

# An Introduction to the Use of Corpora in Teaching and Studying English

James Thomas

## Exploring the BNC with the Word Sketch Engine

### What is a corpus?

A *corpus* (plural *corpora*), in the context of linguistics, is a body of texts that have been assembled in machine readable form for language study. The **British National Corpus** (BNC) represents English through a representative variety of texts created from 1960 to 1994.<sup>\*)</sup> It is a sample of the language — a snapshot. The texts include written language and transcripts of speech. For pedagogical purposes, it is important that it contains attested language, not texts created for instructional purposes. Authentic input is regarded as a better model for students, despite such problems as obscure vocabulary, cultural references and even errors that naturally occur. It is the teacher's task to adapt the task, not the .<sup>†)</sup>

### What is a concordancer?

This is a program that “reads” a corpus and processes the data — it displays the concordances and summarizes the data.

\*) 94 % of texts are in the 1985–1993 range.

†) See pages 8–9 for answers to tasks.

### What use is it to teachers?

As will be demonstrated in these worksheets, useful information can be derived from a corpus. Essentially, you start with a question and have the concordancer assemble the data from which you may derive your answer. If the tasks are suitably adapted to the students' level, they can successfully undertake *discovery learning* tasks themselves. As an approach to language learning, this is referred to as *Data Driven Learning*. Furthermore, since the concordancer locates very focussed sets of sentences, a teacher can use these for creating worksheets that illustrate words and grammar structures, and in activities and exams.

**data** → **information** → **knowledge**

### Logging in

Go to [http://www.sketchengine.co.uk/bnc/reg/reg.cgi/registration\\_form/](http://www.sketchengine.co.uk/bnc/reg/reg.cgi/registration_form/)

and register to use the sampler program. To start using it go to:

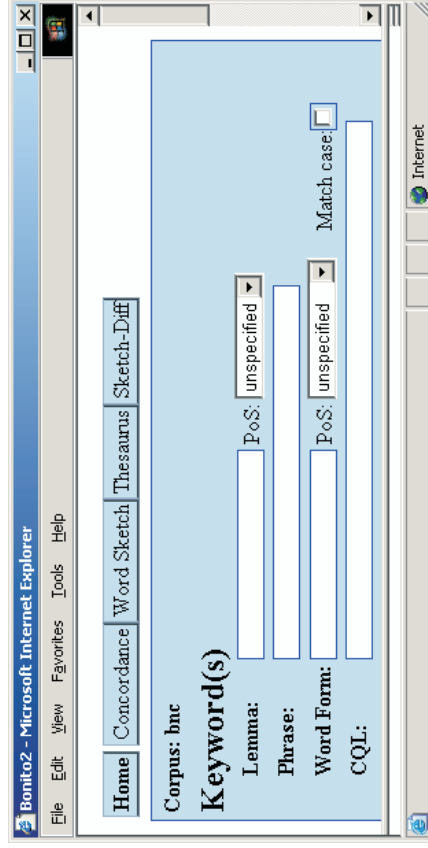
<http://www.sketchengine.co.uk/sampler/>



## Screen One – Top Half: Key Words

This is the principal page for starting a search.  
It consists of:

- » *fields* into which words are typed
- » *drop lists* from which parts of speech are selected
- » *buttons* which navigate to other components of the program or perform actions



When you click the *Make Concordance* button, the program (a **concordancer**) searches the **corpus** for all instances of the **lemma** and presents each of them on a separate line (a **concordance**) with the search item (the **node**) down the centre. This format is referred to as KWIC, *Key Word In Context*.

## Some terms

### Lemma

The set of different forms of a word, such as the inflected forms of a verb, e.g., *sing, sings, sang, sung, singing* are one lemma, *boy, boys* another.

### Phrase

A string of words, e.g., *holding hands, under the bridge, waste not*.

### Word form

A particular form that occurs in a lemma. The verb *to be* has eight word forms.

### CQL

*Corpus Query language* for complex queries requiring formal query syntax, e.g. this

```
[tag = "VMO"] [lemma = "have"] [lemma = "be"]
[word = "hoping"] [tag = "PRP"]
```

searches for any modal verb + perfect continuous of hope followed by a preposition. There are two in the BNC.

### PoS

*Part of Speech*. For example, specify *separate* as a verb or an adjective, if necessary.

## Task

Type 'separate' into the lemma field and do not specify part of speech.

Home | Concordance | Word Sketch | Thesaurus | Sketch-Diff | Frequency | Collocation

KWIC/Sentence | View options | Sample | Filter | Sort | Page 1 of 594 | Go

Corpus: **bnc**  
Hits: 11868  
[conc description](#) | [Next](#) | [Last](#)

**A01** charitable donations in any one tax year . . A **separate** leadlet gives details of that scheme **<p>**  
**A01** you could use the GIFT AID scheme ( see **separate** leadlet ) **<p>**What happens if I do not  
**A01** monthly , you could accumulate the money in a **separate** account and then convert this to a gift  
**A01** Deposited Covenant arrangement . This is quite **separate** from Gift Aid and details can be supplied  
**A03** Delegate for the Armed Forces is conducting a **separate** investigation . **<p><p>**Luis Miguel Solis  
**A03** who actually killed him ' **<p><p>**It is a **separate** tragedy altogether that , in a country where  
**A03** Belam ( LITTLE) whose political aim is a **separate** Tarul homeland ( Eeelan) in North-East  
**A03** British Airways plane landed , they were **separated** from the other passengers , put into a van  
**A04** interpretation and judgement are the subjects of **separate** chapters , where they are considered in  
**A04** **<p><p>**Other civilisations are treated in **separate** studies . One parallel to the scope of Janson  
**A04** Divisions in the book are unusually clear , **separate** chapters being given to topics such as function  
**A04** detail may be the way that the true can be **separated** from the false . William J. Irvine was curator  
**A04** circumstances alter cases . Some artists **separate** their art from their lives , some are articulate  
**A04** London . Divided into national schools , in **separate** publications , the gallery also publishes  
**A04** subsequently developed individual styles and **separate** interests , as is true of most groups of  
**A05** of the De Freitas set , which she firmly **separated** from the serious politics of the Caribbean  
**A05** Jane and Jimmy Ahmed . The encounters are **separated** by the riots and the second is followed  
**A05** have no place in it . ' An early passage **separates** this man , who does not want to be nothing  
**A05** cares of the earth . **<p><p>**Elsewhere Salim **separates** himself from the doers and makers of the  
**A05** history and psychoanalysis , can be found to **separate** . What matters is what happens when the

## Screen Two – Concordances

### Navigation buttons

Home, Concordance, Word Sketch, Thesaurus, Sketch-Diff.

### Process buttons

Frequency: Automatically summarize information.

Collocation: Create lists of collocates.

Sample: Create a random subset

**Sort:** Set the criteria for the order of concordances.

**Filter:** Reduce the found set of concordances according to further criteria.

### Data view

**KWIC/Sentence:** Toggle between KWIC and full sentences.

**View Options:** Set other information (meta-data) which appears with the concordances.

### Concordances

The KWIC format permits two readings, namely *horizontal* in which multiple contexts of the item can be observed, and the *vertical* in which the item's co-text is displayed.

One reading it does not permit, however, is left to right on consecutive lines as we normally read. This is because each concordance comes from a different text or part of a text.

### Task

### Program observations

What happens when you

» click one of the underlined node words (red) ?

» click one of the underlined text identifiers (blue) ?

» click the underlined Next (blue) ?

» click the KWIC/Sentence button? And again?

Click on *View Options*. Try this:

- » Attributes — select *tag* (a three letter PoS code)
- » Structures — select *text*
- » References — select *BNCdoc: date*

Click *Change View Options* button. What do you notice?

### Some observations of *separate*

1. How many occurrences of the lemma *separate* are in the BNC? \_\_\_\_\_
2. Is the full lemma present? \_\_\_\_\_
3. The tags starting *AJ* indicate that in this context the word is a/n \_\_\_\_\_ and the tags starting with *V* are \_\_\_\_\_.
4. Is *separate* more frequently used as an adjective or verb? \_\_\_\_\_
5. Are the nouns after *separate/AJO* singular or plural? \_\_\_\_\_
6. Is *separate + noun* typically preceded by a determiner? \_\_\_\_\_
7. What prepositions follow *separate/V*? \_\_\_\_\_
8. Is *separate/AJO* used both predicatively and attributively? \_\_\_\_\_
9. Are *separating* and *separated* adjectives as well as verb forms? \_\_\_\_\_
10. Do all the passive instances have the same tag? \_\_\_\_\_

11,868 concordances is obviously too much data for the human mind to process and so concordancers offer a range of ways of summarising this data. But before we look at that, let's look at how the initial search can be refined.

### Procedural task

Return to the previous screen (Alt + ←) and enter 'separate' into the lemma field and choose *verb*.

1. Click *Make Concordance*.  
How many concordances are there? \_\_\_\_\_  
Click on the *Frequency* button and then click on the *Node Forms* button. Is the full lemma present? \_\_\_\_\_
2. Return to the first screen and enter 'separate' into the lemma field and choose *adjective*. Click *Make Concordance*. How many concordances are there? \_\_\_\_\_
3. Is *separate* used more frequently as an adjective or verb? \_\_\_\_\_
4. Enter 'separating' into the Word Form field. Choose *adjective*.  
What do you notice about the part of speech of separating? \_\_\_\_\_  
Click *View Options*, choose *ambtag*, and you will see that the machine which assigned the tags was unable to decide — hence ambiguous. Repeat with 'separated'.
5. A phrase using *separate* is 'their separate ways'. Type this into the Phrase field and click *Make Concordance*. How many times does it occur? \_\_\_\_\_ What verb (lemma) almost always precedes it? \_\_\_\_\_ What others do? \_\_\_\_\_  
What typically follows it? \_\_\_\_\_

**Task using the Phrase field**

One often hears *way how + to do something*.

Type 'way how' into Phrase. What do you notice?

What can you observe looking at the following phrases?

» holding hands, under the bridge since, waste not

**Task using the Node Tags button**

Which of these words (WF or Lemma) function as more than one part of speech? Before you consult the corpus, think it through. Find a suitable illustrative sentence for each case.

house	home	target	hold
book	mince	wrong	chip
trust	pilot	trial	second
adult	holiday	found	find

**Task using the Node Forms button**

Which of these words do not have marked plural forms? Before you consult the corpus, think it through. Are words without plural forms always followed by singular verbs?

Find a suitable illustrative sentence for each case.

information	accommodation	experience
sheep	criteria	panic

**Consider these questions before looking****for evidence in the BNC**

1. Is *various* typically followed by a plural noun?
2. Is *whose* only used with people, or can it have inanimate antecedents?

3. What part(s) of the body do we *shrug*?

What do we *clench*?

4. Is *bound* typically followed by an object or a verb phrase?

5. Is *data* typically singular or plural? Is there another form of the word? Is it used frequently?

6. Are both spellings *blond* and *blonde* possible? If so, is there any difference?

7. What typically precedes *namely*? And what typically follows it?

8. Which PoS do *gosh* and *darn* have in common?

9. Is *manage* typically followed by a to–infinitive?

10. Is *found* a lemma, or a word form in the lemma of *find*? Or both?

11. Is *curiouser* the comparative of *curious*?

What word typically accompanies *curiouser*?

12. Is *dived* more common as the past of *dive* than *dove*?

13. What is the difference between *economic* and *economical*?

14. What prepositions most typically follow:

hint	keen	accompany
laugh	good	different

15. The preposition which follows a word can determine which of a word's meanings, or nuances, are intended. Consider the differences between *laugh* plus its various prepositions. Can you find suitable illustrative sentences for each one?

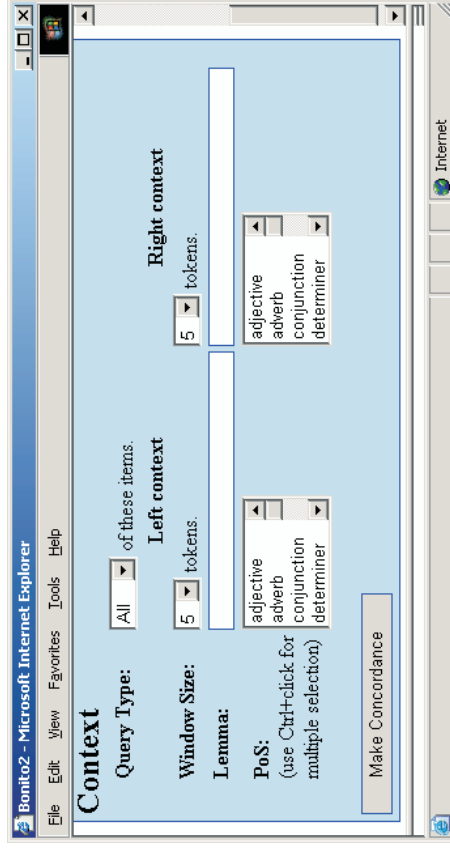
## Screen One – Lower Half: Context

### Context

Context refers to the environment in which a word or phrase is used; the extent (or range) of the environment is determined here by the number of tokens chosen.

### Query Type

When you specify something in the Lemma and/or PoS fields below, you can specify if you want *All*, *Any* or *None* of them to co-occur with the item(s) you have in the top half.



### Task

#### Variable phrases

Pop the question can be entered as a phrase in the top half, but if you want the lemma *pop* followed by the question, enter 'pop' as the lemma

at the top, and in Right Context 'the question'. What do these concordances tell us about this phrase?

Be used to — enter 'be' as the lemma at the top, and in Right Context 'used to'. If you additionally want to see the verb form that follows, choose *verb* in the PoS drop list. Or *noun* if you want to observe that. And don't forget the case of *get used to*.

### False friends

What nouns can be said to be *sympathetic*?

At the top, enter 'sympathetic' as the lemma. In Right Context choose *noun* in the PoS, 1 token. To summarize these 676 concordances, click the *Frequency* button and under First Level, choose *1R* (i.e. first to the right). Similarly: *pathetic*, *actual*.

### Checking a collocation

Type 'homework' into WF or Lemma and 'do' (lemma) in the Left Context. Repeat with 'make'.

Type 'write' into Lemma and 'exam' into Right Context. What do you notice?

Try again with 'make'.

Now type 'exam' into Lemma at the top. In Left Context, choose *verb*.

### Similes, e.g. 'high as a kite', 'sleep like a rock'

(adjective + 'as', or verb + 'like')

For a comprehensive list of potential similes, you can enter 'as a' into Phrase at the top. In the Left Context, choose *adjective*, and in the Right Context, *noun*. Set both left and right to 1 token.



To summarize this, click on *Frequency* and under First Level, click *1L* and under Second Level, *1R*. Select the Second Level radio button.

Investigate specific ones: e.g. enter 'poor as a' into Phrase at the top and in the Right Context chose *noun*, Token 1.

### Delexical verbs

How common is *to bathe* compared with *have a bath*? \_\_\_\_\_

Compare: *to photograph* something vs. *to take a photo* of something

Is *mushroom* a verb as well as a noun? \_\_\_\_\_

Is *to mushroom* the same as *to go mushrooming*? \_\_\_\_\_

### Phrasal verbs

Is *to wear out* a separable phrasal verb? *i. e.*, 'to wear something out', or 'to wear out something'? \_\_\_\_\_

Enter 'wear' in Lemma at the top. Enter 'out' into the Right Context, Tokens 5. Similarly: *help someone out*, *help out someone*.

### Candidates for fixed phrases

How can you tell if a string of words is a fixed phrase or not? Type it into Phrase and see how many you get, and see if there is anything significant in its environment. For example, *on the receiving end*, *far and wide*, *get over it*, *curiouser and curiouiser*.

### Candidates for errors

One often hears "May/Can I have a question".

Type 'have a question' into Phrase. Are any of concordances preceded by a modal? \_\_\_\_\_

For comparison, type 'can' into Lemma and 'I have' into Right Context, tokens 2. Repeat with 'may'.

Even more convincing might be typing 'I have a question' into Phrase. Additionally, you could try 'have' as Lemma and 'a question' in Right Context, tokens 2. What do you conclude?

An excellent source of known problems for Czech speakers of English is "English or Czenglish" (Sparling, 1989) which can now be accessed online at <http://nlp.fi.muni.cz/prjekty/lexdb/czeng.cgi>

### Conclusion

This brings to an end this introduction to using the Word Sketch Engine. Further information about the use of the Word Sketch Engine can be found at:

<http://www.sketchengine.co.uk/Sketch-Engine-User-Guide.htm>

Further information about the use of corpora and concordancers in general can be found in the portal at:

<http://www.fi.muni.cz/~thomas/EAP/concordancers/>

## Answers to Tasks

### Page 1

... not the *language*.

### Page 3

What happens when you

» click one of the underlined node words (red) ?

*Much more context is shown in the bottom frame.*

» click one of the underlined text identifiers (blue) ?

*Metadata about the text is shown in the bottom frame.*

» click the underlined Next (blue) ?

*The next page of concordances is shown.*

» click the KWIC/Sentence button ? And again ?

*This toggles between the KWIC display and full sentences.*

### Page 4

Click Change View Options button. What do you notice?

*You can select what you want to see on the screen, apart from the concordances themselves.*

Some observations of *separate*:

1. How many occurrences of the lemma *separate* are in the BNC? **11,868**

2. Is the full lemma present? **yes**

3. The tags starting *AJ* indicate that in this context the word is **an adjective** and the tags starting with *V* are **verbs**.

4. Is *separate* more frequently used as an adjective or verb? **adjective**

5. Are the nouns after *separate/AJO* singular or plural? **singular**

6. Is *separate + noun* typically preceded by a determiner? **yes**

7. What prepositions follow *separate/V*? **from, by, out, in**

8. Is *separate/AJO* used both predicatively and attributively? **mainly attributively**

9. Are *separating* and *separated* adjectives as well as verb forms? **yes**

10. Do all the passive instances have the same tag? **yes**

Procedural task:

1. Click *Make Concordance*.

How many concordances are there? **4,264 separate as verb**

Click on the *Frequency* button and then click on the *Node Forms* button. Is the full lemma present? **yes**

2. Return to the first screen and enter 'separate' into the lemma field and choose *adjective*. Click *Make Concordance*. How many concordances are there? **7,580**

3. Is *separate* used more frequently as an adjective or verb? **adjective**

4. ...What do you notice about the part of speech of separating?  
Repeat with 'separated'.

*Separating and separated — in these contexts are not always being used adjectivally, so the tagging seems to be wrong. The ambiguous tag setting shows how difficult tagging can be, and not only for automatic, computerized taggers.*

5. A phrase using separate is 'their separate ways'. Type this into the Phrase field and click Make Concordance. How many times does it occur? What verb (lemma) almost always precedes it?

*Their separate ways occurs 46 times and 43 times is preceded by some form of the verb to go.*

### Page 5

Type 'way how' into Phrase. What do you notice?

*Way how occurs only 18 times and never in the structure 'the way how to do something'. This structure is not good English.*

What can you observe looking at the following phrases?

- » under the bridge since

*This appears to belong to the idiom, a lot of water has gone under the bridge since then, though it is not used here in its canonical form.*

- » waste not

*This is typically completed with 'want not' — the full phrase being: waste not, want not.*

Task using the Node Tags button:

*All of these words function as more than one part of speech, the most frequent coupling being noun and verb.*