Main Results Tabs

This document provides additional information about the main results tabs in *The Prime Machine*. For a general overview, see **tPM Help 001 Getting Started**.

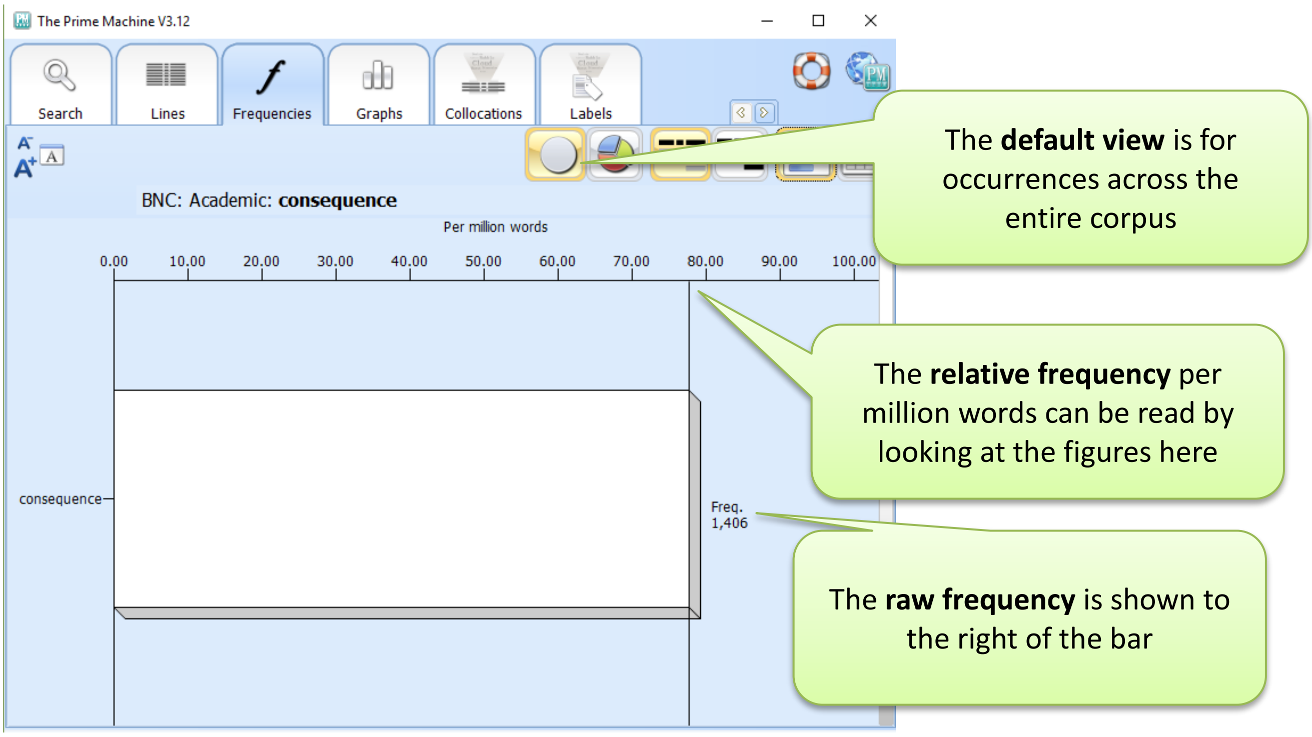
Lines / Cards and Collocations

Information about the Lines/Cards and Collocations results are provided in the Getting Started Guide. The Lines / Cards tab is the most important way to view and analyze corpus data. All the other results tabs are there to help you notice and explore the concordance lines themselves.

Frequencies

Different words (and different word forms of a base word) have very different frequencies, and you can usually find many differences between corpora (based on different genres/registers). The **Frequencies Tab** gives you quick information in the form of charts or tables, so you can see:

- The raw frequency – exactly how many time the word or collocation occurs in the entire corpus;
- The relative frequency – the frequency in the corpus per million words;
- The proportion of instances across the main categories of the corpus.



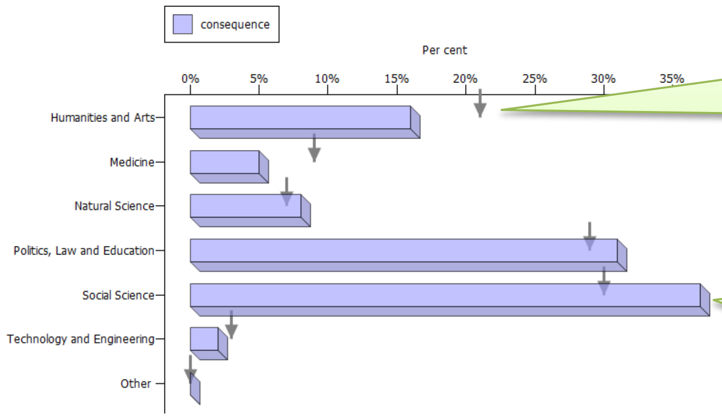
BNC: Academic: **consequence**

The **default view** is for occurrences across the entire corpus

The **relative frequency** per million words can be read by looking at the figures here

The **raw frequency** is shown to the right of the bar

To view a breakdown of proportions of the word or collocation across the main categories, click on the coloured pie chart button.

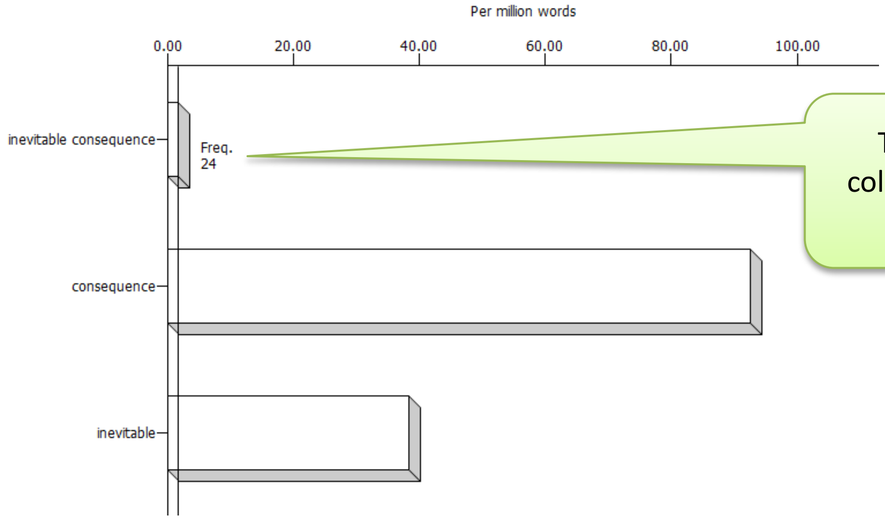


The arrows indicate the proportion of all words in each main category

The bars indicate the proportion of instances of the specific word or collocation

When viewing results for a collocation, it is also possible to compare the frequencies of each individual word in the collocation as well as the collocation as a whole.

To see the frequencies of words in a collocation, click this button. The frequencies for individual words will only be shown if the collocation is stored in the database as a collocation (i.e. *, _ or | searches will not have these data).



The raw frequency of the collocation can be seen to the right of the bar.

As with many of the other charts, you can switch between viewing the data as a graph or as a table by using these buttons.

Graphs

When you complete a search for a word or collocation, square turquoise icons will usually appear in a row at the bottom of the screen. Icons appear according to whether the current search query has a relatively high proportion of instances matching each specific feature. The statistical tests are based on all instances of the word or collocation in the corpus, compared with the "norm" for that corpus.

Clicking one of the icons changes the tab to the Graphs Tab and will display a graph of its specific feature. Features can also be found by clicking on the different groups listed down the left-hand side of the Graphs Tab.

Group	**Feature**	**Values**	**Level**	**Using text structure and rules**	***CLAWS* tags**
Headings TP	Title	Title; Not a title	Sentence	✓	

	Heading	Heading;
		Follows a heading;
		Not a heading
			Sentence	✓	

Position in text[1]
TPS
	Sentence position in text	Text Initial;
		Text Ending;
		Not text initial or text ending
			Sentence	✓	

	Paragraph position in text	First Paragraph;
		Last Paragraph;
		Not first or last paragraph
			Sentence	✓	

	Sentence position in paragraph	First Sentence;
		Last Sentence;
		Not first or last sentence
			Sentence	✓	

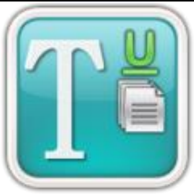
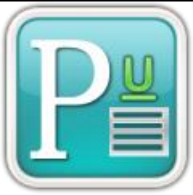
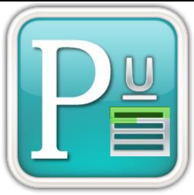
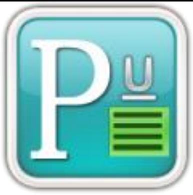
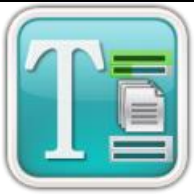
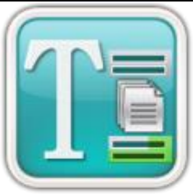
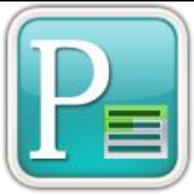
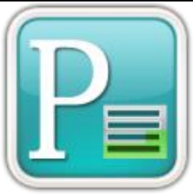
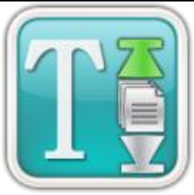
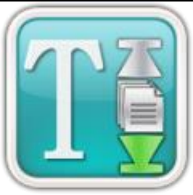
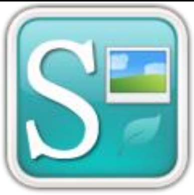
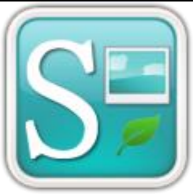
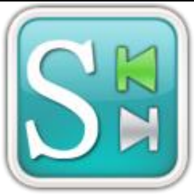
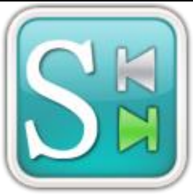
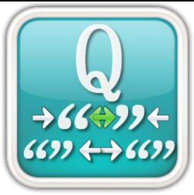
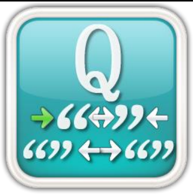
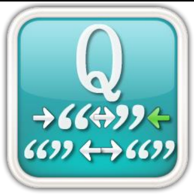
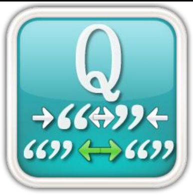
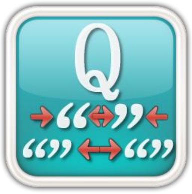
	Word position in sentence	First Fifth; First Third;
		Last Third; Last Fifth;
		Not first or last third
			Word	✓	✓

	Word position in sentence	Theme;
		Rheme;
		(unknown)
			Word	✓	

	Quotations	In Quotations;
		Before Quotations;
		After Quotations;
		Mid-quotation Suspensions;
		No Quotations
			Word	✓	

[1] Not all the values for features in this group are mutually exclusive. For example, words which are in the first fifth of a sentence will also be in the first third. However, paragraphs of one sentence in length and texts of one paragraph or one sentence in length are not included in the calculations for certain tendencies.

Icons representing features related to position

 Text level title	 Paragraph level heading	 Follows a heading	 Tends not to occur in headings at the paragraph level
 Text initial sentences	 Text ending sentences	 Paragraph initial sentences	 Paragraph ending sentences
 Text initial paragraphs	 Text ending paragraphs		
 Word in Theme	 Word in Rheme	 Word in first portion of sentence	 Word in last portion of sentence
 In Quotations	 Before Quotations	 After Quotations	 Mid-quotation Suspensions
 No Quotations			

Graphs Tab: Title

This shows the proportion of concordance lines which are taken from the titles of texts.

Examples from the *BNC: Newspapers* sub-corpus

Only 2.3% of all words in this corpus are part of a title.

Yet three out of ten of the occurrences of the word *sport* are titles. This is partly because one of the newspapers included in this corpus had a regular column called “Sport in Short”.

Tendencies to occur in text titles in the *BNC: Newspapers* sub-corpus.

	Frequency	Title
Tokens in the Newspaper sub-corpus	10,809,050	2.3%
sport	1,268	**30.5%**

Notes:

- Titles appear in bold text on the Cards.
- Not all the texts in all the corpora have titles included, so some data may be missing.
- It is **always** a good idea to look at the concordance lines to see what patterns of priming seem to occur. To filter the concordance lines, click to clear the tick mark against one or more of the features. Then click on the filter or compare buttons.

Examples from the *BNC: Academic* sub-corpus

Only 0.6% of words in this corpus are part of a heading.

Yet 13% of the occurrences of the word *conclusion* are paragraph headings and none of the occurrences of the word *ending* are paragraph headings. Obviously, the heading used for the last section of an academic article is usually *Conclusion,* but it also occurs very frequently within sentences.

Tendencies to be used (or not used) in paragraph headings in the *BNC: Academic* sub-corpus.

	Frequency	Heading
Tokens in the Academic sub-corpus	18,085,284	0.6%
conclusion	2,154	**13.0%**
ending	299	**-**

Notes:

- Additional figures are now shown for the first sentence after a heading (called follows a heading).

Graphs Tab: Text Position (sentence)

This shows the proportion of concordance lines which are text initial (the first sentence in the text), text ending (the last sentence in the text) or somewhere in between.

Examples from the *Hindawi Computer Science* corpus

Less than one in a hundred of all words (0.7%) are part of the first sentence.

Some of the words which frequently do occur in the first sentence of a text give a sense of how changes have occurred and progress has been made. Almost a third of the occurrences of the word *witnessed* are in the first sentence of a text. Other common examples in the first sentence of the text are *worldwide* and *ubiquitous* and words related to change and growth such as *advances, tremendous* and *increasingly.* Obviously, writers sometimes begin academic articles introducing the purpose of their paper, and perhaps they try to introduce the importance of the topic using words like these.

Only a tiny proportion of all words (0.5%) are part of the last sentence.

Yet almost a quarter of the occurrences of the word *hope* are in the last sentence of a text. Other fairly common examples are *intend* and *future*. Words which frequently occur in the very last sentence of a text often give a sense of looking forward to the future.

Tendencies to be used in the first or last sentence of texts in the *Hindawi Computer Science* corpus.

		Frequency	First sentence	Last sentence
Tokens in the entire corpus		9,847,424	0.7%	0.5%
Sense of change / progress / growth	*witnessed*	31	**32.3%**	-
	advances	238	**16.8%**	1.7%
	tremendous	60	**16.7%**	1.7%
	worldwide	94	**16.0%**	1.1%
	increasingly	280	**14.6%**	0.7%
	ubiquitous	241	**12.4%**	-
Sense of looking forward to the future	*intend*	112	-	**14.3%**
	future	2,429	0.8%	**11.5%**
	promising	453	2.4%	**4.2%**

Graphs Tab: Text Position (paragraph)

This shows the proportion of concordance lines which are in the first or last paragraphs of the texts.

Examples from the *Hindawi Computer Science* corpus

Only around 3 in 100 words are part of the first paragraph of texts.

Yet around a quarter of the occurrences of the words *advances* and *increasingly* are in the first paragraph of texts. Other words often used in the first paragraph are *emerging, novel,* and *growing*. These give a sense of how changes have occurred and progress has been made.

Only around 1 in 200 words are part of the last paragraph of texts.

Yet words like *hope* and *future* occur in the last paragraph much more often than that. Words which frequently occur in the last paragraph of a text often give a sense of looking forward to the future.

Tendencies to be used in the first or last paragraph of texts in the *Hindawi Computer Science* corpus.

		Frequency	First paragraph	Last paragraph
Tokens in the entire corpus		9,847,424	3.2%	0.5%
Sense of change / progress / growth	*advances*	238	**26.5%**	1.7%
	increasingly	280	**24.6%**	0.7%
	emerging	232	**22.4%**	1.3%
	novel	1,304	**18.3%**	1.0%
	growing	368	**16.0%**	0.3%
	enhancements	111	4.5%	0.9%
	expansion	323	1.2%	1.2%
	budding	46	4.3%	-
	evolving	188	6.9%	0.5%
	fresh	76	2.6%	-
	unique	963	5.3%	0.3%
Sense of looking forward to the future	*hope*	140	1.4%	**22.1%**
	future	2,429	2.5%	**11.6%**
	anticipate	40	2.5%	2.5%
	expect	303	1.0%	2.3%
	trust	257	1.9%	1.2%

Graphs Tab: Paragraph Position

This shows the proportion of concordance lines which are from the first or last sentences of paragraphs.

Examples from the *Hindawi Biological Science* corpus

Overall, around 1 in 6 of all words are in the first sentence of paragraphs.

Yet more than half of the occurrences of the words *discuss* and *discusses* occur in the first sentence. Other words which occur quite frequently in the first sentence of paragraphs are *summarize, summarizes, focus* and *focuses.*

The word *furthermore* tends to occur quite frequently in the first or last sentence of a paragraph.

Obviously, writers sometimes introduce a paragraph by linking to details which have already been discussed, or by summarizing the focus of the new paragraph. Towards the end of a paragraph, they may use signals like *furthermore* to take their points further.

Tendencies to be used in the first or last sentence of paragraphs in the *Hindawi Biological Science* corpus.

		Frequency	First sentence	Last sentence
Tokens in the entire corpus		23,107,819	18.2%	17.9%
Introducing or linking to a topic.	*discuss*	1,048	**57.0%**	19.0%
	discusses	141	**56.7%**	12.8%
	summarize	345	**50.4%**	17.1%
	summarizes	281	**44.1%**	14.2%
	focuses	342	**38.9%**	15.8%
	focus	1,900	**31.0%**	18.9%
	consider	881	24.0%	18.6%
	considers	108	18.5%	17.6%
	argue	215	19.1%	24.2%
	argues	84	26.2%	16.7%
	center	1,365	16.1%	15.9%
	centers	360	18.6%	18.9%
Signposting	*furthermore*	5,368	**21.9%**	**24.2%**
	besides	1,304	19.1%	21.0%

Note:

- Some paragraphs are only one sentence long, so these results have not been counted.

Graphs Tab: Sentence Position

This shows the proportion of concordance lines where the word is in the first or last portion of the sentence. Sentences are divided into the first fifth, the first third, the last third and the last fifth.

Examples from the *BNC: Academic* sub-corpus

Obviously, around 20% of all words are in the first fifth of the sentence.

Yet more than three quarters of the occurrences of the words *interestingly*, *unfortunately* and *fortunately* are in the first 20% of the sentence. Also, as you might predict, many signposting words like *furthermore*, *moreover*, *firstly*, *lastly*, etc. also occur more than three quarters of the time in the first fifth.

Similarly, around 20% of all words are in the last fifth of the sentence.

Yet more than half of the occurrences of the word *respectively* and three out of ten of the occurrences of *properly* are in the last 20% of the sentence.

Tendencies to be used in the first or last 20% of a sentence in the *BNC: Academic* sub-corpus.

		Frequency	First 20%	Last 20%
Tokens in the sub-corpus		18,085,284	17.7%	22.3%
Attitude	*interestingly*	275	**80.7%**	1.5%
	unfortunately	618	**78.6%**	2.6%
	fortunately	148	**75.0%**	3.4%
Signposting	*furthermore*	1,442	**91.9%**	0.6%
	moreover	1,913	**88.3%**	0.9%
	lastly	132	**84.8%**	6.8%
	firstly	543	**79.2%**	0.2%
Other adverbs	*respectively*	1,303	2.5%	**56.2%**
	properly	1,008	10.7%	**29.4%**

Notes:

- Sentences cannot always be divided into five equal chunks, but one might expect the figures to be close to 20% on average.
- The graphs also show figures for the proportion of occurrences in the first third and last third of the sentence.
- Results may be particularly interesting for collocations, so look out for the icons at the bottom of the screen.

Graphs Tab: Theme/Rheme

This shows the proportion of concordance lines where the word is in the Theme or Rheme of the sentence. The Theme is defined as all the words leading up to the first main verb but not including it. The Rheme is the rest of the sentence.

Examples from the *BNC: Academic* sub-corpus

Less than one in six of all words are in Theme.

Yet more than half of the occurrences of the word *aim* are in the Theme. Also more than 30% of the occurrences of the words *experiment*, *research*, *questionnaire* and *data* are in the Theme. Obviously, *aim*, *experiment*, *research*, *questionnaire* and *data* are often the subject of sentences in academic texts.

The vast majority of the occurrences of the words *difficult* and *likely* are in the Rheme. *Likely* and *difficult* could be used in the subject, but more frequently occur later in the sentence.

Tendencies to be used in Theme or Rheme in the *BNC: Academic* sub-corpus.

		Frequency	Theme	Rheme
Tokens in the sub-corpus		18,085,284	17.2%	78.8%
Matters of academic research	*aim*	1,616	**51.0%**	47.6%
	experiment	1,033	**34.6%**	60.4%
	research	10,338	**31.8%**	45.5%
	questionnaire	488	**31.1%**	63.1%
	data	7,469	**30.6%**	65.4%
Related to certainty	*likely*	6,680	2.9%	**96.7%**
	explanation	1,870	**27.4%**	70.6%
	likely explanation	25	**60.0%**	40.0%
Related to doubt	*unlikely*	1,349	1.6%	**98.1%**
	dubious	139	15.8%	83.5%
Related to difficulty	*difficult*	4,686	3.5%	**95.8%**
	challenging	258	12.8%	82.2%
	tough	126	13.5%	83.3%
Related to practicality	*feasible*	287	5.6%	**92.7%**
	viable	204	11.8%	87.7%

Note:

- This measure relies on automatic part-of-speech tagging and so the results may not be 100% accurate.

Group	**Feature**	**Values**	**Level**	**XML / other encoding**	***CLAWS* tags**
CMVYN group	Complexity	Simple Sentence Projecting Sentence Complex Sentence	Sentence		✓

	Modality	Volition/prediction;
		Permission/possibility/ability;
		Obligation/necessity;
		No modals
			Sentence & Word		✓

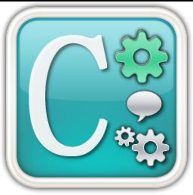
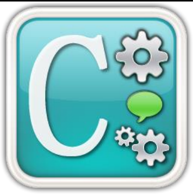
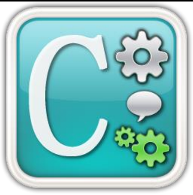
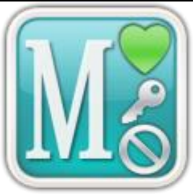
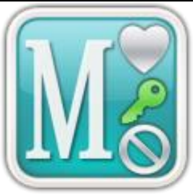
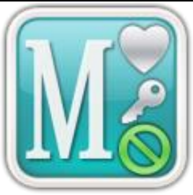
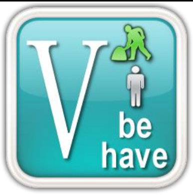
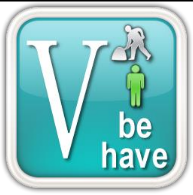
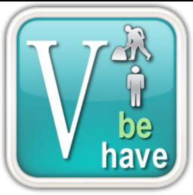
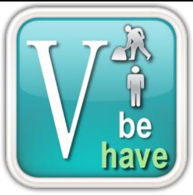
	Voice	Active Voice/Other;
		Passive Voice;
		Basic be;
		Basic have
			Sentence & Word		✓

	Polarity	Is; Is not	Sentence & Word		✓
Det. & Prep. group	Part of Speech	Noun; Proper Noun; Pronoun; Number; Adjective; Verb; Adverb; Other/Unknown	Word		✓

	Determiners	Definite article;
		Possessive;
		Indefinite article;
		No articles
			Word		✓

	Prepositions	Near Prepositions; Not Near Prepositions	Word		✓

Icons representing features related to colligation

 Simple sentences	 Projecting Sentences	 Complex sentences	
 Volition/Prediction modals	 Permission/Possibility/ Ability modals	 Obligation/Necessity modals	
 Active voice/other	 Passive voice	 Basic *to be*	 Basic *to have*

 Positive sentences	 Negative sentences		
 Noun	 Proper Noun	 Pronoun	 Number
 Adjective	 Verb	 Adverb	
 Near definite articles	 Near possessives	 Near indefinite articles.	
 Near prepositions;	 Avoids prepositions.		

Graphs Tab: Complexity

This shows the proportion of concordance lines where the sentence is grammatically complex. A complex sentence includes at least one of the following.

- a subordinating conjunction (e.g. *if*, *because*)
- *as*, *than*, *that* or *whether* as a conjunction

Examples from the *BNC: Newspapers* sub-corpus

Only about one third of all words in the newspaper sub-corpus are in complex sentences.

4 out of 5 of the occurrences of the word *that* are marked as being complex. This is not at all surprising given that *that* is a conjunction.

However, several other words including *indications, stating, argues* and *convince* also occur more than 70% of the time in complex sentences. These words are probably used in complex sentences containing *that.*

The word *survey* occurs 41.1% of the time in complex sentences in newspapers.

Tendencies to be used in Complex or Simple sentences in the *BNC: Academic* sub-corpus.

		Frequency	Complex	Simple
Tokens in the Academic sub-corpus		18,085,284	49.7%	50.3%
	survey	2,226	38.9%	**61.1%**
Other matters of academic research	*project*	3,646	24.5%	**75.5%**
	projects	1,006	37.4%	**62.6%**
	questionnaire	488	28.3%	**71.7%**
	research	10,338	30.2%	**69.8%**
	samples	1,387	32.4%	**67.6%**
	studies	7,155	32.8%	**67.2%**
Related to evidence	*fact*	8,255	**77.6%**	22.4%
	facts	1,953	**59.3%**	40.7%
	detail	1,763	39.2%	**60.8%**
	details	1,349	37.6%	**62.4%**
	statistic	40	47.5%	52.5%
	statistics	1,052	34.7%	**65.3%**
Adjectives	*arguable*	115	**91.3%**	8.7%
	practicable	192	**85.4%**	14.6%
	doubtful	311	**81.0%**	19.0%
	chargeable	136	**77.9%**	22.1%
	unreasonable	386	**76.7%**	23.3%
	achievable	54	51.9%	48.1%
	noticeable	247	50.6%	49.4%
	detectable	170	47.1%	52.9%
	observable	168	46.4%	53.6%
	considerable	3,031	45.3%	**54.7%**
Adverbs	*reasonably*	921	**72.7%**	27.3%
	unreasonably	98	**83.7%**	16.3%
	conclusively	78	**79.5%**	20.5%
	surprisingly	519	40.8%	59.2%
	remarkably	285	43.9%	56.1%
	moderately	140	37.1%	62.9%

Notes:

- In the *BNC: Academic* sub-corpus, the overall balance between complex and simple sentences is roughly equal (50%).
- Figures for **Projecting** sentences are now displayed separately from other forms of **Complex** sentences.

Limitations on the modal groupings included in the software

Groupings from Biber et al. (1999, p. 489)	Groupings in *The Prime Machine*	
	Included	Excluded
Permission/Possibility/Ability:		
can	*can*	
could	*could*	
may	*may*	
might	*might*	
Obligation/Necessity:		
must	*must*	
should	*should*	
have to		*have to*
(had) better		*(had) better*
(have) got to		*(have) got to*
need to	*need to*	
(be) supposed to		*(be) supposed to*
ought to	*ought to*	
Volition/Prediction		
will	*will*	
would	*would*	
shall	*shall*	
be going to		*be going to*
Past time:		
used to		*used to*

Graphs Tab: Modality

This shows the proportion of concordance lines which contain modal verbs within 4 words to the left of the main search word.

Modals are counted in three groups.

- *can*, *could*, *may* and *might*
- *must*, *should*, *need* and *ought*
- *will*, *would* and *shall*

Examples from the *BNC: Academic* sub-corpus

Less than 5% of words in the corpus are near modal verbs.

Yet words like *legitimately, usefully, conceivably* and *easily* are often used with the words *can, could, may* or *might*.

Words like *remembered, noted*, *emphasised* and *stressed* are often used with the words *must, should, need to* or *ought to*. Other words often used with these modals are *carefully* and *surely.*

Words like *suffice*, *cease*, *depend* and *disappear* are often used with the words *will, would* or *shall*. Other words often used with these modals are *examine*, *argue* and *discuss*.

Tendencies to be used with three groups of modal verbs in the *BNC: Academic* sub-corpus.

	Frequency	*can, could, may, might*	*must, should, need, ought*	*will, would, shall*
Tokens in the sub-corpus	18,085,284	2.3%	0.8%	1.5%
legitimately	90	**81.1%**	-	-
usefully	155	**70.3%**	-	1.3%
conceivably	108	**59.3%**	-	1.9%
easily	1783	**39.5%**	0.7%	**3.2%**
remembered	324	3.1%	**40.7%**	**6.8%**
noted	2001	**4.2%**	**17.3%**	1.9%
emphasised	363	1.1%	**16.8%**	-
stressed	681	1.2%	**10.6%**	1.0%
carefully	754	2.0%	**11.1%**	2.9%
surely	700	3.9%	**9.6%**	**8.6%**
suffice	186	**10.2%**	**9.1%**	**50.0%**
cease	252	4.4%	**8.3%**	**37.7%**
depend	1184	**8.4%**	**4.1%**	**35.5%**
disappear	184	8.2%	1.6%	**29.3%**
examine	1487	3.6%	**2.8%**	**22.5%**
argue	1464	**13.0%**	0.3%	**19.4%**
discuss	971	**5.5%**	1.8%	**17.7%**

Notes:

- It is a good idea to look at the concordance lines to see which modal verbs within each group are used most often.
- None of the words given as examples here are always used with modal verbs, but the proportions are higher than those of most other words.

Graphs Tab: Voice

This shows the proportion of concordance lines which are passive voice.

To be counted, passive voice verbs must have a passive auxiliary verb (e.g. *is*, *was*, *got*).

Passive voice is usually associated with formal writing like academic articles, but we can see some differences in the kinds of verbs used in passive voice in newspapers.

Examples from the *BNC: Newspapers* sub-corpus

Less than a quarter of all words are in sentences which are passive voice.

Yet words like *prosecuted*, *remanded*, *discharged*, *rewarded* and *punished* occur more than three quarters of the time in passive voice sentences. These verbs are often associated with police actions and frequently occur in passive voice sentences.

Words like *forgiven*, *tempted*, *understood* and *debated* also occur very frequently in passive voice sentences.

Words describing actions like *clinched* and *jumped* typically do not occur in passive voice sentences. This is also true of words which describe states like *tired*, *worried* and *failed*.

Tendencies to be used (or not used) in passive voice sentences in the *BNC: Newspapers* sub-corpus.

		Frequency	Passive voice	Active voice / other
Tokens in the sub-corpus		10,809,050	22.7%	77.3%
associated with police actions	*prosecuted*	76	**88.2%**	11.8%
	remanded	244	**85.2%**	14.8%
	discharged	119	**82.4%**	17.6%
	rewarded	166	**79.5%**	20.5%
	punished	94	**78.7%**	21.3%
	forgiven	84	**83.3%**	16.7%
	tempted	125	**82.4%**	17.6%
	debated	67	**76.1%**	23.9%
	understood	460	**75.7%**	24.3%
actions	*clinched*	146	6.8%	**93.2%**
	jumped	415	9.2%	**90.8%**
states	*tired*	237	10.1%	**89.9%**
	worried	584	14.6%	**85.4%**
	failed	1,708	17.5%	**82.5%**

Notes:

- Interestingly, the overall proportions in the *BNC: Academic* sub-corpus are much higher with more than one third of sentences in passive voice.
- Figures for sentences with the basic form of *to be* and the basic form of *to have* are now displayed separately from other Active Voice/Other sentences.

Graphs Tab: Polarity

This shows the proportion of concordance lines where the sentence is negative.

Negative sentences contain the word *not*.

Examples from the *BNC: Academic* sub-corpus

Less than 1 in 5 words occur in sentences containing the word not.

Yet words like *watertight*, *invalidate*, *necessarily*, *preclude* and *dissimilar* seem to occur quite frequently in sentences with the word *not*. The word *always* occurs in sentences containing the word *not* more than one third of the time. These frequencies are high given the overall low proportion of negative sentences and given that these words all seem to have strong meanings.

Tendencies to be used in negative sentences in the *BNC: Academic* sub-corpus.

	Frequency	Negative	Positive
Tokens in the sub-corpus	18,085,284	16.9%	83.1%
watertight	15	**73.3%**	26.7%
invalidate	52	**69.2%**	30.8%
necessarily	2,014	**69.1%**	30.9%
preclude	128	**65.6%**	34.4%
dissimilar	114	**64.9%**	35.1%
always	4,879	**34.5%**	65.5%

Graphs Tab: Definite/Indefinite

This shows the proportion of concordance lines where there is an article or possessive within 4 words to the left of the main search word.

They are grouped in this way:

- Definite articles (*the*)
- Possessives (e.g. *my*, *your*, *'s*)
- Indefinite articles (*a*, *an*, *every, no*)

Examples from the *BNC: Academic* sub-corpus

Around a quarter of all words are near definite articles or possessives.

Words ending in "*-est*", usually follow *the*, with 97% or more of the occurrences of the words *biggest*, *widest*, *slightest*, *broadest*, *earliest*, *finest*, *richest*, *largest* and *poorest* near definite articles or possessives.

Less than 1 in 10 of all words are near indefinite articles.

However, words like *lot*, *handful*, *variety* and *dozen* usually follow *a*, with 72% or more of the occurrences near indefinite articles.

Tendencies to be used with two groups of articles and possessives in the *BNC: Academic* sub-corpus.

		Frequency	Definite article or possessive	Indefinite article
Tokens in the sub-corpus		18,085,284	25.5%	8.2%
superlative adjectives	*biggest*	121	**100.0%**	-
	widest	63	**100.0%**	-
	slightest	53	**100.0%**	-
	broadest	62	**98.4%**	-
	earliest	473	**97.7%**	0.2%
	finest	70	**97.1%**	-
	richest	75	**97.3%**	-
	largest	793	**97.2%**	-
	poorest	142	**97.2%**	-
related to quantities	*handful*	123	10.6%	**87.0%**
	lot	719	11.3%	**84.8%**
	variety	2,337	16.3%	**74.2%**
	dozen	123	16.3%	**72.4%**
objects of academic research	*survey*	2,226	**48.4%**	**30.7%**
	questionnaire	488	**40.0%**	**43.0%**

Tendencies to be used with two groups of articles and possessives in the *BNC: Newspapers* sub-corpus.

	Frequency	Definite article or possessive	Indefinite article
Tokens in the sub-corpus	10,809,050	24.7%	9.0%
survey	1096	**41.7%**	**46.2%**
questionnaire	25	28.0%	**60.0%**

Notes:

- Obviously, you can get a sense of how often a word is used as a noun by looking at these figures.
- Figures for Possessives are now shown separately from Definite Articles.

Graphs Tab: Prepositions

This shows the proportion of concordance lines where there is a preposition within 4 words either side of the main search word.

Prepositions include:

- General prepositions (e.g. *at*, *on, by*)
- *for*, *of*, *with* or *without* as prepositions

Examples from the *BNC: Academic* and *BNC: Newspapers* sub-corpora

A little more than half of all words in these corpora are near prepositions.

Yet 99% of the occurrences of the word *spite* are near prepositions while none of the occurrences of the word *despite* are near prepositions.

Sometimes similar words can be quite tricky to use correctly when writing in a foreign language, but a quick search for *despite* vs. *spite* in either of these corpora can show preposition patterns very clearly. We would expect the concordance lines to show us *despite* near verbs and in the phrase "despite the fact". We would also expect to see *spite* used in sentences in the phrase "in spite of".

Examples for tendencies to be used with (or without) prepositions in the *BNC: Academic* sub-corpus.

	Frequency	Prepositions	No prepositions
Tokens in the sub-corpus	18,085,284	58.0%	42.0%
spite	476	**98.7%**	1.3%
despite	2,750	-	**100.0%**

Examples for tendencies to be used with (or without) prepositions in the *BNC: Newspapers* sub-corpus.

	Frequency	Prepositions	No prepositions
Tokens in the sub-corpus	10,809,050	52.3%	47.7%
spite	279	**98.9%**	1.1%
despite	2,261	-	**100.0%**

Examples for tendencies of collocates of *of* to be used with prepositions in the *BNC: Academic* sub-corpus, along with the proportion s accounted for by collocations containing *of.*

		Frequency	Prepositions	*of*	Others
Tokens in the Academic sub-corpus		18,085,284	49.7%		
beginning or presence	*advent*	188	**100.0%**	97.3%	*with*
	outbreak	165	**96.4%**	86.1%	
	commencement	124	**96.0%**	83.9%	*at*
	onset	413	**95.2%**	71.7%	

	presence	2,485	**94.3%**	67.9%	
	aftermath	139	**94.2%**	80.6%	*in*
	dissemination	159	**90.6%**	61.0%	
	conception	1,129	**90.3%**	71.7%	
	beginnings	160	**90.0%**	59.4%	*from*
	adoption	581	**89.5%**	43.5%	
absence or destruction	*irrespective*	361	**100.0%**	100.0%	
	regardless	372	**98.7%**	97.8%	
	absence	2,232	**96.7%**	79.7%	*in*
	abandonment	152	**94.7%**	77.0%	
	breaches	161	**93.8%**	71.4%	
	breach	1,844	**93.5%**	72.9%	*for*
	removal	549	**93.1%**	66.5%	*from*
	demise	181	**91.2%**	58.0%	
	destruction	461	**89.6%**	52.7%	
range or number	*sorts*	425	**98.8%**	86.8%	
	kinds	1,569	**98.6%**	80.4%	
	lots	110	**98.2%**	90.0%	
	kind[2]	4,267	**97.1%**	≥61.5%	
	amounts	762	**96.5%**	48.4%	
	sort	1,858	**95.7%**	69.6%	
	variety	2,337	**95.5%**	85.0%	
	handful	123	**95.1%**	91.9%	
	number	10,648	**94.5%**	84.0%	
	plenty	191	**94.2%**	82.7%	
	proportion	2,484	**93.9%**	77.4%	
	aspects	2,776	**93.6%**	82.9%	
	combination	1,201	**93.3%**	68.7%	*with, by*
	series	2,814	**92.3%**	64.7%	
	subset	139	**92.1%**	71.2%	
	mixture	474	**91.6%**	66.5%	*with*
	parts	2,765	**90.4%**	59.9%	
	aspect	1,391	**90.2%**	75.3%	
	stages	1,290	**89.7%**	42.2%	*at, in*

[2] The proportion of cases collocating with *of* exceeds the figure shown because 39.0% of occurrences are for the collocation *of .. kind,* while 61.5% are for the collocation *kind of.* However, some of these may coincide as in the phrase *of the kind of*.

Group	Feature	Values	Level	Basis

Feeling	Meanings	Positive environment;
		Positive meaning;
		Negative environment;
		Negative meaning;
		Neutral/Unknown
			Word	*CLAWS* tags, UCREL Semantic Tags and Wordlists

Icons representing a tendency for positive or negative meaning

Positive environment;	Positive meaning;	Negative environment;	Negative meaning;

Figures for these features are calculated in the following way:

1. Words (and multiword units) are tagged in the database using UCREL's Semantic Tagging System (Rayson, 2008)
2. 24 Semantic Tags with a strong positive or negative meaning are then used, along with the original wordlists (based on sources mentioned earlier)
   A1.1.2 Damaging And Destroying, *A1.4 Chance, Luck*, etc.
4. Words in a 4 word window either side of these tagged words are marked in the database as being in a positive or negative context.
5. Since words on these wordlists will always be marked, a further flag indicates whether at least one other word in the 4 word window is also on the list.
6. Contingency tables are then used to create lists of words and collocations, following the same procedure as used for other features of lexical priming in *The Prime Machine.*

Group	**Feature**	**Values**	**Level**	**XML / other encoding**	***CLAWS* tags**
Repetition 	Repetition	Same form Same stem Not repeated	Word		

Icons representing a tendency for repetition

 Repetition of the same form	 Repetition of the same stem

Graphs Tab: Repetition

This shows the proportion of concordance lines where the main search word occurs more than twice in the same neighbourhood (one sentence before or after the sentence containing the word).

Labels

Corpus tools usually have a Key Word function, letting you see which words are *key* in a text or group of texts. The Labels Tab shows this kind of information the other way round; it tells you in which kinds of text your chosen word is *key.*

To do this, it compares the frequency of a word (or collocation) in one part of the corpus with its frequency in the rest of the corpus. You begin with a word or collocation – the one you are interested in – and the computer splits the corpus up into hundreds or thousands of pieces, based on the labels (metadata) which are attached to different sections or texts within the corpus.

What could Key Labels tell the user?

- For the teacher
  - Check the kinds of labels to see what kind of examples we're going to pick up;
  - Is it balanced?
- For the language learner student
  - See typical uses
  - Understand "prohibited" use

Text Labels (from Metadata)

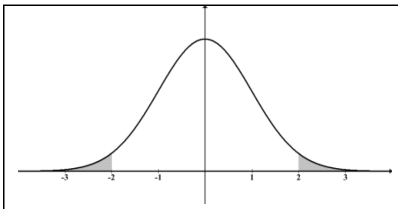
Text labels include the main category of a text (e.g. Fiction or Academic). They may also include information about the publisher, the source or the genre.

Text Labels (from MAT)

Multidimensional analysis is a corpus method which assigns scores to a text for a number of dimensions based on the occurrence of groups of linguistic features.

All of the texts in every online corpus have been processed using MAT (Nini 2014). Labels are added to each corpus text according to the scores for each of the 6 dimensions, following the dimensions used in MAT and Biber (1988).

Dimension	Score	Label
Involved <-> Informational	≥+5.199336	Highly Involved (MAT)
	>+3.71899	Involved (MAT)
	≤ -3.71899	Informational (MAT)
	≤ -5.199336	Highly Informational (MAT)
Narrative <-> Non-Narrative	≥+5.199336	Highly Narrative (MAT)
	>+3.71899	Narrative (MAT)
	≤ -3.71899	Non-Narrative (MAT)
	≤ -5.199336	Highly Non-Narrative (MAT)
Context-Independent <-> Context-Dependent	≥+5.199336	Highly Context-Independent (MAT)
	>+3.71899	Context-Independent (MAT)
	≤ -3.71899	Context-Dependent (MAT)
	≤ -5.199336	Highly Context-Dependent (MAT)
Overtly Persuasive <-> Less Overtly Persuasive	≥+5.199336	Highly Overtly Persuasive (MAT)
	>+3.71899	Overtly Persuasive (MAT)
	≤ -3.71899	Less Overtly Persuasive (MAT)
	≤ -5.199336	Much Less Overtly Persuasive (MAT)
Abstract Information <-> Non-Abstract Information	≥+5.199336	Highly Abstract Information (MAT)
	>+3.71899	Abstract Information (MAT)
	≤ -3.71899	Non-Abstract Information (MAT)
	≤ -5.199336	Highly Non-Abstract Information (MAT)
Online Informational Elaboration <-> Less Online Informational Elaboration	≥+5.199336	Highly Online Informational Elaboration (MAT)
	>+3.71899	Online Informational Elaboration (MAT)
	≤ -3.71899	Less Online Informational Elaboration (MAT)
	≤ -5.199336	Much Less Online Informational Elaboration (MAT)



How Cut-offs Were Determined

When presenting results from Multidimensional Analysis, researchers often interpret the scores on the dimensions in different ways.

In tPM when MAT results are added to the corpus as Labels, the figures ±3.71899 and ±5.199336 were determined in this way:

- It was noted that with z-scores, we can expect 99.7% of the data points to be plus or minus 3 standard deviations from the mean.
- For corpus linguistics, 99.99% is often used for other calculations (e.g. keyness).
- To calculate the standard deviations required for 99.99% and 99.99999% of the data, the NORMSDIST function and Goal Seek operations in Microsoft Excel.

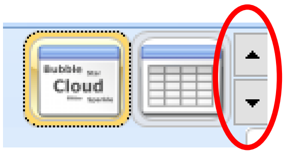
Producer Labels (from Metadata)

The Producer box contains any labels related to the author or speaker. Producer labels provide information about the writer or speaker (e.g. their name, age, gender, etc.). Many corpora do not have detailed information about the author or speaker, but if metadata is available, statistically significant results will be displayed here.

Section Labels (from text formatting and metadata)

The Section box contains any labels related to the Section in which the word or collocation appears. Section labels are the sub-headings of a text (e.g. Abstract, Introduction or Conclusion). If headings and subheadings are available in the corpus, statistically significant results will be displayed here.

Neighbourhood Labels (from UCREL's Semantic Tags)



Neighbourhood labels are based on repeated semantic tags for words and phrases in the current card. It gives a kind of summary of the semantic tags associated with the wider co-text of the specific word or collocation. The arrows in the top right of the screen allow you to adjust the number of repeated tags required for the semantic tag to qualify.

It works in this way:

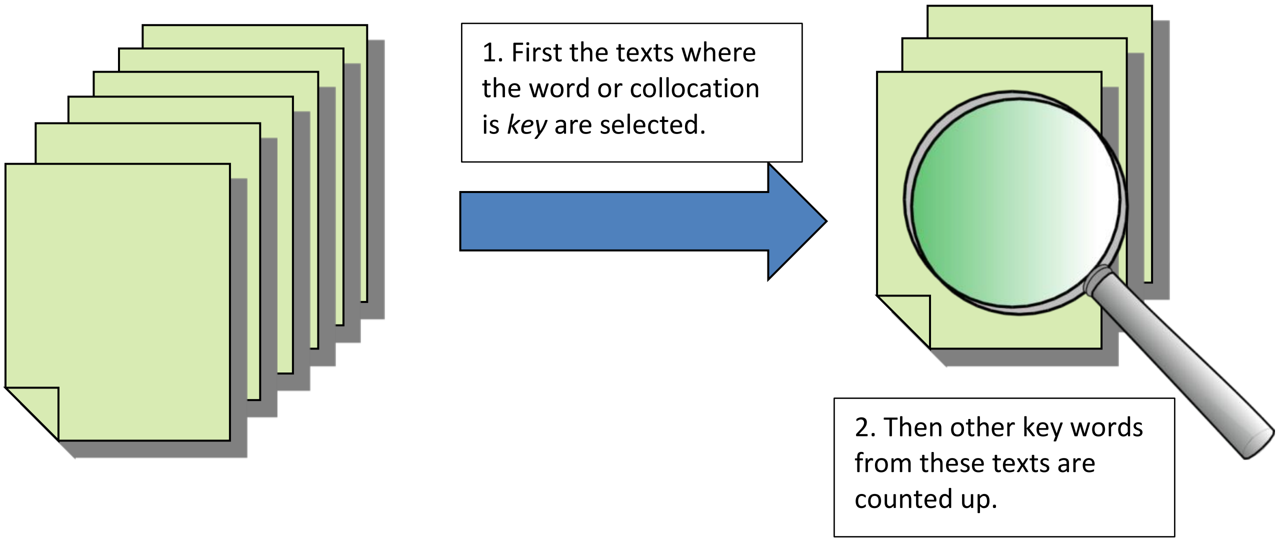
1. Words (and multiword units) are tagged in the database using UCREL's Semantic Tagging System (Rayson, 2008).
2. Links between Semantic Tags and Sentences are then stored in the database
3. Additional links are added to each sentence if the adjacent sentences also contain with the same semantic tags
4. Bonds are established between Semantic Tags and Sentences based on minimums of 2-8 links.

Note:
Some labels have been slightly re-worded to try to make the meanings clear:

Evaluation: True/False & -1 ➔ Evaluation: False

Associates

Key Associates are calculated by finding texts in which the word or collocation is *key*, and then finding other words that are often key in these same texts.



1. First the texts where the word or collocation is *key* are selected.

2. Then other key words from these texts are counted up.

If the corpus has more than one major category, the Key Associates are displayed in boxes according to the categories where the word or collocation you have searched for appears most. These key associates are based on key word calculations using each text in the category, compared with texts in all the other categories within the same corpus.

If a corpus only has one major category, the key words are based on comparing each individual text in the corpus against the British National Corpus (1994) as a reference corpus.

The bars at the top of the boxes of Associate Clouds show the proportion of instances for each category displayed.

References

Biber, D. (1988). *Variation across Speech and Writ*ing. Cambridge: Cambridge University Press.

Nini, A. 2014. *Multidimensional Analysis Tagger 1.1 - Manual*. Retrieved from: http://sites.google.com/site/multidimensionaltagger.

Rayson, P. (2008). "From key words to key semantic domains." *International Journal of Corpus Linguistics* 13(4): 519-549.

Support

The Prime Machine is still undergoing development.
For further information see http://help.theprimemachine.com

Need more details?
If you need more information about any feature of The Prime Machine, or if you have other suggestions and feedback, please complete the short feedback questionnaire, following the link from the home page: www.theprimemachine.net

Last Updated: Friday, October 26, 2018