Semantically-based queries with a joint *BNC/WordNet* database

Mark Davies

Brigham Young University, Provo, Utah

Abstract

The British National Corpus (BNC) *contains a wealth of data about the frequency and distribution of words and phrases in many different registers of English, yet, via the standard interface, there is not explicit way of investigating the semantic relationship between words. On the other hand*, WordNet *contains detailed hierarchies about the semantic relation between hundreds of thousands of lexical items, but it has very limited information about the frequency and distribution of these words. My project employs relational databases to join together these two resources, and allow advanced semantically-based queries of the* BNC. *These include queries that show the relative frequency of all of the synonyms for a given word, which hyponyms (more specific words) or meronyms (part of a whole) of a particular word are more common in the* BNC*, and all of the specific phrases that express more general semantic concepts.*

1. Introduction

Certainly two of the most important linguistic resources for the study of English are the *British National Corpus* (www.natcorp.ox.ac.uk) and *WordNet* (www.cogsci.princeton.edu/~wn/) (Aston and Burnard 1998; Burnard 2002 for the *BNC*; Fellbaum 1998 for *WordNet*). As is well known, the *BNC* contains 100 million words of text from a wide variety of registers, while *WordNet* contains a semantically-organised hierarchical database of hundreds of thousands of lexical relations. Considered from another point of view, we find that although the *BNC* contains detailed frequency and distributional information on lexical items in English, it provides very limited information on the semantic relationships between items (primarily because this was not part of its original scope). *WordNet*, on the other hand, provides precisely the opposite information. It contains a wealth of information on semantic relationships and hierarchies, but says relatively little about frequency or distributional facts regarding these items.

What would be ideal, of course, is to join these two resources together. One can easily imagine the benefits of using the frequency and collocational information from the *BNC*, and joining this together with the extensive semantic hierarchies encoded in *WordNet*. For example, a user could find any of the following:

- which of thousands of different nouns occur with several different synonyms of *bad*, such as *bad idea*, *foul mood*, *wicked witch*, or *evil eye*;

- which terms for parts of the body or parts of a car are most common, including the frequency in different registers;
- which hyponyms (more specific words) of [to walk] are the most common, again with the possibility of comparing frequency counts across registers.

As one can readily appreciate, such queries would be useful for native speakers of English, but they would be even more useful for learners of English, who do not have native-speaker intuitions regarding the relative frequency of given lexical items and collocations. For example, a non-native speaker may not know that *foul mood* and *wicked witch* are quite a bit more common than *severe mood* and *evil witch*. Likewise, a beginning learner would likely have encountered *to walk*, but would have little idea of more specific words for [to walk] like *stagger*, *stroll*, *clomp*, and *pussyfoot*, and even less idea of their frequency and distribution across different registers.

Ideally, each of these types of searches could be done via a simple web-based interface and involve just one simple, quick query. This paper outlines how such as project has in fact been carried out, using relational databases that link together the *BNC* and *WordNet*. To actually use the online corpus, the reader is referred to http://view.byu.edu/, which is available free of charge.

2. The *BNC* in relational database form

As mentioned, in order to be able to link *WordNet* and the *BNC* together, we first have to get the *BNC* into a relational database. We first start with the 4500+ raw text files that compose the *BNC*, which have a linear structure like the following:

(1) […] <w PRP>within <w AV0>even <w AT0>a <w AJ0>small <wNN1>group <w PRF>of <w NN0>people <w PRP>at <w NN1>work<c PUN>, <w EX0>there <w VM0>will <w VBI>be […]

We first strip out all of the headers, and we then place each word/POS pair on separate lines. The final files contain more than 100 million rows of data like the following (in this case we have placed one set of rows to the side of the other to save space):

Table 1: Vertical structure for the *BNC*.

ROW	*POS*	*WORD*		*ROW*	*POS*	*WORD*
50891887	<w PRP>	within		50891893	<w PRP>	at
50891888	<w AV0>	even		50891894	<w NN1>	work
50891889	<w AT0>	a		50891895	<c PUN>	,
50891890	<w AJ0>	small		50891896	<w EX0>	there
50891891	<w NN1>	group		50891897	<w VM0>	will
50891892	<w PRF>	of		50891898	<w VBI>	be
50891893	<w NN0>	people				

Next we import the 100+ million rows of text into MS SQL Server, creating a table of 100+ million rows. Each row contains just three columns: a sequential [ID] number to identify each successive row, a [word] column, and a [POS] column (as in the table above). We then run an SQL command which – for each row in the database – finds the next six words and places these in additional columns of the table. The database then contains 100+ million successive seven word sequences, as in the following table:

Table 2: *N*-grams table.

ID	*WORD1*	*POS1*	*WORD2*	*POS2*	*WORD3*	*POS3*	…	*WORD6*	*POS6*	*WORD7*	*POS7*
50891887	within	PRP	even	AV0	a	AT0	…	of	PRF	people	NN0
50891888	even	AV0	a	AT0	small	AJ0	…	people	NN0	at	PRP
50891889	a	AT0	small	AJ0	group	NN1	…	at	PRP	work	NN1
50891890	small	AJ0	group	NN1	of	PRF	…	work	NN1	,	PUN
50891891	group	NN1	of	PRF	people	NN0	…	,	PUN	there	EX0
50891892	of	PRF	people	NN0	at	PRP	…	there	EX0	will	VM0
50891893	people	NN0	at	PRP	work	NN1	…	will	VM0	be	VBI

This main [7-grams] table can then be converted to specific [x-gram] tables, by collapsing identical rows and placing the number of identical rows as a new column in the database. For example, the following table shows a small fragment of the [3-grams] table with some of the entries for the lemma [break] as a verb. The table contains the [word], [lemma], and [POS] for each unique three word string, and the first column indicates how many times that exact string occurs in the *BNC*.

Table 3: Example of 3-grams where lem1 = BREAK and word2 = THE.

FREQ	*WORD1*	*POS1*	*LEM1*	*WORD2*	*POS2*	*LEM2*	*WORD3*	*POS3*	*LEM3*
106	breaking	VVG	break	the	AT0	the	law	NN1	law
98	break	VVI	break	the	AT0	the	law	NN1	law
56	broke	VVD	break	the	AT0	the	silence	NN1	silence
53	break	VVI	break	the	AT0	the	news	NN1	news
46	broke	VVD	break	the	AT0	the	news	NN1	news
40	break	VVI	break	the	AT0	the	deadlock	NN1	Deadlock
24	broken	VVN	break	the	AT0	the	law	NN1	Law
23	break	VVI	break	the	AT0	the	habit	NN1	Habit

As one might imagine, these tables – although much smaller than the main 100+ million row [7-grams] table – are still quite large. For example, there are more than 800,000 rows in the [1-grams] table (i.e. 800,000+ unique types in the *BNC*). This increases to 4.9 million rows of unique [2-grams], 9.4 million rows of unique [3-grams] and 8.2 million rows for unique [4-grams]. In order to have smaller tables (and therefore faster data retrieval), we limit all [2-gram] and

higher tables to just those *n*-grams that occur two times or more in the *BNC*. If we include even the *n*-grams that occur just once, the tables are much larger – about 11 million rows for [2-grams] and 40 million rows for [3-grams].

At this point it may be profitable to briefly compare our approach to previous work with *n*-grams databases of large corpora. Perhaps the first corpus to use this architecture was the 100 million word *Corpus del Español* (www.corpusdelespanol.org), which I created from 2001/2002. Based on this architecture, there was the subsequent 'Phrases in English' (PIE) website and database that has been created by Bill Fletcher (http://pie.usna.edu). The PIE site is based on the *BNC*, and it allows users to search for sequences of words and/or POS tags, and then see the original context for the matching strings.

Users of our website (http://view.byu.edu/) use simple query syntax to search for the frequency and distribution of strings. For example, to search for all examples of *break* + the + noun, they simply input the following into the search form:

(2) break the [nn*]

(A drop-down list also inserts the part of speech tag, for those who are not familiar with the *BNC* tagset). The web-based script then queries the [3-grams] table and returns the following hits. As with the PIE site, users can click on each of these to see the phrase in context.

Table 4: BREAK (v) THE [noun].

WORD	*FREQ*	*WORD*	*FREQ*	*WORD*	*FREQ*	*WORD*	*FREQ*
law	280	back	16	strike	10	chocolate	5
news	134	world	15	hold	9	connection	5
silence	98	power	14	skin	9	contact	5
rule	82	glass	13	club	8	course	5
deadlock	58	seal	13	impasse	8	health	5
ice	50	camel	12	monotony	8	bond	4
spell	41	code	12	cycle	7	bread	4
habit	35	contract	12	peace	7	company	4
mould	35	door	11	heart	6	consensus	4
chain	34	pattern	11	line	6	diet	4
surface	32	story	11	stillness	6	egg	4
link	26	window	11	stranglehold	6	engagement	4
bank	25	agreement	10	tension	6	fall	4
record	20	journey	10	term	6	chocolate	5

There is one important difference between the PIE site and the database that we have created, however, and this difference deals with coverage. As we have mentioned, our databases contain *all* *n*-grams. The PIE database, on the other hand, is limited to just those *n*-grams that occur three times or more. This may not appear to be overly significant, but in terms of *n*-gram frequency it is quite important. By increasing the coverage to all *n*-grams, we create databases that are

roughly four to five times as large as those that contain only the strings that occur three times or more. In other words, the PIE loses approximately 75-80% of all *n*-grams by including only those strings that occur just three times or more.

A clear example of the importance of including even less common *n*-grams is the following table. This table shows a portion of the [3-grams] for the phrase [BREAK] [THE] [NOUN]. Note the interesting strings like 'break the cohesion', 'break the mood', and 'break the Sabbath'. Each of these potentially adds some insight into the meaning of 'to break', and yet with a more limited database like that of the PIE site, we would not be aware of such strings.

Table 5: Less frequent strings for [BREAK (v) THE [noun]].

activity, ban, barrier, bone, boredom, bound, boundary, circle, cohesion, concentration, conspiracy, country, court, day, dependency, director, enchantment, enigma, filter, fish, force, government, jar, kiss, marriage, mood, organisation, post, pound, promise, regulation, resistance, routine, sabbath, sequence, sound, stone, sunday, thread, top, train, trust, un, union, wheel, wicket, will, yalta

3. *WordNet* in relational database form

Creating the *WordNet* database is somewhat easier than the *BNC* database. There is already a version in relational database form at http://wordnet2sql.infocity.cjb.net/. While this is in MySQL and DB/2 format, it was easily ported over to MS SQL Server, the database used in our project. The database contains the following tables, and it is the interaction of these tables that provides the power behind the SQL queries.

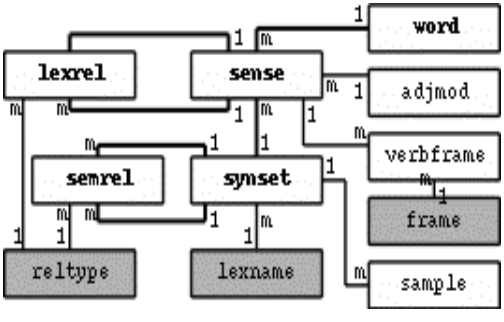

Figure 1: *WordNet* tables.

One of the central database tables is the [sense] table, which contains more than 200,000 entries containing many different 'synsets' or word senses of tens of thousands of words. For example, the following table is just a partial listing of the many different entries for [beat], for several different parts of speech:

Table 6: Synsets of [BEAT] (partial list).

ID	*WORD*	*POS*	*SYNSET*	*LEXFILE*
103065	beat	A	(informal) very tired	adj.all
2475	beat	N	a stroke or blow	noun.act
1401	beat	N	the act of beating to windward	noun.act
26292	beat	N	a regular rate of repetition	noun.attribute
35537	beat	N	(prosody) the accent in a metrical foot of verse	noun. communication
76162	beat	V	wear out completely	verb.body
78744	beat	V	be a mystery or bewildering to	verb.cognition
80925	beat	V	beat through cleverness and wit	verb.competition
80912	beat	V	come out better in a competition, race, or conflict	verb.competition
82481	beat	V	give a beating to	verb.contact
82490	beat	V	hit repeatedly	verb.contact

In the following sections, we will see how this basic 'synsets' table can be used (at times in conjunction with other tables) to find and display to the user all of the synonyms, hypernyms, hyponyms, meronyms, and holonyms for each of the different word senses of a given word, and then how this semantic information is merged with other tables to show the frequency of the semantic concepts in the *BNC*.

4. Basic synonym queries via the web-based interface

Perhaps the most basic use of the *WordNet* databases is to find all of the synonyms for the different senses of a given lexical item. Let us continue with the example of *beat* given above. In order to access the *WordNet* information and look for synonyms, the user would enter the following into the search form:

(3) [=beat].[v*]

The first part of the query string ([=beat]) indicates that we are searching for synonyms of *beat*, while the second part ([v*]) indicates that we are interested just in *beat* as a verb. The following is a screenshot of the search interface:

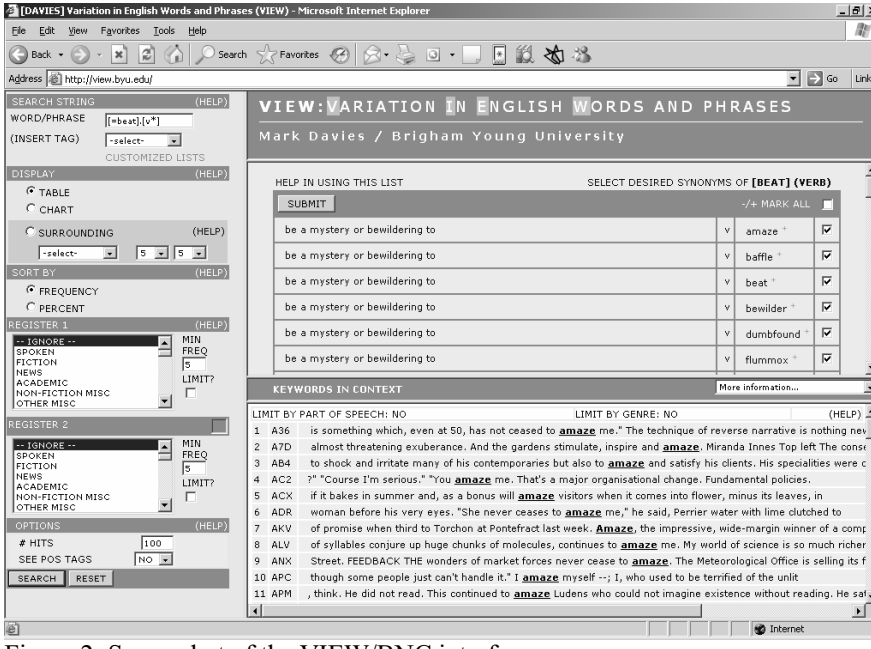

Figure 2: Screenshot of the VIEW/BNC interface.

The user then sees a listing of all of the different synsets for *beat* as a verb, along with all of the other verbs that share each of these meanings. The following is a more detailed view of some of the entries from this results set, which represents the upper right frame seen above:

Table 7: Partial results list for [beat] as a verb.

SUBMIT	+/- MARK ALL		☒
be a mystery or bewildering to	v	amaze	☒
be a mystery or bewildering to	v	baffle	☒
be a mystery or bewildering to	v	Beat	☒
be a mystery or bewildering to	v	bewilder	☒
. . . . .	. .	. . . .	. .
be superior	v	beat	☒
beat through cleverness and wit	v	beat	☒
beat through cleverness and wit	v	circumvent	☒
come out better in a competition, race, or conflict	v	trounce	☒
come out better in a competition, race, or conflict	v	vanquish	☒

If users want to focus in on a particular meaning, they can select just that one entry on the web-based form. For example, suppose that the user is primarily interested in the meaning of *beat* expressing the concept [to be a mystery or bewildering to], as is "*it beats me why she says such things*". After selecting just this one entry, the web-based script then finds all of the other words that express

this concept ("*it beats/puzzles/bewilders/perplexes me*", etc), as well as the frequency of each of these words in the *BNC*. After clicking on any of these words in the lower frame, the user then sees a KWIC display for that word (with the correct part of speech) in the *BNC* (note that in this table there are reduced left and right contexts to fit this printed table):

Table 8: KWIC display for words displayed in Figure 2 above.

TEXT	*LEFT*	*WORD*	*RIGHT*
K5L	children respect ? Respect ? they	puzzle,	what's that ? The head teacher
EW7	— an article which would greatly	puzzle	dog fanciers who had turned to the
CE9	thin air ; it was only then did they	puzzle	and wonder if the dusk had conned
H7H	There were no Pommes Anna to	puzzle	him, but would he find the croûtons
CAB	and stared up at the ceiling trying to	puzzle	it out. Finally he gave up and
FYV	plenty of opportunity to do — to	puzzle	at it, I mean. I puzzle a lot,
BML	the narrative is intended to	puzzle	(is he doing it or dreaming it ?),
CKY	Service lists just 50,000. Dead dogs	puzzle	archaeologists The largest dog

To this point we have considered how the query takes place, from the point of view of the end user. Now let us go somewhat deeper and consider briefly how the query is processed in terms of the underlying *WordNet* and *BNC* databases. The following is one of the key SQL commands, which generates the table seen in the middle frame of Figure 2 – all of the synonyms for each of the synsets of the desired word (in our case [beat] as a [verb]):

(4) select distinct s1, IDs1, w1, c1 from [sense] where IDs1 in (select IDs1 from[sense] where w1 in ('beat') and c1 = 'v') order by ID pos1 asc, lexfile2 asc

Because each synset ('meaning') has a unique ID, the SQL command find all of the other words in the synset database that also have the ID belonging to one of the synsets of [beat]. For example, the following table lists the other lexical items that have ID #78744, which belongs to the synset that expressed the concept [be a mystery or bewildering to]:

Table 9: Lexical items in a synset.

ID1	*WORD*	*ID*	*SYNSET*
112659	amaze	78744	be a mystery or bewildering to
23583	baffle	78744	be a mystery or bewildering to
2187	beat	78744	be a mystery or bewildering to
112656	bewilder	78744	be a mystery or bewildering to
112660	dumbfound	78744	be a mystery or bewildering to
112657	flummox	78744	be a mystery or bewildering to
112253	mystify	78744	be a mystery or bewildering to
112658	nonplus	78744	be a mystery or bewildering to
111840	perplex	78744	be a mystery or bewildering to

5094	pose	78744	be a mystery or bewildering to
32936	puzzle	78744	be a mystery or bewildering to
111301	stupefy	78744	be a mystery or bewildering to
112655	vex	78744	be a mystery or bewildering to

These thirteen words are the ones that appear in the first synset of Table 7. If the user selects this synset, then a subsequent web-based script stores these thirteen words in a temporary database. The script then retrieves these words and inserts them into a query that searches for the frequency of each of these words in the main *BNC* 1-grams table. This provides the output for the table in the lower frame of Figure 2. A final script then finds KWIC-formatted output from the *BNC* for any word selected by the user, as in Table 8 above.

5. Synonym-based collocations

The preceding example demonstrates one of the more basic uses of the *BNC*/*WordNet* database. In this case, we use *WordNet* to find all of the synonyms of a given word, and then use this output to find the frequency of each of these words in the *BNC*. However, this could have also been done manually. In other words, we could have done the following: go to the main *WordNet* site at Princeton (http://www.cogsci.princeton.edu/cgi-bin/webwn), enter in [beat], select the synset [be a mystery or bewildering to], see which other words belong to this synset, copy and paste the first word from the list into a *BNC* interface (e.g. http://sara.natcorp.ox.ac.uk/lookup.html), look at the KWIC display, go to the next of the thirteen words in the list, go through the same process, and so on through each of the thirteen words. With our approach, however, we can carry out this process for any number of synonyms of a given word in just one or two simple steps.

With more complex queries, the advantage of our approach becomes even more pronounced. For example, suppose that a user wants to see all of the collocations involving a synonym of [wicked] followed by a noun. The user simply enters the following into the search form:

(5) [=wicked] [nn*]

This searches for all synonyms of [wicked] from WordNet, followed by any noun. In less than four seconds, the user then sees something similar to the following. (The format on the web interface is somewhat different from the abbreviated listing shown here).

Table 10: Synonyms of [wicked] + NOUN.

PHRASE	*FREQ*	*PHRASE*	*FREQ*
disgusting thing	16	severe shortage	26
disgusting way	5	severe weather	43
distasteful species	5	severe winter	49
evil empire	10	terrible accident	26
evil eye	24	terrible blow	13
evil influence	9	terrible danger	14
foul language	29	terrible feeling	19
foul mood	21	terrible mistake	48
foul play	72	terrible shock	41
foul temper	14	wicked grin	6
severe blow	44	wicked people	13
severe burn	28	wicked thing	27
severe damage	59	wicked way	14
severe drought	30	wicked witch	12
severe illness	23		

Such collocational data can be very useful for a language learner, who is probably unsure of the precise semantic range of each adjective. The type of listing given above, which shows the most common nouns with each of the adjectives, can easily permit the language learner to make inferences about the semantic differences between each of the competing adjectives. For example, s/he would see that *severe illness* occurs but *wicked illness* does not, and that *terrible mistake* is common, whereas *foul mistake* is not.

In terms of text processing, we should note that this type of query would be quite cumbersome with the standard *BNC* interface, and would even by quite difficult with another interface like *BNCweb* or the 'Phrases in English' sites described above, both of which allow searches by part of speech. Again, the difficulty is due to the fact that successive queries would have to be carried out for each synonym of [bad], and the output from each query would then have to be collated together.

6. Related concepts through strings of synonyms

Perhaps an even better example of the power of the database is one that contains a string of synonyms, which express a given semantic concept. For example, suppose that one wants to look for all synonyms of [large] followed by all synonyms of [amount], such as *large sum*, *big amount*, *great measure* and *large total*. Users would simply enter the following into the search form:

(6) [=large] [=amount]

Within two seconds, the user then sees the frequency for each of these phrases from the *BNC*, such as the following. (This is again an abbreviated listing of what would be seen via the web interface).

Table 11: RESULTS: [large] + [amount].

	PHRASE	*FREQUENCY*
1	large (a) amount (n)	793
2	large (a) quantity (n)	488
3	large (a) sum (n)	373
4	large (a) measure (n)	146
5	great (a) quantity (n)	88
6	great (a) amount (n)	71
7	great (a) measure (n)	45
8	great (a) sum (n)	10
9	big (a) amount (n)	6
10	big (a) sum (n)	6
11	large (a) total (n)	2

Imagine if this semantically-based query were carried out with a standard interface. The user would have to input separately each of the synonyms of [large] with each of the synonyms of [amount], which might be on the order of 60 different combinations [10 x 6]. With our interface, it is just one simple query.

7. Semantic hierarchies – hypernyms and hyponyms

WordNet allows us to study much more than just synonyms. For example, we can find all of the words related to a given word whose meaning is more specific [hyponym] or more general [hypernym]. With our joint *BNC*/*WordNet* database, even if the list contains 40-50 words we can quickly find the frequency for all of these words in the *BNC* with just one simple query.

The query syntax is quite simple. The following two symbols are used to extract hyponym and hypernym entries from WordNet, and enter them in as part of the BNC query:

(7) [<word $_x$] more specific words relating to [word $_x$] (hyponyms)
[>word $_x$] more general words relating to [word $_x$] (hypernyms)

For example, suppose that a language learner wants to find out more specific ways of expressing the concept [to walk], involving words like *amble*, *prowl*, *saunter*, and *skulk*, as well as the frequency of each of these words. In order to find these more specific verbs, the user would enter the following into the search interface:

(8) [<walk].[v*]

(Where [<walk] indicates that we are searching for words that have a more narrow [<] meaning than [walk], which are verbs [v*]). Behind the scenes, the web-based script would invoke this following SQL query:

(9) select distinct top 100 w1,c1,s1 from x_sense where IDs1 in (select s1.IDs1 from x_sense as s1 left join x_semrel as r on s1.IDs1 = r.IDs1 left join x_sense as s2 on r.IDs2 = s2.IDs1 where r.relation = 'hypernym' and s2.w1 = 'walk' and s2.c1 = 'v') order by s1 asc

The data would then be output to the user, and s/he would see a listing like the following:

Table 12: More specific terms for [walk] (partial listing).

SYNSET	*WORDS*
take a walk	promenade, stroll
take a walk for one\'s health or to aid digestion, as after a meal	constitutionalise
to go stealthily or furtively	creep, mouse, pussyfoot, sneak, steal
to walk with a lofty proud gait, often in an attempt to impress others	cock, prance, ruffle, sashay, strut, swagger
tread or stomp heavily or roughly	trample, tread
walk (informal)	foot, hoof, hoof it, leg it
walk about	Ambulate
walk as if unable to control one\'s movements	careen, keel, lurch, reel, stagger, swag
walk by dragging one\'s feet	scuffle, shamble, shuffle
walk heavily and firmly, as when weary, or through mud	footslog, pad, plod, slog, tramp, trudge
walk impeded by some physical limitation or injury	Hitch, hobble, limp
walk in one\'s sleep	sleepwalk, somnambulate
walk leisurely	amble, mosey
walk leisurely and with no apparent aim	saunter, stroll
walk on and flatten	tramp down, trample, tread down
walk on one\'s toes	tip, tippytoe, tiptoe
walk or tramp about	shlep, traipse
walk ostentatiously	exhibit, march, parade
walk stealthily	Slink
walk stiffly	Stalk

As before, the user simply selects the synsets that are of interest, and then clicks on ‘Submit’ to see the frequency of each of these words in the *BNC*. The following listing, for example, shows the relative frequency of more specific verbs relating to [walk]:

Table 13: *BNC* frequency counts for hyponyms of [walk] (partial listing).

WORD	*FREQ*	*WORD*	*FREQ*	*WORD*	*FREQ*	*WORD*	*FREQ*
march	1765	trample	286	keel	108	scuffle	46
creep	1481	trudge	257	slink	104	leg it (p)	42
stride	1050	plod	212	stomp	104	bumble	40
stumble	1011	tramp	190	waddle	97	sleepwalk	26
tread	893	saunter	185	slouch	96	hoof	21
stroll	770	foot	183	hike	89	promenade	19
shuffle	687	prowl	182	prance	82	sashay	19
stagger	672	hobble	172	shamble	77	perambulate	13
sneak	452	amble	167	skulk	69	careen	10
trot	450	flounder	160	flounce	60	mosey	10
stalk	422	stump	156	swagger	59	dodder	9
reel	406	totter	155	slog	56	swag	8
lurch	397	tiptoe	151	traipse	51	leg (n) it (p)	6
limp	374	strut	146	toe	49	pussyfoot	4
parade	372	lumber	135	toddle	47	lollop	2

The following is another example of finding more specific lexical items to express a given concept. Suppose that a language learner has forgotten the word [aquarium], but does know that it refers to some type of [tank]. S/he would simply enter the following into the search form:

(10) [<tank]

With one more click, s/he would see the frequency for each of these items:

Table 14: More specific terms for [tank], with frequency counts (partial listing).

SYNSET	*WORDS*	*FREQ*
(German) an armored vehicle or tank	panzer	25
a heater and storage tank to supply heated water	hot-water heater hot-water tank water heater	– – 79
a large gas-tight spherical or cylindrical tank for holding gas to be used as fuel	gas holder gasometer	5 12
a tank for holding gasoline to supply a vehicle	gas tank gasoline tank	10 –
a tank or pool or bowl filled with water for keeping live fish and underwater animals	aquarium fish tank marine museum vivarium	988 95 – 7

8. Finding the frequency of more and less specific words

In addition to looking for more specific or general words relating to a specific concept, it is also possible to use *WordNet* to find the parts of a given whole (meronyms) or see what larger unit a given word is part of (holonyms). This may be useful for language learners as well. Many textbooks will simply give a list of words in a given semantic domain ('body', 'parts of a house'), with a simple one word equivalence from the first language, e.g. [waist/*cinturón*]. The language learner, however, may want to move beyond such a simple list to see a more detailed definition, as well as see which words mean essentially the same thing, and see the frequency of competing items. Again, this would be quite easy with our joint *BNC*/*WordNet* database.

The query syntax is quite straightforward. The following two symbols are used to extract hyponym and hypernym entries from WordNet, and enter them in as part of the *BNC* query:

(11) [**@**whole $_x$] words that are a parts of [whole $_x$] (meronyms)
[**&**part $_x$] words that contain [part $_x$] (holonyms)

For example, to find the meronyms (parts of whole) for the whole = [body], the user would enter the following into the search form:

(12) [@body]

The user then sees a display similar to the following:

Table 15: [BODY] ("Includes part") (partial listing).

SYNSET	*WORDS*
(anatomy) a muscular partition separating the abdominal and thoracic cavities	diaphragm, midriff
a ball-and-socket joint between the head of the humerus and a cavity of the scapula	articulation humeri, shoulder, shoulder joint
a human limb	arm, leg
a protruding abdomen	belly, paunch
any of several muscles of the trunk	serratus, serratus muscles
the angle formed by the inner sides of the legs where they join the human trunk	Crotch, fork
the system of glands that produce endocrine secretions that help to control bodily metabolic activity	endocrine system
the fleshy part of the human body that you sit on	arse, ass, backside, behind, bottom, bum, buns, butt, buttocks, can, derriere, fanny, fundament, hind end, hindquarters, keister, nates, posterior, prat, rear, rear end, rump, seat, stern, tail, tail end, tooshie

As before, the users select the desired synsets, and can then see they frequency for each item in the *BNC*, as in the following:

Table 16: *BNC* frequency counts for parts of body (partial listing).

WORD	*FREQ*	*WORD*	*FREQ*	*WORD*	*FREQ*	*WORD*	*FREQ*
head	37940	waist	1381	torso	251	paunch	56
body	31421	trunk	1123	heads	214	posterior	44
back	20130	rear	1107	fanny	206	butts	42
arm	18933	fork	1018	diaphragm	157	waistline	42
leg	11176	belly	903	prat	119	pressure point	37
shoulder	8203	cavity	534	bum	510	caput	27
middle	5758	bum	510	butt	421	fundament	11
neck	5615	butt	421	haunch	106	tush	11
chest	3743	stern	371	cervix	98	derriere	7
cheek	3228	ass	350	backs	92	serratus	5
tail	3139	backside	311	crotch	73	midsection	4
stomach	2983	rump	311	behind	70	dorsum	2
hip	1791	abdomen	296	midriff	67	necks	2
can	1429	buttock	270	thorax	66	shoulder joint	2

If the user wishes to focus on just one of the synsets, s/he can easily do so. For example, the user may select just the synset [the fleshy part of the human body that you sit on], and would then see just the following:

Table 17: Terms for [the fleshy part of the human body that you sit on].

WORD	*FREQ*	*WORD*	*FREQ*	*WORD*	*FREQ*
seat	10464	stern	371	prat	119
bottom	5848	ass	350	behind	70
tail	3139	backside	311	posterior	44
can	1429	rump	311	butts	42
rear	1107	fanny	206	fundament	11
arse	553	rear (a) end	199	tush	11
bum	510	tail end	122	derriere	7
butt	421				

9. More advanced queries

Most of the examples that we have seen involve single-word queries. For example, we have extracted synonyms, more specific words, less specific words, parts of whole, and whole for parts from WordNet, and have then seen the frequency of each of the words in these lists in the *BNC*. Recall, however, that in Sections 5 and 6, we used the WordNet information as part of a phrase. For example, we searched for [=wicked] [nn*] (all synonyms of [wicked] followed by a noun) or [=large] [=amount] (all synonyms of [large] followed by all synonyms of [amount].

The ability to embed WordNet lists into phrase queries can be carried out for hypernyms, hyponyms, meronyms, and holonyms as well. For example, all of the following queries are possible:

Table 18: More advanced phrase-level queries.

QUERY SYNTAX	*MEANING*	*EXAMPLES*
[av0] [=hot] [<food]	adverb + synonym of [hot] + hyponym of [food]	really hot pizza incredibly spicy chicken
[<eat] the [<food]	Hyponym of [eat] + the + hyponym of [food]	devour the steak munch the chips
br*k* my/his/her [@body]	forms of *break* + [my or his or her] + part meronyms of [body]	broke my nose breaks his ankle

One final extension of the query syntax is the ability for users to create "customized" or "user-defined" lists of words, which they can then re-use countless times in subsequent queries. These are already incorporated into the 100 million word Corpus del Español that I created in 2001 (see http://www.corpusdelespanol.org), and are a new feature in the BNC/VIEW interface. The basic idea is that users can create any number of lists of words that are morphologically, syntactically, or semantically related. They enter this list of words via the web-based interface, and can then re-use the list as part of subsequent queries – ten minutes or ten months later.

For example, if the user wants to focus on British English, s/he could modify the list of [body parts] from the American-based WordNet, and store this list of 80-100 words as his or her own list. Likewise, the user might wish to save just those synonyms of [beat] that refer to music, and then re-use this list. in subsequent queries Finally, in cases where WordNet is somewhat weak in terms of semantic fields, the user might want to create his or her own categories from scratch, such as [negative emotions] or [academic life].

Once created, these customized lists can then be easily incorporated into the standard query syntax. For example, suppose that [Jane.Smith] has created a list called [emotions] with 40-50 words like [happy, sad, excited, relieved], or that a user called [LingProf] has created a list of 80-100 computer-related terms like [mouse, CPU, laptop, web, spreadsheet]. The user then includes a reference to this list as part of the query string, e.g.

(13) [av0] [Jane.Smith:emotions] *really happy, extremely angry*
[vv*] the [LingProf:computers] *clicked the mouse, surfed the Web*

As we can see, the use of a relational database architecture means that we can add any number of levels of annotation or different modules, and then we simply create links between these to allow for very powerful queries.

10. More advanced queries involving register variation

Because the *WordNet* and *BNC* databases are in relational database form, they can easily be joined together with other databases in a way that would otherwise be quite impossible. For example, at http://view.byu.edu/, it is also possible to examine register variation in the *BNC*, with greater ease than with perhaps any other interface.

At the most basic level, one can find the frequency of a given word in all 70 registers of the *BNC* (e.g. for *kick* [v]), and sort the results by the relative frequency in these different registers. More advanced queries allow one to find which words occur with greater or lesser frequency in different registers. For example, in less than three seconds one could find:

- all adjectives that occur much more frequently in tabloid than broadsheet newspapers (e.g. *heartbroken*, *adults-only*, *hunky*, *smashing*);
- all verbs that occur more frequently in the [spoken:courtroom] section of the corpus than in other spoken registers (*harbouring*, *handcuffed*, *ascertained*, *disallowing*), or
- all nouns occurring more frequently in [w_ac_medicine] than in other academic texts (*colitis*, *ulcer*, *biopsy*, *gastrin*).

Because the 'register' databases contain the frequency of each word and *n*-gram in each of the 70 distinct registers of the *BNC*, it would be quite easy to combine this with the *WordNet* database. To follow up on some of the queries already presented in this paper, one could find, for example:

- which phrases with [synonym of *bad*] + [noun] occur more often in spoken than in written English (e.g. perhaps *big problem* or *bad luck*);
- more specific verbs expressing the concept [to walk] that are more common in fiction than in academic writing (e.g. *stagger*, *lurch*, and *trudge*);
- words relating to parts of the body that are more common in academic writing than in fiction or spoken English (e.g. perhaps *endocrine system*, *cervix*, or *thorax*).

One can easily imagine how such a corpus might be of value to both native speakers and to language learners. For example, language learners often encounter long lists of thematically-related vocabulary in a textbook, but there is often little indication of which words occur in more or less formal registers. As a result, they end up using a word like *buttocks* in very informal conversation, or conversely, a word like *fanny* or *bum* in more formal writing. The type of database that we have described would help them to easily check the correct register for a wide range of vocabulary items in different registers with one simple query.

11. Some limitations

In spite of the advantages of this approach, we should end the discussion by briefly considering two potential problems with the joint *BNC*/*WordNet* database. First, we should recognise that the database tends to overgenerate results. For example, [bottom] and [tail] show up as meronyms (parts of the whole) for [body], but probably only a small percentage of the occurrences of these words in the *BNC* refer to [the fleshy part of the human body that you sit on]. There is no way to automatically disambiguate word sense, as has been done manually, for example, with the 1 million word *SEMCOR Corpus* (cf. Landes et al. 1998), which has *WordNet* synset annotation for each word in the one million word *Brown Corpus*. However, by combining the one [body] "slot" in the search string with another word, the necessary disambiguation often occurs naturally. For example, if we add [broken] before [body]/[includes part], then nearly all of the hits do refer to the body: *broken leg*, *broken arm*, *broken neck*.

A second note concerns the content of *WordNet*. Because it is based primarily on American English and the *BNC* of course comes from the UK, there may at times be mismatches between the two. For example, the results from a search of the synonyms of [truck] fails to include [lorry], and [sweets] does not appear as a synonym of [candy]. In spite of the fact that *WordNet* is the most powerful semantically-based corpus in existence, there are still many gaps in its entries, especially to the degree that they do not reflect American English. This is why the 'user-defined' queries described in the previous section are so powerful. The end user can use *WordNet* as a starting point to create customised semantic fields and hierarchies, which can be modified in any number of ways for a particular end use.

12. Conclusion

We believe that the approach discussed here offers real advantages for a semantically-based investigation of the *BNC*. The relational database approach offer endless possibilities in terms of combining together frequency information for words and phrases, register variation, and semantic information. It is simply a matter of having the corpus creator write the correct SQL commands to perform the necessary SQL [JOIN]s from each table and database. One other advantage of this approach is that the queries are very fast. With the 100 million word *Corpus del Español* or the 100 million word joint *BNC*/*WordNet* database described here, even the most complex queries take no more than 3-4 seconds. In summary, our hope is that this database shows how – with the correct database architecture – one can begin to carry out advanced semantically-based, frequency-oriented investigations of large, diverse corpora.

References

Aston, G. and Burnard L. (1998), *The BNC handbook. Exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press.

Burnard, L. (2002), 'The BNC: where did we go wrong?', in: B. Kettemann (ed.) *Teaching and learning by doing corpus analysis*. Amsterdam: Rodopi. 51-70.

Davies, M. (2003), 'Relational *n*-gram databases as a basis for unlimited annotation on very large corpora', in: K. Simov (ed.) *Proceedings from the workshop on shallow processing of large corpora*. Lancaster: Lancaster University. 23-33.

Fellbaum, C. (ed.) (1998), *WordNet: an electronic lexical database*. Cambridge (MA): MIT Press.

Landes S., C. Leacock and R. Tengi (1998), 'Building semantic concordances', in: C. Fellbaum (ed.) *WordNet: an electronic lexical database*. Cambridge (MA): The MIT Press. 199-216.

This page intentionally left blank