DOI: 10.4312/elope.11.1.153-164

ELOPE

Primož Jurko
University of Ljubljana
Slovenia

Target Language Corpus as an Encoding Tool: Collocations in Slovene-English Translator Training

Summary

The opening part of the article discusses reasons for the lukewarm reception of language corpora in the language teaching community. The first reason is the complex syntax and rudimentary user interface of early corpora accessible in the 1990s. The second reason why corpora have witnessed a relatively slow start in language teaching is the fear of the unknown and of an unruly linguistic reality that is often at odds with rules taught at school. The practical part of the article presents a survey conducted among Slovene university students of translation. The survey focused on the effect of using a target language corpus in the course of Slovene into English translation in terms of English collocation. It found that the number of collocation errors in translation can be greatly reduced by competent use of a L2 corpus, which yields a translation with a higher level of idiomaticity.

Key words: translation into L2, collocation, corpus, target language corpus, translator training

Vpliv ciljno-jezičnega korpusa na prevajanje kolokacij pri študentih prevajanja iz slovenščine v angleščino

Povzetek

Članek se v začetku osredotoča na razloge za zadržan sprejem jezikovnih korpusov pri poučevanju jezikov. Kot prvi razlog navaja zapleteno skladnjo in nerodne uporabniške vmesnike zgodnjih korpusov iz 90-ih let. Kot drugi verjeten razlog vidi strah pred neznano jezikovno stvarnostjo, ki je pogosto skregana s šolskimi pravili. Praktični del članka predstavlja raziskavo med slovenskimi študenti prevajanja, ki razgrinja vpliv uporabe angleškega korpusa pri prevajanju v angleščino. Rezultati kažejo, da lahko ustrezna uporaba sodobnega korpusa bistveno zmanjša število napak pri angleških kolokacijah in tako pripomore k višji kvaliteti in idiomatičnosti prevoda.

Ključne besede: prevajanje v tuj jezik, kolokacija, korpus, ciljno-jezični korpus, izobraževanje prevajalcev

UDK 81'25:81'373.72
81'25:378(497.4)

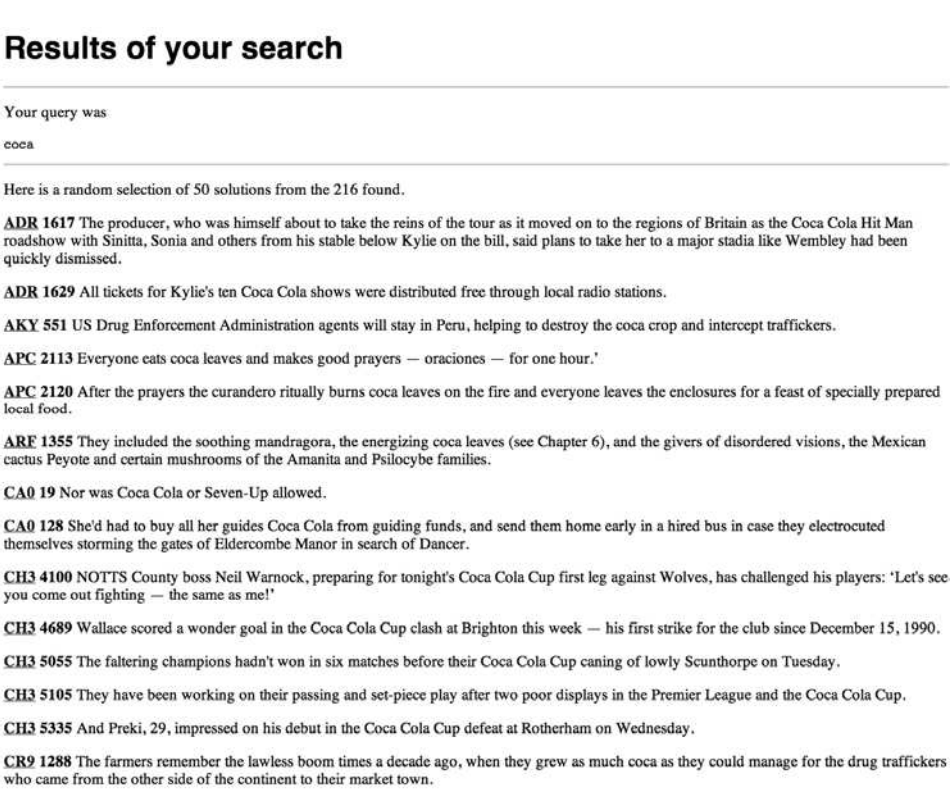
Target Language Corpus as an Encoding Tool: Collocations in Slovene-English Translator Training

1. Corpora past…

The advent of corpus linguistics has in practically every way changed the way we as linguists both deal with and look at language. Access to large quantities of computer-processed texts has proven invaluable in all disciplines of linguistics: rather than merely one of several branches of linguistics, corpus linguistics is seen as a methodology which can be applied to any sphere of linguistics (Mcenery and Wilson 2001, 2). The first and perhaps most obvious linguistic discipline to profit from corpus linguistics was lexicography, with the publication of the revolutionary *Collins COBUILD* EFL dictionary in 1987, edited by J.M. Sinclair, one of the pioneers of corpus linguistics. Other publishers of linguistic reference works soon followed suit, and within less than a decade virtually all contemporary dictionaries were corpus-based, with EFL dictionaries leading the way (e.g., the 4th edition of the *Oxford Advanced Learner's Dictionary of Current English* (Cowie, 1989), the 1st edition of the *Cambridge International Dictionary of English* (Procter, 1995) and the 3rd edition of the *Longman Dictionary of Contemporary English* (Summers, 1995), to name but the most widely used).

The most recent linguistic field to tap into the rapidly growing sphere of corpus linguistics is language teaching. Römer (2009, 84) finds that "corpus linguistics *can* make a difference" and "that it has immense potential to improve pedagogy", but puts it to corpus linguists that they have so far focused on other, arguably higher priority tasks. In her text Römer also noted that corpus linguists have yet to come up with an interface of research and practice that will be sufficiently user-friendly to open the door to a wider acceptance and recognition of corpora as a viable and valuable language teaching tool. Recent developments have shown a marked improvement in precisely this direction, which is attested by high numbers of users of contemporary corpora: the *Corpus of Contemporary American* (aka COCA, Davies 2008) is currently accessed by a massive 40,000 community of unique users each month. The marked surge in numbers of corpora users from hundreds in the early period of corpus history (late 1980s and 1990s) to several ten-thousands today is largely attributable to two developments: one is the growing shift of linguistic focus on lexicological matters, and the other is improved ease of access.

The latter seems to be of particularly high importance, and great effort is directed towards bridging the gap between the wealth and complexity of information stored in corpora, on the one hand, and the needs and expectations of users, on the other. Recent surveys of corpora and corpus tools have shown that a modern user interface "has become more Google-like" (Kilgarriff and Kosem 2012, 49): text-input box, drop-down menus and practically instant results. Such an interface enables the user to gain access to the wide array of information available in the corpus without spending a substantial amount of time just to come to grips with anything beyond the most basic queries. This user-friendliness is a relatively new feature and stands in stark contrast to what corpus users had to live with only a decade or two ago. Users of the early corpora in the 1990s will no doubt remember the painful experience of learning the ropes of the *British National Corpus* (BNC, featuring the old interface available at http://www.natcorp.ox.ac.uk/, see Fig. 1) or its Slovene counterpart *FIDA* (currently superseded by *FidaPLUS* available at http://www.fidaplus.net): the first few days and even weeks of using it were very puzzling, to say the least. The whole enterprise of learning the query syntax of the corpus required one to invest substantial effort and time. Just



Results of your search

Your query was

coca

Here is a random selection of 50 solutions from the 216 found.

ADR 1617 The producer, who was himself about to take the reins of the tour as it moved on to the regions of Britain as the Coca Cola Hit Man roadshow with Sinitta, Sonia and others from his stable below Kylie on the bill, said plans to take her to a major stadia like Wembley had been quickly dismissed.

ADR 1629 All tickets for Kylie's ten Coca Cola shows were distributed free through local radio stations.

AKY 551 US Drug Enforcement Administration agents will stay in Peru, helping to destroy the coca crop and intercept traffickers.

APC 2113 Everyone eats coca leaves and makes good prayers — oraciones — for one hour.'

APC 2120 After the prayers the curandero ritually burns coca leaves on the fire and everyone leaves the enclosures for a feast of specially prepared local food.

ARF 1355 They included the soothing mandragora, the energizing coca leaves (see Chapter 6), and the givers of disordered visions, the Mexican cactus Peyote and certain mushrooms of the Amanita and Psilocybe families.

CA0 19 Nor was Coca Cola or Seven-Up allowed.

CA0 128 She'd had to buy all her guides Coca Cola from guiding funds, and send them home early in a hired bus in case they electrocuted themselves storming the gates of Eldercombe Manor in search of Dancer.

CH3 4100 NOTTS County boss Neil Warnock, preparing for tonight's Coca Cola Cup first leg against Wolves, has challenged his players: 'Let's see you come out fighting — the same as me!'

CH3 4689 Wallace scored a wonder goal in the Coca Cola Cup clash at Brighton this week — his first strike for the club since December 15, 1990.

CH3 5055 The faltering champions hadn't won in six matches before their Coca Cola Cup caning of lowly Scunthorpe on Tuesday.

CH3 5105 They have been working on their passing and set-piece play after two poor displays in the Premier League and the Coca Cola Cup.

CH3 5335 And Preki, 29, impressed on his debut in the Coca Cola Cup defeat at Rotherham on Wednesday.

CR9 1288 The farmers remember the lawless boom times a decade ago, when they grew as much coca as they could manage for the drug traffickers who came from the other side of the continent to their market town.

Figure 1: ***Sample page of hits on the BNC for the query "coca"***

how much effort and time was enough, depended in principle on two factors: one was the users' needs and expectations and the other their previous exposure to computers in their "raw" form, in a manner of speaking. The range of early corpus users extended from highly proficient, trained lexicographers to "just curious what it is" students of linguistics or languages, with several layers of mid-experts in between.

When it comes to classroom activities, it is far from surprising then that the use of early corpora witnessed a slow start. While many (or even most) university teachers were thrilled at the possibility of studying authentic language and discovering language patterns that had up to then been invisible, they were faced with a huge drawback in the form of the highly complex query syntax. For anything beyond the simplest queries, mastering the internal rules of a corpus was a prerequisite. Consider the following example of a relatively simple query in the *FidaPLUS* corpus of Slovene:

#1dež/1-3#1sneg

This particular query returned concordances containing the noun lemma *dež* (rain) appearing anywhere in the span from 1 to 3 words before or after the lemma *sneg* (snow); so, while it is far from complicated in terms of complex queries, it is still highly structured and very cryptic to the untrained eye. This means that early corpora required a level of expertise that was simply too high for many, and even though the broadened horizons of corpus-driven research appealed to teachers and students alike, few of them actually ventured to try more than getting to know the basics of

what corpora are, and more importantly, how they can affect methods of teaching and learning.

Complex syntax and a rough interface are not, however, the only cause of the sluggish start by language corpora in classroom use. There is another put off of corpora in general and this one has to do with the nature of language itself. Or perhaps more to the point, it is related to the ways we teach language and the way language is, which are two very different concepts. Teaching language is invariably subject to simplification and generalization. While it is true that both simplification and generalization are applied to a varying degree on all levels of education in general, there can also be no doubt that without these mechanisms the study of language in full scale would turn into an insurmountable task. Indeed, starting with children's mother tongue acquisition, parents behave like innate language teachers (Bolinger 1981, 165), and simplification in communicating with offspring is practically programmed in our minds. Simplification as a methodology of teaching has strong roots, then, and it is only natural that even at the higher education level we tend to see language as a system of rules. The rules may apply to morphology, syntax, phonology, etc., and although they may change over time, teachers pass them on and students soak them in as eternal truths. However, what corpora reveal is authentic language usage, which often does not play by the rules. Linguistic reality has always been much more intricate than any textbook would have us believe, and modern corpora give it to us (more or less) just as it is. The image of real language is frequently in many ways a distortion of what we learn at school, and distortion leads to discomfort. Linguistic reality is hard to accommodate for both teachers and students: the former may feel threatened by questions from some inquisitive student about a topic that does not comply with the rules taught as school, yet s/he "found it in the corpus". Students, on the other hand, are at first likely to be puzzled by the discovery of data that do not fit into the neatly organized categories of their linguistic knowledge. The discrepancy between corpus-revealed real usage and school-acquired regularity appears to have a particularly strong impact on foreign language learners

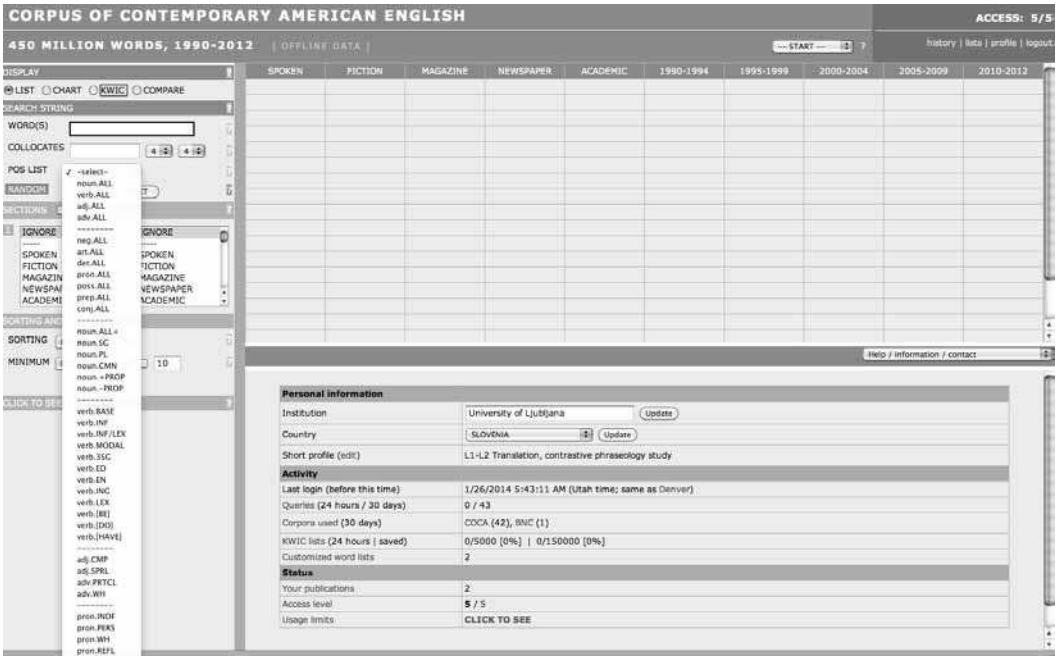


Figure 2: ***COCA interface with POS menu pulled down***

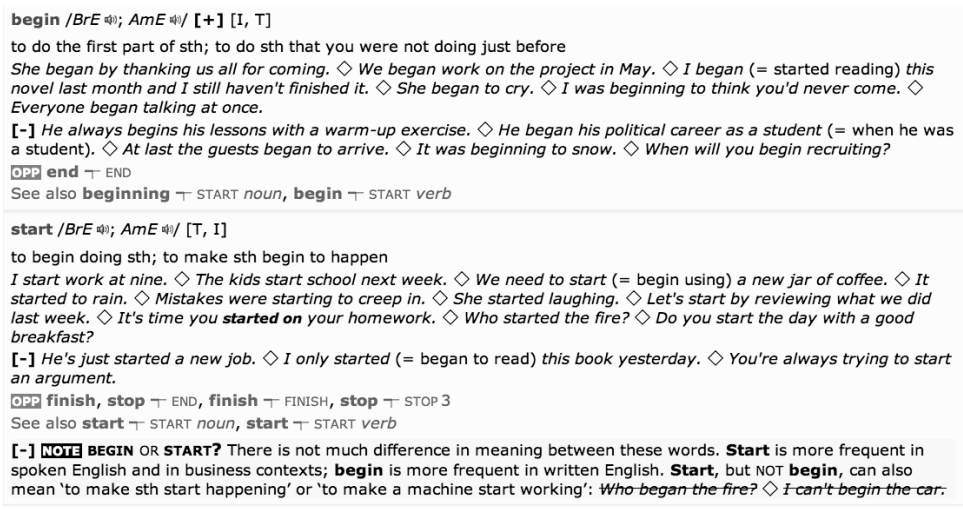
(Granath 2009, 49), who in L2 cannot rely on their language instinct as self-assuredly as they do in their mother tongue.

2. … and present

However, as this paper will show, with proper training and encouragement, students have much to gain from the use of contemporary L2 corpora. As a case in point, a survey based on classroom/home implementation of *COCA* will be presented. *COCA* is a corpus that, since its introduction in 2008, has seen an explosive growth in terms of importance and number of users, which has made it arguably the most popular and widely used English corpus. A detailed presentation of *COCA* and its features would fall beyond the scope of this paper; instead, the main features that facilitate translation of collocations into English are briefly presented below. Suffice it to say that it is freely available on the internet to registered users; once registered, *COCA* saves the entire history of every user's queries.

The first feature allows the user to look for collocates of the keyword within a given span: the interface lets you choose between searching for an exact word form, a lemma or any part of speech (POS). The latter is particularly useful for EFL users in searching for acceptable collocations (like, say, adjectives that can precede a given noun, or prepositions used after adjectives).

One of the greatest lexical problems in translation is dealing with instances of divergent polysemy. The worst case scenario in divergent polysemy for a translator develops when a polysemous lexeme in the source language is rendered by a multitude of lexemes in the target language (for a full account, see Gabrovšek 2005, 120). For instance, let's take the Slovene verb *začeti*, which is in most contexts translatable into English as either *begin* or *start.* But are they interchangeable, i.e. are they full synonyms? In choosing one or the other, most EFL users play it by ear, in a manner of speaking, but how do we get to the bottom of it? In principle, whenever EFL users are of two minds about words with a similar meaning, the first reference work that one thinks of is a thesaurus with meaning discrimination or a dictionary of synonyms. However, with *begin* and *start*, even the most recent one, the *Oxford Learner Thesaurus* (2008) seems to be at a loss for differences in denotation, since one is defined in terms of another: *start* means "*to begin doing sth"* (see Fig. 3).



begin /*BrE*; *AmE*/ **[+]** [I, T]

to do the first part of sth; to do sth that you were not doing just before

She began by thanking us all for coming. ◇ *We began work on the project in May.* ◇ *I began* (= started reading) *this novel last month and I still haven't finished it.* ◇ *She began to cry.* ◇ *I was beginning to think you'd never come.* ◇ *Everyone began talking at once.*

[-] *He always begins his lessons with a warm-up exercise.* ◇ *He began his political career as a student* (= when he was a student). ◇ *At last the guests began to arrive.* ◇ *It was beginning to snow.* ◇ *When will you begin recruiting?*

OPP **end** ┬ END

See also **beginning** ┬ START *noun*, **begin** ┬ START *verb*

start /*BrE*; *AmE*/ [T, I]

to begin doing sth; to make sth begin to happen

I start work at nine. ◇ *The kids start school next week.* ◇ *We need to start* (= begin using) *a new jar of coffee.* ◇ *It started to rain.* ◇ *Mistakes were starting to creep in.* ◇ *She started laughing.* ◇ *Let's start by reviewing what we did last week.* ◇ *It's time you* ***started on*** *your homework.* ◇ *Who started the fire?* ◇ *Do you start the day with a good breakfast?*

[-] *He's just started a new job.* ◇ *I only started* (= began to read) *this book yesterday.* ◇ *You're always trying to start an argument.*

OPP **finish, stop** ┬ END, **finish** ┬ FINISH, **stop** ┬ STOP 3

See also **start** ┬ START *noun*, **start** ┬ START *verb*

[-] NOTE **BEGIN** OR **START?** There is not much difference in meaning between these words. **Start** is more frequent in spoken English and in business contexts; **begin** is more frequent in written English. **Start**, but NOT **begin**, can also mean 'to make sth start happening' or 'to make a machine start working': ~~*Who began the fire?*~~ ◇ ~~*I can't begin the car.*~~

Figure 3: ***begin vs. start in the Oxford Learner's Thesaurus***

The problem of distinguishing between two words with meanings as close as those of the verbs *start* and *begin* frequently boils down to their context, i.e. their collocational behavior. The note at the bottom of Fig. 3 hints at precisely this important difference between the verbs: what objects they can or cannot take.

This is where another feature of *COCA* is invaluable: the comparison. It is intended to compare two words with similar meaning. Selecting the option "COMPARE" in the top left part of the screen offering various display options opens up two comparison boxes, into which the observed words are entered. As above, the "COMPARE" feature also has the option to look for collocates of the two words. The results are displayed with the help of color graphics, which allow the user to evaluate the results with a quick glance at the screen: bright green for word combinations where the use of either "word 1" or "word 2" is exclusive of the other, pale green for combinations where one of the words is markedly preferred, and white for combinations where both words are possible, i.e. for neutral ground, so to speak. So, in order to determine possible contexts for the verbs *begin* and *start*, we looked at the direct objects of the two verbs: a query was built to list nouns that immediately follow the respective verbs (see Fig. 4).

WORD 1 (W1): **START** (1.28)

	WORD	W1	W2	W1/W2	SCORE
1	CRYING	143	1	143.0	111.6
2	BUSINESSES	68	0	136.0	106.1
3	POSITION	56	1	56.0	43.7
4	HAPPENING	26	0	52.0	40.6
5	SEEDS	25	0	50.0	39.0
6	FIRES	24	0	48.0	37.5
7	TROUBLE	21	0	42.0	32.8
8	MENU	20	0	40.0	31.2
9	COMPANIES	20	0	40.0	31.2
10	DATE	20	0	40.0	31.2
11	FAMILIES	19	0	38.0	29.7
12	GAME	17	0	34.0	26.5
13	FEELING	15	0	30.0	23.4
14	BUYING	14	0	28.0	21.9
15	FIGHTS	14	0	28.0	21.9
16	RUNNING	53	2	26.5	20.7
17	TIME	48	2	24.0	18.7
18	PEOPLE	12	0	24.0	18.7
19	WALKING	88	4	22.0	17.2
20	SUPPER	10	0	20.0	15.6
21	SEEDLINGS	10	0	20.0	15.6
22	WARS	18	1	18.0	14.0
23	LISTENING	18	1	18.0	14.0
24	LIVING	17	1	17.0	13.3
25	SHOOTING	96	6	16.0	12.5
26	HUNTING	39	3	13.0	10.1

WORD 2 (W2): **BEGIN** (0.78)

	WORD	W2	W1	W2/W1	SCORE
1	VIDEO	5060	0	10,120.0	12,966.2
2	GRAPHIC	42	0	84.0	107.6
3	DEVELOPMENT	11	0	22.0	28.2
4	PLAY	43	6	7.2	9.2
5	PREPARATIONS	28	4	7.0	9.0
6	OPERATION	21	3	7.0	9.0
7	DELIBERATIONS	17	3	5.7	7.3
8	RESEARCH	11	2	5.5	7.0
9	DEBATE	31	6	5.2	6.6
10	OPERATING	10	2	5.0	6.4
11	DISCUSSIONS	18	4	4.5	5.8
12	SPRING	13	3	4.3	5.6
13	AIRING	25	6	4.2	5.3
14	SERVICE	22	6	3.7	4.7
15	OPERATIONS	36	10	3.6	4.6
16	PRACTICE	13	4	3.3	4.2
17	TALKS	40	17	2.4	3.0
18	PRODUCTION	33	15	2.2	2.8
19	CONSTRUCTION	89	44	2.0	2.6
20	WITH	12	6	2.0	2.6
21	HEARINGS	13	7	1.9	2.4
22	THERAPY	10	6	1.7	2.1
23	TREATMENT	31	19	1.6	2.1
24	NEGOTIATIONS	54	36	1.5	1.9
25	WORK	202	157	1.3	1.6
26	TRADING	18	15	1.2	1.5

Figure 4: ***COCA - comparison of nouns following "start" and "begin"***

A quick check of the results reveals that there are indeed considerable differences between the two verbs in question: many things can only be "started" and fewer can only be "begun". Even advanced EFL users are likely to find in the results something they did not previously know. The comparison feature of *COCA*, then, is particularly valuable in highlighting the differences in collocators of two words that are regarded as quasi synonyms. As will be shown below, students of

translation who used *COCA* in their home assignments were able to show marked improvement in their English collocation output, as compared to their colleagues who did not.

3. *COCA* and students: the survey

To assess the level of improvement in L2 translation of collocations that can be achieved through the use of a target language corpus, a survey was conducted among Slovene university students of translation. The sample group did not access a corpus of English in preparing their translations, while the control group was free to use *COCA*. Here are the technical details:

- all subjects were 3rd year undergraduate students (6th semester)
- A language (mother tongue): Slovene.
- B language (first foreign language): English
- C language (second foreign language): German, French or Italian
- sample group size: 19 students
- control group size: 96 students (in 4 groups of 23, 19, 27 and 27 members).

The survey was carried out as part of the A-B (Slovene into English) translation class. The class took place twice a week and students were required to prepare one translation for each session as a home assignment. The source language texts averaged 1,600 characters in length and varied in difficulty, but were not genre or otherwise marked.

All students received basic information about corpora in general and about *COCA* in particular in the course of their preliminary study. However, this study involved only a few previous hands-on experiences with L2 corpora in terms of classroom activity, which is why they were given a brief account of the most useful features of *COCA*. Particular emphasis was put on distinguishing between googling for word combinations and using *COCA*. The use of popular search engines like *Google, Yahoo* or Microsoft's *Bing* (whichever is set as default in their web browsers) in searching for acceptable English collocations seems to be a perennial favorite among students, indeed, in many it has grown into a habit that they find difficult to kick even after having learned the benefits of using a well-structured corpus. For most people the search engine of choice these days is apparently *Google*, and students seem to be persuaded by the exorbitant numbers of hits for a given search string that this particular word combination is widely used in a given language. What they do not know, however, is that *Google* does not really search all of the World Wide Web in a fraction of a second, but rather performs a calculation and gives you an estimate of how many hits appear to be out there. And what you are given is a VERY rough estimate. In all fairness, every Google user can also get the exact number of hits, but only if they log into their account and change the "Google Instant Predictions" setting to "Never show Instant results". Also, you must set the number of results displayed to 100. What this means in practice is something you can easily try out for yourself: enter the string "to the best of my knowledge" into *Google*'s search box, and it will probably give you anywhere between 150,000,000 and 300,000,000 hits. If you repeat the same search without Instant predictions, you will get the same number, but here comes the truth: say you want to check the results listed from 800-899 and you click on the appropriate link at the bottom of the page: you cannot see the requested results, because *Google* runs out of the results at 632. What is more, if you repeat the search a few minutes later, you are very likely to find that the number has changed, in my case the number of hits was reduced to a mere 414. This is a mere technicality, and there are far more convincing reasons why search engines should not be

considered a quick and handy replacement for a corpus (for a fuller account, visit http://corpus2.byu.edu/coca/help/google_e.asp).

The survey was carried out by comparing students' translations of ten different Slovene texts. The students from the sample group (translation done without using *COCA*) handed in three translations for evaluation at every session, while the control group (translation done with access to *COCA*) handed in six translations over the course of two years. Of course, since the translations were made as a home assignment, this leaves some room for speculation as to whether the students stuck to the agreement of refraining from *COCA* or not. However, the results featured a very consistent array of differences between the two groups, which we believe is largely attributable to the (un-)availability of a target language corpus in the process of translation.

4. Survey results

The data table below shows the distribution of collocation errors (CE) in the translations of the sample group compared to the control group. Data is sorted in descending order according to the 3rd column, i.e. average number of collocation errors per translation in the sample group.

Source language text	**SAMPLE: translation WITHOUT L2 corpus access**		**CONTROL: translation WITH L2 corpus access**		**Ratio CE CTRL/SMPL**
	No. of CE in translations	**Average no. of CE per translation**	**No. of CE in translations**	**Average no. of CE per translation**	
1. JSKD	4,4,3	4.7	0,3,2,2,1,4,0,1,2,3,1,2,	1.75	1 : 2.7
2. Terorist	5,2,5	4	1,3,3,2,1,2,0,1,2,3,1,2,	1.75	1 : 2.3
3. DSKP	3,2,5	3.3	1,2,0,0,1,0,0,0,1,1,1,0	0.58	1 : 5.7
4. SG	1,4,5	3.3	1,0,2,0,0,3,0,1,1,0,0,2	0.83	1 : 4
5. NM	2,5,2	3	0,1,0,0,2,0,0,1,1,0,0,1	0.5	1 : 6
6. Ajdovščina	3,1,1	2.7	0,1,0,0,0,1,0,0,0,2,0,0	0.33	1 : 8
7. Izola	2,3,3	2.7	2,1,2,2,3,1,0,2,1,0,0,3	1.4	1 : 1.9
8. MK 60	2,4,5	2.7	1,3,2,1,0,2,1,0,1,0,2,2	1.25	1 : 2,2
9. Muzeji	3,1,0	2	2,4,2,0,1,3,1,3,0,1,2,1	1.8	1 : 1.1
10. Sujka	0,0,2	0.7	1,1,0,0,0,0,0,1,1,0,0,2	0.5	1 : 1.4
Average CE per translation	2.9		1.04		Average 1 : 3

Table 1: ***Distribution of collocation errors***

4.1 Data analysis

As expected, the above table clearly shows that access to a target language corpus can greatly reduce collocation errors in L2 translation. In translation of all texts, students who used *COCA* performed better than their colleagues who did not. The ratio span between collocation errors per translation made by the control group and the sample group extends from a practically even 1:1.1 (observed in text number 9) to a high of 1:8 (in text number 6). Only 3 out of 10 texts resulted in translations that had collocation error ratios between the control and sample groups lower than 1:2. Translations by students from the sample group on average contained 3 collocation errors, while the control group students averaged 1 collocation error per translation. In other words, students who translate their texts into English with the help of *COCA* are three times less likely to commit a collocation error than those who do not.

4.2 Analysis of selected collocation errors

In addition to the above quantitative analysis, we also looked at the types of collocation that resulted in poor translation. Collocation errors in the translations of both sample and control groups were examined and compared. Although quantitative research was in the foreground of the survey, qualitative inquiry yielded a very interesting insight into the benefits of corpus-aided translation. Consider the following collocation errors made by the "no L2 corpus" group:

ADJ+N

(1) *manjše proslave* - *__minor__/small celebrations
(2) *jubilejna proslava* - *__jubilee__/round anniversary celebration
(3) *upravno središče* - *__civic__/administration center

N+N

(4) *trgovanje z otroci* - *__children__/child trafficking

V+N

(5) *voditi pogovor* - host a *__conversation__/talk (= a public event)
(6) *zaslišati osumljenca* – to *__hear__/interrogate a suspect
(7) *podeliti štipendijo* – to give a *__scholarship__/grant
(8) *zbirati prijave* - *__gather__/welcome applications

ADJ+PREP

(9) *rok prijave je do* – applications are due *__until__/by
(10) *značilen/tipičen za* – characteristic/typical *__for__/of

The Slovene half of the examples (1)-(8) all fall within the scope of *lexical* collocations (Benson et al. 1997), i.e., word combinations of two full lexical parts of speech. On the other end, their (erroneously assumed) translation equivalents in English can be seen as belonging to two discrete

levels on the cline in phraseology (Gabrovšek 2005, 92), one being lexical collocations in examples (1, 2, 5, 6, 7, 8) and the other compounds in (3) **civic/administration center* and (4) **children/child trafficking*. However, to the Slovene learners, the difference between the two levels in terms of *encodability* (ibid., 104) is apparently very subtle, as they are obviously puzzled by both types of word combination.

Most poorly translated collocations in our analysis belong to the general category that Schmitt (2010, 143) calls "strongly linked collocations" (e.g., *densely populated, bated breath*). Such collocations are marked by low overall frequency in native speaker texts (compared to more frequent word combinations like *quick run, big problem*), and recent research into L2-learner text corpora has shown that these collocations are underused by learners (Durrant and Schmitt 2009, 175). This finding is in line with the results of this investigation, which shows that "strongly linked collocations" are a major recurring problem not only in L2 text production, but also in encoding, i.e. L1-L2 translation.

Let us now turn to some collocation errors made by the control group: most of these belong to the N+N or ADJ+N group:

(11) pridelava raznih kultur – cultivation of various ***plants**/crops

(12) vinska klet – ***vine**/wine cellar

(13) varnostni ukrepi - ***safety**/security measures

(14) varnostni ukrepi – security ***actions**/measures

In (11) the student failed to observe the distinction between the nouns *plant* and *crop*, and although this is nothing short of pure speculation, such errors are frequently committed either under time pressure or out of lack of interest, rather than a result of ignoring collocational restrictions. The student was apparently happy with a superordinate term meaning roughly the same in the target language, instead of exploring furhter into the field at hand. The error in (12) is most likely a mental typo, as it is close to impossible for an advanced EFL learner not to know the difference between the nouns *vine* and *wine.* In (13) we are dealing with a case of divergent polysemy of the difficult type (Gabrovšek 2005, 121), where a polysemous item in the source language (Slovene *varnost*) is rendered by a multitude of translation equivalents in the target language (English *security* and *safety*), depending on the context.

There is, however, something that is conspicuously absent from the list of collocation errors in translations made with the help of the COCA: there are no errors involving *grammatical* collocations, i.e. word combinations consisting of a dominant (lexical) word and a preposition or a grammatical structure. Translations of the sample group, on the other hand, contained several non-idiomatic instances of an inadequate preposition following an adjective as witnessed in (9) and (10). This clue seems to suggest that, in dealing with encoding tasks, students of the control group did not rely on their gut feelings, but instead checked the COCA for acceptable ADJ+PREP combinations, which resulted in a zero error count in grammatical collocations. If we give the sample group students the benefit of the doubt and assume that they checked their grammatical collocations in Google, the result is hardly surprising and very misleading: the most popular search engine on the planet says that there are 2.9 million instances of the grammatical collocation *typical *for*!

5. Conclusion

The relatively reluctant acceptance of corpora as a tool suited to the classroom teaching has two main causes. First, the early corpora were apparently not sufficiently user friendly, which meant that few non-expert users were willing to invest the substantial effort and time required to master the corpus in an efficient way. Second, the psychological effect of corpora in general and L2 corpora in particular can be adverse for several users, including both language teachers and students. The former may feel threatened in their beliefs, while the latter can be puzzled by the unrestrained presentation of linguistic reality.

The survey has clearly shown that access to a target language corpus in the process of translation into a foreign language helps translators to significantly reduce the number of collocation errors. Although the sample group was relatively small in size, the survey has corroborated the starting hypothesis of the influence of an L2 corpus on the quality of translation. Slovene graduate students of translation who had access to *COCA* performed markedly better than their peers who did not. On average, *COCA*-aided translations contained only one collocation error per text, while translations performed without access to *COCA* contained three incorrect collocations. Evidence seems to suggest that the use of an L2 corpus in encoding has a particularly beneficial effect on translation in terms of grammatical collocations, but this hypothesis remains to be tested in a future survey, larger in scale.

References

Benson, Morton, Evelyn Benson and Robert Ilson, eds. 1997. *The BBI Dictionary of Word Combinations.* Revised ed. Amsterdam / Philadelphia: John Benjamins.

Bolinger, Dwight. 1981. *Aspects of Language.* New York: Harcourt Brace Jovanovich.

British National Corpus. Accessed January 12, 2014. http://www.natcorp.ox.ac.uk/.

Cowie, Anthony Paul, ed. 1989. *Oxford Advanced Learner's Dictionary of Current English*. 4th ed. Oxford: Oxford University Press.

Davies, Mark. 2008. *The Corpus of Contemporary American English: 450 million words, 1990-present*. Accessed January 12, 2014. http://corpus.byu.edu/coca/.

Durrant, Philip and Norbert Schmitt. 2009. "To what extent do native and non-native writers make use of collocations?" *IRAL: International Review of Applied Linguistics in Language* Teaching 47 (2): 157-77.

FIDAPLUS: Korpus slovenskega jezika. Accessed December 15, 2013. http://www.fidaplus.net/.

Gabrovšek, Dušan. 2005. *Words Galore. Aspects of General and Slovenian-English Contrastive Lexicology*. Ljubljana: Filozofska fakulteta Univerze v Ljubljani.

Granath, Solveig. 2009. "Who benefits from learning how to use corpora?" In *Corpora and Language Teaching*, edited by Karin Aijmer, 47-65. Amsterdam / Philadelphia: John Benjamins.

Kilgariff, Adam, and Iztok Kosem. 2012. "Corpus tools for lexicographers." In *Electronic Lexicography,* edited by Sylviane Granger and Magali Paquot, 31-55. Oxford: Oxford University Press.

Lea, Diana, ed. 2008. *Oxford Learner's Thesaurus.* Oxford: Oxford University Press.

McEnery, Tony, and Andrew Wilson. 2001. *Corpus Linguistics.* Edinburgh: Edinburgh University Press.

Procter, Paul, ed. 1995. *Cambridge International Dictionary of English.* Cambridge: Cambridge University Press.

Römer, Ute. 2009. "Corpus research and practice. What help do teachers need and what can we offer?" In *Corpora and Language Teaching*, edited by Karin Aijmer, 83-98. Amsterdam / Philadelphia: John Benjamins.

Schmitt, Norbert. 2010. *Researching Vocabulary.* Houndmills: Palgrave Macmillan.

Sinclair, John McHenry, ed. 1987. *Collins COBUILD Advanced Learner's English Dictionary*. Glasgow: HarperCollins Publishers.

Summers, Della, ed. 1995. *Longman Dictionary of Contemporary English*. 3rd ed. Harlow: Longman.