SOME BENEFITS OF CORPORA AS A LANGUAGE LEARNING TOOL

By

TATJANA MARJANOVIĆ

University of Banja Luka.

ABSTRACT

What this paper is meant to do is share illustrations and insights into how English learners and teachers alike can benefit from using corpora in their work. Arguments are made for their multifaceted possibilities as grammatical, lexical and discourse pools suitable for discovering ways of the language, be they regularities or idiosyncrasies. The reader will be able to reflect on the great potential of electronic corpora in learning English, and draw on illustrations from a specific online venue where such explorations can take place at the user's convenience – the ideal user being preferably an advanced student taking on the role of a young researcher. Corpus-driven learning is seen as an inspirational, resourceful and intelligent way of exploring English as it accompanies and reinforces traditional styles of teaching and learning. When advanced students embark on this journey, which is both linguistically and cognitively challenging, they will encounter ample linguistic evidence and contextual information offering guidance and precision often greater than that found in textbooks. The application of corpora deserves no less than to stand side-by-side with other tried and tested methods of teaching and learning English at university level.

Keywords: Electronic Corpora, Discovery Learning, Autonomous Learners.

Introduction

What are corpora?

Corpora are the plural of the Latin word corpus, which means body in English. The word is frequently used to designate any representative collection of naturally occurring language specimens, i.e. utterances and texts, stored and accessed as an electronic database. The advent of corpus linguistics in the several past decades has not only revolution arised the study of language, but its effects have spread much further than specialist circles, with electronic corpora becoming more accessible to virtually anybody who takes an interest in language matters.

The aim of this paper is to give a brief introduction to corpora and point to their potential as a language learning resource. References to relevant research in the field will be accompanied by the author's own insights and illustrations assembled from the Corpus of Contemporary American English (COCA). This particular corpus, which contains 425 million words extracted from texts published between 1990 and 2011, is frequently singled out for its enormous size (e.g. Reppen 2010).

> A corpus essentially tells us what language is like, and the main argument in favour of using a corpus is that it is a more reliable guide to language use than native speaker intuition is. [...] For example, native speaker language teachers are often unable to say why a particular phrasing is to be preferred in a particular context to another, and the consequent rather lame rationale 'it just sounds better' is a source of irritation to learners. (Hunston 2002: 20)

If this holds true for native speaker language teachers, what can we say to allay the troubles of their non-native colleagues?

What experts say about corpora and how they can be used by learners and teachers

Corpora have often been noted for allowing incidental learning, which can be described in rather informal terms as stumbling upon a new piece of knowledge by looking at concordance lines (i.e. fragments of text presented in lines on the computer screen with a key word surrounded by more text to the right and left of it). The effect often quoted is that of serendipity, i.e. finding something good or useful by chance, or 'searching a corpus when the agenda is not too firmly fixed', as Hunston (2002: 171) puts it.

Tim Johns, the pioneer of the term 'Data-Driven Learning' (DDL), sees the learner as 'a language detective' and comes up with the slogan 'every student a Sherlock Holmes' (1997: 101). Johns first used DDL with international students at the University of Birmingham.

Bernardini uses the term 'discovery learning' and compares corpus-browsing with the exploration of 'an unknown land' (2002: 166), yielding yet another metaphor, 'the learner as traveller' (2004: 22).

According to Johansson (2010: 38), corpora 'encourage reflection', are suitable for 'the training of inferencing', and 'probably increase both the motivation of the student and the learning effect.'

Hunston (2002: 170) has also remarked that 'students are motivated to remember what they have worked to find out.'

Bernardini (2000) shows how learners can use corpora to investigate both form and meaning, noting that 'large corpora can be treated as pedagogic tools (rather than research tools) for engaging the learners' interests, developing autonomous learning strategies, raising their language consciousness, etc.' (2002: 166)

Moreover, Bernardini is confident that this approach can be beneficial not only to learners but also to their teachers, especially if non-native speakers of the language they teach (which is almost exclusively the case in my country), as it takes off some of the burden of (not) being the know-it-all person in the classroom. It means that teachers do not have to rely exclusively on their often limited intuitions when deciding whether something is acceptable or appropriate in a given context (2004: 28).

Cook (2003: 73) highlights the ever more important role of memory in language learning, which corpus linguistics helped to bring to our attention through its study of collocations (i.e. strings of words occurring together with statistically measurable probability). Collocations can add more strain to the learner's (and teacher's) memory, but perhaps corpora can help to ease some of it.

Greatly useful as corpora beyond any doubt are, no claims have been made that there are no limitations to their use in language learning. For one, we would be ill-advised to let the issues of frequency and conventionality curb creativity in language. Corpora are meant to enrich rather than impoverish language input, and should therefore not be approached in a blindly submissive or authoritarian manner. Ultimately, corpora can only do so much as they are designed to do, and are not to be seen as the only right answer to every single question about language.

It is for that reason that Johansson (2010: 42) sounds a note of caution, making it clear that corpora are no substitute for natural communicative settings and that they certainly cannot replace the teacher in the classroom.

Cook (2001: 65) also recommends that the use of computer corpora, impressive as they are, ought to be no more than a contribution to language teaching.

Finally, we shall agree with Hunston (2002) on two accounts. The first one would be that '[c]oncordance lines present information; they do not interpret it. Interpretation requires the insight and intuition of the observer.' (65) Another important point she makes is that it is generally agreed that this kind of learning is 'most suitable for very advanced learners who are filling in gaps in their knowledge rather than laying down the foundations.' (171)

Corpora and learners: a hands-on approach

Gavioli and Aston (2001: 245) are very pragmatic in preparing the ground for an effective application of corpora by learners: students need both free access and user-friendly software, both of which are nowadays available. Reppen (2010: 1) reassures us that existing corpora are not difficult to use at all, which we have already seen for ourselves.

As for the actual research (the word research is used here in a less scholarly and technical sense to describe any individual quest for information and evidence), it can take several forms.

We can choose to investigate two word-forms of the same lexeme, potentially resulting in different nuances of meaning, especially so when dictionaries may not be illustrative (or precise) enough to truly capture the difference. Hunston remarks that at times dictionaries 'can

Example 1: *trouble v. troubles*

up the butter and salt. My mouth had gone dry, and I had **trouble** getting any air into my lungs. I'd have thought having a guy I

be afraid I was going to get in trouble, I was already getting in **trouble**," Ayers said. "They knew it was going to get worse.

was always trying to find ways around the rules, and getting them both into **trouble**. # Like it or not, though, he was her brother -- like

commission just before graduation, and moved quickly from success to success. He had **trouble** finding work, and wound up temping

a deer to a distant check station. I understood why Finari went through the **trouble** of finding a few acres, but I couldn't figure out how

for Juliette's run from the altar had everything to do with Doug's recent **troubles** -- and his journalist father had taught him never to

As in other tales of nice kids gone wrong, the Gallery pear tree's **troubles** can be traced to a gang of new pals, a new genetic analysis

professors, for example, are likely to be disappointed, because the latest economic **troubles** have not only reduced university hiring

not predicting revolution by 1990. Indeed, the next decade will bring severe economic **troubles** and probably some frightening

All right, Chris, thanks. Well, despite his eccentricity, the legal **troubles**, the scandals, the celebrity excess, Michael Jackson's legacy

be of little help.' (2002: 45)

I certainly learn something new each time I browse the corpora. In this particular case it suddenly dawned on me (but not before I selected the Compare option on my computer screen) that the noun troubles is used to refer to a specialised problem area (e.g. economic/legal/marital troubles), and is often preceded by a possessive determiner (the world's/Asia's/Fred's/the family's troubles). The singular form trouble is much more straightforward, featuring repeatedly in much-used collocations involving a range of verbs such as have/get/find/keep/stay/go through/ask for etc.

It is a well-attested fact that not everything is accounted for in traditional grammars (Cook 2001: 64), so a query may be made to elicit grammatical and syntactic information.

The lemma [eye] presented in this way will run a search on all word-forms: nouns, singular and plural, verbs in all tenses and denominal participles. One of the interesting findings here involves eyed, which scored less than 2,000 occurrences compared to roughly 200,000 for mostly nominal eye and eyes. That the lemma [eye] shows a strong preference for nominal form in English is hardly surprising in itself. However, a huge majority of eyed were straight predicates (see the concordances in the example) rather than participles (e.g. bleary-eyed), which came as a bit of a surprise to me.

Example 2: *[eye]*

to drop over the ferryman's nether regions, steaming in her hands. She **eyed** Reilly very astonishedly indeed. "Dead?" she echoed.

grown men slowed behind him. From the front passenger side, a bony-faced man **eyed** Charlie, then took a drag on his cigarette, and

it necessary to pretend to have found Chip Gainsborough's body?" # Lyle **eyed** his fellow-parachutists, and then, in a throaty whisper,

Eleanor wrote her name. She handed the sheet across to the man. He **eyed** it in silence for a moment - like he'd expected more from

it's lovely to see the girls play together." # Sophie's mother **eyed** her own daughter for a long moment, then let her eyes rest on the

dead leaves on the beech hedge fluttered dryly in a bitter northeast wind. I **eyed** the heavy-laden clouds intently and shivered with

"Well, he's at a conference. Is it urgent?" She **eyed** me nervously. Couldn't these people see that I was just a normal honest

Leeches. Bleeding him ought to get rid of the fever." # She **eyed** the doctor's food-spotted shirt and tobacco-stained hands. "What do

or letter. No prior notice of his plans. Just my father – bleary-**eyed** and sleep-deprived – appearing with a duffel bag at our door. I

To having more children." # Fergie # The beach babe # Ultra-fit Black-**Eyed** Peas singer Fergie, 34, had "a blast" playing Saraghina, A caveat is in order here that it is only the initial 100 concordance lines targeting a particular word or structure that all the conclusions are drawn from. Nevertheless, I am merely trying to put myself in the shoes of an imaginary advanced student, and it is highly likely that s/he may grow too weary running extended searches, which are at this point basically unnecessary. The aim is not to produce a piece of academic writing in line with existing research procedures, but to encourage students to discover the pleasure of refining their knowledge of English in these individual and informal research-type tasks, and confirm or amend (why not even reject?) some of the hypotheses formed earlier through their primary sources, such as reference books or teachers.

Of course, it is always possible to google a word or a wider stretch of words and draw some rule-of-thumb conclusions, but a corpus experience is different as it may be more memorable. For instance, KWIC (i.e. key word in context) search results come complete with different colourings for the node word, which appears in the centre of the computer screen, and the surrounding words that precede and follow it (creating a one-colour-one-part-of-speech pattern). There is an abundance of relevant evidence in one place, so it saves some of the time and energy one would normally have to spend sieving through loads of tangential information on the Internet.

The COCA annotation system may present an additional challenge for the problem-solver, but also an obstacle for a student who is not prepared to spend extra time working out these symbolic representations. However, anyone can run basic word searches since these require the software knowledge of a below-average computer user. The rest might prove just about the right amount of challenge for the overachiever.

The annotation of the kind what [vv*] will yield what followed by any verb, but then the student can focus on any one of the combinations in the list, such as what makes. The example sentences are taken from one such search in which we further availed ourselves of the expanded context command to get complete sentences – just what we needed to work out the meaning potential of the structure under investigation. As you will have noticed by now, concordance lines are not confined to orthographical limits, i.e. do not necessarily end with a full stop or some other punctuation mark.

What can a learner find out from these expanded contexts? Firstly, wh-clefts occur across registers, classified in the corpus into five major categories: spoken, fiction, magazine, newspaper, and academic. Secondly, their communicative role in discourse (i.e. focusing on one constituent exclusively) becomes easier to acknowledge and consolidate. Thirdly, they can be embedded into yet another embedded clause, which can add to their prospective complexity. The list does not end here as an observant student may notice that the marked variety (i.e. with wh-clause in post-verbal position) is not so scarce after all, and that in that case the entire structure is frequently introduced by a pronominal that subject. Furthermore, this may lead to another, albeit tentative, conclusion that such structures are neatly woven into the preceding discourse, no doubt adding to its overall coherence. The main difference seems to be a greater degree of emphasis than

Example 3: *what [vv*]*

What makes it illegal is the currency.

What makes this unique is that there are no legal encumbrances on either party.

That's **what makes it so smart**.

We're riding on a thin film of air, which is what makes the **energy expenditure feasible**.

That's **what makes things sad at our home**.

You're training her that buying things is **what makes you valuable as a girl**.

Maybe **what makes life understandable** is not the events which happen but what happens between them.

That's **what makes you toy with the notion that hard work and personal responsibility - the bedrock virtues of our nation - are for**

Chumps.

What makes journalism so fascinating and biography so interesting is the struggle to answer that single question: 'What's he like?'

What makes the research at the MIT Joint Program unique is the systemic approach of combining economics, science, and policy to look at the probabilities, risks, and impacts of climate change.

that engrained in unmarked wh-clefts.

Clicking for more context may also help to better differentiate between wh-cleft and wh-nominal clauses. The latter represent cases of regular embedding, as demonstrated in the following sentence:

What we know about the brain comes from seeing what happens when it is damaged, or affected in unusual ways.

It is now easy to understand why corpora have proven a great tool for observing detail, whether we lifelong students view it as a welcome addition to our examination kit or just another grievance that we lose sleep over. Of course, as I am writing this paper, I have in mind an advanced student passionate about language matters. Indeed, I cannot imagine such a keen interest residing in students other than those that aspire to native-like English. Whilst some may nurture such aspirations, others may feel perfectly comfortable speaking and writing in an international variety of English.

Example 4: *[p*] so [happen]*

hadn't thought of that- oh." "What?" "It so happens I know just where we can find plenty of office workers. The refugee camps

the genetic play of variation and inheritance in the birthing game. It so happens I have thin feet. No other part of me is thin, but the

George asked me to protect this nephew of his during the war and it so happens the kid had five thousand dollars on him. The kid

nothing to you? No, apparently it does not. Well, it so happens that I think I know the person you mean. He is indeed very arrogant

a soup of 35-footers aboard a vessel only three times bigger? # It so happens they were right: In these post 9-11 times, ever bigger

'd heard, so she telephoned my home again and got me. It so happened David and I had stopped back by the house to look at a new

And although this new psychic element never shed light on my consciousness, it so happened that as soon as those chords broke I

you something to think about for the rest of your life: # It so happened that twenty-some-odd years ago a team of sadistic

to come walking along and will step me into the ground. But it so happened that a shepherd was keeping his flock on the field and

very unpleasant in the way he had spoken to her. # Now it so happened that the principal knew a great deal more about mathematics

An especially keen and aspiring student may wish to probe, among other things, whether a structure is a predominant feature of one register rather than another.

The corpora yielded 85 occurrences of it so happens, and 65 of it so happened, which suggests that the structure is relatively infrequent. However, it may be worth the special attention given to it on at least two accounts. Firstly, it confronts us with its often elusive meaning, which we can now agree is occurring by chance, unexpectedly, or accidentally. Secondly, it unequivocally reveals its modal allegiances by being used as a modal marker at the opening of the clause. Also, it becomes evident that the chunk it so happens is often (although not solely) used to refer to a present situation, while the chunk it so happened is always linked to the past.

Register-wise, 32 of all tokens uncovered in the search are representative of spoken English, 43 of fiction, 47 of journalism (covering both magazines and newspapers), and 28 of academic English. The results suggest no extreme disproportion in the usage of the structure across different registers, although fiction and journalistic prose hold a slight advantage over the other registers.

Example 5: *smoke and collocates*

a shock to some South Koreans as they watched television footage showing black plumes of smoke rising from the village on the

Hardly twenty covered and Castro needed a drink. He emitted a puff of black smoke from the bonnet, which had long forgotten how no

doubt sending out distress calls on the unit's behalf. Pillars of black smoke from mortar rounds rose up everywhere. With no

the street. As he approached Nox, he noticed the narrow black column of smoke that was sluggishly rising from the back of the

But great fires are now burning in the west, columns of dense black smoke rising high up in the morning air. When the breeze allows,

be more exposed to relevant toxicants such as lead paint, pesticides, and secondhand smoke. However, we found that the association

babies with smaller head sizes. Some researchers believe that traffic pollution acts like secondhand smoke or marijuana use,

and contributed to high levels of indoor fine particles. Reduction of exposure to secondhand smoke in public environments is

us what you see." I climbed into the rigging. Clouds of thick smoke covered the fort. A breeze came up. I briefly saw the flag.

The lights were on in several of the neighboring houses already and thick biting smoke rose through their chimneys and then fused

Other searches may be run probing which words a particular word tends to collocate with, or whether a particular word is affected by semantic prosody (i.e. whether it is interpreted in a positive or negative light).

After I had entered smoke in the Word(s) box and [j*] in the Collocates box, the search results revealed in order of frequency the most common adjectives that precede the noun smoke. Black, secondhand and thick topped the list with 577, 285 and 278 tokens, respectively. Of the three, I was least familiar with the adjective secondhand collocating with the noun smoke. To be completely honest, I found myself in disbelief at the collocation's high-profile status in today's English.

Example 6: *omen*

‘s dark lustrous eyes.... I took that as a good **omen**. 'Grape leaves stirred once more above the assembled family,

told the last cowboy. 'An eagle. A good **omen**.' # The old cowboy craned his neck skyward. Coyote

or at least a highly respectable **omen**. CNN's Sean Callebs live from the links with the latest

eyeing him boldly. Hardly a welcome **omen**. As he remounted the crows took off, rising like smoke

really, for the Republicans this may be a bad **omen**. I mean, if this kind of disparity lasts into the

to me the name was a poor jest and an ill **omen**. But Mai's wit was ever on the sharp side.

journey through these mountains of evil **omen**. My wife Helene, sitting beside Silviu and me in the

about half-undressed. I had a terrible **omen** the minute after entering Rosenzweig's world of

honestly absorbed in her. She would be an **omen**, of course! Part of the answer he was after.

Mo.' Leroy adds, 'Sometimes you get an **omen**, and you have to deal with it. 'Meanwhile,

Take the word omen, for example. One hundred concordance lines will dispel any doubts one might have whether the word is associated exclusively with definite signs of impending perish or whether it can be used to predict happier circumstances, too. The results show that it actually collocates with a range of premodifiers, some negative (e.g. *bad, terrible, ill, evil*), some positive (e.g. *good, welcome, highly respectable*), and some standing alone requiring expanded context to guide the reader in either direction.

Conclussion

Hunston (2002: 214-216) concludes her deliberations as to whether corpora have made our lives simpler or more complex by sharing a personal story involving one of her students who wrote in an essay that someone was 'under the influence of Halliday'. The wording struck Hunston as comically odd so she corrected it to 'was influenced by Halliday', convinced that the phrase 'under the influence of' was used freely with inanimate nouns in rather negative contexts (e.g. drugs and alcohol). Nevertheless, she did a little corpus research and was surprised to learn that she was only partly right. The corpora clearly allowed the possibility of the phrase 'under the influence of somebody'. Intrigued, she dug a little deeper, and was able to detect a difference in the verb that seems to be the answer to this dilemma: the phrase 'be under the influence of somebody' has an altogether negative connotation, whereas the phrase 'come under the influence of somebody' can be used in both positive and negative contexts.

The moral of this story, as the author's see it, is that it takes a great deal of perceptive and analytical skills do refine the meaning of a single phrase. It is no small task, and if it took a native speaker, who also happens to be an experienced and highly accomplished linguist, some time and effort to unravel the mystery, imagine how much more demanding it would be for someone who is both a non-expert and a non-native speaker.

However, the author's would like to end this episode on a brighter note by saying that not every corpus search has to lead to a major discovery. It could be something fairly simple but new for the student: one more piece of knowledge – however small and insignificant it may seem – makes all the difference to the one who has just gained it.

Recommendations

The study is aimed at advanced students of English and their teachers, both of whom can utilize it to reflect on the potential of electronic corpora in autonomous learning. Teachers are encouraged to offer assistance and guidance to students in their exploration of the language, whether they purposefully seek to test their hypotheses or merely set out to browse corpora hoping to make some unexpected discoveries. Students will find this kind of informal research rewarding as they will be able to pursue their own linguistic interests with the extra flexibility missing from specifically designed tasks. Students will get a new motivational boost from working at their own pace and seeing what results each individual search will yield. Both students and teachers can draw on the examples in the study for ideas on how to approach electronic corpora and meet their specific needs and interests. Last but not least, the study reassures all parties concerned that electronic corpora are relatively easy to operate, illustrating some of the procedural steps that need to be followed and symbolic representations that accompany them in setting the right parameters for a particular corpus search.

References

[1]. **Bernardini, S. (2000).** 'Systematising serendipity: proposals for concordancing large corpora with language learners.' In Lou Burnard and Tony McEnery (eds.), *Rethinking Language Pedagogy from a Corpus Perspective.* Frankfurt: Peter Lang. 225-234.

[2]. **Bernardini, S. (2002).** 'Exploring new directions for discovery learning.' In Bernhard Kettemann and Georg Marko (eds.), Teaching and learning by doing corpus analysis: *Proceedings of the Fourth International Conference on Teaching and Language Corpora,* Graz 19–24 July, 2000. Language and Computers: Studies in Practical Linguistics 42. Amsterdam and New York: Rodopi. 165-182.

[3]. **Bernardini, S. (2004).** 'Corpora in the classroom: an overview and some reflections on future developments.' In John Sinclair (ed.), *How to use corpora in language teaching.* Amsterdam: John Benjamins. 15-36.

[4]. **Cook, G. (2001).** 'The uses of computerized language corpora: A reply to Ronald Carter.' In David R. Hall and Ann Hewings (eds.), *Innovation in English Language Teaching: A Reader.* London and New York: Routledge. 64-70.

[5]. **Cook, G. (2003).** *Applied Linguistics.* Oxford: Oxford University Press.

[6]. **Gavioli, L., and Aston, G. (2001).** 'Enriching reality: language corpora in language pedagogy.' *ELT Journal* 55(3). Oxford: Oxford University Press. 238-246.

[7]. **Hunston, S. (2002).** *Corpora in Applied Linguistics.* Cambridge: Cambridge University Press.

[8]. **Johansson, S. (2010).** 'Some thoughts on corpora and second-language acquisition.' In Karin Aijmer (ed.), *Corpora and Language Teaching*. Studies in Corpus Linguistics 33. Amsterdam: John Benjamins. 33-44.

[9]. **Johns, T. (1997).** 'Contexts: the background, development and trialling of a concordance-based CALL program.' In Anne Wichmann, Steven Fligelstone, Tony McEnery and Gerry Knowles (eds.), *Teaching and Language Corpora.* Harlow: Addison Wesley Longman. 100-115.

[10]. **Oxford Advanced Learner's Dictionary of Current English. (1995).** Oxford: Oxford University Press.

[11]. **Reppen, R. (2010).** *Using Corpora in the Language Classroom.* Cambridge: Cambridge University Press.

[12]. http://dictionary.reference.com/

[13]. Http://corpus.byu.edu/coca/

ABOUT THE AUTHOR

Tatjana Marjanović is currently working as an Assistant Professor in the English Department at the University of Banja Luka, Bosnia and Herzegovina. She read for her MPhil at the Research Centre for English and Applied Linguistics, University of Cambridge, UK. Her doctoral thesis was an attempt to shed light on a correlation between thematic structure and coherence in English news texts. One of her recent articles was published in Changing English: Studies in Culture and Education by Routledge. She takes a close interest in functional approaches to grammar and discourse analysis.