SURFING AN ENGLISH CORPUS WITH OTHER LEARNERS OF ENGLISH

Nany Setyono Kurnia
Atma Jaya Catholic University of Indonesia
nany.kurnia@atmajaya.ac.id

ABSTRACT

The paper describes some experiences of surfing an English corpus, the Corpus of Contemporary American English (COCA for short) with other learners of English. The challenges expressed by other learners inform us of the problems other users may have that we may not be aware of. Several examples of word-form searches, lemma searches, or part-of speech (POS) searches in both word/phrase queries and collocate queries are shown and explained. We can also limit our word/lemma searches to certain parts-of-speech, although the part-of-speech tagging is not always accurate. On top of that, there are wild-card symbols that we can use in our queries.

The word/phrase query window in COCA enables us to search for just a word-form or a lemma, or a succession of word-forms, lemmas, and parts-of-speech as well as the wildcards, within limits. When we make a collocate query, we look for the co-occurrences of one item (a word, a lemma, a POS or a wildcard) that we put in the collocate window, and the item[s] we put in the word/phrase window. We also specify the distance of the 'collocate' and the word/phrase and the direction, to the right or to the left, or both to the right and left. Unless we change the default, it will search for the 'collocate' within the space of four words to the left and four words to the right of what we put in the word/phrase window.

Searching a corpus together can reveal the misperceptions and difficulties that some learners may have, which if properly handled, can lead us to valuable insight as we learn from each other. We learn from other learners, for instance, that some of them are not aware that even though the frequency result of a collocate query shows just a word or the word-forms of a lemma, it is actually capturing the co-occurrences of that one word or that lemma with the item in the word/phrase window that we can see in the concordance lines. It is therefore advisable to give a reminder to always check the concordance lines. By checking the concordance lines we can judge whether or not the co-occurrences are indeed a case of collocation or just co-occurrences within the specified distance where actually the so-called 'collocate' does not relate to the item in the word/phrase window. While scrutinizing the concordance lines we can also check whether certain expressions are used in the meanings we have in mind. Whereas the ability, speed and accuracy in making these judgments, as can be expected, vary among corpus users, it is worth noting that the lemma-word form concept, POS codes, even some POS themselves are still a challenge for some learners. Examining the concordance lines can also reveal how speakers use the same expression in various ways.

The paper ends with a discussion of the potential usefulness of the concordance lines as raw material that can be selected and edited to give learners additional exposure to what is obtained from other means of learning, hopefully leading to better language acquisition.

Keywords: *COCA, word-form, lemma and POS search, word/phrase and collocate query, challenges, corpus data potentials*

Doing corpus search can be fun as we look for ways to best query the corpus, making the most of the available features of the query syntax, to arrive at the target data and then often gain an extra bonus or two. But doing corpus search with other corpus users can benefit ourselves even more. This paper presents some experiences of a learner of English (someone whose English is far from perfect, me), in finding out more about the English language by using the *Corpus of Contemporary American English* (Davies, 2008-), or COCA for short, with other learners of English. We will look at word-form search, lemma search, and part-of speech search in both collocate and non-collocate queries, with or without using the provided wildcards, in trying to aim better at the target result, keeping in mind the possible alternative or variant written forms of words and multiword expressions.

This paper adopts the following scheme: each section presents a language item or a kind of query as shown in the title of the section and discusses the search, the result, the challenges, if any, and the lesson learned. In other words, we just plunge into the use of a certain query in COCA to find evidence of a certain language item, from one query to another in a one-thing-leads-to-another manner, each time taking a pause to reflect on the experience.

The plural *condolences*

In a game of spotting natural/common multiword expressions used in a text, some learners tried their hands on *sincere condolences* in the sentence *"Again, my sincere condolences, but there's nothing I can do"*. Their query and the result can be seen below:

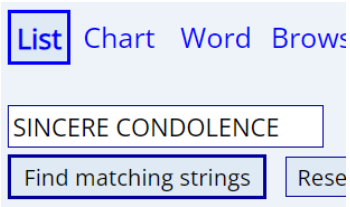


Figure 1. Query 1

ALL FORMS 94	FREQ
SINCERE CONDOLENCES	72
SINCEREST CONDOLENCES	21
SINCERE CONDOLENCE	1
TOTAL	94

Table 1. Result of Query 1

They had learned that we can search for lemmas instead of word-forms and therefore used the capital letters for both *sincere* and *condolence* in Query 1 above. Lemma search like this will include searches for all inflected forms of the lemmas, not just the word-forms *sincere* and *condolence* but also *sincerer*, *sincerest*, and *condolences* appearing in the sequence specified in the query. Result of Query 1 above tells us that there are three strings occurring in the corpus (COCA) that match the query: *sincere condolences*, *sincerest condolences*, and *sincere condolence*. Below are some of the concordance lines:

on its Facebook page. " A true champion of the game. Our **sincere condolences** to the Dombroski family and
Hughes' death is heartbreaking, " Carter said. " I extend my **sincere condolences** to the Hughes family, to ev
me now from Nashville. Thank you for joining us, and our very **sincere condolences** for your loss. MEGAN-B
always remember him -- Brit. Thank you for doing this today. Our **sincere condolences** to his wife Beth, his s
at least six months after her death. # " We express our most **sincere condolences** and deepest sympathy to

' I would like to offer my **sincerest condolences** to Mr. Guerrero's surviving family and friends on the loss
ral agency issued a statement offering its " **sincerest condolences** " to the women's families and saying cri
aunce, and two young children. # " Our **sincerest condolences** go out to Sgt. Brown's family during this dif
heir work in the community. We extend our **sincerest condolences** to S.J.'s family and friends during this

summit meeting between the two Koreas. Clinton expressed **sincere condolence** for Kim's death

Concordance 1. *Sincere(st) condolence(s)*

Had the learners typed in the search window (Figure 1 above) exactly the word-forms from the 'game' text in lowercase, *sincere condolences*, the result would have shown *sincere condolences* only, and the learners would be left none the wiser of the presence of *sincerest condolences* and *sincere condolence* in the corpus.

As it happened, apart from confirming that *sincere condolences* in the 'game' text as well as *sincerest condolences* are likely to be common multiword expressions or, in other words, that the adjective *sincere(st)* collocates with the noun *condolences*, the learners also felt they gained additional information, namely that when expressing our condolences we should use the plural form *condolences*.

The singular/non-count *condolence*

The learners' tentative conclusion above, right or wrong, can inspire another line of inquiry about the singular *condolence*. While *sincere condolence* that occurs once in the one-billion-word COCA maybe a typo, a less natural utterance, or simply a less frequent alternate form, is the singular *condolence* rarely used? If we search for the lemma *condolence*, the singular form is indeed occurring much less frequently than the plural form, but we still can see 416 uses of the singular *condolence* in this large corpus.

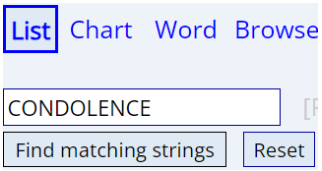
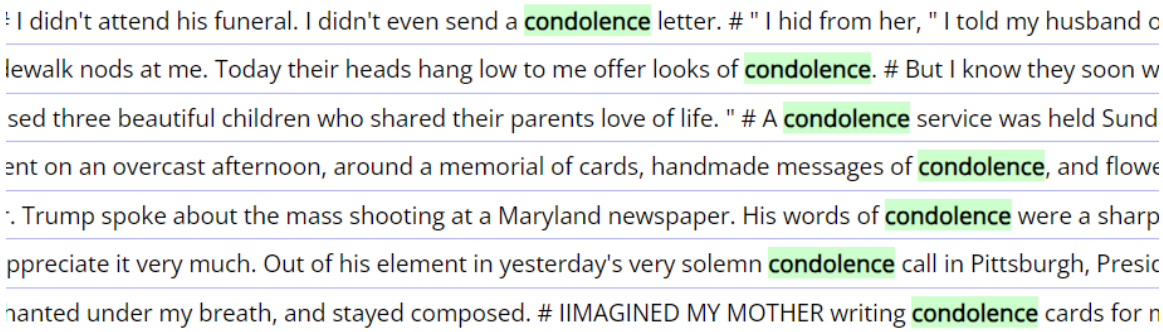


Table 2. Query 2

ALL FORMS	FREQ
CONDOLENCES	2074
CONDOLENCE	416
TOTAL	2490

Table 2. Result of Query 2

A cursory glance at its uses in context (see Concordance 2 below) tells us that the singular *condolence* is often used as modifier: *a condolence letter*, *condolence cards*, *a condolence service*, *words of condolence*, *messages of condolence*, etc.



Concordance 2. *Condolence* as modifier

What started out as an attempt to check the commonness of *sincere condolences* has led to finding out about the frequency of the singular *condolence* compared to the plural one, which in turn led to the use of the singular *condolence* as a modifier.
Another train of thought that we may pursue is whether there are other co-occurrences of the lemma *sincere* and *condolence* in the corpus where those two lemmas collocate even though they are not adjacent to one another.

The *sincere – condolence* collocate query

To find occurrences of *sincere - condolence* collocation, we can do Query 3 below that will capture all forms of the lemma *condolence* occurring within four words to the right of, so not only immediately after, the lemma *sincere*. In other words, unlike Query 1, this query also includes finding instances where the lemma *condolence* occurs not immediately after the lemma *sincere*.

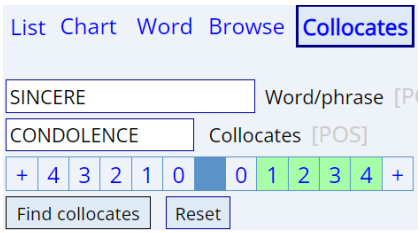


Figure 3. Query 3

	FREQ
CONDOLENCES	97
CONDOLENCE	1
TOTAL	98

Table 3. Result of Query 3

This collocate query resulted in four more instances than the phrase search in Query 1 (a total of 98 in Table 3 vs. a total of 94 in Table 1), and those four additional instances are:

a. where there are intervening words between *sincere* and *condolences*, with *sincere* undeniably a modifier of *condolences*, and therefore definitely our target items*:*

The Japanese prime minister offered his, quote, " **Sincere** and everlasting **condolences**. "

My most **sincere** and personal heartfelt **condolences** go out to the families and loved ones

b. where there is a word or words between *sincere* and *condolences*, and *sincere* is likely but not guaranteed to be a modifier of *condolences* too:
People offer their **sincere** well-wishes, **condolences** and prayers to the victim's families,

c. where *sincere* does not relate to *condolences*, despite their co-occurrence within the specified span, hence not our target item, or what is commonly referred to as 'noise'[1]:
My **sincere** sympathy and my **condolences.**

The query tool, however, also allows us to search for the lemma *condolence* that occurs to the left of the lemma *sincere*, such as in Query 4 below,

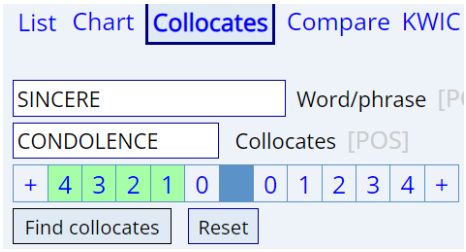


Figure 4. Query 4

which returns just one co-occurrence of *condolences* and *sincere*:

I'm sure you **condolences** are appreciated as **sincere**.

As *sincere* here does relate to *condolences*, this co-occurrence is not classified as 'noise'. The only note we can make here is that *you* is likely to be a typo, possibly meant to be *your*.

We have seen the advantage of using the collocate query for capturing expressions containing more than one word where the words are not always adjacent to one another. In the following section we will look at a query that searches for collocates both to the left and to the right of the word/phrase in the words/phrases window in one go. The default is four words to the left up to four words to the right, but we can go as far as nine words to the right or/and left.

The elusive *sentiment*

As a language learner, when we speak or write, we sometimes find ourselves grappling for words that feel to be on the tip of our tongue, but keep eluding us. Visiting (collocation) dictionaries can be a quick solution, but this is not always the case. Suppose we want to express that we strongly and wholeheartedly agree with an opinion or a stance, and feel that the expression "*second the…*(something)" will be right on the mark, but as this expression apparently has not been firmly instilled in our productive vocabulary, we cannot nail down what this 'something' is. Corpus data drawn from a collocate query can be of help:

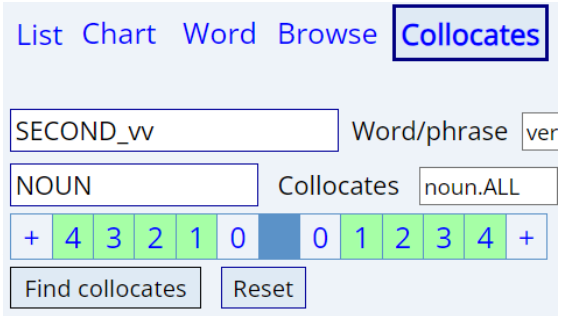


Figure 5. Query 5

COLLOCATES SECOND VERB See also

+ NOUN		NEW WORD	?
90	7.55	motion	
30	4.04	comment	
29	6.95	nomination	
23	6.12	recommendation	
22	2.43	idea	
22	5.60	emotion	
20	3.52	support	
19	5.27	notion	
15	6.38	sentiment	
12	5.01	suggestion	

Table 4. Some results of Query 5

[1] "unwanted hits" or "incorrect hits" (Rühlemann 2019, p. 8 and p. 101)

Query 5 looks for nouns that occur within the span of four words to the left to four words to the right of the verb lemma *second*, as theoretically it is possible to have constructions where the noun that is the object a verb *second* appears both before or after the verb.

Looking at the result of Query 5, partly shown in Table 4, we spot the noun *sentiment*, Bingo! This is exactly what we have been wanting to say. Even though the columns in Table 4 above do not come with headings, we can figure out ourselves that the first column is the frequency of the co-occurrences or the number of instances where the noun co-occurs with verb lemma *second*, the second column is the Mutual Information Score, the fourth column is the noun, and we can click the far-right column to see the co-occurrences of the noun and the verb *second*. The table informs us that there are fifteen instances of *sentiment* occurring within the specified span or distance from the verb lemma *second* (specified in the query, that is) and the Mutual Information Score (MI) of 6.38 suggests that *sentiment* is likely to be a strong collocate of the verb lemma *second*, as MI >3 is considered an indicator of strong collocation. Below are the fifteen co-occurrences:

strictly opposed to it. " # A Mason man, Mark Pine, **seconded the sentiment** that the project would harm the community's
ould he lose, Carolla says he might shut down his show, a **sentiment seconded online by** others. Podcasting has been arou
ow Testing and Choice are Undermining Education, " 2010). # Nussbaum **seconds this sentiment** (and anticipates Arum an
of discussion on exactly how to phrase things. # I just want to **second this sentiment**. I served on a jury once, and we paid
ıo have ultimately squeezed into this office are finding the conversation therapeutic. # Clark **seconds the sentiment** about
ıdying Gale Crater's terrain. Pete Theisinger, the mission's project manager, **seconded the sentiment** for going slow: " We h
eiterated Wednesday he plans to do that without a tax increase, a **sentiment seconded a few** hours later by Republicans on
NOT true? Losing battle, my friend.... losing battle. # I **second this sentiment**, everyone (and I mean everyone!) knew Rmone
Barack Obama, for all his faults, has done. " # I **second the sentiment**, but please keep quiet about our Kenyan president.
ot) under discussion in this bloody awful bit of self-indulgent bullshit. I **second the sentiment** above. I wish NDPR would just
n't be reached. " # Joy Zinoman, Studio's artistic director, **seconds these sentiments**. She was so taken with McCraney's wo
William Barlow notes: pre-formatted table In her autobiography, gospel great Mahalia Jackson **seconds this sentiment**, like
r friend and fellow teacher, Linda LaRue, 54 and also of Newark, **seconded those sentiments** and added: " His passion spe
dad can do for them or give to them. # That **sentiment** is **seconded by J.D.** Roth, who runs the website Get Rich Slowly. " I g
e way of keeping this team together, " the source said, a **sentiment seconded Friday by** Gibbs. # " I really do think that this

Concordance 3. Verb lemma *second* + noun *sentiment* co-occurrences

As we can see in Concordance 3, *sentiment* occurs both to the right and to the left of the verb lemma *second*, justifying the use of both left and right span in Query 5. In four out of these fifteen co-occurrences, the noun *sentiment* occurs before or to the left of the verb lemma *second*: one is a passive construction *That sentiment is seconded by J.D. Roth* (the line before the last) and the other three show the noun *sentiment* followed by a reduced relative clause: *a sentiment seconded online by others* (second line from the top), *a sentiment seconded a few hours later by Republicans* (seventh line from the top) and *a sentiment seconded Friday by Gibbs* (the last line). The remaining lines are active constructions where the verb lemma *second* is followed by the noun lemma *sentiment* as the verb's direct object.

All had seemed spick and span regarding the advantages of the collocate query until a fellow corpus user expressed reluctance to use the collocate query because it resulted in just one word. To be fair, the collocate query here can only be for one item, as we can see in the collocates window of Query 3 and Query 4, the lemma *condolence*, and the part-of-speech NOUN in Query 5. The frequency results too are indeed in the form of one word. The one-word results, however, convey that those one-words co-occur with the expression we put in the word/phrase window above the collocates window and we can see their co-occurrences in the concordance lines (Concordance 3 and the sample lines in the previous section). What is extracted from the corpus are the co-occurrences, not the occurrences of only the one word shown in the frequency result.

The importance of checking the concordances cannot be understated. We have seen how from examining the concordance lines, we found, for example, the modifier use of the singular *condolence* such as in *condolence letter/service*, and the co-occurrence of *sincere* and *condolences* that is a 'noise' as they do not collocate (*My sincere sympathy and my condolences*). As Susan Hunston says (2022a: 47):

> One of the most basic techniques in corpus linguistics is the scrutiny of sets of concordance lines. Although most corpus studies start from a quantitative perspective – obtaining a frequency list of

words, for example – they often end with the qualitative investigation of individual words or phrases using concordance lines.

This experience with the reluctance to use the collocate query shows that having fellow corpus users air or share their concerns, perceptions, impressions, or ideas about certain aspects of the corpus query can lead to clarifications that are potentially beneficial for all users. It is therefore important that while practicing doing the searches corpus users are encouraged to express their views on the task, so these views can be discussed such as in this case what the collocate query and its result actually is.

If at all and the amazing wild card (*)

Sometimes language learners are not struggling with a forgotten part of an expression they want to use, such as in the case of *second – sentiment* collocation detailed in the previous section, but rather with an insecurity feeling about using a certain expression. If we find occurrences of the expression being used with the same meaning and in seemingly similar situations in a corpus of good speakers of the language, we can feel more confident. There is some sort of "safety in numbers" (Mair 2002: 122), when we find the expressions occurring frequently in such a way in a corpus. Take, for example, we want to say/write *It seldom happens, if at all*. but we are not sure whether such use of *if at all* is alright. We search in COCA for *if at all* and find 1954 occurrences, a few of them are:

ǂ He wasn't around much, **if at all**, when I was growing up. I had to deal with teasing
oads. They then trot out poorly analyzed (**if at all**) statistics (with arbitrary start dates) to show th
ing. But this will likely affect, **if at all** within the relevant period, only the wealthiest societies. Sim
use but only a subset of use, **if at all**. # For example, monitoring can be optional for voting syste
estigations and prosecutions, aren't pursuing P2P downloaders particularly vigorously, **if at all** (
, which are but remotely - **if at all** - related to the field of software research and development, ar
es. What does it mean and how (**if at all**) is it going to affect me personally? # Sincerely, Uninsure
State rallies more rioters. # Avoid likely neighborhoods **if at all** possible a few days post election

Concordance 4. *If at all*

We may be curious about how often *if at all* is used at the end of a sentence or a clause, compared to other uses, we can check what follows *if at all* by using the wild card symbol * after *if at all* and a space:

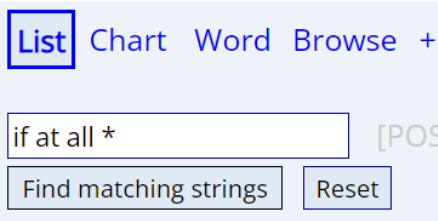


Figure 6. Query 6

ALL FORMS	FREQ
IF AT ALL ,	588
IF AT ALL .	581
IF AT ALL POSSIBLE	410
IF AT ALL)	108
IF AT ALL ?	66

Table 5. Top Five of Query 6 Result

We can see in Table 5 above the five top (most frequent) strings consisting of *if at all* followed by something, and at least three out of these five top ones are *if at all* used at the end of a sentence or a clause (followed immediately by a full-stop, a comma, and a question mark). This kind of data gives us learners confidence in using the expression at the end of a sentence or clause.

For learners who are yet to familiarize themselves with the expression *if at all*, concordance lines such as those shown in Concordance 4 above and Concordance 5 below can expose them to natural uses of the expression.

flip a switch and make it alright. This will take years to overcome, **if at all .** There is real serious talk
2005, and, I wonder how the current economy has impacted the plan, **if at all .** Also, the plan was c
the Jewsmedia. That's why you won't often hear such positive songs, **if at all .** They promote Satani
all the options out there, people don't have to commit as quickly, **if at all .** Many men and women a
come to realize that the cops probably won't show up for an hour, **if at all .** And, I don't blame the c

Concordance 5. *If at all* followed by a full-stop.

In the last section, the 'Epilogue', we will return to the discussion of repeated exposure and the potential usefulness of concordances extracted from a corpus.

***Us learners*: searching for apposition**

The string *us learners* appears in the previous section: "This kind of data gives **us learners** confidence in…". If we are interested in finding uses of apposition other than *learners* following *us*, we can try, for instance, using the query *us* followed by a Part-of-Speech, namely a plural noun, shown in Figure 7 below:

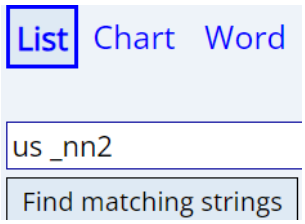


Figure 7. Query 7

	ALL FORMS	FREQ		ALL FORMS	FREQ
1	US TROOPS	1038	9	US HUMANS	280
2	US CITIZENS	999	10	US GIRLS	276
3	US AIRWAYS	914	11	US COMPANIES	261
4	US FORCES	872	12	US CITIES	233
5	US OFFICIALS	824	13	US MARINES	230
6	US SOLDIERS	448	14	US THINGS	219
7	US KIDS	377	15	US CHILDREN	216
8	US WOMEN	329	16	US GUYS	214

Table 6. Top 16 of Query 7 result

The top (most frequent) 16 strings listed as the result of Query 7 can be seen in Table 6 above. Even a brief glimpse is enough to suspect that many of the strings in Table 6 above are not our target string, the *us* likely to be an abbreviation of 'United States', *US troops*, US *forces, US airways, US cities* and so on:

Do you support this decision by the President to withdraw **US troops**? SENATOR-JEFF-MERKLEY: So i
and it just might suit President Trump's agenda of getting **US troops** out of the peninsula where the
ogations we came to learn that he had been " embedded with the **US troops** serving in Afghanistan.

, everyone, survived. The first pictures of **US Airways** Flight 1549 reveal... You think that you're g
human performance investigation on the crash of **US Airways** Flight 1549. - Water landing. - Ca

Israel, during secret briefings, as a potential threat to **US forces**. # Netanyahu has been further e
homes and congregating in the centre of the city as the **US forces** advanced from all sides. They h

While taxi services in major **US cities** are usually reliable and efficient, taxi services in many foreigr
research mission. In older Eastern **US cities**, nine times as much natural gas is leaking out of pipeli

Concordance 6. *Us* an abbreviation of United States

To aim better at our target string, we can stipulate in our search that the word *us* we are looking for is a personal pronoun:

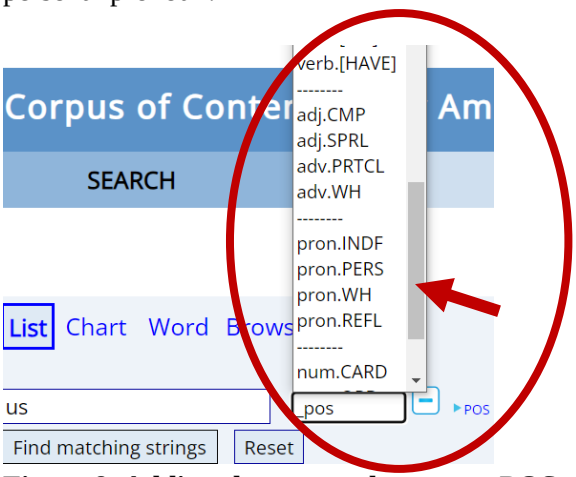


Figure 8. Adding the personal pronoun POS to the word-form *us*

Figure 9 below shows the query, aiming at strings consisting of the personal pronoun *us* followed immediately by a plural noun, and Table 7 presents the top (most frequent) sixteen strings:

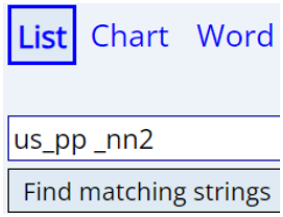


Figure 9. Query 8

_	ALL FORMS	FREQ	_	ALL FORMS	FREQ
1	US KIDS	360	9	US CITIZENS	125
2	US HUMANS	277	10	US STORIES	122
3	US GIRLS	267	11	US SINNERS	119
4	US THINGS	216	12	US CHILDREN	116
5	US GUYS	209	13	US TICKETS	106
6	US WOMEN	194	14	US BROTHERS	103
7	US AMERICANS	189	15	US NAMES	80
8	US MEN	154	16	US QUESTIONS	80

Table 7. Top 16 of Query 8 result

Without going into a thorough analysis of these sixteen strings in Table 7, we can roughly group them into three categories, strings that are likely to be 'noise'(non-target strings), strings which are likely to be our target strings (*us* followed by an appositive of *us*), and two strings with mixed results, some 'noise' and some target items:

1. Strings numbers 4 *us things*, 10 *us stories*, 13 *us tickets*, 15 *us names* and 16 *us questions* are likely to be 'noise' because the plural nouns are not appositives describing the preceding personal pronoun *us:*

the bags? You guys win a contest? No, people just like giving **us things**. Have you talked to MJ? S

. Sometimes she sends more letters and some money, and Papa can now buy **us things**. Somet

She did? She was always telling **us stories** of when you first moved in. How excited you were.

irits can get. So our aunt told **us stories** and then her daughters actually inherited the place. I

And that's just the start. I also got **us tickets** to the 76ers game! No way! Hey, feels good to s

the trip yet. My husband bought the four of **us tickets** to LA. He booked the trip on a whim

spending thousands, thousands of hours with him, he would always ask **us questions**, like,

They started cursing at us, calling **us names**, telling us that we're not from here.

Concordance 7. Plural nouns no apposition for *us* – 'noise'

2. The rest of the strings, except *us citizens* and *us brothers* (point 3 below), seem to mostly represent the target item:

She was kind of lonesome there, you know, just **us kids** and her. So when anybody passed by

by robots, then in order for **us humans** to stay on top of the food chain, we need to develop skill

The detention officer sat behind his desk. He was not watching **us girls**. He was reading a newsp

our female counterparts. Of course, most of **us guys** don't want to be quite as sociable as the gals.

sad. **Us women** aren't taken seriously when we express how uncomfortable we feel when it comes

t is a commonplace among **us Americans** that we live a fast-paced life. We're always busy. It is

EORY # It is a commonplace among **us Americans** that we live a fast-paced life. We're always busy.

it of times gone by. How detached **us men** of culture have become from our ancestors' knowledge.

spread a lie. Is that why you did it then? To teach **us sinners a lesson**? If I was a more pious Quak

I don't understand why they are being so mean to **us children**. Don't they know how much we love

Concordance 8. Plural nouns appositive of *us* – target items

.

3. The string *us citizens* consists of both target items, *us* being a pronoun and *citizens* an appositive of *us*, and non-target items ('noise') where the *us* is apparently mistakenly tagged as personal pronouns while actually is the abbreviation *US* for *United States*. The two lines below show, respectively, the non-target string and the target string:

We have no problems with Mexicans or Hispanics, as long as they become US citizens and pay th

the insurance companies, the credit card companies and will lighten the load for us citizens that

Concordance 9. Non-target *US citizens* string and target *us citizens* string

The string *us brothers* appears in two different constructions, one construction where *brothers* is an appositive of *us* (target string), the other construction where *brothers* is a complement of *us* ('noise'):

We turned in early. It was us brothers to the tent and my parents to the van. Franny, becau

ns, my brothers? How can you to call us brothers and... to refuse to help us? At least, don't

Concordance 10. Target *us brothers* string and non-target *us brothers* string

This experience with apposition search shows that what looks like a simple query may end up in less simple data. In addition, the use of more specific parts-of-speech in the query prompted some learners to remark that they were not familiar with the Part-of-Speech (POS) tags or codes in the drop-down menu such as shown in Figure 8 above.

It is recommendable therefore to urge users surfing corpora to familiarize themselves with the part-of-speech tags in use, perhaps even with English Part-of-Speech itself in the first place.

Name calling: the lure of pattern

The last line of non-appositive examples in Concordance 7 shows a case where the string *us names* co-occurs with the verb *calling*:

They started cursing at us, calling **us names**, telling us that we're not from here.

What immediately strikes us when we look at the concordance lines for the string *us names* is its frequent co-occurrences with the verb lemma *call*:

lunatics smoking marijuana in their apartments who treat us mean, calling **us names**. # My mother has asth
a good way! Okay, why do you keep calling **us names**? I am just busting balls. That's what we do in the
robably they might do the same thing if there was an ad that called **us names** and our faces are on it, that th
ng behind him on the motorcycle. " So far they're only calling **us names**, but once their bomb- tossing buddie
don't I remember your names? - You didn't give **us names**. - You gave us feelings. - I wish I had a different
vded around our news car. They started cursing at us, calling **us names**, telling us that we're not from here. \
us hints about what happens in the film, it does give **us names** -- confirming that Mikkelsen's character is ca
ospectors, lumberjacks, fishermen, homesteaders, and others who bequeathed to **us names** reflecting the la
debate about that. That's far better than their calling **us names** and our calling them names. So I hope it's th
ative values can do and it scared the liberals so much that they called **us names** and demonized us at every

Concordance 11. Frequent 'verb lemma _call_ & _us names_' co-occurrences

This is hardly surprising given the fact that *call somebody names* is an English expression, as attested in these dictionary entries:

> **call somebody names**
> to use offensive words about somebody
> *Stop calling me names!*
> (https://www.oxfordlearnersdictionaries.com/definition/english/call_1#call_idmg_6)
> **call somebody names**
> to use unpleasant words to describe someone in order to insult or upset them
> *The other kids used to call me names.*
> (https://www.ldoceonline.com/dictionary/call-somebody-names)

So what started as checking the concordance of the string *us names* has led us to see a pattern where the verb lemma *call* co-occurs with noun *names*. If we shift our focus to searching the co-occurrences of the verb lemma *call* with the word *names* we can find evidence in COCA of the occurrences of the expression described in the dictionaries above. Table 8 below shows that there are 2427 *names* occurring within four words to the right of the verb lemma *call*.

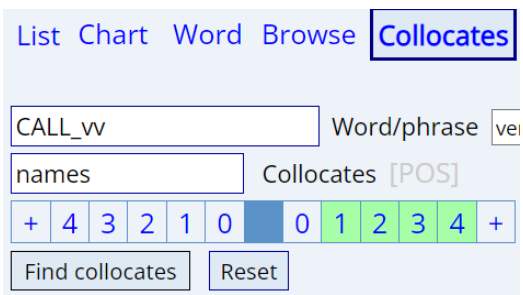
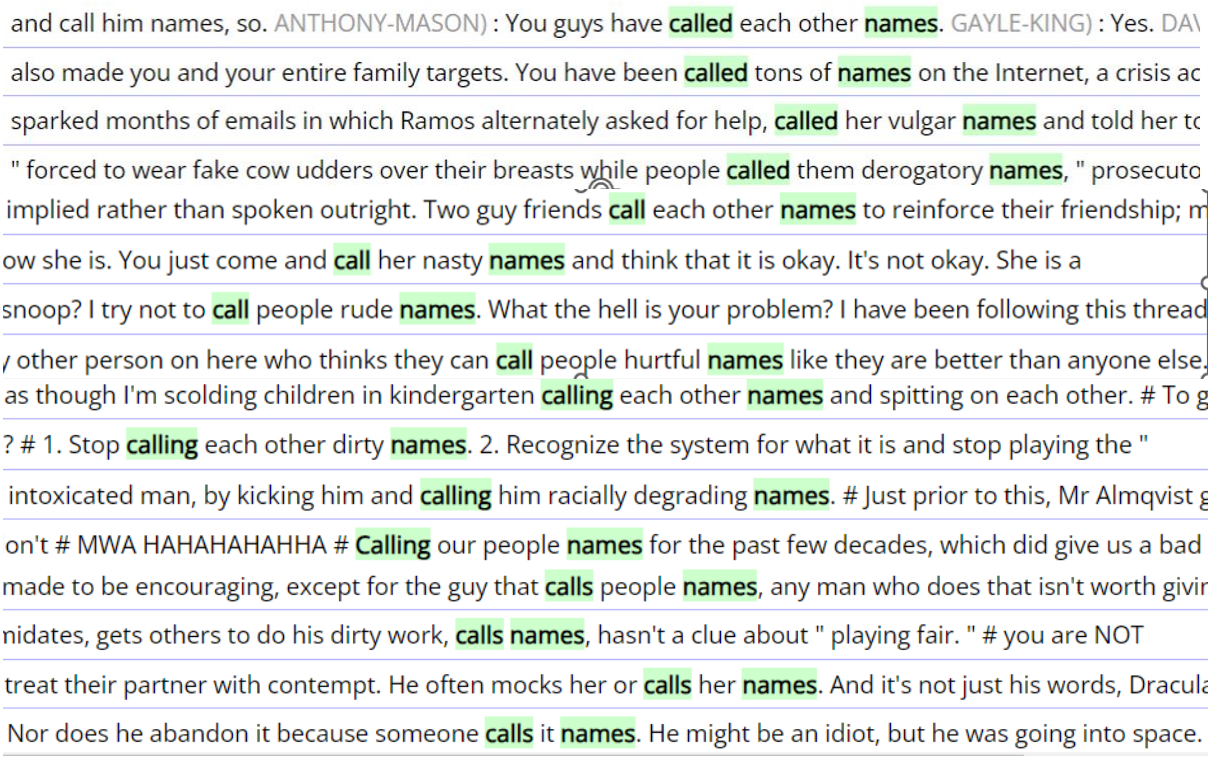


Figure 10. Query 9

RE-USE WORDS	FREQ +
NAMES	2387

Table 8. Query 9 Result

Concordance 12 shows some of the concordance lines:



Concordance 12. A few instances of *call somebody names* in COCA

The earlier 'verb *call – us name* string' co-occurrences in Concordance 11 have given us a hint at the likely high frequency of the 'verb *call – names*' co-occurrences, leading us to a query in Figure 10 above, resulting in 2387 co-occurrences of *names* and the verb lemma *call*, a few of them displayed in Concordance 12.

Both Concordance 12 and Concordance 11 illustrate how corpus data in the form of concordance lines can show patterning in language with clarity because of repeated co-occurrences, in this case the repeated co-occurrences of the verb *call* and *names* or *us names*. To quote Susan Hunston (2010: 152), "A pattern is essentially repetition." or what she referred to in a later version as "an observed regularity" of various kinds (Hunston 2022b: 140). In her 2022 book, a whole chapter, viz. chapter 3, is devoted to a discussion on discerning this repetition or regularity in concordance lines, as hinted by the title of the chapter: "Learning from a corpus. Finding pattern in concordance lines" (Hunston 2022a: 47).

In the few lines we see in Concordance 12, where the focus is on the co-occurrences of the verb *call* and *names*, we see pronouns/nouns other than *us*, namely t*hem*, *her*, *him*, *it*, as well as *people* and *each other*. We also see further qualification of the noun *names* such as *derogatory*, *vulgar*, *racially degrading*, *hurtful*, *rude*, *nasty* and *tons of*. If we scrutinize more concordance lines, who knows what we will find. The extracted concordance lines show speakers' colourful realizations of the expression *call somebody names* in natural settings and this should remind us corpus users of the debt we owe to those who give us access to large corpora like COCA which in turn allow us to get exposed to these colourful realizations. In later sections entitled *Last but not least: Joy* and *Epilogue* the issues of corpus size and corpus data potential will be discussed.

The serendipitous *bastardise*

In the section of apposition search above, we have talked about the need to separate the *United States* abbreviation *US* from the pronoun *us* and this brings us to the issue of American English and British English differences in spelling. Most learners of English know that some words have different spellings in American English and in British English but may not be aware of this fact while doing searches in an English corpus. Past experience has shown that in large English corpora, even if they are British English corpora or American English corpora, American and British spellings can both occur, albeit to different degrees, such as *rumor* and *rumours*, *organize* and *organise*. If we want to capture all occurrences of words (or expressions that contain words) with variant forms of spellings, we may want to cover the different spellings in our queries, or in the case of using a corpus of, say, American English, we need to at least look for the occurrences of the target words in American spelling.

There was an occasion when dealing with a text that contained the following: *"At the parties we played dance music. You couldn't* ***bastardise*** *a Dylan song into dance music,…"*, learners were supposed to find all occurrences of the verb lemma *bastardise* in COCA. Some reported no occurrences at all, which is true, as when we search for the lemma *bastardise* the result says there is no match. Others reported the presence of four *bastardise* in COCA, which is also true, when we look for the form *bastardise* in COCA. If we make a query in COCA for the lemma *bastardize*, however, we get 84 *bastardized*, 24 *bastardize* and 17 *bastardizing*.

While searching for the verb lemma *bastardise* might have been intended as a reminder to be alert of the possibility of variant spellings, by sheer serendipity it also led us to be aware that a lemma search that returns no-match result does not always mean that the lemma does not occur in the corpus. In the case discussed above, the lemma *bastardise* returns no matches, but the form *bastardise* occurs four times, the form *bastardised* 11 times, with no occurrences of the forms *bastardises and bastardising*. Apparently, as kindly explained by Prof. Mark Davies (personal communication), the lemma *bastardise* is not in the lexicon of the tagger that has been used to tag the words in COCA.

To reiterate, if we aim to capture all occurrences of an expression in a corpus, we have to make sure that we search for variant spellings of the words making up the expression where applicable. We look, say, for both *judgement* and *judgment*, not just one of them. Secondly, if a lemma search finds no-match, it is advisable to search for the word-forms that belong to the lemma, just in case this particular lemma is not in the corpus tagger's lemma list.

To hyphenate or not to hyphenate

Differences in written form are not limited to British English - American English spelling variation. Some multiword expressions, and also some compound words, have more than one written form. Take for example the expression *all or nothing*. Even dictionaries write it differently:

> all or nothing ***idiom***
> **relates to doing something either completely or not at all:**
> *She either loves you or hates you - it's all or nothing with her.*
> *The government has rejected the all-or-nothing* ***approach*** *in favour of a compromise solution.*
> (https://dictionary.cambridge.org/dictionary/english/all-or-nothing)

> **all-or-nothing** ***adjective***
> used to describe two extreme situations that are the only possible ones
> *an all-or-nothing decision* (= one that could either be very good or very bad)
> (https://www.oxfordlearnersdictionaries.com/definition/english/all-or-nothing?q=all-or-nothing)

Both the hyphenated *all-or-nothing* and the non-hyphenated *all or nothing* occur in COCA, with frequencies of occurrence 372 and 551, respectively:

; that we needed to stay alive. # The **all-or-nothing** aspect of startups was not something we w
. The men are less willing to take the **all-or-nothing** positions for fear of being called sexist. # R
it has to be so extreme and **all-or-nothing**. There's lots of UGC on the web that isn't infringing,
ear. I don't agree that it's **all-or-nothing** as you do. I think a balance is possible and preferable
or adaptation - it seems to me that setting up an **all or nothing** paradigm in choosing betweer
hought seems to embrace the idea that antipiracy efforts are **all or nothing**. You have folks lik
et away with it is absolutely cheating. # its **all or nothing**. either you follow rules or you do nt.
interests based in realism; it's the cultists with an **all or nothing** stance. I think some of them

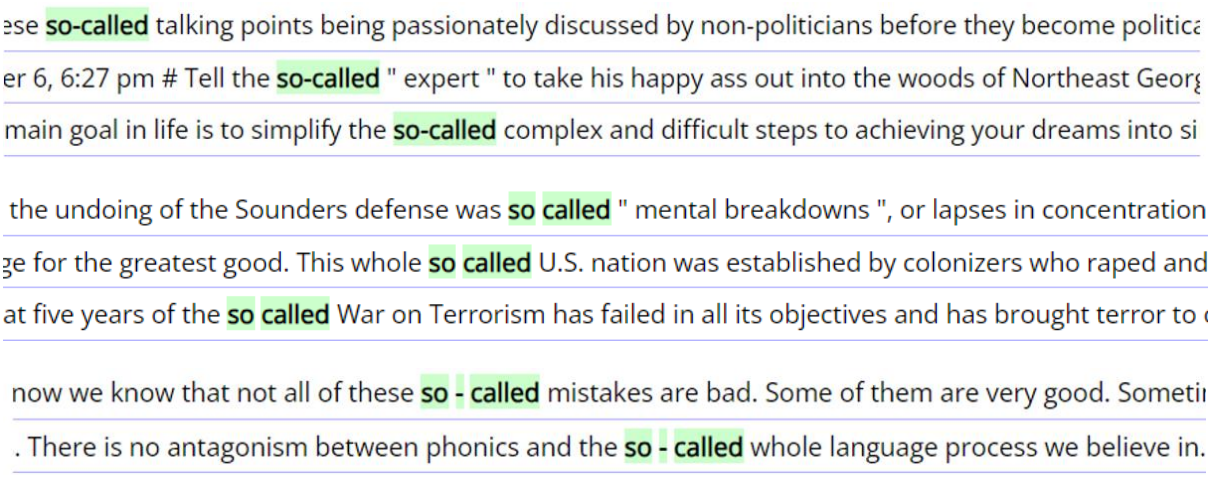
Concordance 13. A few instances of *all-or-nothing* and *all or nothing* in COCA

Another example is the adjective *so-called*:

so-called adjective
1. : commonly named
   the *so-called* pocket veto
2. : falsely or improperly so named
   deceived by a *so-called* friend

(https://www.merriam-webster.com/dictionary/so-called)

In COCA there are three written forms of this adjective, *so-called* occurring 29840 times, *so called* occurring 3610 times and *so – called* 11 times; Concordance 14 below shows some instances of the three variant written forms:



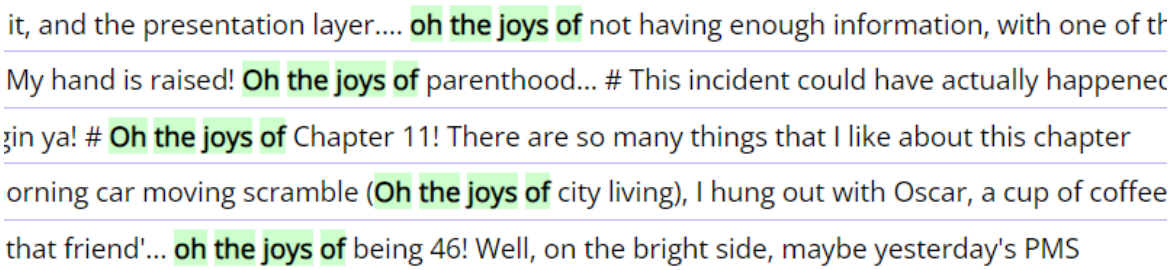
Concordance 14. A tiny sample of *So-called* in 3 alternative written forms in COCA

These two examples, *all-or-nothing* and *so-called*, can hopefully serve as a reminder to always consider the possibility of variant writing forms when making corpus queries.

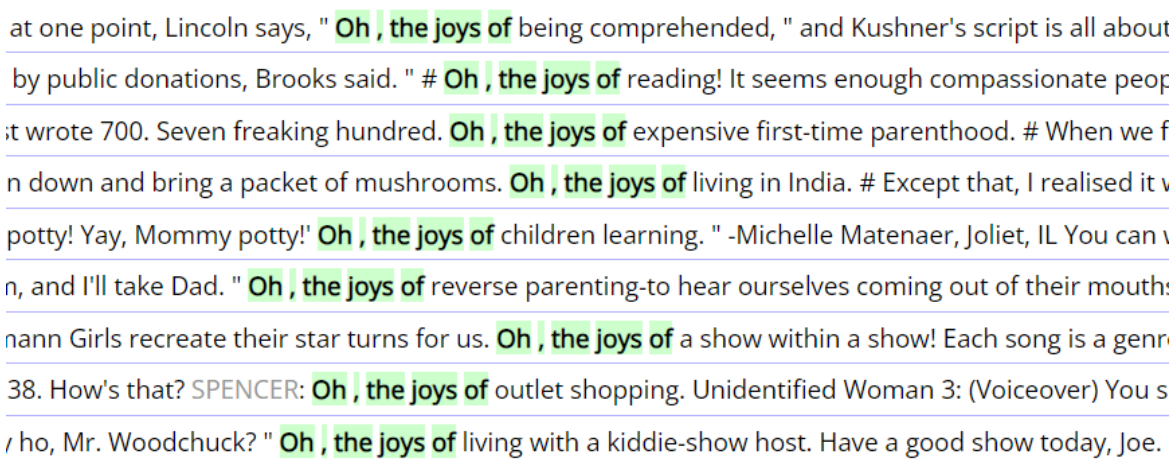
Last but not least: *Joy*

Sometimes the stimulus for corpus search comes from everyday conversations. In a WhatsApp message somebody was sharing his frustrations when reading thesis drafts submitted by students. And a British lady responded, perhaps in a British ironic way of commiserating, "Oh the joys of teaching!😉 😊".

In COCA we can find occurrences of both *Oh the joys of ….* as well as *Oh , the joys of …*,[2] showing that it is likely to be a natural expression in English, and some, if not many or most, carry a meaning that is closer to "…the pains of…":



[2] There is a space after *Oh* before the comma, following the guideline "If you want to include punctuation in the search, you need to separate it from the word before or after."



Concordance 15. ….*the joys of*…

Recorded multiword/phrasal expressions like this had some learners expressed their disbelief that speakers of English from various parts of the world and various walks of life, use (almost) exactly the same multiword expressions. The extracted corpus data, however, tells us that that's the way the cookie crumbles. The fact that (competent) speakers of English have in their lexicon a large number of natural multiword expressions/chunks/strings of various forms can indeed be seen as 'bad news' for learners, as it will be a heavy learning load to aim at acquiring such a huge lexicon.

The availability of large corpora has allowed us to show more and more convincing evidence of (long) multiword expressions used by speakers of the language. Sinclair, Jones and Daley (2004) show that the expression *fit into place* did not occur at all in 2- and 20-million-word corpora, despite the fact that the three words making up the expression are frequent words. It was found occurring six times in a 200-million-word corpus, and in four out of the six occurrences collocating with the word *jigsaw*, making it a four-word expression. Size does matter. In an era where available corpora are huge in size, exposing learners to various multiword expressions and their varied, natural uses extracted from a corpus is quite feasible, and this in turn can raise learners' awareness of the pervasiveness of phraseology.

Looking back at the days in the early 1980s when some lexicographers struggled as they had to contend with a corpus of only several million words and how in just about 20 years afterwards available corpora had gone so big in size, Adam Kilgarriff and Michael Rundell (2002) say, "In a single generation, we have gone from famine to feast." And indeed now, nearing the end of the first quarter of the 21st century, we enjoy the presence of billions-word corpora generously made accessible to not only privileged or target groups but also the general public.

EPILOGUE

It has been discussed so far that using an English corpus with other learners of English can lead to interesting additional learning: getting to know more about the kinds of struggles or the different perceptions of other learners, as well as obtaining unexpected findings about both the language and the corpus.

On top of that, we have also seen how a large corpus can provide abundant evidence of uses of certain expressions in the language by various speakers. This kind of evidence can be a valuable source of repeated exposure to users. Paul Nation has always emphasized language learners' need for repeated exposure (Nation 2013, 2019, among others) and the invaluable role of extensive reading for providing such an exposure: "It is hard to learn another language if you do not have sufficient input to get repetitions that you need for learning." (Nation 2019: p. 20). I often wonder if, in addition to extensive reading, the fragmented lines of certain utterances, be it a clause or a phrase in their surrounding natural co-texts shown in the concordance lines of corpus search results, can be a short-cut, albeit not perfect, to the much-needed repetitions.

As Elena Tognini-Bonelli (2001) remarks, a corpus is read vertically, in that the concordance lines extracted from the corpus are those found across the texts in the corpus, lines that match the construction specified in the corpus query. If we rely on reading English texts in natural settings, it is difficult to get exposed to instant repetitions of, say for example, *condolence letter/call, messages/words of condolence, if at all*, or *call* somebody/*people/each other (nasty/derogatory) names* in its varied uses. A query in a (large)

corpus can extract the target occurrences with just a few clicks, providing raw material for the repeated exposure learners could benefit from.

The first step of course is extracting the relevant/target data from a corpus that minimizes the occurrences of *noise* (non-target items) and we have discussed so far some of the challenges learners or teachers may encounter. The major challenge, however, seems to be the selection of lines that is comprehensible to the learner. Stephen Krashen (Krashen 1981, 2004; Krashen, Lee & Lao 2017) has long proposed the enormous importance of 'comprehensible input' for the repeated exposure to lead to acquisition. In some cases, for comprehensibility purposes, the concordance lines from a corpus may need to be edited. Nevertheless, as a starting point, the raw material for repeated exposure can be obtained from a corpus.

REFERENCES

Davies, Mark. (2008-) *The Corpus of Contemporary American English (COCA)*. Available online at https://www.english-corpora.org/coca/.

Hunston, Susan. (2010). How can a corpus be used to explore patterns. In Anne O'Keeffe & Michael McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 152-166). New York: Routledge.

Hunston, Susan. (2022a). *Corpora in applied linguistics* (2nd ed.). Cambridge: CUP.

Hunston, Susan. (2022b). How can a corpus be used to explore patterns. In Anne O'Keeffe & Michael McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (2nd ed., pp. 140-154). New York: Routledge.

Kilgarriff, Adam, & Rundell, Michael. (2002)). Word Sketches - the Modern Lexicographer's Tool. *MED Magazine, 2,* https://macmillandictionaries.com/MED-Magazine/November2002/2002-Word-Sketches.htm. First published in 'Proceedings of the Tenth EURALEX International Congress, Copenhagen, Denmark', August 13–17, 2002. Eds. Anna Braasch & Claus Povlsen (Vol. II. pp. 807-818).

Krashen, Stephen D. (1981). *Second language acquisition and second language learning*. Oxford: Pergamon Press.

Krashen, Stephen D. (2004). *The power of reading: Insights from the research* (2nd ed.). Portsmouth: Heinemann.

Krashen, Stephen D, Lee, Sy-Ying, & Lao, Christy (2017). *Comprehensible and compelling: The causes and effects of free voluntary reading*. Santa Barbara / Denver: ABC-CLIO.

Mair, Christian. (2002). Empowering non-native speakers: the hidden surplus value of corpora in continental English departments. In Bernhard Kettemann & 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*. Amsterdam - New York: Rodopi.

Nation, I.S.P. (2013). *What should every ESL teacher know?* Seoul: Compass Publishing.

Nation, Paul. (2019). An interview with Professor Paul Nation by Hannah McCulloch. *ETAS Journal, 36*(2), 19-20.

Rühlemann, Christoph. (2019). *Corpus linguistics for pragmatics: A guide for research*. London and New York: Routledge.

Sinclair, John, Jones, Susan, & Daley, Robert (edited by Ramesh Krishnamurthy). (2004). *English collocation studies: the OSTI report*. London: Continuum.

Tognini-Bonelli, Elena. (2001). *Corpus linguistics at work*. Amsterdam: John Benjamins.