Semantically-Based, Learner-Oriented Queries with the 400+ Million Word Corpus of Contemporary American English

Mark Davies

Abstract: The architecture and interface for the Corpus of Contemporary American English (COCA) allow learners of English to carry out a wide range of semantically-oriented queries, including: 1) quick and easy collocates searches 2) comparison of collocates of two words (e.g. small/little) 3) comparison of collocates in different genres (e.g. collocates of "chair" in fiction and academic) 4) use of integrated thesaurus (entries for 60,000+ words) to see frequency of all synonyms (including by genre) and to create more powerful queries (e.g. all synonyms of "beautiful" + a synonym of "woman") and 5) customized wordlists (including hundreds or thousands of words in a semantic domain).

Keywords: English corpus, semantic acquisition.

1. Introduction

One of fundamental problems facing language learners is of course to acquire the semantic and pragmatic knowledge shared by native speakers of the target language. This involves such things as knowing:

- what words co-occur with a given word or phrase, which of course relates to native speakers' knowledge of what the word means and how it is used (i.e. "you can tell a lot about a word by the other words that it hangs out with")
- how the meaning and use of a word differs between genres
- the difference between related words, in terms of meaning and use
- how all of the words in a particular semantic field are related, in terms of frequency and distribution in different genres

(See Schmitt 2000, Nation 20001, and Gardner 2007).

Corpora architectures and interfaces differ widely in terms of how much attention they pay to providing tools to answer questions such as these. It seems that sometimes these architectures and interfaces are oriented much more towards the interests of computer scientists and computational linguists than they are towards language learners.

In this paper, we will focus on how language learners can use the new Corpus of Contemporary American English (COCA)[1] to carry out and use semantically-based queries such as those listed above. As we discuss COCA[2], we will compare it to the four other architectures for large corpora that are currently available online for language learners:

1. Corpus Query Processor (CQP), as exemplified by its implementation in Sketch Engine (www.sketchengine.co.uk) (hereafter Sketch Engine)
2. Corpus Query Processor (CQP), as exemplified by its implementation in BNCweb (bncweb.lancs.ac.uk) (hereafter BNCweb)
3. VISL / CorpusEye (corp.hum.sdu.dk) (hereafter VISL)
4. Phrases in English (pie.usna.edu) (hereafter PIE)

BNCweb and PIE have only one corpus available – the British National Corpus (BNC), while Sketch Engine and VISL have several corpora – although no large corpora from the United States.

As we will see, both Sketch Engine and BNCweb offer fairly rich semantically-oriented queries. However, the range of semantically-oriented queries that are available with the architecture used for COCA is unique, and it is the only architecture that allows language learners to answer all of the types of issues shown above.

2. The composition of the corpus

Before discussing how the COCA architecture and interface can address this wide range of semantically-oriented queries from a learners' perspective, we should first briefly discuss the composition of the corpus, since of course the corpus data is only as good as the textual corpus on which it is based. For example, if we create a corpus that is based on just web pages and/or newspapers (the easiest types of materials to collect), then we will get a very skewed view of a given language. Ideally, we would want equal samplings from a number of widely divergent genres and registers, from genres as informal as spoken to genres as formal as academic, with a number of genres in between (cf. Biber *et al.* 1998).

In the Corpus of Contemporary American English (COCA), the corpus is divided almost equally between spoken, fiction, popular magazines, newspapers, and academic journals (see Davies 2009b for a more complete overview of the textual corpus). This composition holds for the corpus overall, as well as for each year in the corpus. As of August 2009, there are more than 160,000 texts in the corpus, and they come from a variety of sources:

- Spoken: (83 million words) Transcripts of unscripted conversation from more than 150 different TV and radio programs (examples: *All Things Considered* (NPR), *Newshour* (PBS), *Good Morning America* (ABC), *Today Show* (NBC), *60 Minutes* (CBS), *Hannity and Colmes* (Fox), *Jerry Springer*, *Oprah*, etc.).
- Fiction: (79 million words) Short stories and plays from literary magazines, children's magazines, popular magazines, first chapters of first edition books 1990-present, and movie scripts.
- Popular Magazines: (84 million words) Nearly 100 different magazines, with a good mix (overall, and by year) between specific domains (news, health, home and gardening, women, financial, religion, sports, etc.). A few examples are *Time*, *Men's Health*, *Good Housekeeping*, *Cosmopolitan*, *Fortune*, *Christian Century*, *Sports Illustrated*, etc.
- Newspapers: (79 million words) Ten newspapers from across the US, including: *USA Today*, *New York Times*, *Atlanta Journal Constitution*, *San Francisco Chronicle*, etc. There is also a good mix between different sections of the newspaper, such as local news, opinion, sports, financial, etc.
- Academic Journals: (79 million words) Nearly 100 different peer-reviewed journals. These were selected to cover the entire range of the Library of Congress classification system (e.g. a certain percentage from B (philosophy, psychology, religion), D (world history), K (education), T (technology), etc.), both overall and by number of words per year.

There is no corpus of any language that is this large and which allows for a genre distribution this diverse. The British National Corpus has very good genre distribution, but is less than one fourth the size of COCA. The Cobuild / Bank of English corpus is somewhat larger than COCA (about 520 million words), but it is heavily weighted towards easily-available newspapers and contains very little spoken, fiction, or academic. The Oxford English Corpus is even larger, but it is based mainly on material from websites. In order to provide the best semantic and pragmatic information, we should have texts from a wide range of genres.

[1] www.americancorpus.org

[2] In this paper we refer to the "COCA" architecture and interface, as though it were particular to that one corpus. In reality, this architecture and interface have also been applied to a number of other textual corpora, such as the BYU-BNC, the TIME Corpus of Historical American English, the Corpus del Español, and the Corpus do Português. In this paper, however, we will focus on just the Corpus of Contemporary American English, and all of the examples are taken from that one corpus. Those who are interested in the more technical aspects of the corpus architecture and interface might consult Davies (2005) and Davies (2009a) for descriptions of earlier versions, and Davies (2009b) for a technical discussion of the current version.

For a corpus this size, this range of genres is uniquely available in the Corpus of Contemporary American English.

3. Collocates: the size effect

As we will see, BNCweb and Sketch Engine provide useful insight into word meaning via powerful collocate-based queries of the BNC, as does COCA. At 100 million words, the implementation of the BNC in these two architectures seems like it would be quite adequate in terms of size for just about any collocate-based query. However, while 100 million words was huge when it was released in the early 1990s, it is beginning to look increasingly small, especially for collocate-based queries with lower-frequency words. Let us briefly look at just a few examples.

The following compares the collocates sets for given words in COCA and the BNC. The bolded number in the [BNC] and [COCA] columns shows the number of collocates that have a frequency of at least five tokens, within the specified span of words (e.g. 2 left, 0 right). The number in parentheses shows the overall frequency of the node word in the two corpora (e.g. the noun lemma *click* occurs 445 times in the BNC and 3145 times in COCA).

Table 1. Number of collocates in COCA and the BNC

node word (PoS)	collocate PoS / span	BNC **# collocates** (node freq)	COCA **# collocates** (node freq)
click (n)	adj 2L / 0R	**5** (445) double, sharp, loud	**26** (3145) loud, audible, double, sharp
nibble (v)	noun 0L / 3R	**2** (244) ear, bait	**28** (1194) edges, grass, ear, lip
crumbled (adj)	noun 0L / 3R	**0** (27) ---	**25** (446) cheese, bacon, bread, cornbread
serenely (adv)	verb 4L / 4R	**0** (83) ---	**30** (308) smile, float, gaze, glide

The node words were not selected on the basis of those that looked particularly good in COCA as opposed to the BNC. We simply looked for words with an overall frequency at a particular range in COCA, and then searched to see how many different collocates had a frequency of at least five tokens (within the specified span of words) in COCA and the BNC. As one can see, the difference is quite striking. Even though the 400+ million word COCA only has a little more than four times the number of words at the 100 million word BNC, it provides much richer collocational data for these lower frequency words. For example, a word like *serenely* occurs only about four times as frequently in COCA as the BNC (which is to be expected in a corpus about four times the size), yet when one looks at moderately frequent collocates (five tokens or more), the difference is much more striking. In a smaller corpus like the BNC, the raw frequency of a node word might appear to be quite robust, but when it comes to finding collocates, the data often is too meager. As a result, in order to gain insight into the meaning and use of these lower frequency words, it seems clear that we need more than the standard 100 million word corpora of past decades.

4. Collocates: basic queries

Turning now to query types, perhaps the most basic type of information that a corpus architecture ought to be able to provide – in terms of meaning and usage – is collocational information. As the well-known saying in corpus linguistics points out, "you can tell a lot about a word by the other words that it hangs out with".

The VISL and PIE architectures can produce collocates for specific sequences of words (e.g. two word strings like *fast* [noun]), but they cannot find, for example, all nouns "near" *fast*. Sketch Engine and BNCweb are the two architectures, in addition to COCA, that do allow quick and easy access to full collocational information. With all three architectures, it is possible to define the collocates span (e.g. 2 words left or 5 words to the right of the node word), and to limit the collocates to a particular part of speech. In all three cases, the architectures are also quite fast – 1–3 seconds for a moderately frequent word like *catch* in the BNC (~15,000 tokens).

With COCA, for example, to find the most frequent nouns within three words after the verb *catch*, users would enter *[catch].[v*]* into WORD(S) (all forms of *catch* as a verb), *[nn*]* into COLLOCATES, and set the span to [0]–[3] (0 words to the left of *catch*, but up to three words to the right). They would then see results like the following:

Table 2. Collocates display (noun collocates of the verb *break*)

COLLOCATE	TOTAL		SPOK	FIC	MAG	NEWS	ACAD
HEART	1718		482	736	235	219	46
LAW	1478		739	134	196	335	74
NEWS	1379		616	211	224	225	103
SILENCE	992		174	540	119	91	68
RECORD	975		199	35	218	512	11
RULES	957		181	184	227	225	140
GROUND	884		114	96	228	325	121
STORY	786		519	54	94	94	25
TIME	780		164	248	193	115	60
WAR	710		206	87	153	153	111

The chart shown here is somewhat more complicated than a typical collocates chart. We see the frequency for each collocate in each of the five genres (and users can also see the frequency in each five year block since the early 1990s), but it is possible to also just see the overall frequency. The differences in collocates between different genres is something that we will return to in Section 7. In addition, we see here the raw frequency in each genre (color-coded by frequency per million words), but it also possible to see the actual normalized frequency in the chart as well.

COCA, Sketch Engine, and BNCweb also allow users to sort the collocates by "relevance" using Mutual Information score (MI) or other statistical tests. In addition, with each of these architectures it is possible to limit the results to just those collocates with a certain frequency or above a certain Mutual Information score. For example, the following are the most frequent collocates of the verb lemma *break* from COCA where the collocate occurs in the span [4 left / 4 right] at least 20 times, and the results are ranked by Mutual Information score.

Table 3. Collocates display: sorted by Mutual Information score

COLLOCATE	TOTAL		ALL	%	MI
LOGJAM	74		178	41.57	8.04
DEADLOCK	122		464	26.29	7.38
MONOTONY	70		346	20.23	7.00
COLLARBONE	74		445	16.63	6.72
STRANGLEHOLD	42		280	15.00	6.57
TABOOS	52		545	9.54	5.92
IMPASSE	78		881	8.85	5.81
SCUFFLE	30		354	8.47	5.75
BURGLARS	34		419	8.11	5.69
STALEMATE	61		806	7.57	5.58
LEVEES	64		861	7.43	5.56
BARRIER	394		5543	7.11	5.49

[TOTAL] shows the number of times that the collocate appears within the indicated span, [ALL] is the total number of tokens for that word in the corpus (e.g. 114 total occurrences of *red-handed* in the corpus, with or without *catch*), [%] is the percentage of tokens that occur in the span near *catch*, and [MI] is the Mutual Information score.

5. Collocates: more advanced queries

As we have seen, COCA, Sketch Engine, and BNCweb all allow for basic collocates functionality. What is the difference, then, between these architectures? The first difference is the range of node word and collocates pairs that the architectures allow. While Sketch Engine and BNCweb allow users to limit by part of speech and word form for the collocates, they are somewhat more limited than COCA, which allows all of the following:

Table 4. Types of collocate-based searches with COCA

NODE	COLLOCATES	SPAN (L/R)	EXPLANATION	SORT BY GROUP BY	EXAMPLES
LAUGH.[N*]	*	5/5	Any words within five words of the noun *laugh*	Percentage Collocates	hearty, scornful
[THICK]	[nn*]	0/4	A form of *thick* followed by a noun	Frequency Collocates	glasses, smoke
[LOOK] INTO	[nn*]	0/6	Nouns after a form of *look* + *into*	Frequency Collocates	eyes, future
[EYE]	clos*	5/5	Words starting with *clos** within five words of a	Frequency Both words	closed // eye closing // eyes

			form of *eye*		
[FEEL] LIKE	[*vvg*]	0/4	A form of *feel* followed by a gerund	Frequency Collocates	crying, taking
FIND	time	0/4	*Find* followed by *time*	Frequency Collocates	time
WORK/JOB	hard/tough/difficult	4/0	*Work* or *job* preceded by *hard* or *tough* or *difficult*	Frequency Both words	hard // work tough // job
[=PUBLISH]	[n*]	0/4	Nouns after a synonym of *publish*	Frequency Both words	publish // book issue // statement print // money
[=EXPENSIVE]	[[jones:clothes]]	0/5	Synonym of *expensive* followed by a form of a word in the *clothes* list created by *jones*	Frequency Both words	expensive / shoes, pricey // shirt
[=BOY]	[=happy]	5/5	Synonym of *happy* near a synonym of *boy*	Frequency Both words	happy // child, delighted // boy

With COCA, it is basically possible to look for "anything near anything else", including all synonyms of a given word, or any word in large "customized lists" that are created by the users via the web interface. The ability to look for "anything 'near' anything else" allows for very complex collocate-based queries, and the range of queries is unique to the COCA interface.

6. Collocates: word comparisons

COCA and Sketch Engine are the two architectures that allow for a powerful variation on regular collocates-based queries, one which involves comparison between words. Researchers have recognized the value of corpora in using collocates to tease apart slight differences between near-synonyms (e.g. *small* and *little*), or to provide insight into culturally-defined differences between two terms (e.g. *girls* and *boys*) (see, for example, Sinclair 1991 or Stubbs 1996). The architecture of COCA allows users to carry out searches such as this quickly and easily, by comparing the collocates of two contrasting words or lemmas. For example, to compare the collocates of the adjectives *utter* and *sheer*, a user would simply select COMPARE WORDS, then enter *utter* in one search field and *sheer* in the other, and then select [nn*] as for CONTEXT. Finally, s/he might specify that the first word (*utter* or *sheer*) should occur at least 20 times with the given noun. The user would then see the following:

Table 5. Word comparisons (*utter* / *sheer* [NN*])

WORD 1 (W1): **UTTER** (0.30)

WORD	W1	W2	W1/W2	SCORE
DARKNESS	42	0	84.0	276.8
FAILURE	30	0	60.0	197.7
DESTRUCTION	26	0	52.0	171.3
DISREGARD	23	0	46.0	151.6
CONTEMPT	35	1	35.0	115.3
ABSENCE	15	0	30.0	98.8
AMAZEMENT	25	1	25.0	82.4
FOOL	12	0	24.0	79.1
DESOLATION	11	0	22.0	72.5
SILENCE	63	3	21.0	69.2

WORD 2 (W2): **SHEER** (3.29)

WORD	W2	W1	W2/W1	SCORE
NUMBER	226	0	452.0	137.2
VOLUME	187	0	374.0	113.5
FORCE	144	0	288.0	87.4
SIZE	241	1	241.0	73.1
WEIGHT	64	0	128.0	38.8
SCALE	62	0	124.0	37.6
LUCK	53	0	106.0	32.2
MAGNITUDE	51	0	102.0	31.0
AMOUNT	37	0	74.0	22.5
PLEASURE	73	1	73.0	22.2

This table shows that *sheer* occurs about 3.29 as common overall in the corpus as *utter*, and therefore all other things being example, it ought to occur about 3–4 times as frequently with any collocate than does *utter*. Conversely, *utter* should only occur about .30 times for each occurrence of *sheer*. In the case of *amazement*, however (#7 on the left side of the table), the collocate occurs 25 times as frequently with *utter* as with *sheer*, which is about 79 times as frequent as we would otherwise expect (taking into account its overall frequency, discussed above). In the case of *pleasure*, on the other hand (the last entry for *sheer*), it occurs about 22 times as frequently with *sheer* than we would otherwise expect. As one can readily see, the collocates for *utter* tend to be much more negative than those for *sheer*, and this points out an interesting semantic distinction that most non-native speakers of English would not otherwise be aware of (and perhaps not many native speakers either).

The following table provides a few additional examples of word comparisons that can be done with the corpus. [A] and [B] refer to the two words being compared, [Collocate PoS] shows the part of speech of the collocates, and the rightmost two columns show the collocates that occur with either [A] or [B] much more than the overall frequency of either of these two words would suggest.

Table 6. Examples of word comparisons

[A]	[B]	Collocate PoS	Collocates with [A]	Collocates with [B]
[BOY]	[GIRL]	[j*]	growing, rude	sexy, working
DEMOCRATS	REPUBLICANS	[j*]	open-minded, fun	mean-spirited, greedy
CLINTON	BUSH	[v*]	confessed, groped, inhale	assure, deploying, stumbles
SMALL	LITTLE	[nn*]	amount, fee	while, luck
[ROB].[V*]	[STEAL].[V*]	[nn*]	bank, store	cars, money

In summary, the simple yet quick word comparisons that are possible with COCA would be of value to many different types of users (and similar word contrasts are possible with Sketch Engine as well). Linguists can quickly contrast synonyms and language learners can move beyond simple thesauruses to see more in-depth differences between words. And using COCA, even those interested in using corpora to investigate cultural studies, political science, and other social sciences (cf. Stubbs 1996) can quickly and easily compare how contrasting words (*Bush/Clinton*, *Democrats/Republicans*, *men/women*, *Christians/Muslims*) are used in contemporary American English.

7. Collocates: register differences

There is one type of collocates-based search that is possible only with COCA, and that is the comparison of collocates in two different sections of the corpus (such as genres or time periods) to see how word sense is a function of genre, or how a given word is changing in meaning over time. In this section, we will focus on genre-based differences.

Sketch Engine, BNCweb, and COCA all allow users to limit the query to one section of the corpus, such as Fiction or Newspapers-Sports. In COCA, however, it is possible to compare across two different sections. For example, the following table compares the collocates of *chair* in FICTION and ACADEMIC, and clearly shows the very different word senses in the two sections:

Table 7. Comparison of collocates by section

SEC 1: ACADEMIC

WORD	SEC1	SEC2	PM1	PM2	RATIO
DEAN	25	2	0.34	0.03	11.91
BOARD	76	8	1.04	0.11	9.05
COLLEGE	25	3	0.34	0.04	7.94
SECTION	39	5	0.53	0.07	7.43
COUNCIL	14	2	0.19	0.03	6.67
CONFERENCE	19	3	0.26	0.04	6.04
COMMITTEE	145	23	1.99	0.33	6.01

SEC 2: FICTION

WORD	SEC1	SEC2	PM1	PM2	RATIO
KITCHEN	197	2	2.83	0.03	103.35
LEATHER	209	3	3.00	0.04	73.10
LAWN	185	3	2.66	0.04	64.70
EYES	107	2	1.54	0.03	56.13
WINDOW	156	4	2.24	0.05	40.92
SWIVEL	137	4	1.97	0.05	35.94
ARMS	170	5	2.44	0.07	35.67

As can be seen, the collocates of *chair* that occur much more in ACADEMIC than FICTION are *dean*, *board*, *college*, etc., while those in FICTION but not ACADEMIC are *kitchen*, *leather*, *lawn*, etc. The tables show the frequency of each collocate with *chair* in the two sections (e.g. 197 tokens of *kitchen* near *chair* in fiction but only 3 tokens of *kitchen* near *chair* in academic). These then are converted to tokens per million words in the two sections (2.83 in FICTION, .03 in ACADEMIC), and the ratio figure (103.35) is the ratio of the normalized tokens per million figures for the two sections. As can be seen in this table, the data clearly show that in academic texts, *chair* refers to the position on a committee, whereas in fiction texts it refers to the piece of furniture.

With some modifications, the implementations of the BNC in BNCweb and Sketch Engine could conceivably allow for the same type of cross-genre comparisons, because of the way in which the BNC has been carefully constructed and annotated for genre and sub-genre. On the other hand, it would likely be very difficult to do this with "UK Web as Corpus" (ukWaC) corpus on Sketch Engine, because the architecture does not distinguish as clearly which web "genre" the texts belong to. Of all of the different corpus architectures, COCA is unique in the way in which it shows the relationship between genre and word sense.

8. Basic synonyms-based queries

To this point we have focused on collocates, which are of course one of the best ways of getting some sense of the meaning and use of words and phrases. However, there are two other powerful tools that are part of the COCA architecture and interface, and which are unique to this corpus.

The first feature relates to the integrated thesaurus in COCA. A standard printed thesaurus would show the following synonyms for *fast*: *quick*, *immediate*, *sharp*, *brief*, *sudden*, *rapid*, *swift*, *high-speed*, *abrupt*, *brisk*, *short-lived*, *speedy*, *fleeting*, *momentary*, *hasty*, *prompt*, and *hurried*. Obviously, however, some of these words are more frequent than others, and they would have a different distribution in different genres. Without this information, however, inexperienced language learners might end up sounding strange if they use *fleeting* or *momentary* much more than *fast* or *quick*. Language learners might also sound strange if they over-use a synonym in a genre where it is not appropriate, such as academic writing or in conversation.

COCA has an integrated thesaurus with entries for more than 60,000 synsets, which allows for powerful synonym-based queries. For example, users can enter a simple query like *[=fast].[j*]* (*fast* as an adjective), and then see the following (this is just a partial listing of all of the synonyms):

Table 8. Synonyms list (partial listing for the adjective *fast*)

	SYNONYM	TOTAL		SPOK	FIC	MAG	NEWS	ACAD
1	QUICK [S]	29005		6820	8057	6997	4993	2138
3	IMMEDIATE [S]	15606		2497	1445	3105	2948	5611
5	BRIEF [S]	15206		1845	3600	2790	2371	4600
7	FAST [S]	12713		1960	2524	4131	2736	1362
8	SUDDEN [S]	11345		978	5569	2125	1321	1352
10	RAPID [S]	10394		756	890	2377	1558	4813
14	SWIFT [S]	2817		393	907	714	510	293
15	HIGH-SPEED [S]	2671		269	124	1064	787	427
17	ABRUPT [S]	1886		115	610	408	278	475
18	BRISK [S]	1702		86	563	594	365	94
19	SHORT-LIVED [S]	1467		78	151	457	342	439
20	SPEEDY [S]	1388		184	196	493	353	162
21	FLEETING [S]	1330		93	472	366	191	208
22	MOMENTARY [S]	1068		51	504	185	92	236
24	HASTY [S]	940		77	346	182	160	175
27	PROMPT [S]	709		80	70	152	117	290

This table (which is about the most complex one that the user might see – most tables would be much more simple) contains a wealth of information. It shows all of the matching synonyms for the adjective *fast* in the thesaurus, along with their overall frequency and the frequency in each of the five main genres. Sketch Engine also has a "thesaurus-like" feature, but there are at least three important differences. The most basic difference is that the list of words are not really synonyms per se, but rather words with shared collocates. For example, in Sketch Engine for the adjective *fast* it shows *slow*, *quiet*, *dangerous*, etc.

9. Comparison of synonyms across genres

A second difference between COCA and Sketch Engine is that COCA is the only corpus architecture that allows learners to see the frequency of all synonyms in the different genres, as are shown in the table above. With this information, users can see which synonyms are more formal or informal, and thus appropriate for different styles of speech. For example, the table above shows that *brisk*, *speedy*, and *hasty* are relatively less common in academic writing, but that *immediate*, *constant*, and *prompt* are relatively more common in that genre. Such information allows language learners to begin to develop some sense of which synonyms are most appropriate for a given target genre.

It is also possible to directly query to corpus to ask "which synonyms are more common in one genre than another?" For example, users could easily compare the synonyms of *smart* in newspapers vs. academic writing, by simply entering *[=smart]* for the word, and then selecting Newspapers for Section 1 and Academic for Section 2. They would then see that *ritzy*, *nifty*, *brainy*, *stylish*, *glitzy*, *chic*, and *trendy* are more common in newspapers, and that *intelligent*, *keen*, *clever*, and *shrewd* are more common in academic. Another example would be synonyms of *strong* in fiction and academic. In fiction, the synonyms are *beefy*, *burly*, *strapping*, *spicy*, *brawny*, and *pungent* (relating to people and foods), whereas in academic they are *effective*, *deep-seated*, *clear-cut*, *compelling*, *robust*, and *persuasive* – most of which refer to arguments.

10. Comparing collocates across a range of synonyms

A third and final difference with Sketch Engine, and one that adds real power to the synonyms feature in COCA, is the ability to include synonyms as part of more complex queries. For example, users can enter a query like *[=fast].[j*] [nn*]*, which would yield the following results. (These are just a handful of the more than 400 matching strings, and they are grouped by lemma (e.g. *fix* = *fix* and *fixes*), and include the frequency of that lemma string).

Table 9. Collocates of all synonyms of the adjective *fast*

SYNONYM	**collocate**	**frequency**
FAST	food	1178
QUICK	break	679
FAST	track	648

SYNONYM	**collocate**	**frequency**
IMMEDIATE	impact	224
RAPID	expansion	221
FAST	facts	205

QUICK	look	637
QUICK	question	529
QUICK	fix	511
BRIEF	moment	465
BRIEF	period	363
HIGH-SPEED	Internet	333
FAST	lane	275
IMMEDIATE	threat	195
SUDDEN	change	192
RAPID	pace	191
RAPID	succession	188
SHARP	decline	185
SHARP	increase	149

The power of a list like this is that users can quickly see which collocates occur with each of the synonyms. For example, language learner would see that native speakers talk and write about *brief moments*, *rapid succession*, *high-speed Internet*, *fast lane*, *sharp decline*, *immediate impact*, and *quick look*, but probably not *fast moments*, *brief succession*, *rapid Internet*, *sharp lane*, *quick impact*, or *rapid look*. Sketch Engine also allows users to click on words in the list of synonyms and to compare two words at a time, but COCA is the only corpus architecture that allows users to see all collocates with all synonyms at once. Such functionality allows users to move far beyond a typical thesaurus to compare competing words.

11. "Synonym chains"

Before leaving the topic of synonyms, we might mention one other very useful feature that is uniquely available via the COCA architecture and interface. As one can see in Table 8 above, each of the synonyms of a given word have an "[S]" after the word. Users can click on this to go from one synonym set to another, via a "synonym chain", and thus see an entire web of related words. For example, if users search for *beautiful*, they will see 18 synonyms, including *exquisite*. They can then click on the [S] after *exquisite* to see the synonyms of that word, including *delicate*, and from there to *sensitive*, and then to *mild*. And as before, for each of these synonym sets (as in Table 7), they can see a frequency-ranked listing of the synonyms, as well as see in which genre they are most common. All of this allows users to quickly and easily investigate a "web" of interrelated concepts and meanings, via a few clicks of the mouse.

12. Customized lists

One last feature of note related to semantically-based searches with the COCA architecture and interface is the ability to create one's own customized wordlists, and then seamlessly integrate these into the query syntax. There are two ways of

creating these lists. First, users can save a subset of the words or phrases from an existing search. For example, they could search for the synonyms of *beautiful*, or *crash*, or *money*, and then save just the synonyms that are of interest to them. Similarly, they can find the collocates of a given word, and then save some of these collocates in their own wordlist. They could simply create from scratch a wordlist, such as emotions (*sad*, *happy*, *worried*, *ecstatic*, etc.), colors (*blue*, *green*, *red*, etc.), or parts of clothing (*shirt*, *blouse*, *suspenders*, *hat*, etc.). In any of these cases, they simply create a name for the list and store it via the web interface under their chosen username.

These customized wordlists are saved in a database on the server, and can then be used a day, week, or year later as part of another query. For example, if a user *lingprof* creates a list for words related to emotions, s/he can then use these words as part of the query: *[r*] [lingprof:emotions] that*, to retrieve strings like *pretty worried that*, *quite sad that*, *extremely perturbed that*, etc. Likewise, these customized lists can be used as part of a collocates search. For example, the user *lingprof* might create a second list named *familyMember* (with *mother*, *mom*, *brother*, *uncle*, etc.), and then search for any *familyMember* within six words of one of the *emotions* words, e.g. *her aunt was quite happy to see that*, *when Dad is as angry as that*, *they were excited that Mom could be there*, etc. Again, the ability to incorporate user-defined lists as part of the query, as well as the basic corpus architecture, allows users to carry out quite complex semantically-oriented queries on the corpus. And again, this feature is not available with any other corpus architecture and interfaces.

13. Conclusion

As mentioned, one of the fundamental problems for language learners is the acquisition of the meaning and use of words and phrases. In order to do this efficiently, learners need to be able to quickly and easily find the collocates for a given word or phrase, see how the meaning and usage differs across registers, compare sets of collocates for two words to see differences in meaning between the words, compare collocates across a wide range of synonyms (hopefully all at one time), and see how the word compares in frequency and genre distribution with all related synonyms. Many corpus architectures – which are created by computational linguists or computer scientists – are oriented much more towards syntactic structure (parsing, complex regular expressions, etc.). Relatively few have semantically-oriented features like these, which are of real use to language learners. As we have seen, BNCweb does simple collocates quite well, Sketch Engine adds in a number of other features, but the COCA architecture and interface is perhaps the most advanced in terms of all of these different types of semantically-oriented queries.

References

Biber, D., Johansson, S., Leech, G., Conrad, S., and E. Finnegan (1999). *Longman Grammar of Spoken and Written English*. London: Longman.

Davies, M. (2005). "The advantage of using relational databases for large corpora: speed, advanced queries, and unlimited annotation". *International Journal of Corpus Linguistics* 10: 301–28.

- (2009a). "Relational databases as a robust architecture for the analysis of word frequency". In: D. Archer (ed.). *What's in a Wordlist?: Investigating Word Frequency and Keyword Extraction*. London: Ashgate: 53–68.

- (2009b). "The 385+ Million Word Corpus of Contemporary American English (1990–2008+): Design, Architecture, and Linguistic Insights". *International Journal of Corpus Linguistics* 14: 159–190.

Gardner, D. (2007). "Validating the construct of Word in applied corpus-based vocabulary research: A critical survey". *Applied Linguistics* 28: 241–265.

Nation, I.S.P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Schmitt, N. (2000). *Vocabulary in language teaching*. Cambridge: Cambridge University Press.

Sinclair, J. McH. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Stubbs, M. (1996). *Text and Corpus Analysis*. Oxford: Blackwell.

The Case of the Czech National Corpus: its Design and History

František Čermák

Abstract: A brief survey of needs and problems that led to establishment of a corpus institute and, subsequently, to the build-up of Czech National Corpus project are offered. These are followed by a presentation of a strategy adopted, data acquisition and their division, later on, into a number of corpora. Along the way, a number of problems had to be dealt with, out of which some attention is paid here to that of methodology, specifically such that would enable a representative shape of the contemporary corpora. Finally, a survey of existing corpora is presented and some open questions noted.

Keywords: Czech National Corpus, language data, methodology, corpus design, corpus representativeness.

1. General Remarks

Linguists have always suffered from data insufficiency, although they have only rarely admitted that this was the case. Reliable language data are a prerequisite and usual precondition for any information and subsequent conclusions that linguists are likely to draw, just as in any other science. Working with data has always been the mainstream in linguistics and Chomsky's stubborn and irrational contempt for any data hardly invalidates this general data necessity, a fact generally acknowledged. Despite of what he has mistakenly thought ("corpus data are skewed") it has always been quite clear that no one is able, for example, to write a dictionary from introspection only, i.e. the only approach he has subscribed to.

However, linguists may not have always been aware that they lack more data and reliable information being satisfied with what they had, a fact which is being gradually revealed only now, with new and better linguistic output based on and supported by better data. The old illustrious linguists like Otto Jespersen, a grand old man of English linguistics before the war, had been able to collect some 300 000 manual citation slips that he based all his grammars and books on. Today there is just no one willing to follow in his steps: having a corpus he/she does not have to. It used to be prohibitively expensive and time-consuming to collect large amounts of data manually, an experience familiar to anyone who has worked with citation slips from lexical and other archives trying to compile a dictionary or even, in the case of a student trying to write an essay required by his/her professor. To create such an archive of some 10–