Abstract

In the course English Linguistics 1, in the first semester of the Aarhus University BA in English, students are introduced to the British National Corpus (BNC), which is "a 100 million word collection of samples of written and spoken language from a wide range of sources, designed to represent a wide cross-section of British English, both spoken and written, from the late twentieth century".

In my presentation, I will show how we use the BNC in English Linguistics 1, and in doing so, I hope to illustrate how useful the BNC can be to both students and teachers of English, especially (but not only) those of us who are not native speakers. We use the Brigham Young University interface to the BNC (http://corpus.byu.edu/bnc/), and one of the advantages of this is that the same interface can also be used to access the 520-million-word Corpus of Contemporary American English (COCA, http://corpus.byu.edu/coca/), which thus allows comparisons between the two major variants of present-day English.
1. Access to the BNC

Go to http://corpus.byu.edu/bnc/, the Brigham Young University interface with the British National Corpus (BNC). Here you can enter your search directly on the home page.

You will almost certainly have to register to use the web site, as only a few searches are possible without registration. Registration is free, and it does not generate spam. Even with registration, however, you may not get very many searches from within Aarhus University (as all of AU counts as one internet address), but it should be different from outside AU.

The reason we use BNC and other corpora is that they provide a way of finding "real life" examples instead of inventing them or looking in dictionaries. There is of course absolutely nothing wrong with dictionaries or with invented examples (especially not if they are checked by native speakers), but we would like our students to have many different tools at their disposal. Linguists often use corpora in their research, and not only (but also) when people question whether native speakers "really" do use a particular syntactic structure (cf. e.g. Vikner 2015, 2016, Wood & Vikner 2013).

2. Simple searches and wildcards

There is a search box labelled "word(s)", right under "search string", at the top left of the main BNC page, and here you can type your search. Type in e.g then and press the search button (or just press "enter").

(1) then

Result: all cases of exactly this word in BNC, 153,261 in total.

There are two wildcard characters: ? is exactly one character. * is one or more characters.

(2) a. th?n
b. th*n

Result: then, than, thin, ...
Result: then, than, thin, thrown, thirteen, thermoregulation ...

(Both wildcards may also be useful e.g. for solving cross-word puzzles.)

For one of the written exercises in the course English Linguistics 1, students need to find out the properties of three given verbs (e.g. whether a given verb is transitive, ditransitive or intransitive). Let us therefore look at the verb visit, which was part of this exercise a couple of years ago.

(3) visit

Result: all cases of exactly this word in BNC, 11,966 in total.

3. Word classes

Now you can scroll down and look at the example sentences containing visit. To see the example sentences from a particular search, you may have to click the word you searched.

You may notice that many of the examples are examples of visit as a noun rather than as a verb. E.g.

(4) A panic-stricken Julie immediately cancelled a planned visit to the Cannes Film Festival ...
In order to find only examples of the verb *visit*, you need to restrict your search to the appropriate word class. Click on the question mark immediately right of the search box to get more information. This gets you to the page [http://corpus.byu.edu/bnc/help/syntax_e.asp](http://corpus.byu.edu/bnc/help/syntax_e.asp), which is extremely useful.

Right now we are interested in word class = "part of speech" = PoS. The full set of PoS tags is available on a different page, [http://ucrel.lancs.ac.uk/claws7tags.html](http://ucrel.lancs.ac.uk/claws7tags.html). These tags can be very detailed:

(5) PPGE nominal possessive personal pronoun (e.g. mine, yours)
PPH1 3rd person sing. neuter personal pronoun (it)
PPHO1 3rd person sing. objective personal pronoun (him, her)
PPHO2 3rd person plural objective personal pronoun (them)
PPHS1 3rd person sing. subjective personal pronoun (he, she)
PPHS2 3rd person plural subjective personal pronoun (they)
PPIO1 1st person sing. objective personal pronoun (me)
PPIO2 1st person plural objective personal pronoun (us)
PPIS1 1st person sing. subjective personal pronoun (I)
PPIS2 1st person plural subjective personal pronoun (we)
PPX1 singular reflexive personal pronoun (e.g. yourself, itself)
PPX2 plural reflexive personal pronoun (e.g. yourselves, themselves)
PPY 2nd person personal pronoun (you)

PoS tags can be combined with the wildcards from before, and this is actually something that you can get the BNC to do for you: Click "POS LIST", which is three lines below "search string", at the middle of the left edge of the main BNC page. This will give you a pull down menu of some of the combinations between PoS tags and wildcards. Here is a list of the most interesting such combinations:

(6) [at*] any article
[c*] any conjunction
[cs*] any subordinating conjunction
[d*] any determiner (excl. articles)
[i*] any preposition
[j*] any adjective
[m*] any number
[n*] any noun
[nn*] any common noun
[np*] any proper noun
[p*] any pronoun
[pn1*] any indefinite/quantifying pronoun
[pnq*] any *wh-*pronoun
[pp?o*] any clearly accusative pronoun
[pp?s*] any clearly nominative pronoun

[ppx*] any reflexive pronoun
[r*] any adverb
[v*] any verb
[v?d*] any simple past tense verb form
[v?g*] any present participle (-ing-form)
[v?n*] any past participle verb form
[v?z*] any 3.sg. verb form
[vb*] any form of *be*
[vd*] any form of *do*
[vh*] any form of *have*
[vm*] any modal auxiliary
[vv*] any main verb (= any lexical verb)
[y*] any punctuation sign ( , ; . : ? ! # etc.)

Here is how to specify the word class of a particular word that we are looking for:

(7) [visit.[v*]] Result: 4794 cases of the verb *visit*.

Like (3), (7) still looks up the word *visit*, but the [v*] part says that we are only interested in verbs, and the dot says that it is *visit* which should be a verb.

(With a space instead of a dot, we would have searched for the word *visit followed* by a verb. With spaces around it, a dot is just a full stop. Try them and see.)
4. Exact word vs. lemma

The result of (7) was that there are 4794 examples of the verb visit in BNC. However, when you look at the examples, you may notice that all of the examples contain the verb visit in exactly that form. There are no examples at all containing visited, visiting, or visits. To remedy that, we need to put [] around visit:

(8)  [visit].[v*]  Result: 4794 visit
       4086 visited
       1741 visiting
       370 visits

The difference between visit and [visit] in the search box is the difference between "exact word" and "lemma" (Danish: 'opslagsord').

The result is a choice between four different searches, of the four verbal forms visit, visited, visiting, and visits, with a total of 10991 results. Now click one of these four forms to see the examples.

5. The actual example sentences

Once you have a list of results, clicking on the reference letters/numbers to the left of the example will reveal more of the context, and also the source of the example.

It is often interesting to note whether the example is from spoken or written language, or from formal or informal language, etc. You can see the exact genre at the left edge, ranging from sermons, S_sermon, to e-mails, W_email. You can also restrict searches to one or more of these genres, right below the search button.

I referred above to a written exercise in English Linguistics 1, where students need to find out the properties of three given verbs (e.g. whether a given verb is transitive, ditransitive or intransitive). As part of this, students have to submit examples sentences that illustrate these properties. If the verb in question had been visit, part of a student answer could have looked like the following:

<table>
<thead>
<tr>
<th>visit: V, [ ___ NP] (transitive)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(89) a. Only 55 per cent of registered voters visited the ballot box in June 1987. (BNC)</td>
</tr>
<tr>
<td>b. Derry is no stranger to Inchcolm, having visited the island many times as a boy. (BNC)</td>
</tr>
<tr>
<td>c. Patrick Taylor has visited the 400 places he describes in the pocket-sized The Gardener's Guide to Britain. (BNC)</td>
</tr>
</tbody>
</table>

Here we see that visit requires one and only one object. In the first example, the object is [the ballot box], in the second [the island], and in the third [the 400 places he describes in the pocket-sized The Gardener's Guide to Britain].

Time adverbials like [in June 1987] or [many times], on the other hand, are optional.
6. Synonyms
Let us now have a look at synonyms. The following will find the most common combinations of red + a noun:

(9)  \[\text{red \[n^*\]}\] Result: 436 Red Cross, 224 red wine, 218 red hair, ...

However, if you want to look for many different colours at the same time, you need to try

(10)  \[\text{\[=color\] \[n^*\]}\] Result: 1100 white paper, 822 blue eyes, 737 white house, ...

With this feature, the BNC can also be used to find synonyms:

(11)  \[\text{\[=blue\]}\] Result: ..., 3241 sad, 1822 unhappy, 1570 depressed, ...
      ... 211 turquoise, ... 135 indigo, 103 azure, ...

7. Alternatives in the search string
If you want to see which articles are used with series, use \[\text{[at^*]}\]:

(12) \[\text{[at^*]} \text{series}\] Result: 6149 a series, 1218 the series, 5 no series, ...

This shows that \text{series} can be used after a singular determiner, but it does not guarantee that it can also be used with a plural one. Instead we can use \[\text{[d^*]}\], which refers to determiners:

(13) \[\text{[d^*]} \text{series}\] Result: 271 this series, ... 11 these series, ...

However, now the articles from before are not included. We can get them both on one page with a /:

(14) \[\text{[at^*]/[d^*]} \text{series}\] Result: 6149 a series, 1218 the series, 271 this series, ...
      ... 11 these series, ... 5 no series, ...

However, / can only combine expressions of the same type, i.e. either \text{/this} or \[\text{[at^*]/[d^*]}\], but not \text{/[d^*]} (where it just ignores the second expression).
8. UK vs. US spelling (BNC vs. COCA)

The / can also be used with the words themselves:

(15) colour/color Result: 11135 colour, 112 color.

At this point, we might be interested in comparing British English to American English. To do so, open the Corpus of Contemporary American English (COCA, http://corpus.byu.edu/coca/) in a different window.

Whereas the BNC has been fixed at 100 million words since 1994, COCA is growing every year. Right now it is 520 million words, and it is thus "the largest corpus of American English currently available". COCA was created by Mark Davies, Professor of Corpus Linguistics at Brigham Young University, who also created the interface with BNC that we are using right now. BNC itself was a collaboration between six parties (Oxford University Press, Longman, W. & R. Chambers, University of Oxford, Lancaster University, and the British Library).

Here is the result of the BNC/COCA comparison:

(16) colour/color BNC: 11135 colour, 112 color
COCA: 1256 colour, 91779 color

This is a good way of testing how solid such UK/US differences really are:

(17) travelling/traveling BNC: 3318 travelling, 12 traveling
COCA: 832 travelling, 15066 traveling

(18) syllabi/syllabuses BNC: 11 syllabi, 231 syllabuses
COCA: 347 syllabi, 27 syllabuses

(19) generalization/generalisation BNC: 269 generalization, 150 generalisation (!)
COCA: 1783 generalization, 19 generalisation

9. UK vs. US lexical choice

We can also compare more syntactic choices in the two corpora:

(20) [comparison] with/to BNC: 1203 comparison with, 232 comparison to
COCA: 2780 comparison with, 2369 comparison to

(21) [compare] [pp*] with/to BNC: 245 compare it with, 95 compare it to (!)
COCA: 741 compare it with, 2119 compare it to
10. UK vs. US syntax: negation of *I am*

Let us look at the different options that speakers have when they want to produce a negated declarative clauses with *I am*, e.g. *I am not going* or *I am not sure*:

(22)

<table>
<thead>
<tr>
<th></th>
<th>BNC (100 million words)</th>
<th>COCA (520 million words)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>AM NOT</strong></td>
<td>I + am + not + going/sure</td>
<td>15.7 % (964 cases)</td>
</tr>
<tr>
<td><strong>AMN'T</strong></td>
<td>I + am + n't + going/sure</td>
<td>0.0 % (0 cases)</td>
</tr>
<tr>
<td>'M NOT</td>
<td>I + 'm + not + going/sure</td>
<td>80.7 % (4948 cases)</td>
</tr>
<tr>
<td><strong>AREN'T</strong></td>
<td>I + are + n't + going/sure</td>
<td>0.1 % (6 cases)</td>
</tr>
<tr>
<td><strong>AIN'T</strong></td>
<td>I + ai + n't + going/sure</td>
<td>3.5 % (217 cases)</td>
</tr>
<tr>
<td><strong>(TOTAL)</strong></td>
<td></td>
<td>(100.0 %) (6135 cases)</td>
</tr>
</tbody>
</table>

(Search e.g. [I’m not [v?g*]/[j*]]

*I'm = I + 'm, ain't = ai + n't, going etc. = [v?g*], sure etc. = [j*])

- There is slightly more contraction in COCA than in BNC.
- In both corpora, *I'm not going* is the main option with contraction.
- In both corpora, *I ain't going* is also possible, but fairly limited.
- In both corpora, the options *I am't going* or *I aren't going* seem to be impossible:

Consider now what happens in negated questions with *I am*, e.g. *Am I not going?* or *Am I not lucky?:

(23)

<table>
<thead>
<tr>
<th></th>
<th>BNC (100 million words)</th>
<th>COCA (520 million words)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>AM NOT</strong></td>
<td>am + I + not + going/lucky</td>
<td>42.2 % (19 cases)</td>
</tr>
<tr>
<td><strong>AMN'T</strong></td>
<td>am + n't + I + going/lucky</td>
<td>0.0 % (0 cases)</td>
</tr>
<tr>
<td><strong>AREN'T</strong></td>
<td>are + n't + I + going/lucky</td>
<td>46.7 % (21 cases)</td>
</tr>
<tr>
<td><strong>AIN'T</strong></td>
<td>ai + n't + I + going/lucky</td>
<td>11.1 % (5 cases)</td>
</tr>
<tr>
<td><strong>(TOTAL)</strong></td>
<td></td>
<td>(100.0 %) (45 cases)</td>
</tr>
</tbody>
</table>

- There is much less contraction in these questions than in the declarative clauses above.
- There is much less contraction in COCA than in BNC.
- In both corpora, *Aren't I going?* is the main option with contraction, even though *aren't* was not possible in declarative clauses with *I*.
- In both corpora, *Ain't I going?* is also possible, but fairly limited, especially in COCA.
- In both corpora, the option *Amn't I going?* seems to be impossible:
11. UK vs. US syntax: questions with *have*

Both in BNC and in COCA, we find three different ways of asking questions with *have*.

**Have you any money?** (the 'old-fashioned' one):

(24) a. Well that's when I said to him, I said *have you* any outside help, and he said well really, no. (BNC, conversation, 1992)
b. *Have you* any sense of what it's going to be like for you when she's not here? (COCA, CBS SunMorning, 1997)

Quirk et al. (1985:131): (24) "is now somewhat uncommon, particularly in the past tense".
Huddleston & Pullum (2002:1139): (24) tend "to sound relatively formal or old-fashioned".

**Have you got any money?**

(25) a. And *have you got* any message for this year's winner? (BNC, broadcast news, 1990)
b. *Have you got* any idea how much training that took? (COCA, NBC Today, 2009)

**Do you have any money?**

(26) a. *Do you have* any deposit at all? (BNC, conversation, 1991)
b. *Do you have* any thoughts on the style of wedding dress she may go for? (COCA, CBS Early show, 2011).

Here are the results from the two corpora:

<table>
<thead>
<tr>
<th></th>
<th>BNC</th>
<th>COCA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(100 million words)</td>
<td>(520 million words)</td>
</tr>
<tr>
<td>'OLD' CONSTR.</td>
<td>have + SUBJ + any</td>
<td>19.1 % (489 cases)</td>
</tr>
<tr>
<td>got CONSTR.</td>
<td>have + SUBJ + got + any</td>
<td>38.5 % (985 cases)</td>
</tr>
<tr>
<td>do CONSTR.</td>
<td>do + SUBJ +have + any</td>
<td>42.3 % (1082 cases)</td>
</tr>
<tr>
<td>(TOTAL)</td>
<td>(99.9 %) (2556 cases)</td>
<td>(99.9 %) (13197 cases)</td>
</tr>
</tbody>
</table>

(Search e.g. [[have] [ppy]/[pp?s*] got the/a/an/no/every/this/these/that/those/any/some] the subject = [pp?/s*]/[ppy], where [pp?/s*] = /he/she/we/they  and  [ppy] = you, the object = the/a/an/no/every/this/these/that/those/any/some)

Bearing in mind that (i) the searches are only for pronominal subjects (except *it*), and (ii) the two corpora are not directly comparable, and (iii) we have not tested for statistical significance, the above figures do show some differences between the two corpora:

- The 'old' construction (*have you any money?*) is much more frequent in the BNC than in COCA.
- The *do*-construction is the most frequent, not just in COCA but also in BNC.
- COCA favours *do*-support almost to the exclusion of the other two constructions.
- BNC is more varied and there appear to be differences depending on the object.

Let us now try to narrow down exactly how the object type makes a difference. Depending on the object (e.g. abstract *conversation* vs. concrete *money*), the interpretation of *have* is either dynamic or stative/possessive, which again makes a difference e.g. when it comes to progressive tense:

(28) a. DYNAMIC    *We're* having a conversation. (BNC, spoken 1993)
b. STATIVE    *We're* having some money.
Using the criterion of being compatible with a progressive tense of have, as in (28)a, I found a number of objects that require dynamic have: affair/party/fight/conversation/effect/nightmare/row.

Using these objects, we can now see that dynamic have is the same in both corpora, i.e. 100% of the questions have do you have:

<table>
<thead>
<tr>
<th></th>
<th>BNC (100 million words)</th>
<th>COCA (520 million words)</th>
</tr>
</thead>
<tbody>
<tr>
<td>‘OLD’.</td>
<td>0.0% (0 cases)</td>
<td>0.0% (0 cases)</td>
</tr>
<tr>
<td>got</td>
<td>0.0% (0 cases)</td>
<td>0.0% (0 cases)</td>
</tr>
<tr>
<td>do</td>
<td>100.0% (8 cases)</td>
<td>100.0% (55 cases)</td>
</tr>
<tr>
<td>(TOTAL)</td>
<td>(100.0%) (8 cases)</td>
<td>(100.0%) (55 cases)</td>
</tr>
</tbody>
</table>

To get the stative/possessive version of have only, I used concrete objects: money/books/brothers/sisters/apples/flowers.

We can now see that stative/possessive have is very different in the two corpora:

<table>
<thead>
<tr>
<th></th>
<th>BNC (100 million words)</th>
<th>COCA (520 million words)</th>
</tr>
</thead>
<tbody>
<tr>
<td>‘OLD’.</td>
<td>25.6% (11 cases)</td>
<td>0.0% (0 cases)</td>
</tr>
<tr>
<td>got</td>
<td>48.8% (21 cases)</td>
<td>4.4% (2 cases)</td>
</tr>
<tr>
<td>do</td>
<td>25.6% (11 cases)</td>
<td>95.6% (43 cases)</td>
</tr>
<tr>
<td>(TOTAL)</td>
<td>(100.0%) (43 cases)</td>
<td>(100.0%) (45 cases)</td>
</tr>
</tbody>
</table>

We can now refine our conclusions, taking the dynamic/stative difference into account:

- The do-construction is the only option with dynamic have in both BNC and COCA.
- The do-construction is almost the only option in COCA both with dynamic and stative/possessive have.
- BNC is very varied with stative/possessive have.
- The ‘old’ construction (have you any money?) is not found at all in COCA, and in BNC it is only found with stative/possessive have (but there, it is fairly robust).
- The got-construction is the dominant option with stative/possessive have in BNC.

12. Conclusion
You have seen we use BNC and COCA in English Linguistics 1, and in more advanced courses in English linguistics. I hope also to have shown you how useful the two corpora can be to both students and teachers of English, especially (but not only) those of us who are not native speakers.

In the Wikipedia-article on BNC (https://en.wikipedia.org/wiki/British_National_Corpus), it says
Useful only in studying grammatical patterns which have distinctive lexical correlates

Despite being an excellent source of lexical information, the BNC can only really be used to study a limited set of grammatical patterns, particularly those which have distinctive lexical correlates. While it is easy enough to find all the occurrences of "enjoy", and to sort them according to the part-of-speech category of the following word, it is impossible to find all cases of verbs followed by a gerund, since the SARA index does not include part-of-speech categories such as "all verbs" or "all V-ing forms".

This may very well be true for the "old" interface, but in the interface that I have demonstrated today (http://corpus.byu.edu/bnc/), this criticism is not well-founded at all (as was demonstrated above).

However, we always have to keep in mind that

- Although the BNC contains 100 million words, it is still a finite language sample (of British usage) and all the possibilities that exist in the language need not be represented. The same goes for the 520 million words in COCA.
- The word class tagging has been done automatically, i.e. by a machine and not by a linguist. The corpus is not foolproof and the computer won't think for you.

One such example of a mistake in tagging became clear last autumn, when we told the students to describe the verb *sparkle*. Because of this, we might use it for the next few years:

(31)  \( \text{sparkle.}[v^*] \)  
Result: 0 cases of *sparkle* as a verb form.

However, students were told to look for \([\text{sparkle}]\), not for *sparkle*, and so if they had done what they were told, they would have been fine:

(32)  \( \text{[sparkle].}[v^*] \)  
Result: 182 cases of *sparkling*, 165 cases of *sparkled*.

It is a mistake in tagging on the part of the BNC, however. There clearly are cases of *sparkle* as a verb form:

(33)  \( \text{they/to sparkle} \)  
Result: 11 cases of *to sparkle*, 3 cases of *they sparkle*.

References


