

Discovering English with Sketch Engine

a corpus approach to studying English

James Thomas

Dear Students,

On the following pages, you will find the first practical section of a book that I am writing for you, among others. There is a long introductory section which precedes this which will be available soon.

Working through this book has a number of goals. You should ...

1. learn a lot of language
2. learn a lot about language
3. learn how to extract linguistic information from linguistic data
4. be able to solve many language quandaries that you encounter in your work as teachers.

Register for Sketch Engine here: http://bit.ly/muni_rego. Although some parts of this software can be used without registration, much of what you need to do requires registration. The information on the registration pages needs to be read as you click through it. Choose a memorable password.

The structure of the book is a step by step guide to using Sketch Engine. There are hundreds of questions posed which need to be investigated online as your Sketch Engine skills expand. It is therefore best worked through cover to cover with Sketch Engine open and a notebook at hand. Every aspect of the software allows you to ask new questions or interpret data beyond the scope of the given question. It is therefore expected that the reader will be constantly going from the known to the unknown at many levels, and enjoying a rich learning experience.

Some KAA students are familiar with Sketch Engine, but this book is designed to take you in different directions. Arguably, the most important thing you will gain from this experience is learning what language questions can be asked.

Some parts of this book have been used in various training sessions and courses, but this is its first outing as a continuous course. Therefore, any feedback you have will be greatly appreciated. You can fill in a quick form here: http://bit.ly/DESKE_feedback115

Without further ado, let the fun begin!

James Thomas

1. Words and phrases

Three levels of querying

Most of the language questions and discussions in this book can be addressed in most of the corpora available in Sketch Engine. But as we are focussing on the core features of British English, DESKE mainly uses the **British National Corpus (Tree Tagger)** for its discovery tasks.

Open it by clicking on its name in the table on the Home Page⁶. It may be necessary to click the **All Corpora** button to see it the first time you visit the page. Once a corpus has been used, it appears in the top table of your Homepage, which shows the ten corpora you have most recently used. Clicking on the name of a corpus, takes you to a Query form with three levels:

1. The top level is **Query types** where we search for words and phrases, sometimes like an internet search. The top of the top is Simple Query, and we will explore this first.
2. The second level lets us filter a top level query within a **Context**. This means that we can search for something in the context of other words and/or parts of speech, e.g. searching for *dictionary* in the context of *look* and a preposition. See Question 156 p.103. Search for *kith* in the context of *kin* as we saw in **Error! Reference source not found.**
3. The third type allows us to specify the **Text Types** in which the word and/or context items occur, e.g. in spoken language, in academic texts, in legal texts, spoken by women.

Asking questions simply

A good research question is one that you can envisage finding an answer to – never just ask the question. Always imagine the possible answers and how you would find them out. (Wray and Bloomer 2012:1)

Simple Query field

We will click on the **Query types** button to toggle between the Simple Query and the full form from p.63 onwards. When

you type a word or phrase into the Simple Query field and click Make Concordance, it searches the corpus and returns page(s) of the search results.

Let's ask our first question.



Simple query:

[Query types](#) [Context](#) [Text types](#)

⁶ <https://the.sketchengine.co.uk/auth/corpora/>

Question 1 Does *whose* only refer to people or does it also refer to non-living things?

Before answering any of these questions, make a note of what your intuition tells you.

Type *whose* into Simple Query, click **Make Concordance**, and look at the nouns and pronouns on the left of the node to see what it refers to. Here is a sample of five lines.

anti-science movement of the 1970s **whose** antecedents lay in t
so urgent as the self-employed guy, **whose** gonna need money
s of the cast were wary of the actor **whose** leading lady, a class
he street, and moved into a cottage **whose** roof did not leak. T
lesson from a Dane, Otto Jespersen, **whose** papers on language

Anyone who is under the impression that *whose* is limited to people only, as *who* is, tends to use *which* or *that* instead of *whose*. This is wrong.

Question 2 Do we say people who or people that?

Simply type one into Simple Query and the other into another Simple Query in another browser tab. Click between them to compare your findings.

Given that one of them is about ten times more frequent than the other, we'll take that as the pattern of normal usage. For further information, look down the concordance page at what follows *people that* – this is revealing.

Question 3 The following three expressions refer to the same thing: *pass away*, *bite the dust*, *kick the bucket*. What do their various frequencies tell us?

Do the numbers surprise you? Why might there be such a range of results? Observe how the verbs conjugate in each one.

What meaning do the three share? Do you know other wordings for this concept? Do we need so many?

Per million (HPM) is a number that appears beside the raw frequency of hits at the top of the concordance page. It is most useful when comparing searches in different corpora and subcorpora. In ukWaC, a corpus of over 1.5 billion words of English, *while holidaying* occurs 44 times, which is still 0.0 per million. In enTenTen, a much larger corpus, *while holidaying* occurs 287 times, and this is still 0.0 per million. Expressing the number of hits per million is referred to as NORMALISATION. See the Online Glossary.

Question 4 **Do problems *lie*? If so, what is meant by this?**

It often happens that we meet items of language that surprise us, and asking the thousands of native speakers whose language has been sampled in this corpus can not only answer the question, but provide other quite useful usage information, serendipitously.

In the Simple Query field type the two base forms (lemmas): *problem lie*. What data is returned? What sorts of things precede and follow this pair of words? What information do you now have?

Here are several other quirky collocations that may be equally revealing:

shed light, burning issue, fall pregnant, lose interest

Compare their HPMs in different corpora.

Question 5 **Do people respond to something with the phrase *that's a moot point* to mean that they don't believe what was just said?**

Before doing anything, what does your intuition tell you?

Type *moot point* into Simple Query and be prepared for some very clear patterns. Is it ever used in the plural? What does its position relative to the full stops and commas tell us? What about the verb? Is it in the past often enough to be called a pattern? Is it used with modal verbs? Is there any reported speech? Does the language appear to be mostly spoken or written?



Moot Hall, Keswick UK

Question 6 **Can you say *handsome woman* in English?**

You can say anything. Try the question, *Do you say handsome woman* in English?

Students are routinely taught that *handsome woman* is not 'good' English, but when you type this collocation into the Simple Query field, be prepared to find that it does indeed occur. At issue is not the collocation, but the semantics of *handsome*. An internet image search of various adjectives that describe men as physically attractive yield quite different pages of thumbnails, e.g. *handsome man, sexy man, gorgeous man*. To find more such adjectives, click on Thesaurus in the left panel and enter *handsome*.

The features of *handsome*, although most common in men, also appear in a certain type of woman, namely a handsome one. We revisit this question towards the end of this book (p.184).

Remember that data has to be interpreted before it becomes information, thus going beyond simple frequency counts is an important step.

We are now going to work through one example. In the process, we observe some features of Sketch Engine, start to see how language is the source of the data from which we draw conclusions about patterns of normal usage and how these are sometimes exploited by language users. We also consider the relevance of these things to language learning.

Question 7 **Does anyone really say, *while holidaying*?**

Perhaps a language teacher made it up to demonstrate something, or perhaps it appeared in translation software. If you are ever suspicious of a construction, even in your own writing, don't hesitate to check it in a corpus.

Remember that the BNC is more than 25 years old – it is sometimes fairer to ask, *Did people say ...?* rather than *Do people say ...?*

Enter the phrase into Simple Query and click the Make Concordance button.



The data is very revealing. The screenshot below shows all five concordances of *while holidaying*. Don't worry if yours looks different – in the section below, View Options (p. 39), we learn how to customise the appearance of concordance pages.

How many times does *while holidaying* occur in the BNC? In what sort of texts? Is it ever at the beginning of a sentence? What precedes it? What immediately follows it?

Query while, holidaying 5 > GDEX 5 (0.0 per million)	
1 W_news_script	Gloucestershire Royal Hospital, died 10 days ago while holidaying in Egypt. The
2 W_newsp_tabloid	feared they would die after being snatched while holidaying in France. Th
3 W_newsp_other_report	married and guide dog Peter was best `man'. While holidaying in Scotland t
4 W_newsp_brdsht_nat_misc	was playing whist, and two years later, while holidaying with an uncle
5 W_newsp_tabloid	fend off the local romeos from his daughter while holidaying on the encha

Our search item occurs five times in the whole BNC of 100 million words. Five of anything is very small as the 0.0 per million tells us. This means that for every million words in the corpus, the item occurs this often. The word *while* occurs 54,791 times in the BNC which is 484.9 times per million words. This throws some more light on the significance of *while holidaying*.

Would you expect the alternative construction, *while on holiday(s)* to occur more frequently? Try it. Bear in mind that the first recorded use of *holiday* as a verb is as recent as 1869¹⁵, which makes it less entrenched in the language than most vocabulary.

It would be fair to conclude that *while holidaying* is possible, but not probable.

Given that there are millions if not billions of words written and spoken in English every day, the paltry 100 million words of the BNC really is a tiny sample. Chances are that if it occurs five times in the BNC, it occurs thousands of times in authentic usage every day¹⁶.

The probability that *while* and *holidaying* would occur together (contiguously) would be the same as the probability that any two words would co-occur if:

- all words were equally frequent in the language or the corpus, and if
- their parts of speech did not influence their likelihood of co-occurrence, e.g. adjective + noun is probable in English whereas adjective + adverb is not. In fact, we will find (p.130) that adjective + adverb is not rare⁷ – it is just that there is only one salient case when the adverb post-modifies the adjective.

GDEX in the name of the grey bar across the top of any corpus search. It is the abbreviation for Good Dictionary Example. As various operations are performed on search results, e.g. sorting, sampling, filtering, the grey bar expands with breadcrumbs. **You can therefore return the data to previous states by clicking on them.**

Returning to our concordance, the text to the left of the NODE (the search item), usually indicates that something happened while holidaying; the text on the right usually says where, as can be expected in unmarked word order (see below). In one case only, it starts a sentence.

When you click on the node, a wider context box pops up as in this screenshot. Click on the arrow head at the top to close it.

⌵

< [expand left](#) many mourners turned up to say a final goodbye to nurse, Sharon Hill, that some had to stand outside Gloucester crematorium for the service. Sharon, who worked at Gloucestershire Royal Hospital, died 10 days ago **while holidaying** in Egypt. The tourist bus in which she was travelling was fired on by terrorists. The mourners were led by Sharon's parents, John and Sheila Hill. After the service, the rector gave his reaction [expand right >](#)

⁷ ske.li/bnc_adj_adv_nodeforms (UK)

Click on the blue letter(s) and number(s) to the left of a concordance line to have this metadata pop up. If there is a lot of metadata, it will scroll vertically.

bncdoc.id	K24
bncdoc.year	1993
bncdoc.title	[Central television news scripts]
bncdoc.info	[Central television news scripts]. Sample containing about 44589 words of material written to be spoken (domain: leisure)
Text availability	Ownership has not been claimed
Publication date	1985-1993
Text type	Written-to-be-spoken
David Lee's classification	W_news_script

Unmarked and marked features of language

Deviations from the normal, standard, typical way of doing something are said to be *marked*. There must be a situation which allows for choice; the choice may be conscious or subconscious. Wearing red shoes with your suit instead of the black ones you usually wear is a marked choice, and is likely to have an effect especially on those familiar with your typical attire. What constrains your choice?

In English, unmarked word order revolves around Subject Verb Object. The verb and the object form the predicate and this us something new about the subject.

Corp ex. 3 Thus **he met Catharine Hanson**, a handsome woman of 29 and immediately led her to his bedchamber.

Corp ex. 4 **Canano introduced me to two handsome women**, one of whom was his mistress.

Subject Verb Complement is also standard since the verbs in this structure are copular verbs, which are among the most frequently used, e.g. *be, appear, seem, looks like*.

Corp ex. 5 Nona Thorp was a handsome woman but she had withdrawn into herself, ...

Marked word order sees some restructuring that does more than provide new information about the subject.

Corp ex. 6 ... by far the handsomest women **we** have seen in France were born and reared in America.

Corp ex. 7 I thought what a very **handsome woman** she was.

Vocabulary choices often display degrees of markedness: the CONNOTATIONS of a word or phrase often motivate choosing *bright* over *clever*, when describing a student. Troponyms also express degrees of markedness, for example, in the series,

go → walk → stroll → amble

each one is more marked than the one before. Each one contains an additional element of meaning: if someone is walking, it does not mean that they are ambling, but if they are ambling, they are surely walking. Analogous issues arise when we explore hypernyms (p.110).

Question 8

Would you expect each of the troponyms to be less frequent than the one to the left of it?

Use Simple Query to observe the occurrences of these words. As they all appear in various parts of speech, the results will be mixed.

	go	walk	stroll	amble
Frequencies				
Hits pm				

What are some of the reasons for this wide range of hits?

Question 9

Would you expect that these frequencies impact on the order in which native speaker children *acquire* them (FLA), or learners *learn* them (SLA)?

FLA questions can be investigated using the CHILDES corpus, which is among the Sketch Engine corpora. Corpus study of SLA is mostly conducted via learner corpora. See the Online Glossary for information.

The normal, standard, typical way of doing or saying something is obviously more frequent than any of the marked variations on the linguistic menu. This means that language learners have more exposure to them and are primed for unmarked forms first.

These verbs of motion are followed up on p.72 when we have some more skills under our belts.

Question 10

One swallow does not a summer make. What is this curious word order?

Is this a proverb or an idiom? Where does it come from? Why is the verb at the end? Under the influence of German? Which meaning of *make* does this structure realise? Is this marked word order a pattern of normal usage?

The BNC does not have enough data for this archaism. Make a Simple Query search in LEXCMI for *does not a*. The indefinite article blocks the verb from appearing in its unmarked position. There are 123 hits, 79 of which have the structure with noun phrases: *NP does not a NP make*. Use your browser's Find: *make* to see the prevalence and position of *make* on the concordance page. The pattern is clear. Right sort them to see the lexical patterns as well. For example,



- Corp ex. 8 However, chutzpah alone does not a good film make.
- Corp ex. 9 Of course, one battle does not a war make, ...
- Corp ex. 10 One election loss **doesn't a failure make**.
- Corp ex. 11 Having a girlfriend **doesn't a straight man make!**

See p.33 for forming queries with apostrophes.

But why not: *One swallow makes not a summer?* What is the function of *does* in our idiom: negation or emphasis? See the paradigm under Periphrasis (p.93) for more on the roles of *do*.

In Shakespeare's day, the periphrastic *do* had not yet settled into its modern day usage. *Romeo and Juliet* was written around 1595. See what they were saying:

- | | | |
|--------|---|--------------------|
| JULIET | <i>O think'st thou we shall ever meet again?</i> | (do you think ...) |
| ROMEO | <i>I doubt it not; and all these woes shall serve</i> | (I don't doubt it) |
| | <i>For sweet discourses in our time to come.</i> | |

Many expressions in English also come from the King James Bible (1611) and retain their echo of times past.

Man doth not live by bread alone. Deuteronomy 8:2-3, Matthew 4.4, Luke 4.4.

But does our idiom date from this period?

Let us consider this earlier form: Search LEXCMI for *doth not a*. The seven examples contain contemporary references, e.g. *girlband*, *scouts*. What does this suggest? There is only one example of *doth not a* in the BNC, also with contemporary references.

- Corp ex. 12 A: A little smoke and a couple of pills doth not a junkie make.
 B: Yeth Shakespeare, you're right. It dothn't. Oth courth noth -

Speaker B is clearly amused by what he or she takes as Shakespearean English. But why all the *'th's*?

For comparison, search for the unmarked form, *does not make a*, in LEXCMI. It seems that many of the NP make + NPs are delexical verbs, e.g. *make a profit*, *make a contribution*.

Does our target **structure**, NP does not a NP make, have its own underlying meaning? Could it mean: indicate, constitute, justify the label of..., or ... ?

⁸. The concordance is here: ske.li/lexcmi_doesnt_a_make (UK)

In searching for answers, we have found many questions. Che sarà, sarà! Guided discovery is journey. *One swallow does not a summer make* is an underlying principle in pattern hunting: only when many instances flock in are we able to declare something a pattern of normal usage.

Constrained options

The corollary of Kilgarriff's statement quoted on p.1, *Language is never, ever, ever random*, is that every choice we make is constrained by a continuum of factors that range from the purely linguistic **grammatical conventions?** to the pragmatic considerations that are embedded in the context of the culture.

The sentence you have just read took much longer to write than to read. It is a long sentence containing a number of infrequent words sometimes in standard constructions, sometimes not. As a native speaker, such elements as *make a choice, factors range from A to B, context of culture*, placed no demands on my writing. Structuring the sentence with three that clauses, however, certainly took some juggling and a string of choices were made to achieve this result.

Halliday's view is that "users select from options that arise in the environment of other options, and the power of language resides in its organization as a huge network of interrelated choices" (Systemic functional grammar, 2013)

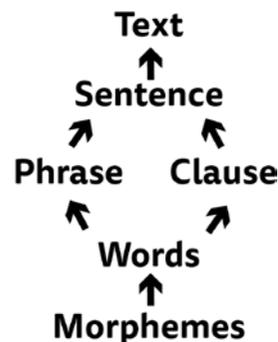
Michael Lewis' *The Lexical Approach* (1993) divides grammar for pedagogical purposes into Patterns, Facts and Choices¹⁷. **Patterns** are the regularities within a language such as morphological forms (plural s, past -ed), unmarked word order (SVOMPT) and the gerund following prepositions. **Facts**, on the other hand, relate to individual items. For example,

- the use of the subjunctive after certain words (mandative subjunctive)
- irregular forms of about 200 mostly high frequency verbs
- the comparative and superlative forms of some high frequency adjectives
- the irregular plurals of some mostly low frequency nouns
- some verbs are most frequently used in the passive
- words have different meanings which are mostly realised by their patterns¹⁸
- cognates might have quite a different meaning or use in other languages

All of these have to be learned as language facts.

Choice presents the greatest challenge for non-native speakers because selecting the most probable form from possible forms involves weighing up the *constraints* on each one. There are more options as we move up the Hierarchy of Language, as seen in this diagram.

This could be thought of as a fuzzy area of language, but in fact, as the options are whittled away through a process of elimination, only few remain. After all, every communicative act has a perlocutionary force, such as persuading, convincing, enlightening or getting someone to do something, and each of these has its own lexico-grammatical realisations: never ever random. Perlocutionary force is an aspect of Austin's (1911–1960) influential theory of speech acts, a cornerstone of the field of pragmatics. Another choice we make is referring to Shakespeare as Shakespeare, Will, the Bard, a man, he, the playwright? Consider what motivates someone to choose one at the time of speaking or writing. Jean Aitchison (1989) refers to the *psychological trigger* that performs this choice in native speakers. Non-native speakers, however, draw on whatever criteria they have acquired, through general exposure and direct study, and thus the choice is made more consciously. A corpus query shows the choices that have been made.



Other Simple Query searches

We can now work through some more questions that can be asked and answered using Simple Query only. Note that **lemmas** entered into the Simple Query field search for all the inflected and conjugated forms that we saw in the Inflection side of the Morphology Taxonomy (p. 68), whereas

word	Freq
P N is welcome	118
P N is welcomed	40
P N is welcoming	6
P N is Welcome	1
P N IS WELCOME	1

word forms you enter return that word form only. For example, searching for *be welcome* returns the combinations of all eight forms of *BE* plus all three forms of *WELCOME*, whereas *is welcome*, searches for all forms of

is, of which there is only one, and all forms of *welcome*, as can be seen in this screenshot. To see the frequency list for BE WELCOME, open this permalink⁹.

Question 11 **Is open door an adjective + noun or a verb + noun? Or something else?**

The minimal use of inflection in contemporary English guarantees the high proportion of words that undergo conversion. Here we have an example of a pair of words that look like a typical verb + noun combination, but which turn out to be used in a variety of POS combinations.

As we saw with *be welcome*, Simple Query searches for all the word forms of all lemmas regardless of part of speech.

There are 412 hits of *open door* (3.6 per million) in the BNC. It is mainly an adjective + noun as in Corp ex. 13. The following example shows a compound adjective describing *policy*. We then see it is a verb and its object, conjugated and declined respectively. This literal use of *to open a door* contrasts with its figurative use in the last example. Interesting, isn't it?

⁹ ske.li/bnc_be_welcome

- Corp ex. 13 To the right, a partially **opened door** led to a connecting office.
- Corp ex. 14 Harding, the man who implemented BNFL's **open door** policy has said goodbye.
- Corp ex. 15 ... he would prowl the ward, **opening doors**, locking me in a bathroom.¹⁰
- Corp ex. 16 Our album **opened doors** for us all over the world.

These variations are problematic for error tagging software: be prepared for unexpected, unreliable results.

Question 12 **I've come across *boldly go* a few times and wonder if it is more than a collocation.**

Type the phrase into Simple Query. There are not very many hits, which tells us something. There are a few words in the contexts which give a good clue as to the source of this phrase.

Does GO conjugate or is the word form part of a fixed phrase?

It is interesting to find that GO is very rarely preceded by an adverb that describes the *going*. Open this permalink¹¹ and look at the list; it was generated using the Frequency tool, which we study on p. 52. For more information about *boldly go*, see its very own entry in Wikipedia¹⁹.

Question 13 **I've heard the phrase, *Just when you thought it was safe to ...* and I wonder how it might be used. Is it ironic?**

Let's search for *just when you* and see where it leads. In fact, the raw unsorted data doesn't easily lead us anywhere. Click on Sort Right in the left panel and scroll down the page. Who said language wasn't patterned?

Question 14 **I've also heard *in my point of view*. I thought it was *from...***

In the BNC there is one example with *in* and 69 with *from*. Let that be a lesson to you! A single instance cannot be pattern of normal usage.

Now try an internet search for *in my point of view*. What do you learn?

Forming Simple Query searches is similar to querying Google and Yahoo, etc, but the results are very different because the internet is a vast, random collection of documents of many types changing moment by moment. This is in stark contrast to a corpus designed for language study analysis. Thus the internet is able to say what is possible in English while a corpus search tells us what is probable. This contrast appears in classroom language too: compare *Can you say this in English?* with *Do you say this in English?* Let that be a lesson to you too!

¹⁰ See here for other examples of this pattern: ske.li/iq

¹¹ ske.li/bnc_ly_go (FI)

Question 15 **How is the word *germane* used?**

Resist the temptation to look this word up in dictionary, at least until we have worked through the steps below. In any case, there are questions to be asked that a dictionary will not answer.

Search for *germane* in the BNC using Simple Query. It is indeed a rare word.

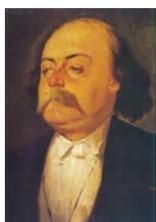
It is mostly followed by *to*. Search now for *germane to* in Simple Query¹². Is *to* a preposition or an infinitive marker when used with *germane*? Or both?

Which verb precedes *germane*? Always, often, sometimes?

What sorts of things are the subjects?

By this stage, you have probably inferred the meaning and worked out the pattern that *germane* is used in.

How would you express this notion in other words?



People's needs for precision and concision vary. These needs certainly influence the choices we make when writing and speaking. It can take a long time to weigh up the suitability pros and cons of particular expression. The author, Gustav Flaubert, made such high demands on himself, or went to such great pains to find *le mot juste*, that his overall output was much smaller than that of many writers of the period. *le mot juste* means the most precise and appropriate wording for a specific concept in a particular context, for which English does not have an expression. We borrowed it from French. The absence of a word or phrase for a concept is referred to as a lexical gap. This cannot be sought in a corpus!

Question 16 **Does *curiouser* occur more frequently than *more curious*?**

Is frequency an adequate criterion for determining patterns of normal usage?

As with *boldly go*, scanning the context provides a good hint as to how *curiouser* slipped into the language and how it is used. Another example of this is *never walk alone*, which may be sage advice in some circumstances, but the phrase is as culturally loaded as *curiouser* and *boldly go*. What can you discover about this?

The formal aspect of language we observe in *curiouser* relates to what we know about the formation of the comparative and superlative in English, as we explore on p. 103.

Question 17 **Is *Pacific Ocean* written with this capitalisation? And is it always preceded by *the*?**

¹² ske.li/bnc_germane_to (UK)

And what about *Czech Republic*, *East Timor*, and other such compound country names. Are they always used with the definite article? Can you determine a pattern here?

Apostrophes in English

Welcome to punctuation, a right royal pain in corpus work, and none more so than the apostrophe. In English, it is used for contractions e.g. *they're*, *who'd*. and possessives *Megan 's*, *sufferer 's*. Behind the scenes, corpora store them as two separate words.

To search for words with apostrophes, it is therefore necessary to separate the two parts, e.g. *pianist 's*, *they 'd*, *hangover 's*. To complicate matters even further, the full negative contraction is separated from the base word, e.g. *do n't*, *could n't*, *wo n't*.

Full stops, exclamation marks, question marks, colons and semicolons are treated as separate tokens which is why the number of tokens in a corpus exceeds the number of words, as seen in the list of corpora on the homepage.

Question 18 Does the phrase *who'd have thought* stand alone? Or is it a part of a sentence? If so, how is it followed up?

Enter the search into Simple Query. Try it in various corpora.

Simple query:

Question 19 Is it the norm to use apostrophes with decades? Search for both *1970s* and *1970's* in the Simple Query field.

It is necessary to include the space: *1970 s*, *1970 's*. As with wrong spelling, other instances of incorrect usage appear in corpora because they occur in the data that was included. But as we are primarily interested in identifying patterns of normal usage, *1970's* occurring 98 times vs. the 3,046 occurrences of *1970s* unequivocally answers the question.

tion makes obvious. **It's the** 1960s, and Alice Shel
view." And he did. " **It's the** 1970's," Dedopulos be

Question 20 What's the difference between ...

If this is a chunk, we should be able to observe plenty of examples in corpora. It may exist in both the contracted and full forms. Corp ex. 17 is a request for information, whereas Corp ex. 18 is a formula used in English jokes.

Corp ex. 17 What is the difference between information and data?

Corp ex. 18 What's the difference between a prostitute and a cockerel?

Question 21 **Is *why* followed by both *can't* and *cannot*?**

For example, do we say *Why can't I come too?* and *Why cannot I come too?*

Once again, before answering, what do you expect the answer to be?

Simple query:

Operators available in Simple Query

Operators such as **question marks** and **asterisks** are used in searches for various substitutions. Unfortunately, these work differently in other search fields as we shall see.

Question mark: *d?g* finds lemmas that begin with *d*, followed by one letter followed by *g*. Since Simple Query is a lemma search, the results will include inflected forms, for example, *dog*, *dogged*, *dags*. Click on Node Forms in the left panel to summarise the findings. *z??k* allows two letters between them, while *m???y* three. *???n?tion*, etcetera.

Question 22 **Are *wh-* words followed by both *can't* and *cannot*?**

What pattern emerges from this observation?

Question 23 **Is the spelling of words with either *-ise* or *-ize* a difference between UK and US English?**

As the LEXCMI corpus has 100 million words of US English, the question mark is very valuable in such "bi-lingual" corpora.

Select verbs that might have this spelling variation and search for them with the question mark in place of the alternating letter, for example,

reali?e *comprimi?e* *organi?e* *televi?e* *bowdleri?e* *epitomi?e* *ostraci?e*

Corpora provide data that can indicate such tendencies, but it is necessary to read further afield. Interestingly, this spelling difference is not related to a US-UK divide, rather Cambridge prefers *ise* while Oxford prefers *ize*, making both acceptable in UK English. (More on this at this [link²⁰](#)). There are some words which only have one correct spelling. For example, *summarise*, *prize*.

Do these spelling tendencies extend to other parts of speech?

For example, search for *televi?ion*, *reali?ation*.

Question 24 **Do you love me**

We cannot use a question mark in a question as it functions as an 'operator'. If you want to find out how people follow up a question, enter it without the punctuation.

We study in Question 230 how punctuation can be included in the more complex query types called CQL.

Question 25 **What is your favo?rite**

This is an instructive question, especially when the results are sorted to the right.

Asterisk: *d*g* finds lemmas that begin with *d* and end with *g*. Any number of letters may appear in between. For example, *during*, *drug*, *deteriorating*. Try also combinations such as *t*n*tion* and list the node forms.

Let's see if what the BNC can tell us anything about rare words. Even though it is a small corpus by today's standards, no publically available English corpus has better metadata, so even though it does not contain the most recent additions to English vocabulary, for many searches, it remains the corpus of choice.

Question 26 **What is the pattern that *cahoots* operates in?**

Type it into Simple Query. With so few lines to look at, the pattern is clear. Then see [here](#)¹³ for a Frequency listing – we start making such lists on p. 52.

Quantitative data should lead to qualitative information. Intuition will play a part in answering your question. A look at the 16 lines of *cahoots* may also tell you something about its meaning and usage.

Quantity also varies greatly due to the size of the corpus. The almost 2.4 million words of English as spoken in London and recorded in the LOC corpus (2011), will provide far more focussed data for its tenor, field and mode than the NMC (2008), a general English corpus many thousands of times larger captured from the internet – horses for courses.

Question 27 **How ecstatic is life in London?**

Another question you'd never thought to ask! The word *ecstasy* occurs in the LEC 15 times and 438 times in NMC. Which of the word's meanings appear in the two corpora?

The answers will also depend on the patterns in the context and co-text, and on the extent to which the data corresponds to the choices you are weighing up. The raw frequency of patterns with rare words differs significantly from those of common words and obviously, the bigger the corpus, the greater the chance they will appear. The mere existence of an item does not make it significant, as we saw on p.31 when discussing the internet as a source of language information.

Question 28 **What other wayward items might appear in the BNC?**

Search the corpus for the following items. To learn more about their source, click on Text Types under Frequency in the left panel. This is putting metadata to good

¹³ ske.li/bnc_cahoots_freq (UK)

use and indicates the degrees to which some things are possible and probable in language.

accomodation, seperate, asshole, old fart, cotton candy, spag bol, would of done, alot

Speaking of these oddballs which do appear in the BNC, let us now consider some that do not.

zine, blog, road rage, regift, emoticon, sexaholism, snail mail, vertically challenged, metrosexual, merkin, sunshower, website, fess up, shirkaholic, bling.

Are there any items here that you cannot live without?

The basic reason why words do not appear in the BNC is that they are not used in the texts included in the corpus. As obvious as that sounds, it is important to bear in mind. Some words are technical words and have limited use, others are rarely used, others are very slangy, new to the language or outside core English. But the most important reason is they were not in the right place at the right time. Remember, there are no texts in the BNC after 1994. Such is the nature of a synchronic corpus.

Try searching for some of these words in larger, more recent corpora such as EnTenTen, and corpora with a wider geographical spread, such as LEXCMI.

Question 29

Did *tweet* meaning something before the social medium, Twitter?



Compare the use of *tweet* in the BNC and the NMC, two corpora of British English of approximately the same size. Does the meaning from the older BNC still appear in the more recent NMC? Is Twitter's logo related to either meaning of *tweet*?

Question 30

What do we with those words like *thingamabob*?

In the BNC, type *thinga** into Simple Query field to get the concordances. Click on Sort Node in the left panel to group them. Click on the second one which is *thingamabob* preceded by *a thinggroup*. Clicking on the node opens the wider context box as we saw on p.25. This particular example is of interest because it is someone talking about this linguistic phenomenon and lists 30 examples.

Can you tell if these words are most typically used in speech or in writing?

What parts of speech are they?

Do they refer to people as well as things?

Corp ex. 19

I went to that little shop on the corner, you know the one, bought by Miss Thingambob when poor old Mr Whatshisname went to Australia or somewhere with his asthma.

To study items that are not in the BNC, try the much larger and more recent corpus, ukWaC, or en-TenTen. To make a Simple Query in another corpus, you can use the search field in the top right of Sketch Engine interface as in this screenshot. The dropdown list of corpora contains the most recent ten that you have accessed.



For recently coined expressions, see **New words in English** at the BBC/British Council site²¹.

Question 31 **How do company names compare in the two corpora?**

e.g. Google, Amazon, General Motors, McDonalds, Kodak, Apple

Learning language from language

This brings to an end our introduction to the Simple Query field, which we will be returning to throughout the book. By this stage, we have used the Simple Query field to obtain language data that can answer questions about many aspects of language.

We have looked at the meaning and use of specific words and phrases, the grammatical and lexical company they keep, compared the frequencies of near synonyms, inferred the meaning of phrases from multiple contexts, explored the cultural relevance of some phrases, compared the significance of both patterns of normal usage and of rare items. This is why we need to distinguish possible from probable language. We have investigated some linguistic minutiae such as articles, apostrophes and the morphology of adjectives.

In the process, we met a number of general linguistic concepts that are important to language users generally, and are valuable foundation constructs in DATA-DRIVEN LEARNING. The over-riding concept here is frequency from which flows patterns of normal usage, marked and unmarked use of language, norms and exploitations, collocation and anything else that comes under Firth's notion of *knowing a word by the company it keeps*.

Not unrelated to frequency is the issue of corpus size as well as the criteria for its design. We found that corpora contain a wealth of sentences that illustrate our findings and that they are as alive as the real world in which they were created. In the process, we learnt some of the basics of using Sketch Engine and familiarised ourselves with the main parts of the interface. And in accordance with the title of this book, we are finding the answers to language questions through guided discovery tasks.

Most of this has been gleaned from the BNC. Remember that most of the questions can be asked of any corpora, including those you make yourself. Sketch Engine has tools for building your own corpora so if you would like to make a corpus now, so that you can ask the following questions of your own data and compare them with corpora provided by Sketch Engine, turn to p. 178 and make one.

We interrupt this section to bring you some important information about customising the interface. We will return you to the other features of top level queries as soon as possible.

2. Taking ownership of your concordance page

In this section, we learn how to customise the concordance pages and what language information can be derived from different presentations of the data.

Regarding the phrase, *take ownership of something*, it was mostly used referred to something tangible: after a transaction something becomes yours. However, it occurs only four times in the BNC compared with 167 hits of *take possession*, 14 of which are *take possession of goods*, 13 of which are tagged W_ac_polit_law_edu in David Lee's Classification¹⁴. In the more current and much larger LEXCMI corpus, *take ownership* yields 1,083 hits from British English and 14 of American English.

Question 32 **What do we *take ownership of* nowadays?**

We ask this question because its answer tells us how its usage has changed since the BNC was assembled. We find that tangible items are outnumbered by intangible items in this list of collocates that occur to the right.¹⁵

It becomes clear that taking ownership of intangible abstractions such as *responsibility*, *issue*, *development* and *learning* involves a sense of being in control through mental/cognitive processes. How far is this from obtaining a thing?

It is now time we took ownership of our concordance pages and customised them to suit our needs. To do so we click on View Options in the left panel.

View Options: Top

Click on the View Options button. It takes you to a new screen in the main panel that allows you to change the concordance view. In the Top Section of View Options, the features are determined by the metadata in the corpus, not by Sketch Engine. These screenshots are from the BNC – other corpora have different attributes and the Top Section varies accordingly.

¹⁴ ske.li/bnc_take_possession_goods (UK)

¹⁵ ske.li/lexcml_take_possession_coll_r4 (UK)

Attributes

Attributes	Structures	References
<input checked="" type="checkbox"/> word	<sp>	Token number
<input type="checkbox"/> tag	<spkr>	Document number
<input type="checkbox"/> lempos	<stage>	bnndoc.id
<input type="checkbox"/> lemma	<u>	bnndoc.author
<input type="checkbox"/> word (lowercase)	<event>	bnndoc.title
<input type="checkbox"/> lemma (lowercase)	<gap>	bnndoc.info
	<loc>	Text availability
	<pause>	Text type
	<shift>	Publication date
	<trunc>	David Lee's classification
	<unclear>	Domain for written corpus texts
	<vocal>	Domain for context-governed spoken material
	<g>	Medium for written corpus texts
		<input type="checkbox"/> References up

Display attributes

For each token

KWIC tokens only

The Attributes column allows you to change the default display.

Selecting the other attributes shows, for example, the nodes' POS-tags and lemmas.

Lempos is a portmanteau word: lemma + POS (part of speech). Click on **Corpus Info** in the top part of the left panel to see a list of lempos suffixes.

Under Display Attributes, choose between displaying them for the node only ("KWIC tokens only") or for every word in the concordance line ("For each token").

KWIC stands for Key Word in Context. This is an unfortunate term, since the search item is elsewhere referred to as the *NODE*, a perfectly reasonable term. Furthermore, key words in a text are those that carry the bulk of its meaning.

Showing tags can be used to find out why a corpus line has unexpectedly matched a query: a common culprit is an incorrect POS-tag. In this 'broken' screenshot, the node is marked as a past participle in all cases (VVN), despite some being adjectives (AJO).

1	heard of anyone dying from a broken VVN ankle - that's if it is broken
2	and explain this. I think it is broken VVN because it says also er Manchester
3	Scotland, followed by a day of broken VVN cloud. OUTLOOK FOR THE FOLLOWING
4	`But I can't leave. My jeep's broken VVN down.' `Precisely,' he said curtly
5	soaked through and spiked with broken VVN glass. There was a gummy label
6	over the weekend. Four cars were broken VVN into at the town's golf club
7	PA Dental jobs threatened by ` broken VVN promises' By Bryan Christie,
8	Suspension : If dirt was merely broken VVN up into small particles cleaning
9	finds the breeze blowing through broken VVN windows. It ruffles the torn

Structures

The Structures column allows you to change from the default display to show the beginning and end tags for structures such as sentences, paragraphs and documents.

One of the most useful is <g> which stands for *glue*. With this selected, punctuation will not be separated from letters and numbers by a space on concordance pages. However, it is still necessary to separate them when you form a query, as when making queries with apostrophes as we saw on p.33.

References

The References column allows you to select the type(s) of information about the source texts which appears in blue to the left of each concordance line, as in the *give it me* screenshot

The items in the list are those included in the metadata when the corpus was created. It is usual to select one, but more is possible, in which case you use CTRL click. It is also possible to select none of them, also using CTRL click.

David Lee's classification

This is only available for the BNC but is generally found to be more useful than the original system. See the Online Glossary for more information.

1	W_ac_soc_s...	was . I thought I bet they still wo n't give it me . I 'm the only applicant and if it 's
2	S_conv	No , I do n't mind . No , you 've gotta give it me back , I 'm gonna . Oh . Hope she looks
3	S_intervie...	Monday , she used to borrow it , forget to give it me back and . I could always ask me mother
4	W_fict_prose	belonged to my mother , " she said . " She gave it me before she died . " Jill turned it over
5	W_fict_prose	always , and when you come home , you can give it me back on our wedding day . " She put the

Question 33 **Which of these variations of this ditransitive are most frequent? Do they display any patterns in their register distribution?**

Give it me, give me it, give it to me.

Enter these in the Simple Query field one at a time. Does anything else emerge, e.g. punctuation before and after, modal verbs.

What other ditransitive verbs might be candidates for this variation? And different indirect objects?

Question 34 **Is *moot point* genre specific?**

We studied some aspects of this phrase in Question 5. Now with David Lee's Classification, the metadata adds another piece to the puzzle. See this link¹⁶.

View Options: Middle

Page Size

The Page Size box allows you to specify the page width and length for the display. KWIC Context size allows you to specify the width of the concordance window in number of characters.

Page size (number of lines):

KWIC Context size (number of characters):

Sort good dictionary examples

Number of lines to be sorted:

Although dramatically increasing the Page Size may slow down the initial retrieval of the concordance,

¹⁶ ske.li/bnc_moot_point (UK)

having many lines on one page reveals patterns as soon as it is sorted. 1,000 is good, if you have nothing against scrolling!

Jump to ...

If the number of hits exceeds the page size you set, a **Jump to** droplist appears at the top of the page

after it has been sorted. This is very handy for getting past punctuation quickly, as well as going to a specific letter, or where you have sorted by References, to a particular reference.



Sort Good Dictionary Examples

A set of criteria for selecting illustrative sentences is available here: *Choosing Illustrative Sentences*¹⁷. These criteria have been automated using algorithms that provide *good dictionary examples*, a.k.a. GDEX. Selecting this in View Options and specifying the number of lines instructs the program how many "good examples" to put at the top of the first concordance page. This is of great value to language teachers and learners, as well as for the lexicographers for whom it was originally programmed.

If you make the number of GDEX lines the same as Page size, your first page of concordances contains all the "good examples".

View Options: Bottom

Sentence copying

Selecting the checkbox for **one-click sentence copying** adds icons to the right of each line. Clicking on that icon copies the

- Icon for one-click sentence copying
- Allow multiple lines selection
- Do not use Flash for copying to clipboard
- Checkbox for selecting lines
- Show line numbers
- Shorten long references

¹⁷ http://bit.ly/maelt-ill_sents

whole sentence to the computer's clipboard. You can then paste it into a word processor document, database, web page, etc.

Allow multiple lines selection: this adds icons for copying more than one line at a time as in the screenshot below.

Query why, ca, n't 442 > GDEX 442 > Random sample 10 (0.1 per million)		
1	W_fict_prose	carrying what looked like a huge steel club. `Why can't the idiots pray somewhere else?' he said 
2	W_misc	might ask themselves, `If he can do it, why can't I?'. Now they go and ask him why he is 
3	W_non_ac_s...	happened to say to the footman, I say: `Why can't I be like some of them girls!' `He say: 
4	S_conv	them. Just a little bit more. Gently Tim. Why can't we leave them in that water? Well what 
5	S_conv	back please can I have another cheque book. Why can't you ask for two cheque books at once? Well 
6	W_newsp_ot...	are allowed to divorce their parents, so why can't the public divorce bad politicians? Animal 
7	W_pop_lore	mind works makes things much easier, so why can't we welcome someone - like a home visitor 
8	W_fict_prose	bright, are you? If I'm allowed to have food, why can't I have it in the same cell as Elaine? 
9	S_intervie...	wasn't particularly everyday language. And why can't we have a communion service in everyday 
10	W_ac_soc_s...	got them and why can't they have this and why can't they do this... it's hard, it's really hard 

Checkbox for selecting lines This works with the concordance lines on the current page only. If you need to select lines from more, it is necessary to make the Page size longer. Once you select a line, a new button appears in the left panel under Filter: **Selected lines**. Click on this to shorten your page to the selected lines only. This is very useful if you want to make a screenshot, print out or **Save** a manually selected set of lines.

Show line numbers Select this to see each line numbered, as in the screenshot here. This is very helpful when you need to refer to a specific line as often happens in lectures and classrooms. It is also useful when mentioning a line in an article that contains a screenshot.

Shorten long references Select this to show only the beginning of long references, as in this screenshot. This is especially the case with corpora drawn from the web, e.g. ukWaC. The full metadata appears when the mouse hovers over it.

XML template for one-click copying is a feature used for specific projects only.

Save and Change Options

This button takes you back to your concordance page with your changes implemented.

Special note about copying sentences

When you copy sentences, it is often best to paste them into a simple text editor such as Notepad first. This strips away formatting such as background colour, red node word(s), thereby cleaning it for use in other programs. It may also be necessary to remove spaces between the node and the words either side of it, and before punctuation. In some word processors, Paste Special pastes text without any formatting, which makes the text editor step unnecessary.

We interrupted the Query types to bring you some important information about customising the interface. We will return to the rest of the top level queries after we do one more thing, Menu Items.

3. What's on the menu?

The following menu items are used extensively to manipulate the results of a search. They appear in the left panel once you do a search. You can place it at the top instead: click Menu Position at the bottom of the panel. This may be preferred when working on a narrow monitor.

Save

Click on Save to save the current data as a file on your computer. You can specify whether the output is text or xml, how many pages, whether a heading is included, whether the lines are numbered, whether the KWIC are aligned in the output as well as the maximum number of lines. Where it saves depends on your browser settings.

KWIC/Sentence

This lets you toggle between the standard KWIC concordance view (which appears by default) and full sentence view.

If you selected **Icon for one sentence copying** in View Options, the full sentence will be copied whether in KWIC or in Sentence view.

Sort

As we have already seen, sorting turns data into information that reveal patterns of normal usage. The examples we have looked at so far have mostly returned relatively small concordances. More frequent words and phrases yield a great deal of data and it may be necessary to use the Sample button (p.50), make searches more specific using the Filters (p.95 ff), or use other tools such as Collocation (p.62) and Word Sketches (p.154).

The five buttons under Sort in the left panel perform the following automated sorts.

Left and right

The default order of lines on a concordance page after a search is Document IDs – this cannot be expected to reveal any patterns in the language. As we saw above, the right sorting of *just when you* (Question 13) was very revealing. Right Sorting sorts the words on the right of the node alphabetically, which reveal patterns and paradigms in what follows the node, e.g. punctuation, bound prepositions, wh- words, to-infinitive, words that form compounds with the node. Left Sorting reveals patterns before the node, e.g. words forming compounds, prepositions, punctuation. This is the company that words keep.

Save
View options
KWIC
Sentence
Sort
Left
Right
Node
References
Shuffle
Sample
Filter
Overlaps
1st hit in doc
Selected lines
Unselected
Frequency
Node tags
Node forms
Doc IDs
Text Types
Collocations
ConcDesc
Visualize

Question 35 **What things are said to be *unthinkable*?**

Sometimes we come across a word in the course of our normal reading and we would like to know more about it. However, in reading it is likely to appear only once. To get a feel for how it works generally or variously in the language, we need some more examples.

Do a Simple Query search for *unthinkable*. Left sort the concordance page. Are there any adverbs describing the adjective?

What verbs and verb phrases are most prominent? Patterns of normal usage emerge.

What is the relationship between the verb BE and SEEM? Why use the latter instead of *be*?

Another significant colligation is *the*. How often does this occur directly before *unthinkable*? This is easy to calculate if line numbering is turned on. What does this tell us about our target word? What other adjectives are used in this way?

Now right sort the page and look down the list. What typically follows it? There are not many nouns, which is unusual for an adjective, is it not? Here are the lines with *unthinkable + noun*¹⁸. There is a general air of doom and gloom in these 25 lines. This is referred to as negative semantic prosody. Since *unthinkable* is not usually followed by nouns, what sort of adjective is it usually? Is this compatible with the high frequency of the verb *be* preceding it?

Click on **Right 352** in the GDEX bar at the top of the concordance page to go back to the full set, right sorted.

Query **unthinkable** 352 > GDEX 352 > Sort **Right** 352 > Positive filter 25 > Sort **Right** 25 (0.2 per million)

In the right sorted set, what follows *unthinkable*? There must be some significance to the punctuation, given that it accounts for almost half. What grammar words follow, and what are their roles?

Are the prepositions bound to *unthinkable* (bound prepositions) or do they start prepositional phrases (free prepositions) functioning as adverbials?

¹⁸ ske.li/bnc_unthinkable_noun (UK)

Question 36 **Do marriages break up or down?**

Search for *marriage break* in Simple Query to get a concordance page with both words inflected. Click Sort => Right in the left panel and see the patterns that emerge.

Compare related nouns such as *relationship, friendship, union, affair*. Does any pattern emerge?

Another approach is to make Simple Query searches for the two phrasal verbs, *break up* and *break down*. When they are intransitive, left sorting will help identify the subjects.

Question 37 **Does *whatsoever* typically come at the beginning, middle or end of a sentence? And is it noticeably used in negative contexts?**

Make a Simple Query search. Click the button to right sort the words alphabetically.

If you have your page length set at 1,000, the patterns that emerge from sorting are easy to recognise. If you have line numbering turned on, counting numbers of each phenomenon is simplified.

What do we learn from the observation that more than half of the occurrences of *whatsoever* are followed by punctuation?

This pattern of normal usage is too significant to be ignored. In *Lexical Priming* (2005), Hoey speaks of words having typical positions in sentences.

What chunks emerge from a Left Sort¹⁹ of *whatsoever*?

You might compare *whatsoever* with *at all* (see also Question 47).

Question 38 **Which is correct when listing things in prose, *first* or *firstly*?**

Search for both words and observe them in both KWIC and Sentence views. Sort left. Since we find both *first* and *firstly*, it is difficult to arrive at a descriptive, data-driven decision. How should we deal with such mixed messages? In fact, it is a moot point among stylists, purists, prescriptivists, and the like.

Question 39 **In the name of language observation, what can we discover about the word *hamper* from a left sort?**

What part of speech is it usually?

Does the verb form prefer passive or active? If *hamper* is typically used in passive voice, a Right Sort should confirm this. Does it? By what means?

¹⁹ ske.li/bnc_whatsoever_left_colls

What things are *hampered* and by who(m) or what?

What adverbs typically qualify it?

Does it appear to have positive or negative CONNOTATIONS?

Question 40 **What words in the context of *hamper* the noun indicate a different meaning from the verb form of the word?**

As a noun, *hamper* has a totally different meaning, as *food*, *picnic* and *Christmas* suggest.

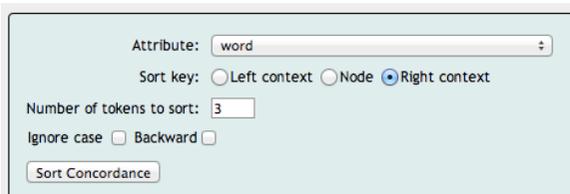
When a word operates in more than one part of speech, a process known as CONVERSION, the two words very often have basically the same meaning, e.g. *abuse*, *rubbish*, *perfect*. But there are also many cases where the word has a different meaning, and this is referred to as POLYSEMY, e.g. *dog*, *address*, *count*, *mushroom*, *goldfish*, *hamper*. To complicate matters even further, there is some overlap between conversion and polysemy. Learners need to pay attention to such features of English vocabulary (Schmitt 2010), and anyone studying concordances needs to bear this in mind. We have already seen that corpus software can be confused by these features, e.g. *broken* on p.40.

Question 41 **Why does *wont* return so many concordance lines?**

There are two basic reasons for this. Perform a search for this word without an apostrophe. Sort to the left and observe some patterns of normal usage, sort to the right and observe some more. The patterns are mostly colligations²⁰.

In some cases *wont* is a typo, but in most cases it is a word in its own right. Can you infer its meaning? And can you describe the relationship between the grammar patterns and its meanings and uses?

Sorting Node, References and Shuffle



The screenshot shows a search interface with the following options:

- Attribute: word
- Sort key: Left context Node Right context
- Number of tokens to sort: 3
- Ignore case Backward
- Sort Concordance button

Earlier, we investigated *open door*. As we know, Simple Query items that are lemmas are conjugated and declined. To group all the possible forms of a search such as *open door*, sort the **Node**. We immediately see that of the 412 hits in the BNC, the first 314 are in their BASE FORMS.

Sort **References** rearranges the lines according to the metadata you selected in View Options. David Lee's classification shows that 13 hits of *open door* are spoken and over half are from fiction.

²⁰ ske.li/bnc_wont (UK)

Given that spoken language constitutes only c.10% of the BNC, raw frequencies are not a fair representation. For the studies in register and stylistics that corpora provide data for, these numbers need to be normalised. See the Online Glossary for more on normalisation.

As mentioned above, a concordance first appears in order of document IDs. We use **Shuffle** to arbitrarily order the order of lines so as to avoid the bias which may emerge from only looking at the first portion.

Simple and Complex Sorting

The automated sorts discussed so far are single level only: alphabetical. If we wanted to sort, for example, firstly by David Lee's Classification and then alphabetically to the left, we need to use the Multi-level sort. In fact, this was used for the *moot point* example mentioned above. See the result of the sorting here²¹.

When you click the Sort button in the left panel, you are presented with both a Simple Sort form as pictured above, and a Multilevel Sort form as pictured below.

To customise your sort, click on Sort itself. The Simple Sort has a dropdown list of Attributes derived from the metadata of the corpus.

Question 42 **Are references to the Irish political situation spelled with a capital T? Perhaps even two?**

Simple Query: *the troubles*. Click Sort. Simple sort attribute: Word.

Sort key: node. And leave **Ignore case** unchecked.

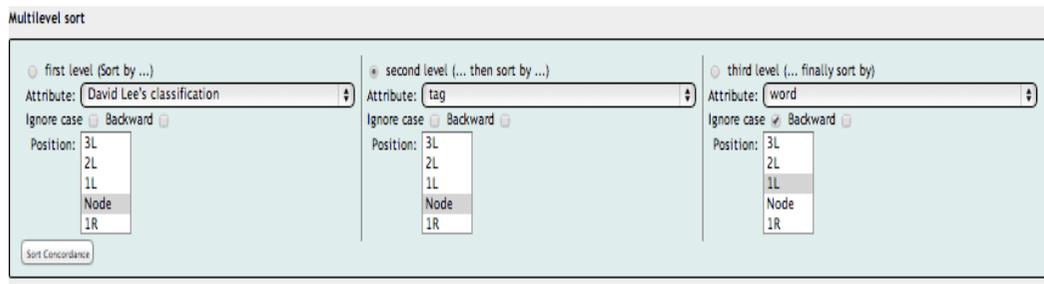
Question 43 **Is *opentypically* a verb or adjective with *door* in fiction?**

To answer this question, our First level must be the domain. Use David Lee's Classification. Position: Node. The second level will be tag, also Node. To have the lines sorted in alphabetical order, make the third level word, 1R. Making it 1L will show patterns in the determiners, which was not the question, but is not uninteresting.

The page will be thus sorted. W_fict_prose starts at about Line 50. To see the tags, it is necessary to go to View Options.

²¹ ske.li/bnc_moot_point (UK)

Selecting *backward* shows the element in reverse alphabetical order, e.g. David Lee's Classification listing starts with W_items (written) instead of S_items (spoken).



Question 44 **In which text types are *The Troubles* discussed?**

Using the Multi-level Sort, make the first level word – node. Second level Domain for written corpus texts – node. The capitalised node starts halfway down the page which can be marked using the Select Lines checkbox (in View Options).

Sample

Most corpora are samples of the language they were designed to represent as it is not usually possible to gather all the examples of a text type. An exception to this is a corpus of a specific text type, e.g. all Shakespeare's plays in one corpus when it was designed to study the language of Shakespeare's plays. We are generally satisfied with observations made from a sample, as is the case in many scientific fields.

It is not unusual for thousands of corpus lines to be returned for a search. Analysing the first lines whether the first 100 or the first 1,000, is not advised since the default order is the document ID numbers in the corpus. To get a random sample, use the Sample button.

TIP It makes sense to make the sample size the same as your page length.

You can specify the size of the sample (i.e. the number of lines) or use the default of 250. For example, if you search for *play*(verb) in the BNC and decide that you do not want to analyse 37,632 lines, use the sample tool to reduce it to a manageable number.

Question 45 **Does *be bound to* refer to anything else than something like 'being tied with rope'?**

Let's look at *be bound to*. Take a sample of 250. Right sort and patterns appear.

Do you see any instances of something being literally *bound*, e.g. with a rope?

Is there any sense of being emotionally bound to someone?

Perhaps most striking in the concordance is the verb: *is bound to* + verb, as this frequency list reveals²². What can observe about its subjects²³?

- Corp ex. 20 well sooner or later er this mixture of languages is gonna spoil the English language isn't it? **Bound to**. Yes it's **bound to**. No what the they'll do, they'll put all the words that are used into a computer er from all these things won't they?
- Corp ex. 21 Breakfast **was bound to** arrive early enough for them to have time to talk afterwards.
- Corp ex. 22 A cake plate so temptingly situated **is bound to** be fought over.
- Compare what happens when the object of *to* is a noun phrase and when *to* is the infinitive marker.
- These examples reveal the very different jobs that words do, and how the meanings of words are influenced by the words around them.

Question 46 **I've heard people say that they are *open to* various things. Any things at all?**

When you search for *be open to* in Simple Query, you potentially get all eight forms of BE and all four forms of OPEN. Let's try it. There are too many hits for a human to process. First, make a Sample which is the same as the page length you set in View Options. Then in the left panel click on SORT NODE. To answer your question about the things that people are open to, click on SORT RIGHT.

Question 47 **Does *at all* reinforce both positive and negative things? And where in the sentence is it typically found?**

Use the same procedures as in the previous question.

Question 48 **Does *put* typically have the structure: someone puts something somewhere?**

Being such an extremely frequent verb, a sample will help you start to answer this question on typicality.

Filter

After performing a search, Filter allows you to remove lines that contain recurring irrelevant elements. It allows you to specify constraints on the context of your search results to retrieve a subset of your concordance. Click on the help icon to open the page-specific help.

²² [ske.li/bnc_be_bound_to_freq](#) (UK)

²³ [ske.li/bnc_subj_be_bound_to](#) (UK)

Question 49 **Do we prefer *turkey breast* to *breast of turkey*?**

This could be an important question for someone translating a menu into English, but such questions would be better answered using a corpus of English menus. See DIY corpora from p.178. It is also possible that modern culinary movements express themselves differently from the BNC days.

Search for *turkey* in Simple Query. Then click filter in the left panel. Type *breast* into the field as in the screenshot and leave all other settings as they are.

Try a similar search for *Mother of God* vs. *God's mother*, as seen on a label in an art gallery.

Question 50 **How often does *Ireland* occur in the context of *the troubles*?**

Search again for *the troubles*. Then click Filter. In the Simple Query field, type *Ireland* or even *Ireland\ Irish*, to make the subset. Of course, such filtering may answer a more sophisticated question than *How many?*

Question 51 **How are *The Troubles* represented in spoken language?**

Once we see enough instances of something within a search, we can constrain the data by selecting that aspect of it in Filter.

Click on filter. Beside the Simple Query, there is the Text Types button. Click it and select the relevant aspect, and then at the bottom Filter Concordance.

At the top of the Filter form, there is also the possibility of excluding a feature. Select the negative radio button. Excluding 1975 to 1993 clearly indicates that *the troubles* had not yet taken on its association with Ireland. One can't help wondering if the non-Irish associations of this phrase are not a substitution for the more general idea of *the problems* someone has with something.

Frequency

Multilevel Frequency Distribution is another tool that summarises data. It generates a variety of multi-word lists, depending on the Attributes you select. These lists reveal the horizontal, syntagmatic company that your Node keeps.

BUNDLES	strings of word forms only
SYNTAGMS	strings of POS only
HYBRID FORMS	strings of POS tags and words and/or lemmas
WLTS	word form + lemma + POS tag (of the same word)

After you search for a word or phrase, click on Frequency in the left panel.

Question 52 **How do sentences start with a word like *Given*?**

Enter *Given* into the word form field and select Match Case. Click Make Concordance then Frequency.

To see *Given* plus one word, set First Level as in the screenshot above, and Second Level select 1R, i.e. one to the right. Click Make Frequency List.

To add the next word, select 2R in the Third Level, etc. And we find that sentences start with *Given* like this.

Given the nature of... *Given the fact that ...*
Given the importance of... *Given the choice, ...*

Click on the **P** on the left to see the actual concordances.

The question might also be answered with parts of speech instead of words. Select the Attribute, *tag*, in the second level to find that determiners are the most frequent item in the second slot. Tag in the third slot unsurprisingly reveals that nouns and adjectives follow determiners.

	word	word	word	Freq
p/n	Given	that	the	137
p/n	Given	this	,	22
p/n	Given	that	there	21
p/n	Given	the	choice	17
p/n	Given	that	this	17
p/n	Given	that	it	17
p/n	Given	the	nature	16
p/n	Given	that	we	16
p/n	Given	the	importance	14
p/n	Given	the	current	14

Selecting the appropriate attribute for each column is important for answering your question. Note that First Level refers to the column of data that appears on the left of the table generated.

Question 53 **Is *Given that* looking back to what has already been said, or forward to what is going to be said?**

These are two quite different functions of *Given that* and both are represented in these concordances.

When a word, often a pronoun, refers to something already mentioned, this is referred to as anaphora. The item of specific reference is referred to as the antecedent and the referring item is a pro-form: bold and underlined respectively in the following examples.

Corp ex. 23 The **petite blonde star** refused to go in his car and drove her own vehicle to his central

London offices.

When the pro-form precedes the antecedent, this is referred to as cataphora.

Corp ex. 24 Though she had no car of her own just now, **Charlotte** had been driving, and driving well, for more than four years.

In the case of *Given that*, the anaphoric *that* is stressed when spoken and marked with a comma in the written form, as in Corp ex. 25 and Corp ex. 26. It is unstressed when cataphoric, as in Corp ex. 27 and Corp ex. 28.

Corp ex. 25 Well, it seems that taxis are considered "public transport" for the purposes of this exercise. **Given that**, as a group, taxi drivers are the most reckless, aggressive, and inconsiderate road users, they should have been the first, not the last, to be banned. (NMC)

Corp ex. 26 ... so the increase in water storage will need to be somewhere between 10-15 percent. **Given that**, the debate looks set to intensify, not evaporate. (NMC)

Corp ex. 27 **Given that** the drive is very long, it is essential to spend a few days on this amazing journey. (NMC)

Corp ex. 28 **Given that** the number of different platforms that IE is not even available for is increasing as time passes ... (NMC)

Question 54 What patterns does *unthinkable* and other such adjectives work in?

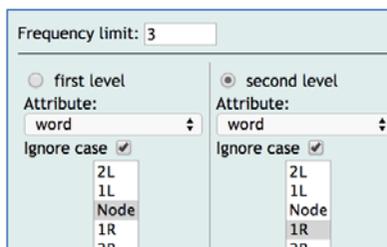
Make a Simple Query search for *un*able*. We will now use the Frequency tool to investigate its colligation patterns.

It is not necessary to look at the > 22k concordance lines. Click Frequency in the left panel and make a First level list only: node – tags.

This list shows that the to-infinitive marker, *to*, is the most frequent POS after these adjectives. Click on the P beside TO to see the concordance page and then click on Node forms to see which words are followed by *to*.

The second and third most frequent POS after these adjectives is NN and SENT. One of these suggests predicative adjectives and the other attributive. Which is which and why?

The fourth tag on the list, IN – preposition, does not include *to* as the infinitive marker. Click on the P beside IN. Make a two-level frequency list of the adjectives followed by prepositions, as shown in the screenshot. Set the Frequency limit to 2 or 3. Then sort by clicking on the columns headings – the left column to put the adjectives in alphabetical order, the right column heading to put the prepositions in order.



We have made some observations about a set of adjectives formed with the affixes, *un-* and *-able*, concerning some of their colligation patterns. Perhaps the same patterns apply to all adjectives, other sets of adjectives. There's only one way to find out. Actually, that last sentence is a fixed phrase in English (with over 1,000 hits in enTenTen) – there is always more than one way to find something out.

Question 55 **How does the collocation *problem lie* work?**

We met this collocation in Question 4. Using frequency, we can now show what patterns follow it. It is quite instructive.²⁴

Question 56 **How does *arresting officer* come to have more than one meaning?**

It is worth observing the POS in compounds, as we saw with *open door*. The variety of combinations often needs to be borne in mind when interpreting concordances.

We can generate a word family, lemma, tag (WLT) table by setting all of the levels at Node but with different Attributes.

Here is the result of *arresting* as found through a Simple Query. It is interesting to see it almost as often an adjective (JJ) as an -ing verb (VVB)²⁵.

	<u>word</u>	<u>lemma</u>	<u>tag</u>	<u>Freq</u>		
<input type="checkbox"/>	P N	arresting	arrest	VVG	123	<div style="width: 100%; height: 10px; background-color: blue;"></div>
<input type="checkbox"/>	P N	arresting	arresting	JJ	104	<div style="width: 90%; height: 10px; background-color: blue;"></div>
<input type="checkbox"/>	P N	arresting	arresting	NP	3	<div style="width: 3%; height: 10px; background-color: blue;"></div>

Click on the **P** to see the contexts. Here are six of the 104 adjective (JJ) sentences. The polysemy of this adjective is clear from these examples. These were Selected using the check boxes, and Saved as a text file, and then imported into this Word document, complete with the angle brackets and spaces as shown.

Corp ex. 29 This area is the Coulin Forest, ... its mountains forming a compact group but individually having distinction of character and outline, some being of <**arresting**> appearance.

Corp ex. 30 Mark usually achieved this by thinking out an <**arresting**> beginning, nearly always of the same type, asking his congregation to imagine themselves standing gazing at the Pyramids or the Acropolis ...

Corp ex. 31 Ways of <**arresting**> decay also included covering eyes and mouth with carefully shaped

²⁴ ske.li/bnc_problem_lie_bundles (FI)

²⁵ ske.li/bnc_arresting_wlt

pieces of jade and inserting jade plugs into other orifices of the body.

Corp ex. 32 Mrs Spurling added: But this <arresting > image takes no account of the fact that Freya's was essentially an imaginative achievement. ...

Corp ex. 33 She swung around and tapped the <arresting > officer on the breastplate.

The word is closely related to the French, *arrêter*, meaning *to stop*. As a positive adjective, it suggests that something was impressive enough to make the viewer stop in their tracks – at least metaphorically. What else does it suggest in these examples?

To see what immediately follows the Node, choose 1R under second level. 1R generates a list of items that occur one to the right of the node, e.g. words, lemmas, tags. 5L, generates a list of items that occur five to the left of the node. Once you have determined the parameters for your columns, click Make Frequency List.

Question 57 Would you consider *well within* a collocation?

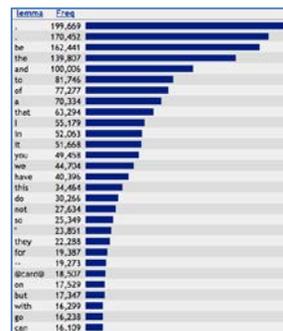
If *well within* is not in your active vocabulary, the little piece of research we undertake here should raise your awareness of it.

A Simple Query search returns a fairly small number of hits. Sorting to the left and again to the right gives some idea of how this pair of words works. Does it match your criteria for a collocation?

Using Frequency, however, reveals bundles that indicate that it is a semi-fixed phrase: the elements are finite but they mix and match in a variety of ways. Perhaps you can diagram its template: what **types** of elements precede and follow it? See *Template* in the Online Glossary.

Zipf's law

As frequency graphs often show, relatively few items make up the bulk of the whole. This statistical phenomenon is known as Zipf's Law. In essence it states that the frequency of any word is inversely proportional to its rank in a frequency table (Johns). This thumbnail screenshot demonstrates this tendency, as do many other such graphs in this book, regardless of the feature of language being investigated. See the Online Glossary for more information on this. This particular screenshot is the top of a list of all lemmas in the TED corpus²⁶.



²⁶ ske.li/ted_top_lemmas_freq (UK)

Question 58 **Is the phrase *a long way* complete? Does it occur at the end of an information unit?**

Search for the phrase, *a long way* and click on Frequency. Select the columns appropriate to your investigation. This screenshot shows us that the phrase, *a long way*, is most typically followed by (a) one of three prepositions, (b) a full stop indicating that it is common at the end of an information unit – sentence or clause, and (c) it is also followed by *away*, which is not insignificant.

word	lemma	Freq
p/n a long way to		336
p/n a long way from		279
p/n a long way off		122
p/n a long way .		109
p/n a long way away		103

We further find that *a long way away* is mostly followed by a full stop, comma or *and*, confirming its final-position status²⁷. Click on the **P** beside *a long way to go*. And then Frequency again. This time, First level – 2L, Second level – 1L, Third level – node. The word *still* is particularly salient before this phrase.

Question 59 **Is it true that *practice* is the noun and *practise* the verb?**

Frequency is an invaluable tool for observing Firth's *company a word keeps*. Search the BNC for *practi?e* and then ask Fre-

lempos	Freq
practice-n	20231
practise-v	3636
practice-v	315

quency to create a list of Node forms using the Attribute, **lempos**. It should look like this, which means that it is true. But what about mysterious 315? More tagging errors or is there another explanation?

Across the pond, the answer to the *practice vs. practise* question is different. Use the 'am' English part of LEXCMI to discover what happens in the US.

DOC. LANGVARIETY

am

br

ie

Select All

Question 60 **According to everyone, *according to me* is wrong. Is it? According to whom?**

²⁷ ske.li/bnc_a_long_way_freq

You
couldn't
be
wronger

Who decides what is right and wrong? Where do rules come from? Cambridge Dictionaries Online has a usage note about *according to me* here²², for example. Where do they get their authority from? What is the difference between their authority and yours?

Perform a Simple Query for *according to*, then in Frequency, select R1 to find out what follows it. Does this answer the question? Or another question?

See also p. 113 for more on *according to*.

Question 61

Is *thereby* always following by -ing forms?

After performing a Simple Query on *thereby*, click Frequency and make the first level the node and the second the tag of all the items that immediately follow it, as shown in the screenshot.

<input type="radio"/> first level	<input checked="" type="radio"/> second level
Attribute: word	Attribute: tag
Ignore case <input checked="" type="checkbox"/>	Ignore case <input checked="" type="checkbox"/>

The result makes it quite clear that verbs are the most frequent elements af-

ter *thereby*, and the most frequent of these is the -ing form. To learn more about the patterns of normal usage of *thereby*, use the Frequency listings putting *thereby* in different columns (levels) and observing what goes before and after it, thereby listing the bundles in which it occurs.

Node tags

Clicking on the Node Tags button in the left panel shows a table of the POS tags for the search items.

Question 62

I thought *lend* was the verb and *loan* was the noun.

But I hear *loan* as a verb more and more often these days. After searching for *loan* in Simple Query, click Node Tags. Try this in both the BNC (1994) and ukWaC (2008) to see if a change of use is in evidence.

Question 63

What is the difference between *be interested* and *be interesting*?

In Simple Query, type *be interest.** Click Node tags to see the combinations of POS tags that result. Click on the P beside the most frequent ones to see their concordances.

Question 64

Is *friend* a verb?

In the BNC, try a Simple Query on *friend* then click Node Tags or Show Tags in View Options. Open a new tab in your browser and repeat these steps using more contemporary corpora. The following examples all come from enTenTen.

Corp ex. 34

Come for food, drinks and to meet all the people you've tweeted, **friend**ed or (gasp!)

emailed who are in town for the conference.

Corp ex. 35 ... you want to be actively **friending** and directly connecting with 3-5 people a week that have influence in your industry.

The following examples are not verbs, but are derived from the new usage of *to friend*. Corp ex. 37 goes one step further using the negative prefix as well.

Corp ex. 36 The Host Community on LiveJournal was hosting a **friending** meme.

Corp ex. 37 Perhaps this will help us develop a theory of the entire cycle of **friending** and unfriending. The next example seems to say it all, really!

Corp ex. 38 Researchers flew under Facebook's radar, automated **friending** random strangers and THEN **friended** friends of those that **friended** the socialbots.

Question 65 **Is *will* a lexical verb in addition to being a future marker?**

Search for *will* in Simple Query. Then click Node tags: MD indicates its modal status, while tags starting with V indicate lexical verbs. Having found that it is a lexical verb, how does it work? Is there one dominant V tag? Why? What are its typical subjects, objects, complements and its typical grammar patterns?

Is this worth adding to your active/productive vocabulary?

Question 66 **Are the typical subjects of *couldn't bear to* nouns or pronouns? And are there any typical verbs following *to*?**

Remember to separate: *couldn't*. Here is one of many examples of this chunk in the BNC.

Corp ex. 39 Jessamy prized her privacy far too much, she **couldn't bear to** have a pack of journalists prying into every corner of her life.

Question 67 **Who, apart from journalists, *pry*?**

Search for this word in the BNC and use a simple webpage search (CTRL F) to find *journalist* on the page. A left sort will indicate other subjects of this verb, and much more, as certain phrases appear.

Question 68 **Is *pry* only followed by *into*? Can it be followed by an object instead?**

Could it be a phrasal verb? And/or a transitive verb? As it is such an infrequent verb in the BNC, a simple right sort will answer these questions.

In the process of answering these '*prying*' questions, the connotations of the word appear from its context. David Lee's Classification (View Options) reveals the one text type in which this word is most significant.

Question 69 **Pack o? When *pack* is not a short form of *packet*, what are the most common things that come in packs?**

Search for *pack of* in Simple Query and use Frequency to investigate the 1R column.

Do you think the author of the 'Jessamy sentence' associates journalists with *lies*, *wolves*, etc.?

Question 70 **Why might the author have chosen *prying into* instead of *look into*, *investigate*, *put under the microscope*, etc?**

This is a typical choice we make when we speak and write. Using Simple Query and Frequency as described above demonstrates the negative connotations of *to pry*²⁸.

The cumulative effect of *couldn't bear*, *pack* and *pry* is an example of so-called LEXICAL SUPPORT. When we are speaking and writing, one of the criteria we apply to selecting a word from among its synonyms is its connotations: words with similar connotations lend support to each other to create the desired rhetorical effect. See, for example, Corp ex. 25.

The difference between connotation and lexical support is that the former is a property of a word: the consistent understanding of a word's connotations across a language community can be shown through word association tasks²⁹. Speakers and writers achieve their rhetorical effect by co-selecting words with similar connotations, and this is lexical support.

Serendipity

Finding something valuable unintentionally. It is not unusual to find something of great interest in a dictionary when you are looking for something else. In Question 66 we were looking at the objects of *couldn't bear* and within four or five questions have ended up discovering some rather unrelated but valuable features of English and language in general. These serendipitous findings are as common in corpus work as they are in dictionary searches, and other areas of life. The index of this book lists several other references to serendipity.

Node forms

The next questions deal with prefixes and suffixes

Question 71 **Which words employ these meaning-laden suffixes?**

-cide, *-phobia*, *-phyte*, *-dactyl?* *-mania?*, *-proof*, *-gate*, *-esque*

²⁸ ske.li/lexcmi_pryinto_freq4r

²⁹ Paul Baker on the new use of gay

After a search involving operators (p.34), click on Node Forms to produce the list of word forms.

Simple query:

It is worth observing:

- what proportion of the word forms occur once only. See *nonce words* (p.85).
- if they look like someone is playing with, or exploiting, language.

Question 72 **Are words with the feminine suffix, -ess, fading from use?**

The word *actress* seems to be fading from use as most male and female thespians refer to themselves as actors. What other nouns end with the feminine marker, -ess? Can we conclude that using it is a marked choice these days? A comparison of old and new corpora may be instructive.

Is the use of the -ess ending a marked choice that people still make?

Question 73 **Which words that end in -sis form their plural with -ses?**

Once you have built a list of these, can you observe the text types that they are most typical of?

Doc IDs

Staying with the 58K hits of *-sis* (BNC), click on Doc IDs. A list of codes for each text in the corpus that contains words ending in *-sis* is generated. Given the Greek origin of words ending in *-sis* and given the fact that Graeco-Latin words figure significantly in academic prose, it is not surprising to find that the texts with the most uses of such words are academic in nature.

Click on **P** beside one of them and then click on the blue reference at the left of the concordance. The yellow metadata detail frame pops up.

Text Types

Click on the Text Types button to see a summary of the metadata for a particular search. For example, the academic nature of the *-sis* words is confirmed: small proportions of spoken language and written texts are tagged as Arts domain, Imaginative domain. At the bottom of Text Types, you can specify David Lee's

DAVID LEE'S CLASSIFICATION

W_fict_prose|W_fict_poe

[Documentation](#)

Classification categories which more specifically reveal the domains and sub-domains. To select more than one of David Lee's Classifications, use the vertical bar as in this screenshot.

See p.112 for more information about Text Types.

Question 74 ***Blench* seems like a perfectly normal English word, but one does not come across it very often.**

In what sort of texts does it appear?

Question 75 ***Harmony* is mainly a technical term in music. How do we use it elsewhere?**

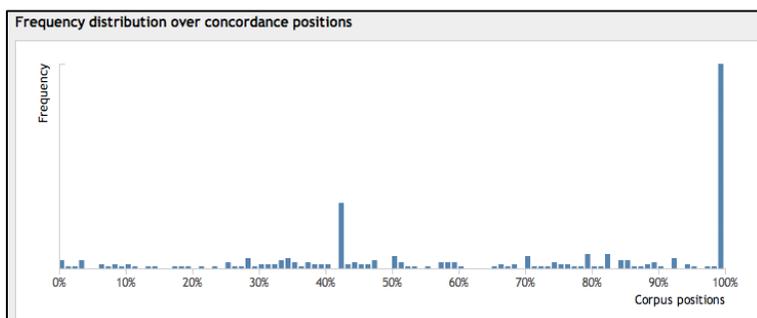
Find *harmony* via Simple Query. Click Text Types and see that according to David Lee's Classification it is mostly in the humanities. Clicking on **P** (positive) shows that it is mostly used in music contexts. To answer your question, click on **N** (negative) beside *W_humanities_arts*. This will show the examples of *harmony* that are not in this domain. There are still references to music, but it also answers the 'elsewhere' question.

Collocation

The next button in the Left Panel is collocation. This topic is dealt with at considerable length starting on p.143.

Visualize

The visualisation graph divides a whole corpus into 100 chunks of equal size. These chunks are not related to specific documents, domains, genres etc. Rather, it is necessary to think of the corpus not



as a collection of documents but as c.100 million words that have been drawn from the documents of the corpus.

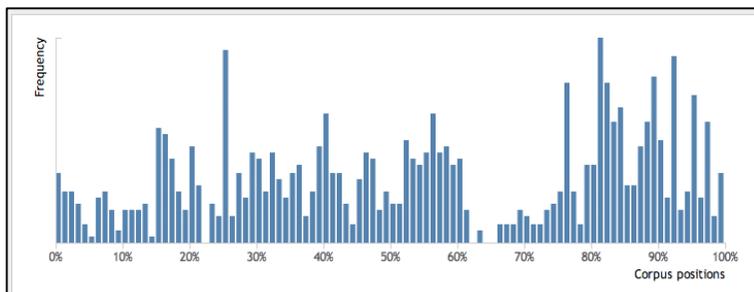
The word *eagle* occurs 1,000 times (according to Word List³⁰) and the graph of its distribution,

³⁰ ske.li/bnc_lemma_1000

as seen in this screenshot, indicates that it occurs significantly at about 42% and very significantly at 100%. Its other occurrences across the corpus are trivial.

By contrast, the second graph depicts the word *shared*, which also occurs 1,000 times – it is evenly distributed across the whole corpus.

Visualise reveals that some parts of a corpus are quite specific. It is usually caused by the fact that in these parts there are some specific documents in which specific words are used. The Visualize tool appears several more times in the book, demonstrating further applications of this data representation.



Question 76 **Compare the distributions of *blench* and other word forms that also occur exactly 21 times in the BNC.**

chromaticism cannibalistic lowbrow symbolist tone-units self-love ill-educated
cash-strapped rail-coaches shoebox **tweet** subliminally polysyllabic edification

4. Top Level Queries continued

We interrupted First Level Queries (from p. 21) to provide some important information about the Left Panel buttons. Via the Simple Query searches we learnt about customising our View Options (p.39) and then we proceeded with the Menu Items (p.45).

The screenshot shows a 'Simple query' interface with the following elements:

- A text input field for the query and a 'Make Concordance' button.
- Links for 'Query types', 'Context', and 'Text types'.
- 'Query type' radio buttons: simple (selected), lemma, phrase, word, character, and CQL.
- Fields for 'Lemma:', 'Phrase:', and 'Word Form:', each with a 'PoS:' dropdown menu (set to 'unspecified').
- A 'match case' checkbox.
- A 'CQL:' field and a 'Default attribute:' dropdown menu (set to 'word').
- A 'Tagset summary' link.
- 'Make Concordance' and 'Clear All' buttons at the bottom.

TIP It is not necessary to click any of the radio buttons beside Query type – it's enough to click in your field of choice.

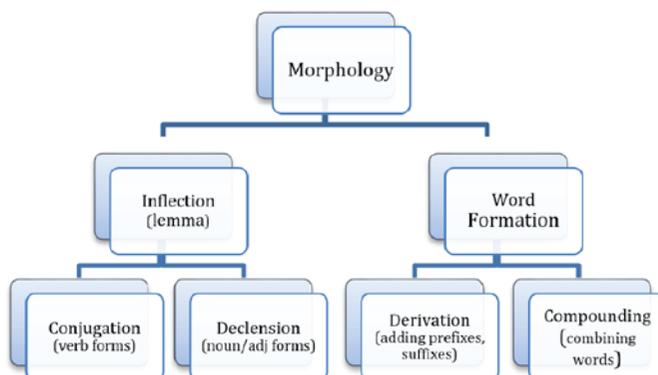
We now return you to the rest of the top level queries, as shown in this screenshot. For this, we need to have a grasp of the lower branches of the Hierarchy of Language which come under the heading of morphology.

Morphology

The taxonomy of morphology accounts for the processes that words undergo grammatically and lexically as can be seen in the left and right sections of the morphology taxonomy respectively. Its metalanguage is used throughout this book as many types of searches and post-processing involve these concepts.

Conjugation is the name of the process that verbs undergo when they change their forms to indicate person, number, tense, e.g. compute, computes, computed, computing.

Declension is the name of the process that nouns, adjectives, adverbs and determiners undergo. Nouns change their forms for number (*computer* –



computers; corpus – corpora), adjectives and adverbs for degree (comparative, superlative) and determiners such as *this those his their* for various combinations of number, person.

Derivation is the name of the process that involves adding affixes (prefixes and suffixes) to words. In comparison with the above two inflectional processes, derivation creates different meanings and/or changes the part of speech which gives it a different syntactic role.

Compounding sees the combination of two or more words to make a new meaning. Sometimes they are written as one word, they are hyphenated or they remain as two words. Nevertheless, *mobile phone, ballet dancer* and *boarder guard*, for example, each expresses a single concept.

There is another word formation process which involves the derivation of one lexeme from another (e.g. the verb *father* from the noun *father*) without any overt morphological change. This phenomenon is labelled variously: conversion, zero-derivation and zero-affixation. (See Carstairs-McCarthy 2002).

A set of words with the same stem or root that are the product of both inflectional and word formation processes is known as a Word Family.

Part of Speech Tags

A part of speech tag is a code that indicates a word's part of speech in the context in which it appears. The word *like*, for example, is tagged as a preposition (IN), a lexical verb (VV), an adjective (JJ) and a plural noun (NN) in the BNC, which the following four extracts show respectively.

- Corp ex. 40 Europeans have turned whales into a sacred animal, **like** the Hindu cow.
- Corp ex. 41 Where appropriate, we **like** to take an active role.
- Corp ex. 42 Methinks it is **like** a weasel.
- Corp ex. 43 You give me anyone's phone line and mail for a month and I'll tell you exactly who he is, how old he is, his **likes** and dislikes, his character, his worries, what he had for breakfast and supper, what time he wakes up.

The metadata in this corpus tells us that this last sentence is from *Esquire*, 1992, long before Edward ³¹Snowden's NSA revelations. It also seems to predate the preference for non-sexist language, as *anyone* here is assumed to be *he*. Nowadays we are more likely to use *they*.



Notice that the four *like* tags above, indicate more than POS. It is just the first element of the tag that identifies the POS. The following elements provide 'sub information'. For example, VV indicates a lexical verb, VM a modal verb and VVZ is the third person singular (Z) of a lexical verb.

³¹ <https://openclipart.org/>

Given that *like* occurs 144,481 times in this corpus, it is helpful to have some criteria to filter our searches. It is not usually necessary to show the tags, but Sketch Engine can be set to display the part of speech tags of every word if required. For example, when 'show all tags' is on (see p.40), the format is:

He /PP not /RB only /RB looks /VVZ **like** /IN me /PP , / , he /PP feels /VVZ **like** /IN mine /NN . /SENT

You might like to imagine the scenario in which this sentence occurred.

Part of speech (POS) tags are used by humans when we create searches for words functioning in a particular part of speech, and for identifying patterns in the grammatical company that words keep. As we will see time and again, they are used by concordancing software to group language phenomena in many ways.

Tagging is almost always done computationally using man-made algorithms: it is therefore unreasonable to expect 100% accuracy. Language is a very complicated business and not even linguists always agree on the part of speech of words in context. There are many tagging errors. For example, in the BNC, *broken* is never tagged as an adjective despite 150 occurrences of *broken glass*, 72 of *broken heart* and 26 of *broken promises*. See a 'broken. screenshot on p.40. It is not considered a lemma which means that it does not have a word sketch. Another example appeared when studying *cause* as a verb followed by a *to infinitive*. Most of the *causes* are in the phrase, *have cause to*, where it is a noun³².

Corp ex. 44 Later we shall have cause to doubt its accuracy and validity.

Corp ex. 45 Well, I wish he did have cause to be jealous of me.

The 18 examples of *likest* are tagged as superlative adjectives in the LEXCMI corpus. They are, in fact, all second person singular forms from an earlier period before English verbs shed most of their inflections.

Corp ex. 46 Out of my doors, knave, if thou likest it not.

This could be a useful sentence to learn as a chunk – it's far more elegant than today's, *If you don't like it, piss off, prat*. Corp ex. 46 is from *Thomas Lord Cromwell*, a play bearing Shakespeare's name but generally considered apocryphal.

The BNC was originally tagged with a tagset called CLAWS and was retagged by the University of Lancaster using an extended CLAWS tagset. Information about the so-called BNC2 is available on their website³³, which describes related activities such as other tagsets, tokenisation, ambiguity. Everything except when BNC2 was released!

³² http://ske.li/bnc_cause_to_vb.

³³ http://ucrel.lancs.ac.uk/bnc2/bnc2postag_manual.htm

However, most corpora use another tagset called the Penn Treebank Tagset, using software called the Tree Tagger (TT). Sketch Engine team recently retagged the BNC with PTB, which is the only version used in this book. A comparison of these two tagsets is available on the Versatile site.

Another layer of metadata that Sketch Engine adds when compiling a corpus is the lemma of each word. The most familiar uses of lemmas are headwords in dictionaries. *Lemma* can be seen in (**Error! Reference source not found.**) which we will take a quick look now at in preparation for the notes on lemmatizing.

Lemma

go	82713
going	62387
went	45631
gone	18000
goes	14414
Go	3939
Going	1061
Gone	250
Went	145
Goes	128
GOING	41
GO	15
goest	5
GONE	3
goer	2
gOing	1
gO	1
WENT	1
GOES	1

(*Stubbs, 2002*) also stresses that “word is ambiguous and we need to distinguish between word-forms and lemmas, because we need to differentiate between units of texts and units of the vocabulary of a language”. (p.25)

The LEMMA **GO** consists of the five WORD FORMS, *go, goes, going, went, gone*. This is another stratum of data that is available in Sketch Engine corpora, and is therefore an important choice when creating a query: are you searching for all forms of GO, or just the WORD FORM *go*?

Nouns also have lemmas and word forms, though declension is not a particularly notable feature of English

– it mostly distinguishes singular and plural forms, and possessives. Nouns in highly inflected languages have many word forms that express number and case. Czech, for example, has seven cases, two numbers singular and plural plus some historical relics, such as *pluralia tantum* and dual. This gives Czech nouns a potential for fourteen word forms.

English also has some historical relics in words deriving from Greek and Latin. For example, behind the word form *criteria* (3,798), we find its base form (lemma) CRITERION; ditto *bacteria, media, millennia, hypotheses, analyses, algae*.

Perhaps more interesting is the noun GO and its plural *goes*, which both appear in our first corpus example.

Corp ex. 47 Choddo had a **go** at a few and scored wins after just two **goes**.

When you search for a lemma, you get all of its word forms. **Error! Reference source not found.** shows the word forms of GO, and their frequencies in the BNC resulting from a lemma search not limited to the verb.

Such an observation might pique the linguist's curiosity as to whether the base form of verbs is always significantly more frequent than the other word forms. On finding that it is, the data needs to be interpreted to account for why this might be the case.

Thus when a corpus is compiled, one of the main machine processing tasks is *tagging* and *lemmatizing*, which involves putting every word in every text on a separate line accompanied by its lemma and POS tag. In this book, this is thus referred to as WLT.

The lemma field is case-sensitive. For example, *grant* (n) (BNC: 6,121 hits) vs. *Grant* (BNC 2,412 hits); *centre* (BNC 22,250) vs *Centre* (BNC 6,930). But capitalisation does not entirely guarantee a proper noun due to sentence capitalisation. Proper nouns can be specified using tags in CQL, as we see on p.124.

We prefer Simple Query to Lemma and Word Form when we don't have much knowledge about a word. It is often a good starting point, like a general internet search.

Question 77 **Is *holiday* used as a verb more today than in the BNC days?**

In the lemma field, type *holiday* and select verb from the POS droplist. Try this in both BNC and ukWaC. Because of the very different sizes of these two corpora, compare your findings using **per million**. What do you find?

Furthermore, Node forms reveals a preference for one form.

Visualize reveals its quite equal distribution across the corpus. Do any of these observations shed further light on Question 7?

Question 78 **Compare *thatcher* and *Thatcher* in the BNC.**

Given the period in English history that the BNC covers, it is not surprising that the Iron Lady is well-represented.

Inferring meaning of unknown words in a text is much discussed in language teaching literature. Language learners are often tasked with inferring the meaning of a word which occurs once in a long text. Compare this with what we have been inferring from multiple examples in short extracts from thousands of native speakers.

Question 79 **If you are not familiar with the common noun, *thatcher*, can you infer its meaning from these contexts?**

Can you deduce the verb from the noun *thatcher*? Search for *thatch** and right sort and you'll be well on the way to answering the question.

Question 80 **What can we learn about the word *slouch*?**

If you do not know this word, resist the temptation to look it up before doing the following. It could be an interesting voyage of discovery inferring its meanings and uses.

Do a lemma search without specifying part of speech. Then ...

1. Click on Node Forms and see if all the word forms are represented.
2. Click on Node Tags to make some other useful observations.
3. Click on Doc IDs and check that the word is not used disproportionately in one document, which does happen, especially with uncommon words.
4. See what Visualise offers in terms of spread across the corpus.
5. Click on Text Types to see in what genres *slouch* is most common?

Remember that if you have your page length set to at least the number of hits (p.41), you can scroll down and see quite a few patterns at once.

Question 81 **What other patterns of normal usage can we infer about *slouch* from a left sort?**

- Is there anything of interest about the subject when *slouch* is a verb?
- Only people? Both males and females slouch?
- Does *slouch* ever start a sentence?
- Is it used in both positive and negative contexts?

If *being no slouch* is a positive thing, what does this tell us about *slouch* itself?

Do any patterns of normal usage emerge about *slouch* from a right sort?

- Is it followed by any particular prepositions?
- Are the prepositional phrases places indicating where people *slouch*?
- Are any of the prepositions part of a phrasal verb with *slouch*?
- What is a *slouch hat*?
- Is the verb a troponym of (a) standing or (b) moving?

Question 82 **Do we *elaborate something*, or *elaborate on/about/oversomething*?**

Are there other COLLIGATION patterns to be aware of, e.g. preference for passive?

Search for the lemma and specify *verb*. Sort to the right as that is where the patterns in our data need to be observed³⁴. Beware of the tagging errors here – this word is also an adjective, in which case its pronunciation is different³⁵.

To answer the question about *elaborating something*, it is necessary to look for S V O patterns: who elaborates what? It is worth noting all of the objects and to see if they form semantic sets. In the section on Clustering (p.160) we look into this idea of grouping words into semantic sets that occur in a particular syntactic slot.

Left sorts do a similar job. Verbs are preceded by other elements of the verb phrase, such as modals and auxiliaries. This grouping of the auxiliaries reveals the word's use in passive voice, perfect aspect, to-infinitive form, thanks to the PERIPHRASTIC nature of English.

- Corp ex. 48 To **elaborate** a little on the definition I gave earlier, an 'edition' is any number of books, small or large printed from one setting of type.
- Corp ex. 49 Her answer was crisp, and she did not **elaborate** any further.
- Corp ex. 50 This point will be **elaborated** further in the discussion of 'crime' in the next section.
- Corp ex. 51 Although he discusses and exemplifies the other maxims, Grice does not **elaborate** on the simple instruction 'Be relevant'.
- Corp ex. 52 William Stukeley further **elaborated** the legend in the eighteenth century when he proposed that ...

These corpus examples also represent the polysemy of this verb. Examining its collocations and colligations even in this tiny sample, the associated patterns with each meaning can be gleaned. These lexicogrammatical patterns are at the core of WORD TEMPLATES as we explore later.

The high frequency of *and* suggests a pattern of normal usage when another verb precedes it frequently, e.g. *developed and elaborated* occurs five times.

Further discussion of *elaborate* can be found at the MAELT website³⁶.

Question 83 **Are both *centre* and *center* used as verbs?**

Does the BNC represent both spellings? Use LEXCMI for a better sample of US and UK varieties? When did American reform its spelling?

Search for the lemma and then click on Node Tags in the Left Panel to see what its parts of speech are.

What are the patterns of normal usage of the verb, *centre*? Is it used in both transitive and intransitive structures? What are its typical subjects and objects? Are there any salient prepositions?

³⁴ ske.li/bnc_elaborate_vb_colls_r

³⁵ http://bit.ly/howjsay_elaborate

³⁶ http://bit.ly/maelt_elaborate

Verbs of Motion

Verbs of motion, such as the troponyms of *go*, are intransitive, i.e. Subject + Verb, so when investigating the valencies of these verbs, we are primarily interested in the subjects. This information is part of the semantics of the verbs and it provides a structure for using them. For example, are the verbs performed by people only?

Here are some more words for *ways of walking*:

tramp, ramble, rove, tiptoe, lope, wander, swan, flounce, mince, pace, step, tread, trudge, meander.

Such a list raises such questions as:

- What are the differences between them?
- What use are they? Who uses them and when?
- Why are there so many ways of walking?
- Are there as many ways of doing other things?

We look at approaches to these questions under componential analysis (p.76), word templates (p.161) and sketch differences (p.168).

Question 84 **Do only people lope, meander, tread and mince?**

If such verbs have other subjects, do they mean something quite different, or are there common elements of meaning?

Also of interest to verbs of motion are the adverbs that describe how they are performed and more significantly in valency studies, the spatial adverbials that express where the verb takes place, the range and the destination.

Many verbs of motion are also used transitively. This means that the pattern is Subject Verb Object (S V O). To find some examples, search for some of these verbs + *the*, e.g. *walk the*. You can expect to find S V O, which by definition, is not a verb of motion. They will have a different meaning and/or usage.

These verbs often combine with adverbs and prepositions to make phrasal verbs, which are discussed in many places up throughout this book. See the Index. They express a wide range of transitivity.

These verbs also occur in idiomatic phrases, e.g. *walk tall, walk the plank, walk the beat*. Match these three phrases with these three glosses: (a) a form of execution, (b) patrolling, (c) pride. Now find idiomatic usages yourself.

Many verbs of motion also occur in delexical structures, such as *take a stroll*, in which the verb is "noured". And many of the noun forms of these verbs are used in various compounds and expressions such as a *walkover, strollathon, runaway, bush walk, protest march*.

Question 85 **Are there any patterns of normal usage in the adverbs that describe different semantic groups of verbs of motion?**

Compare the adverbs used with verbs of walking with those of travelling. Or those of Body-Internal Motion (Levin 1993: 49: 261)²³

fidget, gyrate, rock, squirm, twitch, wriggle.

Question 86 **Can you think of some other series of troponyms that start with *go*, not in a walking sense?**

Series of troponyms and other semantic relationships are given at Wordnet²⁴.

You could make a hierarchical diagram or mindmap with *go* as the headword.

By now we have arrived at some answers to o. And hopefully much more. Let us now move on.

Phrase

The definition of a phrase for this search field is 'a group of adjacent (a.k.a. *contiguous*) word forms in a fixed order'. The definition of *phrase* in general parlance is not as clear. According to Oxford Dictionaries³⁷, a phrase is:

1. A small group of words standing together as a conceptual unit, typically forming a component of a clause.
 - 1.1 An idiomatic or short pithy expression.

The following definition comes from the *Glossary of Linguistic Terms*²⁵:

A phrase is a syntactic structure that consists of more than one word but lacks the subject-predicate organization of a clause.

So we can see that the definitions of *phrase* are more general than its definition for the Sketch Engine search field.

A search for *go fish* in the phrase field returns these four lines only in the New Model Corpus (NMC),

Leachman. He played checkers, zilch, and **go fish** with her. He rested some more and then I to
 10: 36 PM: "I had to laugh at your comment, **go fish** . Yes, we were a happy bunch in the 70 ' s we
 RTITO - the (fashion) bash not to be missed! **go fish** background information: i was " crowned " M
 ies, some Nick TV , we played checkers and **go fish** in which he kicked my behind in both games.

³⁷ http://bit.ly/oxdict_phrase

Corp ex. 53 I had to laugh at your comment, **go fish**. Yes, we were a happy bunch in the 70's weren't we?

We know that searching in the Simple Query field returns the inflected forms of both items: *go fish* returns 20. However, when *go* conjugates and *fish* declines, we no longer have the fixed phrase, *go fish*.

By the way, is the apostrophe correct with 70's? (See p. 33). It is important to remember that incorrect, non-standard, deviant, erroneous forms, call them what you will, occur in the language that we are exposed to, our INPUT.

The phrase field is also case-sensitive: *Go fish* returns seven lines, and *Go Fish* returns four.

We have just seen that searching for *go fish* in the Simple Query field produces different results from the phrase field. But in both cases the two lemmas are adjacent. If we want to search for words that might appear between them, we use the Context search which is described from page 95 onwards.

Question 87 **I wonder if *I had to laugh* is a fixed phrase in English.**

o opens with the phrase, *I had to laugh*. This could arouse one's suspicions, pique one's curiosity.

The BNC seems to think so, but only in a few text types. It usually stands alone. What can you find out about this phrase?

Question 88 ***I wonder if* / *I wonder whether*?**

Both of these are bountiful in the BNC but not equally so. Can you identify any systematic usage in one over the other?

Question 89 **Does *I wonder* open an indirect question that requires subject-verb inversion?**

For example, *What time is it?* inverts after an introductory clause: *I don't know what time it is*. Look in the BNC at *I wonder* and see if this inversion is standard. A sample of 1,000 right sorted lines is quite revealing.

Wh- words:

who

whose

what

when

where

why

Question 90 **Do the subject and verb typically change places in indirect questions?**

To answer this question convincingly, study the different *wh*-words and the word order of their clauses.

Search in the Phrase field with: *Do you know wh.**

Here is just one example exhibiting this inverted word order.

Corp ex. 54 Do you know **whose** button this is?
 After performing the search, set First level as node. Change the attribute in the Second and Third levels to Tag³⁸. Pronouns certainly dominate.
 Click on the **P** beside any of the lines to see sentence examples.

Question 91 **Is the phrase *in mixed company* euphemism, code, a cultural reference?**

Mixed company was referred to earlier in the context of fig-leaved language (p.9):

Victorians and vicars birdwatching in mixed company blushed when they spotted one.

Search for it in the BNC using the phrase field. The small number of hits is indicative of its limited use in English. Reading either side of the node may lead to recognising its reference to social acceptability.

Question 92 **To return briefly to Margaret Thatcher, does her nickname appear in the BNC?**

Is it used as frequently? Do *iron lady* and *Iron Lady* return the same results?

Before you look, try to predict the answers. Check it in the Phrase field, then using Sentence instead of KWIC view, compare what is said about the lady with what is said about the nickname itself.

Question 93 **Is the '*way how* to do something', a structure in English?**

Before considering the structure, consider the collocation: how frequent are the two words individually? Search for them and write them under Frequency here.

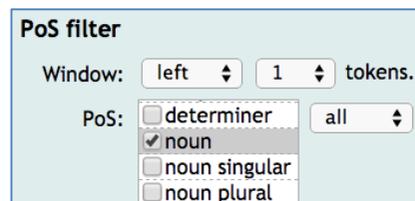
Word 1	Frequency	Word 2	Frequency	Co-occurrence
way		how		

Before searching for the phrase, *way how*, think about how frequently you would expect them to co-occur.

Type it into the Phrase Field and see how often this pair co-occurs. Then see how often it is followed by a *to infinitive*. Is this a pattern of normal usage?

³⁸ ske.li/bnc_doyouknow_tag_tag (UK)

It is worth noting that other 'way-like' words do not occur in front of *how* either. Try some of the following: *strategy, system, approach, plan, method, technique + how*. The nouns that do precede *how* can be seen here³⁹. This search was filtered as shown in this screenshot.



Question 94 ***more than likely* is perfectly standard English. In what situations?**

In what sort of texts is it used?

Is it used just as a reaction, or in the course of text?

What does it express the speaker's attitude towards?

Question 95 **Can we say more than unlikely?**

We can say anything. But do we? Significantly? In specific contexts? If not, what are some alternative wordings?

Question 96 **Are there any patterns of normal usage of *more than*?**

Search in the Phrase Field and form bundles using Frequency 1R. And 2R. Select Ignore case.

Fixed phrases, chunks and bundles

The explosion of interest among linguists in multi-word units in the last twenty years has seen an explosion of terminology. The innovating linguist has sometimes used a known term in their own specific way or created a new one. The following componential analysis of three of these terms is unlikely to meet with unanimous acceptance among linguists: it can be seen as an attempt to distinguish them as a basis for further work, and to clarify how they are used in this book.

³⁹ ske.li/bnc_1_n_how (UK)

MULTI-WORD UNITS	Phrase	Chunk	Bundle
semantically whole	yes	yes	not necessarily
flexibility of words	permits variation	fixed	word forms only
length	depends on meaning	depends on meaning	fixed (n-gram)
storage	holistic	holistic	not stored
phonology	usually one nucleus	usually one nucleus	not pronounced as units

Native speakers of any language have vast stores of multi-word units (MWU): using them contributes significantly to fluency and idiomaticity. They are not tagged as MWUs in corpora which means that we cannot derive them from corpora automatically. Rather, searches reveal them as patterns that hover around specific words and combinations.

Question 97 **What can you discover about the following chunks?**

What we need to know about chunks is how they fit into streams of language. When they appear in KWIC format, and sorted left and/or right the company they keep emerges quite clearly. Furthermore, when observed in the BNC, the metadata will indicate the registers and domains in which they occur.

Some of these will be more fruitfully sought in larger corpora such as ukWaC and enTenTen.

every so often	if truth be told	only a matter of time
age before beauty	never walk alone	alive and well
by and large	all very well	be it on your own head
clean and tidy	gloom and doom	from cradle to grave
crux of the matter	food for thought	given half a chance

Learners of English would do well to prioritize such chunks in their study plans. This involves observing them in their contexts and getting a feeling for how they might meet a need, especially when there is no direct equivalent in the first language.

Question 98 Does the phrase, *once in a lifetime*, function as an adjective?

It occurs 23 times in the BNC, followed by nouns such as *affair, event, find, holiday, opportunity*. How is it tagged?

Sentence stems

At the heart of an English sentence is a clause which expresses its message. In addition to message-bearing clauses, sentences also contain single and multi-word adverbials, e.g. *although, be that as it may*, that express relationships between sentences and across text.

In some linguistic work, such organisational language is referred to as sentence stems. They make important contributions to text on the levels of discourse, stylistics and pragmatics. These are often chunks expressing temporal and causal relationships in the discourse.

Apart from single word items, they are mostly prefabricated chunks, which ensures that they are well-attested in corpora. With the Phrase Field being case sensitive, these sentence-initial chunks can easily be found. For example,

Remember that	Given that
It may be the case that	It should be noted that
There is no doubt that	It is clear (and other adjectives) that
It appears/seems that	It has been suggested that
Having said that	This does not mean that
Despite the fact that	It was agreed that

Question 99 How are the above sentence stems followed up?

As we asked in Question 97, what are their collocation and colligation patterns?

It is valuable to have such chunks at your fingertips, but knowing how they are typically followed up is also part of knowing them. Look them up in the Phrase Field and use Frequency to see what follows them at various levels. It is sometimes surprising to find that even the most general of these are followed by quite a limited set of items.

Make sure you are showing the David Lee's Classification as well. There are observations to be made.

Word Form

Unlike Simple Query and Lemma searches which search for lemmas as well as word forms, words sought via the Word Form field are returned only as entered. This is particularly useful when investigating the use of a word form that is also part of a lemma, for example, the noun *going* is a word worth studying separately from its lemma *go*.

Question 100 **Are both *though* and *although* used to start sentences? Equally?**

Perhaps you already know the answer to this through intuition, or through being taught it. Match Case is useful for identifying words at the start a sentence.

Word Form: <input type="text" value="Though Although"/>	PoS: <input type="text" value="unspecified"/>	<input checked="" type="checkbox"/> match case
---	---	--

You don't have to process tens of thousands of lines. Rather, click on **Node Forms** in the left panel.

Question 101 **Is it true that *and* and *but* cannot start sentences?**

These two questions deserve deeper answers than *yes* or *no* – check the Text Types as well. Also, compare the hits per million of those that start sentences with those that don't.

Question 102 **Is the plural *mouse* the same for the animal and the computer peripheral?**

If *mouses* occurs in corpora at all, check the text types as well as context and co-text. The frequency of your findings with its co-text should be decisive.

The CHILDES corpora are used for the study of first language acquisition. Even though young children are unlikely to ever hear the word *mouses*, at least not from their caregivers, once they start to form plurals according to the paradigm that they figure out, they do form the plural as if it were regular. This is after a period of having used the irregular form correctly, based on imitation.

See the English CHILDES corpus⁴⁰ for the instances of *mouses* sorted according to the age of the children. They also form it correctly: search *mice* in the Word Form field. Compare the ages of speakers of *mouses* and *mice*.

Speaking of *childs*, do children form this plural according to the rules also?

Children go through the same language acquisition stages with irregular verbs: they use the correct forms first, then *goed*, *comed*, *taked*, then they realise that there are

⁴⁰ ske.li/childes_mouses_age

both regular and irregular forms. For more on this a well-documented phenomenon in FLA, see Jean Berko and her wug tests, which has many citations on the internet²⁶.

Question 103 **Have you ever wondered if some surnames are also common nouns?**

For example: Thatcher, Butcher, Baker, Teacher, Hooper, Cooper, Blower, Archer, Brewer, Parker. This must pique your curiosity about the origin of surnames.

Some readers may be familiar with Bill Posters. Open this link⁴¹ and identify some of the patterns of normal usage around the node. You might wonder if Bill is a person, etc.

If the collocation *pique curiosity* piques your curiosity, see this page via the permalink⁴².

Countability

The issue of countability of nouns looms large⁴³ in learning English as a foreign or second language, although it seems that few native speakers are aware of the concept. In the first of the next two corpus sentences, *scholarship* refers to academic pursuit and has no plural form, whereas in the second it is a financial award and there can be more than one of them.

Corp ex. 55 Fine art, architecture, **scholarship**, literature, music and piety jostled for attention alongside hunting, feasting, jousting, politics, diplomacy and war. (ukWaC)

Corp ex. 56 Lawrence won a **scholarship** to Nottingham High school, leaving at the age of 16 to work as a clerk in a local factory. (ukWaC)

Another example is the word **experience**. In the first of the next two sentences, it refers to knowledge of the world and has no plural form, whereas in the second it is an event of which there can be more than one. Note that in other languages, these two notions have quite distinct forms, e.g. Czech has *zážitek* and *zkušenost*, while German has *Erfahrung* and *Erlebnis*.

Corp ex. 57 I am a political refugee with personal **experience** of being the target of hate speech ... (ukWaC)

Corp ex. 58 British gay men reported having at least one **experience** of unprotected sex last year, ... (ukWaC)

Countability is an important factor in the use of determiners including the definite and indefinite articles. We shall not investigate this here.

Mass nouns such as *water*, *toothpaste*, *money*, *homework* and *coffee* only have plural forms when they refer to discrete entities, however, not all of them do. Can you find some illustrative sentences in corpora comparing these uses?

⁴¹ ske.li/ett_bill_posters (UK)

⁴² ske.li/bnc_pique_curiosity

⁴³ ske.li/bnc_loom_large

We therefore use quantifying nouns that do take the plural if necessary, e.g. *tube of toothpaste, bar of soap, cup of coffee, piece of homework*.

English also has expressions of quantity such as *array, bunch, group, heap, multitude* and *range*, which are typically followed by *of*.

All of these **nouns + of** can easily be sought in the Simple Query field and a frequency listing will show the items that follow them. It will show any items that meet the criteria, but human intuition will distinguish *bars of chocolate* and *soap* from *bars of music* and *bars of a cage*.

Question 104 Are the 'quantity nouns' that follow *of* in the singular or plural?

Here are two examples. Note that in the first, *multitude are* (plural) and in the second *a multitude binds* (singular).

Corp ex. 59 A **multitude** of medical conditions **are** due to being overweight, which in many people may simply be the product of self-indulgence rather than food addiction.

Corp ex. 60 ..., yet a **multitude** of non-verbal signals **binds** them together in a group.

Use the Word Form field to investigate some more examples of these words as they are used in the singular. They can also be used in the plural, but then agreement is not contentious.

Some nouns are used mainly in the plural. To search for the plural form only, use Word Form.

Question 105 Are singular nouns always used with a determiner? Do they have singular verb agreement?

Singular nouns are nouns which have no plural in at least one of their senses. For example, *ability, brainchild, crux, dearth, epitome, fray, grounding, heartbeat, impasse*, and etc. through the alphabet. See Hunston and Francis²⁷.

1. Choose some of the nouns listed above and search for them in the Simple Query field.
2. Then look at the Node Forms – we are expecting they will reveal a high proportion of the singular form. Make a note of the frequencies.

3. Click on the **P** beside the singular form. To isolate those preceded by determiners, click on **Filter**. Here we need to enter the CQL query, as in the screenshot here, [tag = "DT"] under Query Types, and set the range -1, 0, as in this screenshot. In the Context section (p.95), we will see an alternative way of filtering such a query.
4. Compare this frequency with the one you noted. To see which determiners are used, use Frequency again - 1L + node. Is there one that is salient, in that is, it stands out from the rest?
5. To see the bundle in which this appears, use several levels in Frequency, e.g. 1L Node 1R. Are there any salient bundles? It's Zipfian!

Filter: positive negative
 Selected token: first last
 Search Span: from to
 Simple query:
 Query type: simple lemma p
 Lemma:
 Phrase:
 Word Form:
 Character:
 CQL:
 Filter Concordance

This was a long journey, but the destination was not the only aim. There are no shortcuts for the pioneer – only once the trail has been blazed can the search for answers be shortened.

Character

While this field is of particular relevance to languages such as Chinese and Japanese, it serves English in some useful ways too.

Brackets and parentheses

Bracketed or parenthetical information is often an admission of irrelevance or downgrading: by putting a comment in brackets, the author is saying, *well it's not germane to my point, but it occurred to me while I was writing and I thought you might be interested ...*

In contemporary writing, therefore, there is very little information in brackets (even footnotes rarely contain information). This makes reading a very smooth process indeed. It places a demand on authors to work everything germane into the regular syntax of well-crafted sentences, or to leave some material out. In the following sentence from the BNC, the author uses parentheses twice and for different purposes. Curiously, she gives some examples of parenthetical information in parentheses!

*An **aside** is a dramatic device in which a character speaks to the audience. By convention the audience is to realize that the character's speech is unheard by the other characters on stage. (Wikipedia)*

Corp ex. 61 His stimulus material was a film and he particularly noted the expression of parenthetical information (asides, or personal expression of feeling) in the recall of such events (something which does not appear in recall of verbally presented stories).

The second use is more of an aside, something that she mentions in the first parentheses!

Question 106 How much information typically appears in brackets?

Type a left bracket into the Character Field. Make a sizeable Sample and Sort to the right.

What appears in brackets? How many words? Are there any clauses, i.e. containing a finite verb? Give some thought to the reasons for putting such things in brackets. And observe the text types of your findings.

Consider carefully the use of brackets in your own writing.

@ a.k.a. at

If you are looking for email addresses, enter @ and click Node Forms. These are somewhat more frequent in more recent corpora than the BNC.

Question 107 How early did people have email addresses?

Starting with the BNC, enter @ into the Character Field. Use the Select Lines check boxes to separate the email address from the other uses of @. And in View Options, choose Publication Date.

Question 108 How is the at symbol used when not in email addresses?

Search for it in the character field. Sort node.

Question 109 Are capital letters used after colons and semicolons?

Type these punctuation marks into the character field, perform searches and then use Frequency to make a list of at least 1R.

Question 110 Is it true that *above-mentioned* (with or without a hyphen) and *afore-mentioned* are used in written and spoken language respectively?**Question 111 How do answers begin?**

Type a question mark into the character field. This contrasts with Question 24: Do you love me (p. 34), where we could not use a question mark in the Simple Query field. There are more ways of killing a cat, as the old saying goes.

Corp ex. 62 ... there are more ways of **killing a cat** than skinning it ...

Corp ex. 63 Truly, as the old saying goes - 'there are more ways of killing a cat than choking it with cream!'

Operators used in fields other than Simple Query

The following operators work in these fields: Lemma, Phrase, Word form. Some of them behave differently in CQL, as we will investigate later (p.124ff).

full stop	before, during or after a word, a full stop finds that word with any letter in place of the full stop.
	In the Word Form field, b.g gives <i>big bag beg bog bug</i> . Click on Node Forms to obtain the list. This is useful for pronunciation teaching in generating lists of words that can be used in minimal pairs activities.
	In the Lemma field, .rick gives all word forms of the lemmas that have these four letters starting in second place, e.g. <i>brick, tricks, pricked, tricking</i> . This typically offers words that rhyme. ²⁸ Note that in Simple Query, a question mark provides this list.
asterisk	after a full stop finds any number of letters, including none, in place of the full stop. In the Lemma field, b.*g gives 153,551 hits ⁴⁴ , including <i>broadcasting, brainstorming, bing, beg</i> . In the Word Form field, it returns 264,766 hits.
	.*ship finds words ending with <i>ship</i>
	oxy.* finds words beginning with <i>oxy</i>
	.*advantag.* finds words containing <i>advantag</i> of which there are 27 different ones. Use Node Forms to create this list.
backslash	To search for a full stop, use the backslash and the full stop. In computing, this is referred to as <i>escaping</i> .
	Dr\ finds <i>Dr.</i> only. Without the backslash, you also get <i>Dr.</i> , along with hundreds of other three-letter combinations starting with <i>dr.</i>

And now for a quick comparison:

Simple Query	?ock ⇒ 22,695	does not work with a full stop
Lemma	.ock ⇒ 22,685	does not work with a question mark
CQL	".ock" ⇒ 22,685 lemma attribute	does not work with a question mark

⁴⁴ http://ske.li/bnc_b.*g

Affixes

The full stop operator is like a blank tile in Scrabble – it can stand in for any letter. Used alone it is particularly useful for alternative spellings. In combination with the asterisk however, we discover what prefixes and suffixes attach to stems. And as we have already seen many times, some things will be very typical and others less so. Given our creative use of language, many novel words are created in this way, as we see among the less frequent ones. Here, for example, are six of the dozens of recent coinages, a.k.a. neologisms, starting with *faux* that occur once only in enTenTen.

*faux-modest, faux-sarcasm, faux-sophistication,
faux-patriotic, faux-lesbian, faux-deferential*

A word that only occurs once in a corpus is referred to as a hapax legomenon.

Question 112 Which words beginning with *faux* occur in the BNC?

Search **faux* in Simple Query or **faux* in other fields. What does the BNC data indicate about the recent history of this prefix?

Question 113 Did any of these words catch on and enter the language?

The biggest corpus in Sketch Engine is enTenTen. Search for them there. And there is always the world wide web.

By the way, do words *catch on*? What do you think of this collocation?

Question 114 Look for them in corpora more recent than the BNC, e.g. enTenTen, ukWaC.

Nonce words are created for a specific occasion and cannot be expected to be met again – they often go unrecorded. In corpus terms, many nonce words are also hapax legomena. According to Zipf's Law, between 40% and 60% of words in a corpus can be expected to occur only once. We will use the Word List tool (p.173) to search for words that occur once only.

New words in a language may start their lives as nonce words. These can only be open class words.

Question 115 What are some examples of the suffix *-ity* indicating the state or quality of being the adjective (LGSWE, pp 232-3)?

Try **ity*. This gives all the words that end in *-ity*. Those which are the abstractions of adjectives are typically singular nouns (see also Question 106). Nevertheless, performing this search in the Lemma field returns plural forms as well, e.g. *opportunities, eccentricities*. Are these also abstract forms of adjectives?

Question 116 **Is *dis* the only prefix that our titular word, *discovering*, takes?**

Question 117 **Search for *. *cover* in the Word Form field. What are the differences between the words that you find? Do their contexts and cotexts provide any clues?**

Question 118 **How do other prefixes fare?**

Try *mega.** in the Word Form field in the BNC. Click on Node Forms – the list has *Megan* at the top while the following occur only once:

megafirm, megafabulously, megamoney, megamum, megamouth

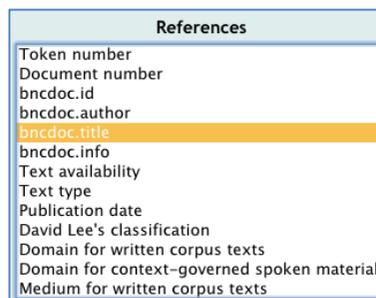
You might also try *hyper.**, *over.**

Question 119 **Given *that Megan* is a fairly uncommon name, how did it get such a high score?**

How many *Megans* do you know or know of? Wikipedia's list of *Notable Megans*²⁹ might jog your memory.

Clicking on Text Types and David Lee's Classification shows that more than half come from fiction. Click on the **P**.

To find out what books they come from, select *bnc.doc title* in View Options' References as shown in the screenshot.



Question 120 **Can we identify any language change through suffixes?**

The masculine form of words used to refer not only to men, but to mankind, and was therefore considered gender neutral. There has been a shift away from this and person has become the suffix replacing man. For example, *chairperson* is now preferred over *chairman*, regardless of the gender of the person occupying that position. Search various corpora to find both of these endings and compare them.

Similar investigations could be conducted with the new, politically correct suffix *challenged*.

Discovering English. Does English need discovering? Many linguists, of the armchair and empirical varieties, have been writing for over a century about *English as she is spoke*³⁰ and writ. Now that English has achieved lingua franca status, describing her every curve and nuance has virtually become an industry. This serves as a great impetus to linguistic departments in universities – their findings are filtered down to a vast community of language professionals by a relatively small band of journal and book publishers. Since the work of teachers is above all to teach, and of translators is to translate, language professionals are very glad of the work that linguists do. As Gleiser recently wrote, in defence of scientists:

Behind the scientist's complicated formulas, the tables of data obtained from experiments, and the technical jargon, you will find a person eagerly trying to transcend the immediate boundaries of life, driven by an unstoppable desire to reach some deeper truth. (2005)

Nevertheless, of great interest to the current book is treating "the learner as a research worker whose learning needs to be driven by access to linguistic data and whose role is to identify, classify and generalize that data" as Tim John wrote almost a quarter of a century ago. For this he coined the widely-used term, data-driven learning. Thus, we are less interested in the findings of linguists for the consumption of teachers, learners and translators than in the process of discovery which allows learners to create knowledge for themselves.

Piaget (1896–1980) said: [pic]

Each time one prematurely teaches a child something he could have discovered for himself, the child is kept from inventing it and consequently from understanding it completely.

As elegant as this may be, the support of a teacher can enable learners to achieve things that they never would alone. This is part of what Vygotsky (1894–1936) meant by his Zone of Proximal Development. And guided discovery is a logical consequence of this co-operation between the teacher as *the guide on the side* (King 1993), and learners. Teachers facilitate learning rather than transmit information. Johns again: the task of the learner is to "discover" the foreign language, and that the task of the language teacher is to provide a context in which the learner can develop strategies for discovery - strategies through which he or she can "learn how to learn".

To return to our question: Does English need discovering? In our work so far, *Discovering English with Sketch Engine* has been enabling you to obtain language data that you can interpret to answer set questions with the help of instructions for the Sketch Engine's tools as they are introduced step by step. The process is intended to develop your awareness of language as a linguistic object, provide you with information about English that may be new to you, and above all to equip you with the skills to find answers to your own questions. It is hoped that knowledge is being created within you every step of the way. This is what needs discovering.

On with the show!

Question 121 **Are there any affixes making verb forms out of *friend*?**

In the BNC, put **friend.** in the lemma field and choose verb from the POS droplist. Then click on Node Forms.

Repeat this search in the enTenTen corpus. How do the node form lists compare? Do you find that the standard form still has a strong place in contemporary English?

Question 122 **It emerges that *be-* is often used as a prefix. Does it have a meaning, or perhaps a function?**

Is this the case with both nouns and verbs? Perform a search and use Node Tags.

Now search for *be.** in the lemma field and select verb. Make a frequency list with lemma as the attribute.

The words of interest are those which are obviously *be + free morpheme*, e.g. *be-friend, bewitch, bejewel*.

The data is not the answer – it is a starting point, as we have seen in the more or less 100 questions so far. The question asks about the meaning of the prefix.

Question 123 **Why do people say things like *not uninteresting*? Is it not enough to say *is interesting*?**

We find that *not uninteresting* occurs once only in the BNC and 162 times in en-TenTen⁴⁵. Using *.**, however, we find that *not un...* is quite well-represented, even in the BNC⁴⁶. Is this negative particle + negative prefix a pattern of normal usage? Does it belong to any particular text types? If not, can any other explanation be found in the data?

We will now observe how this operator can be used between words.

Question 124 **What's the correct form: *no holds barred* or *no holes barred*?**

Search for *no.* barred* in Simple Query.

Furthermore, can you infer the meaning and/or function of the phrase from its context and contexts?

Question 125 **I've heard something that sounds ironic: *in the looks department*. Apart from actual departments, what *departments* are referred to?**

Try a search for *in the.* department*, then click on Node Forms.

⁴⁵ [ske.li/ett_not_uninteresting](#) (UK)

⁴⁶ [ske.li/bnc_not_un_freq](#) (UK)

As we have already noted a number of times, patterns of normal usage emerge at the top of the lists and graphs, while the exploitations of the language are inevitably lower down. Hanks' *Patterns of Norms and Exploitations* (2013) is a useful dichotomy. Some examples in the BNC: *the brain department, the dick department, the butchering department* are at the bottom in terms of frequency. There is even one sighting in the NMC of *in the irony department!* Is that eponymous or just plain ironic?

An interesting facet of language that corpus studies have unearthed, is the influence word meanings have on each other when they habitually co-occur.

Firth noted in his work on phonology, there are many ways in which a sound can be influenced by its environment – he referred to this as prosody³¹. By analogy, some of his followers applied his term to meaning (semantics) – thus SEMANTIC PROSODY. For example, the objects of the verb *to harbour* (BNC 434: 3.8 per million) include *grudge, resentment, bacteria, terrorist, hatred, prejudice*. Adjectives in the vicinity include *ineffective, deep-seated, irrational, hostile, illegal*. This negativity around the verb is its semantic prosody. A list of collocates⁴⁷ of *harbour* is quite revealing, as we will study in greater detail from p.143 onwards.

Question 126 **What semantic prosody emerges when the phrase *or just plain* is used?**

Search for this in the Phrase field, and make a Frequency list of the words 1R. They are either nouns or adjectives, which means that the POS of *plain* must vary here. To find out how it has been tagged, click on Node tags or use lempos as an attribute.

And what do you notice about the nouns and adjectives in terms of their prosody or connotations?

Question 127 **Is it true that more English words have *k* as the third letter than as the first?**

This may seem like a bizarre question, but it concerns a finding by psychologists Kahneman and Tversky (1973) about the way humans process information³². Before we investigate it, what do you think the answer is?

To find words with k as their third letter, type *..k.** into the lemma field.

To find words that start with k, type *k.** into the lemma field.

With the results of both, use Frequency: choose attribute lemma, Ignore case and set the Frequency limit to 1. The first thing we notice is the Zipfian tendency for a small number of words to have a great number of occurrences.

⁴⁷ ske.li/bnc_harbour_v_colls (UK)

To find out how many such words there are, it is necessary to keep clicking Next till you get to the end, and multiply the number of pages by the number of lines per page. There are 79 pages of lemmas with k as their third letter⁴⁸.

Unfortunately, there is no 'Go Last' button that would jump to the end of the list in a single click. However, along the way, you will observe some interesting words and non-words³³!

To exclude hapax legomena, make the Frequency limit 2. In the case of words starting with k, this reduces the number of pages from 43 to 20⁴⁹, in accordance with Zipf's Law.

The answer to the actual question is yes, contrary to the intuition of most of their subjects. Remember that computers process information differently from people.

Vertical bar

The vertical bar can be placed between items so that the concordancer searches for them at the same time. In regular expressions it means OR.

Question 128 **Can *amid* and *amidst* be used interchangeably?**

Use the vertical bar between them: **amid|amidst**. Make a Sample, investigate their text types, etc. Can you find any systematic difference between them? Whatever answer you arrive at will influence how you choose to use one or the other.

Whilst investigating this pair, observe their semantic prosody. It is of surprising interest. Once again, a list of collocates⁵⁰ is revealing.

Question 129 **Are the same prepositions used with synonyms?**

struggle|battle|fight searches for all of these items. This particular search reveals interesting patterns in preposition use. Since all three are both nouns and verbs, it is worth choosing PoS when making the query. Various frequency listings are useful here.

Question 130 **Can the vertical bar be used to look up British and American spellings at the same time?**

Try color|colour and centre|center, for example. Try such searches in both the BNC and LEXCMI and see which corpus is the more British!

⁴⁸ ske.li/bnc_k3_lastpage (UK)

⁴⁹ See where the hapax legomena begin: ske.li/bnc_k_words_p20 (UK)

⁵⁰ ske.li/bnc_amid_colls (UK)

Corpus Query Language – a brief introduction

Until now, we have been searching corpora by typing into the Simple Query, lemma, phrase and character fields. In the background, Sketch Engine converts these into Corpus Query Language (CQL) queries, which you can see by clicking on ConcDesc at the bottom of the left panel. This screenshot shows the CQL that was generated for Question 131.

Corpus: British National Corpus (TreeTagger)		
Operation	Parameters	Num of Hits
Query	lc,[lemma="color colour"]	16371

CQL is not only an alternative way of forming searches, but it has many features that enable very specific searches. In many cases there is no alternative to a CQL query. Here are some questions that simple CQL queries can provide data for.

Before we start, check that you are using BNC (TreeTagger). The BNC without TreeTagger uses CLAWS with which the following steps are incompatible.

Question 131 **How many sentences are there in the corpus?**

The Penn TreeBank tag, "SENT", which marks full stops, question marks and exclamation marks, invites a wide range of questions, as we are about to see.

Knowing the number of sentences in a corpus, especially a specialised corpus or a subcorpus, allows calculations of average sentence length, which is an important stylistic consideration.

Type the query, as shown in the screenshot: the tag in capitals between non-curly quotation marks; choose tag as the Default Attribute. Then use some of the tools to check the reliability of the tagging. For example, are decimal points counted as full stops?

CQL: "SENT"	Default attribute: tag
-------------	------------------------

Question 132 **What are some common ways of starting sentences?**

This might not be a question that keeps you awake at night, but the observations that follow from it can lead to great things. It is not uninteresting to compare the use of stance adverbials, disjuncts, conjuncts, discourse markers, backchannels and sentence stems across different corpora, different text types and different contexts. The data can also be used for questions concerning given and new information, theme and rheme³⁴.

I	do	n't	know	2,369
That	's	right	.	1,311
Yeah	.	Yeah	.	1,165
I	do	n't	think	1,065
I	think	it	's	588
Thank	you	very	much	551
Mm	.	Mm	.	535
That	's	right	,	469

This asks the software to generate concordances of sentence punctuation, thus full stops, question marks and exclamation marks becomes the node and new sentences follow.

It is then simply a matter of using the Frequency tool to make lists of words or BUNDLES up to four TOKENS long. Some of the tokens in these four positions are not words – sometimes they are punctuation and we often see the negative particle separated from its stem, as in *do n't*, which makes it two tokens. Click the **P** to see the sentences. This screenshot shows a Frequency list of some spoken sentence beginnings.

Quite different and informative data emerges when we process the data differently. For example, you can:

1. make a sample.
2. include the Node in the Frequency list.
3. select different attributes, e.g. a Frequency list of tags generates SYNTAGMS, as we saw under Frequency on p.52.
4. experiment with the Frequency limit at the top of the Multilevel frequency distribution.
5. The default sort is according to frequency. Click the column headings and sort by the chosen attribute.

Question 133 Does English permit numbers at the beginning of sentences?

In other words, are sentence initial numbers written as words as the style manuals tell us? After the sentence punctuation tag, we add the tag for cardinal numbers, then make a node forms list. Enter this into the CQL field and select TAG as the default attribute.

"SENT" "CD"

To use the data to answer the question, It is necessary to click on some of the "P"s in the list to see why some of the numbers are not written as words.

Question 134 Is it true that past participles are only used in compound forms, that is, with auxiliaries?

This question surprises a lot of people and a lot of people benefit from knowing the answer. To search for past participles, we need "VVN" in the CQL field. Set the Default Attribute to *tag*.

We then look to see if there are any instances without *be* (passive voice and continuous aspect) or *have* (perfect aspect and causative). The Sample button can be put to good use here as there are over 2.5 million past participles in the BNC.

Alternatively, use the Frequency tool to generate columns to the left of the past participles.

Do the concordances contain the lemma *have* for perfect and *be* for passive? Does the past participle perform any other function(s)? There are so many examples of bad tagging that it is necessary to use the data to answer the question manually. The Select Lines check boxes are useful in such situations (see p.42 – View Options: Bottom).

Periphrasis

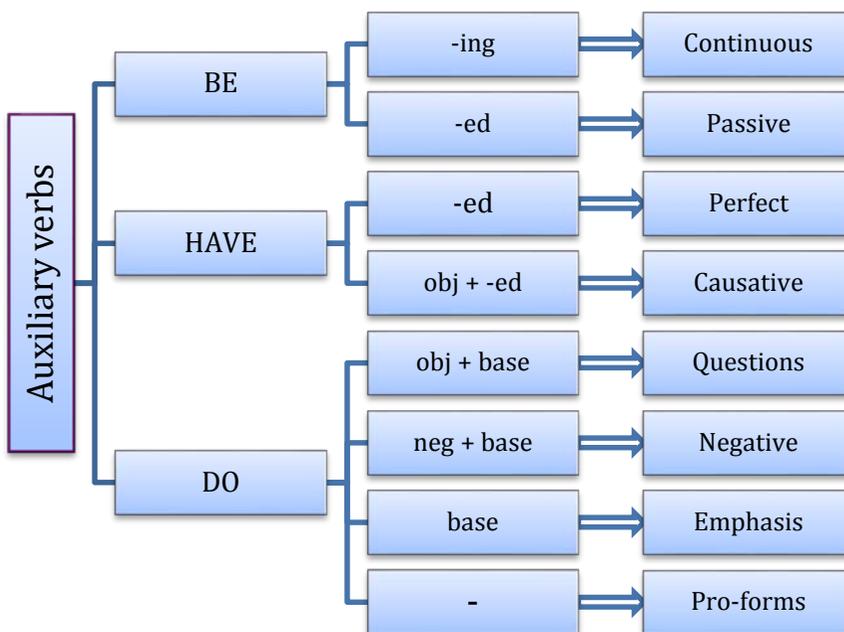
Periphrasis is when grammatical functions are expressed by auxiliary words in combination with content-bearing lexical words, as is the case in analytic languages. Chinese is a prime example: it has no verb forms other than the base form and nouns do not have plural or possessive forms. In contrast, synthetic languages express grammatical function such as tense and aspect with inflections. See the Italian example below. However, as David Crystal notes, most languages have features of analytic and synthetic languages (2006:370).

For example, in English, the comparative and superlative forms of adjectives and adverbs are mostly inflected with *-er*, *-est*, but the use of *more* and *most* are examples of periphrasis. Delexical verb structures also exemplify the periphrastic nature of English because the verbs are semantically light or even empty. We investigate these structures on p. 107.

From a linguistic standpoint, English has two tenses only, past and present, as we have already noted. From a pedagogical standpoint, however, tense refers to all of the expressions of TIME: **continuous** and **perfect**, which are technically referred to as aspects, and the **passive**, which is a voice, contrasting with the unmarked form, **active** voice and which does not involve an auxiliary. Unmarked aspect is referred to as **simple**.

*"Time is a game played
beautifully by chil-
dren." Heraclitus*

This chart depicts the use of the two verbs that are used as auxiliaries to form aspect and voice. It also shows the periphrastic role of *do* in creating other language constructions.



In Italian, the *passato prossimo* is formed with the verb *avere* (italicized in the table below). This periphrastic structure is quite similar in both meaning and form to English present perfect. Italian's *imperfetto* however, is formed with endings (underlined) and expresses the English *used to / would always*, a periphrastic realization.

Future time in English is expressed in ways that reflect or respect our uncertainty about things which have not yet happened. Because there is no inflection to indicate the future, it is mostly expressed periphrastically with *will*, *shall*, *going to* and *be to*. Italian has a set of endings (in bold) to mark the future.

Pronouns	Presente	Passato prossimo	Imperfetto	Futuro semplice
io	scrivo	<i>ho</i> scritto	scrive <u>vo</u>	scriver ò
tu	scrivi	<i>hai</i> scritto	scrive <u>vi</u>	scriver ai
egli	scrive	<i>ha</i> scritto	scrive <u>va</u>	scriver à
noi	scriviamo	<i>abbiamo</i> scritto	scrive <u>vamo</u>	scriver emo
voi	scrivete	<i>avete</i> scritto	scrive <u>vate</u>	scriver ete
essi	scrivono	<i>hanno</i> scritto	scrive <u>vano</u>	scriver anno