>TRAVEL </7> agencies {S3 starts typing} where you go book your ticket OR erm (.) what what the article is mostly talking about ONLINE erm (1) {S3 stops typing} using websites as channels (.) e:rm (1) the channel conflict strategy matrix? (3) offers basically FOUR (2) core com- erm core strategies? (.) {S3 is leaving through papers} which can he:lp erm (.) to ada:pt or or (1) adjust to an (.) to the use of new channel. (2) so the first one is compete? (2) which is basically saying that it it is important to use the channel effectively because everyone else i- (.) is using it? (1) so you need to be optimal in in the use of your channel <5> for instance </5> (.)
EDcon496:486	S1:	<5>i'll but i'll do that </5><fast> i'll like guys i'm a little bit confused here you know i'm a student too </fast> (.) i'm <6> learning </6><7> i'm </7> in the <1> learning </1> process help <2> me out </2>
EDcon496:501	S1:	= <loud>help me out </loud> help me out guys. (.) help me out. (.) whoever helps me out will get a cookie and a date with me. (.)
EDcon496:708	S1:	pretty lady with the blue eyes. (.) we can't help it (.) i'm gonna go to class i think.
EDcon521:1185	S4:	<un>xx </un> that but the thing is that hh they can't (do) drugs (1) e:r (.) or if they're alcoholics they can't drink alco- alcohol (.) they have to be sober and (.) straight up (.) and er and er (.) a lot of people choose (1) to rather do drugs because they're drug addicts (.) than (.) get the (help) (.) and those are the ones that we have? (.) in stock<7>holm </7> (.)
EDint328:338	S1:	can i help you about @
EDint328:366	S3:	<7>when they </7> were young at school? (.) we used to help them.
EDint331:495	S1:	<4>because it is </4> it is the language that is going to (.) help us live with other nations (.)
EDsed31:103	S10:	and the seminar i think e:r will help (.) to (.) er learn about the other culture? (.)
EDsed31:225	S1:	great. so you'll help me as a native.<9> @@@@ </9> hh (.)
EDsed31:475	S6:	<9>later on </9> or (.) [S22] can also help me <7> in this </7>
EDsed31:538	S16:	fo:r the same reason (.) the other italian girls (have) said (.) because e:r (.) <smacks lips> e:r for example older people er hh er are very: er friendly and they: er HELP us er with everything but er students are erm quite cold er and especially in the courses they're (.)
EDsed31:1243	S11:	you don't have to to ask er for help (.)

Corpora of English as a Lingua Franca

ELFA
English as a Lingua Franca in Academic Settings

ELFA 2008—The Corpus of English as a Lingua Franca in Academic Settings.

http://www.helsinki.fi/elfa

- 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%.
- **Explore ELFA:** http://metashare.csc.fi/repository/browse/elfa-corpus/b0a50844086d11e68302005056be118ec040f2484984409ab1b22ec303278d96/

Corpora of English as a Lingua Franca:

The Asian Corpus of English

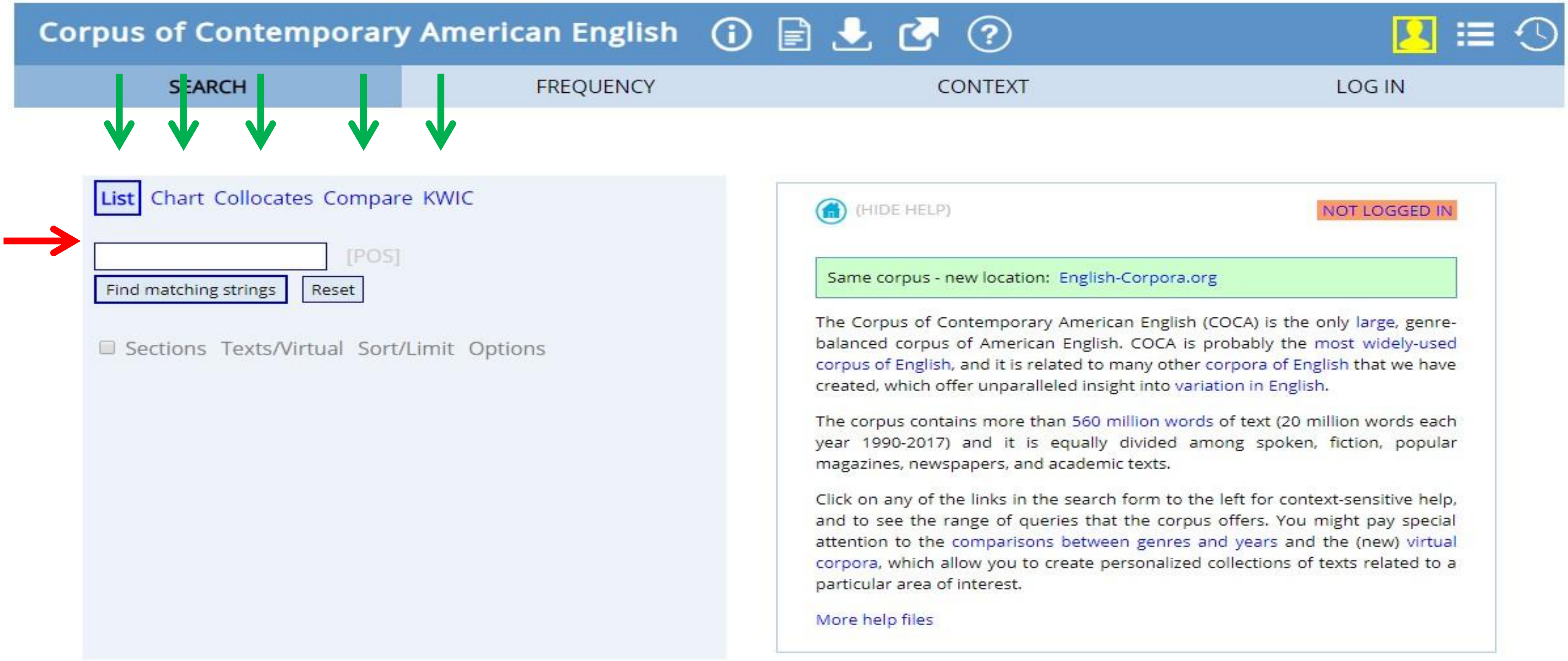
http://corpus.eduhk.hk/ace/index.html

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

Corpus platforms:
BYU (Brigham Young University)

https://corpus.byu.edu/overview.asp

This online platform includes **16 corpora of English** (British, American, Canadian, English for Specific Purposes).



Same corpus - new location: English-Corpora.org

The Corpus of Contemporary American English (COCA) is the only large, genre-balanced corpus of American English. COCA is probably the most widely-used corpus of English, and it is related to many other corpora of English that we have created, which offer unparalleled insight into variation in English.

The corpus contains more than 560 million words of text (20 million words each year 1990-2017) and it is equally divided among spoken, fiction, popular magazines, newspapers, and academic texts.

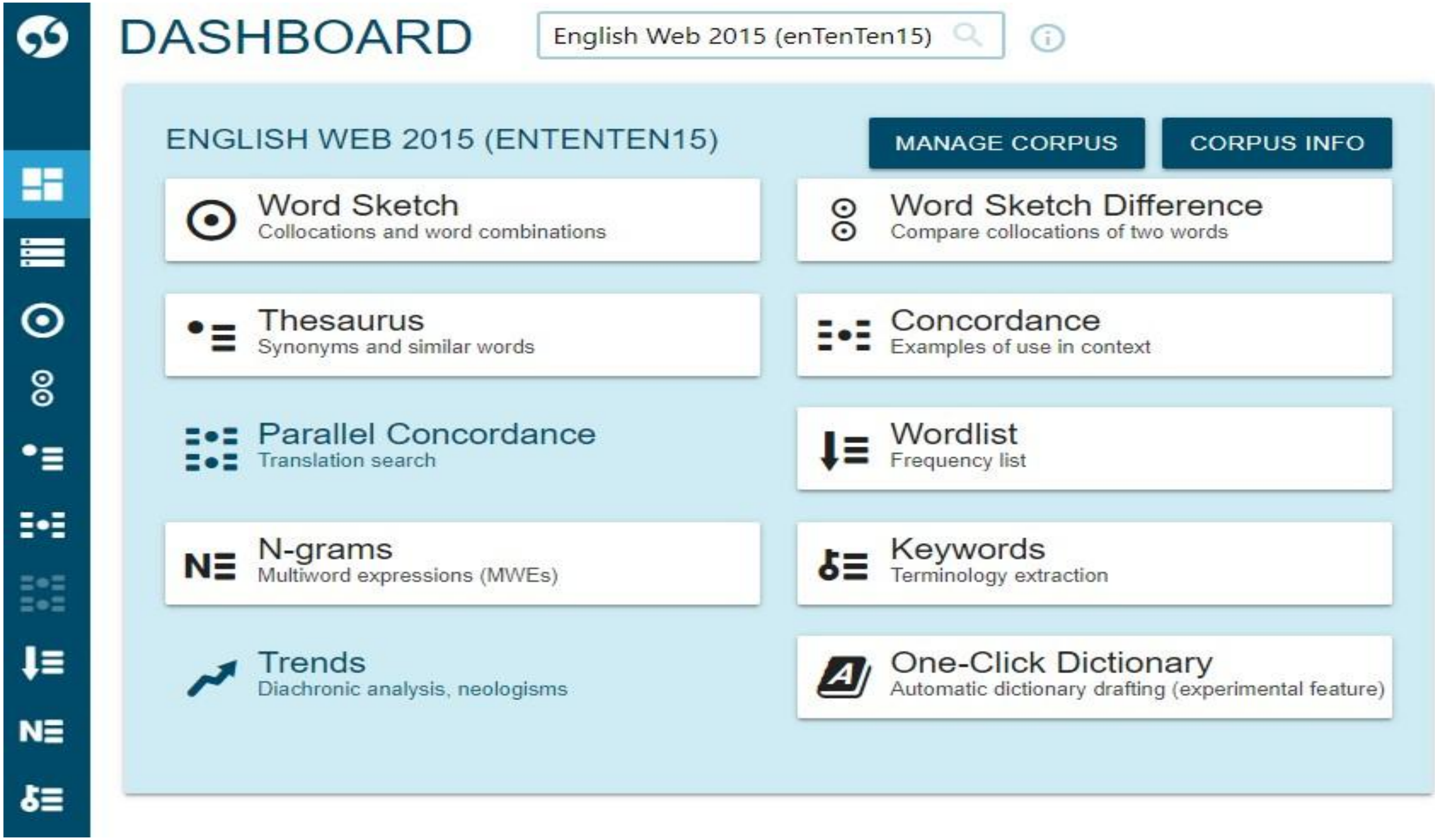
Click on any of the links in the search form to the left for context-sensitive help, and to see the range of queries that the corpus offers. You might pay special attention to the comparisons between genres and years and the (new) virtual corpora, which allow you to create personalized collections of texts related to a particular area of interest.

More help files

Corpus platforms: SketchEngine

https://www.sketchengine.eu

This online platform contains **about 500 corpora** in **more than 90 languages.**



DASHBOARD

English Web 2015 (enTenTen15)

ENGLISH WEB 2015 (ENTENTEN15)

MANAGE CORPUS

CORPUS INFO

Word Sketch
Collocations and word combinations

Word Sketch Difference
Compare collocations of two words

Thesaurus
Synonyms and similar words

Concordance
Examples of use in context

Parallel Concordance
Translation search

Wordlist
Frequency list

N-grams
Multiword expressions (MWEs)

Keywords
Terminology extraction

Trends
Diachronic analysis, neologisms

One-Click Dictionary
Automatic dictionary drafting (experimental feature)

Data Driven Learning (DDL)

Pedagogical implications of language corpora use

- Corpora are tools with which learners can engage directly for guided or autonomous learning
- The direct use of corpora for learning purposes has become known as **Data-Driven Learning (DDL)**
- **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).
- It is an approach where real language data are investigated by learners, and **learner-centered activities** focus on **language discovery**

Observe the following extract from the ELF Asian Corpus. What features of non-standard English (ELF) use do you notice?

S1: i think start from your first er statement (.) yeah
S2: [first name1]
S1: you can talk about anything from er malaysia
S3: any any any topic will do i i think
S2: i'm interested in (.) the (1) usage of english in your country how widely is english used [first name1]
S1: oh (.) in my country as you know we use and learn english as a foreign language yeah
S2: does this begin at the secondary school level
S1: no you know erm only people from the city have opportunity to study english
S3: hm
S1: i mean the people who live far away from the city they don't have any opportunity to study english
S3: do do you mean english is not used as a second language it's a foreign language
S1: yeah it's foreign <1>language</1> yeah (.) some student they start er learning english just only at the <2>university</2>

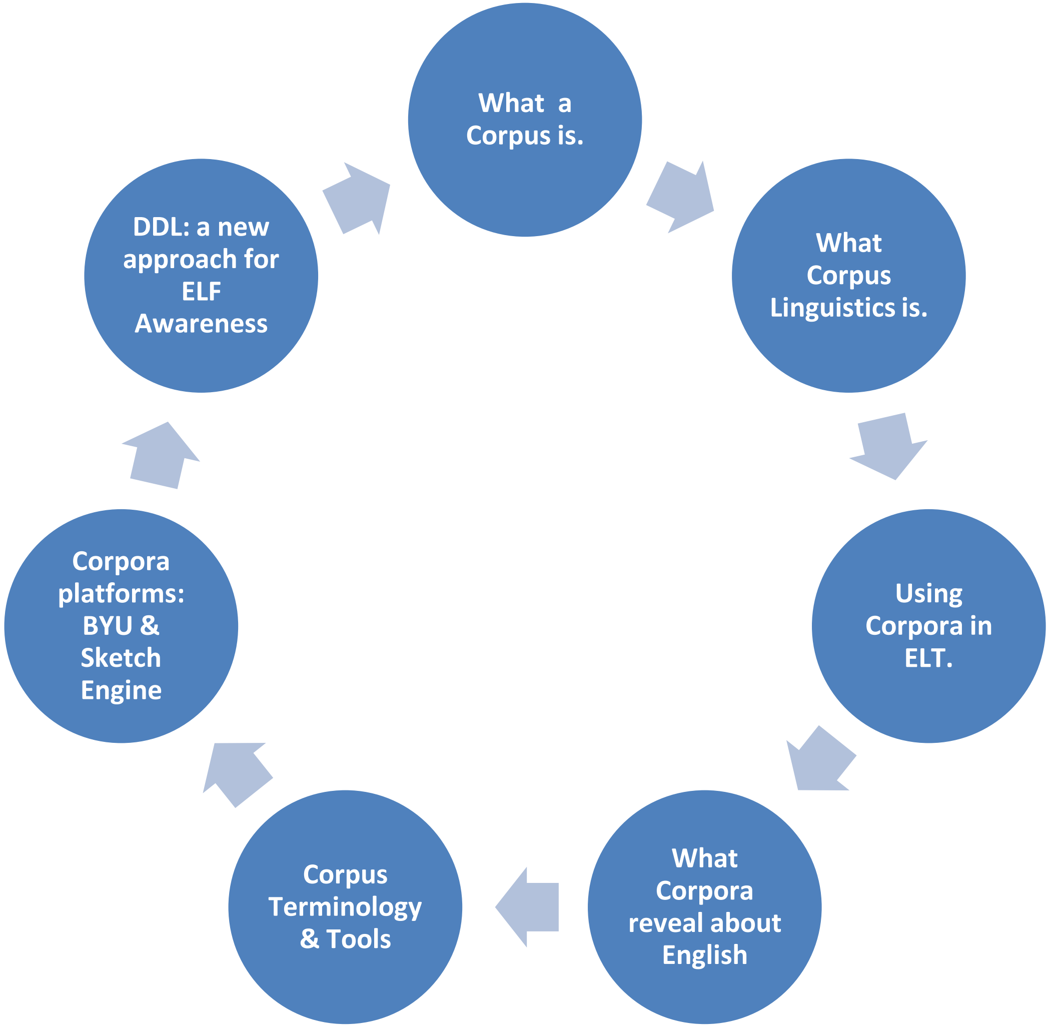
ACTIVITY 4

Corpora for language learning can be a very useful tool for exploring authentic uses of English

- How 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

Answer the above questions and report your thoughts in the **FORUM**

What have you learnt about Language Corpora?


What a Corpus is.
What Corpus Linguistics is.
Using Corpora in ELT.
What Corpora reveal about English
Corpus Terminology & Tools
Corpora platforms: BYU & Sketch Engine
DDL: a new approach for ELF Awareness

Erasmus+

ENRICH

English as a Lingua Franca practices for inclusive multilingual classrooms

ΙΔΡΥΜΑ ΚΡΑΤΙΚΩΝ ΥΠΟΤΡΟΦΙΩΝ IKY

The ENRICH Project is funded
with the support of the Erasmus+ programme of the European Union.

Grant Agreement: 2018-1-EL01-KA201-047894

The European Commission support for the production of this publication does not constitute an endorsement of the contents which reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.