BNCweb Intro

Jean Mark Gawron

Linguistics 503 San Diego State University

March 21, 2016

Outline

- Intro
- 2 Changing context window
- Sorting
- 4 Distributions
- Downloads
- 6 Collocations
- Dispersion



BNCweb

BNCweb is a pretty intuitive web-based interface to BNC you can use with your web browser.

BNCweb in CompLing Lab

http://bulba.sdsu.edu/bncweb

Requirements

Password and account name

Simple queries

Queries

introvertedness [no matches]

introverted

Results

Your query "introverted" returned 84 hits in 69 different texts.

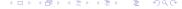
- 1 A06 1335 But besides this more obvious point, there are subtler connections between voice and body: Cicely Berry observes that an introverted and thoughtful person often finds more difficulty in speaking and does not carry the
 - person often finds more difficulty in speaking and does not carry the thought through into the physical process of making speech.

 A18 429 Razumikhin himself may or may not have come from the country, but he is
- 2 A18 429 Razumikhin himself may or may not have come from the country, but he is certainly a member of the floating, unbelonging population of students and ex-students, and he records in simple puzzlement that Raskolnikov has been growing increasingly moody and suspicious and introverted; he has no time for anything, people are always in his way, and yet he lies about and does nothing a confirming echo of Raskolnikov on his bed telling Nastasya the maid that he is working, by which he means thinking.

Multiword queries and tokenization I

Multi word queries ok. Special characters must be escaped (preceded with "\"). Comma (","), question mark ("?"), plus ("+"), colon (":"), at-sign ("@"), various kinds of braces, all special. Quotation mark NOT speciali

```
talk of the town Good query
very very pretty one hit
very, very pretty syntax error
very , very pretty still a syntax error
very \, very pretty correct query, 5 results
" pretty " 7 hits, all pretty in scare quotes
```



Intro

Lemmas, intervening stuff I

```
{kick}
Intervening stuff
 day \rangle \rangle 2 \rangle \rangle night
                           day followed by night within a 2 word window
                           (excludes by day and by night)
 night \langle\langle 2 \langle\langle day \rangle\rangle\rangle
                           day followed by night within a 2 word window,
                           but night will be the highlighted word
 day \langle \langle 2 \rangle \rangle night
                           day and night within 2 words in either order
```

All instances of all lemmas with the form kick

 $\{kick/V\}$ curly braces signal lemma, $\{kick/N\}$ {kick/V} are different lemmas

Intro

Lemmas

Lemmas, intervening stuff II

Plus and Star this $(_{A})^* \{day\}$ this followed by any form of day with any number of adjectives intervening this $(_{A})+ {day}$ this followed by any form of day with at least one adjective intervening this $+ \{day\}$ this followed by any form of day with exactly one word intervening this $++ \{day\}$ this followed by any form of day with exactly two words intervening this + _NN1 this followed by exactly one word followed by a signular noun (note no curly braces to use BNC C5 tag).



Intro

Lemma/partof speech/sentences boundary

Query

Intro

```
{kick/V} <s> bucket_NN1
```

"All tokens of the LEMMA *kick* and the singular noun *bucket* occurring within a single sentence (in either order)"

Results

```
Your query "{kick/V} <s> bucket_NN1" returned 24 hits in 17 different texts
```

```
No
    Filename
               At any speed, in any gear on the mile straight there is enough power to bury you
    A6W1120
               hard into the thin bucket seats: every quick-fire gearchange though the massively
               solid Borg-Warner box kicks at the back end.
    AC4 2431 Jinny was so startled that she nearly kicked the bucket over.
```

- ATE 787 ''Did you think I'd kicked the bucket, Ma?''

Combining ideas

 $\{kick/V\}\ the\ (_{\{A\}})^*\ bucket_NN1$



More BNCweb info

Hoffmann et al. (2008)
Reading
Book extract
Getting started
Quick tutorial

ISBN 978-3-631-56315-1



Changing context window Sorting Distributions Downloads Collocations Dispersion References

Looking at more context

Clicking on the highlighted target word takes you to a new screen in which more context is shown.



67 The five were previously imprisoned from June until October 1990 for allegedly organizing a political party — all parties are prohibited.

68 The releases occurred shortly before two AI representatives arrived in Swaziland for talks with the government.

69 Mauritania

70 In April AI called on the Mauritanian Government to investigate reports that up to 200 political prisoners had died or been killed in military or police custody—some executed without trial and many others as a result of torture. 71 The victims were among some 3,000 black Mauritanians arrested late last year. 72 The authorities claimed they were conspiring to overthrow the government, which is dominated by a different community, the Moors, but offered no evidence to substantiate this claim.

13 It took AI several months to compile details of the killings, but prisoners released in March 1991 provided information and themselves bore scars from torture. 14 In a few cases, prisoners are reported to have been deliberately executed without trial. 75 Most of the reported deaths, however, were due to torture in both military barracks and police stations.

76 Torture of political detainees in Mauritania has been routine since 1986, but it has never before been used on such a scale.

77 After the March 1991 releases, over 70 detainees were still held and were apparently due to be brought to trial in April.



Scrolling through context

Clicking on the left and right arrow will scroll back or forward through the context window. Here we click right.



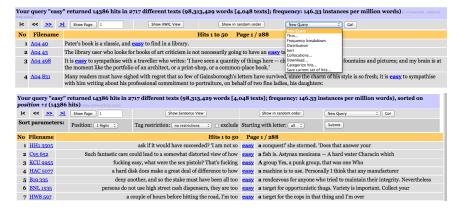
Scrolling through context

Clicking on the left and right arrow will scroll back or forward through the context window. Here we click right.



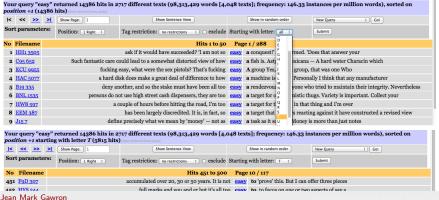
Sorting

Lexicographers like to sort their results to put similar contexts together.



Sorting parameters

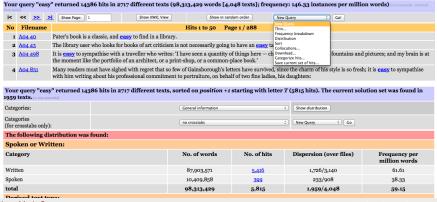
The sorted window has a new set of sort parameters that can be set. Position 1 right is the default, and often most useful, but we may want to start at a particular letter. Here we choose "t" and show part of p. 10 of the result





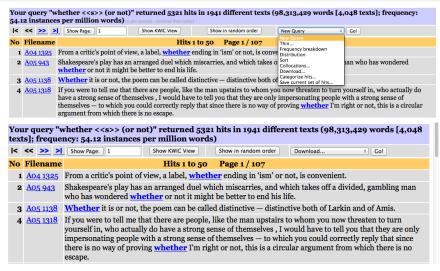
Simple distribution

To look at various distributions, pull down the main search results action menu and click on distribution. The Adj *easy* is much mor ecommon in written than spoken language.



Changing context window Sorting Distributions Downloads Collocations Dispersion References

Downloading results







gawron

Download form

Download concordance

Enter name for the downloaded file:

Output format options	
Choose operating system on which you will be working with the file:	UNIX (incl. OS X) ‡
Print codes (numbers) or full values for metatextual categories:*	full values :
Mark query result in sentence (format: <<< result >>>):	yes ‡
Size of context:	1 <s>-unit ‡</s>
Download both tagged and untagged version of your results:*	yes ‡
Write information about order of categories at the beginning of file:*	no :
Format of output: KWIC or list:*	List ‡
Include corpus positions (required for re-import)*	Yes :
Include URL to context display*	Yes :

Default results (1 example)

1 A04 1325 From a critic 's point of view , a label , <<< whether >>> ending in

Abbreviated results

- 1 A04 1325 From a critic 's point of view , a label , whether ending in ' ism ' or not , is convenient .
- 2 A05 943 Shakespeare 's play has an arranged duel which miscarries , and which takes off a divided , gambling man who has wondered whether or not it might be better to end his life .
- 3 AO5 1138 Whether it is or not , the poem can be called distinctive distinctive both of Larkin and of Amis .
- 4 A05 1318 If you were to tell me that there are people , like the man upstairs to whom you now threaten to turn yourself in , who actually do have a strong sense of themselves , I would have to tell you that they are only impersonating people with a strong sense of themselves to which you could correctly reply that since there is no way of proving whether I 'm right or not , this is a circular argument from which there is no escape .

Collocates

Definition

The collocates of a word w are the words that co-occur with w most often. [First pass]

Definition

The collocates of a word w are the words that, statistically, are most strongly associated with w.

Requesting collocates



Collocation settings window

BNC Collocation Settings		
Calculate over sentence boundaries:		no ;
Include lemma information:		no :
Maximum window span:		+/- [5 :
	Submit	

Example: powerful collocates

Collocation parameters:	on parameters:			
Information:	collocations	Statistics:	Log-likelihood ‡	
Collocation window span:	1 Left : 3 Right :	Basis:	whole BNC :	
Freq(node, collocate) at least:	5 💠	Freq(collocate) at least:	5 .	
Filter results by:	Specific collocate:	and/or tag: any adjective :	Submit changed parameters CGol	

There are 9846 different types in your collocation database for "[word="powerful"%c]". (Your query "powerful" returned 7062 hits in 1919 different texts) and account to the control of th

No.	Word	Total No. in whole BNC	Expected collocate frequency	Observed collocate frequency	In No. of texts	Log-likelihood value
1	most	97,999	35.401	647	448	2549.7677
2	<u>a</u>	2,164,238	781.812	2301	1160	1989.3687
3	more	209,333	75.620	<u>671</u>	472	1749.4068
4	tool	2,180	0.788	25	64	537-7234
5	very	119,400	43.132	<u>260</u>	200	501.9577
6	influence	10,324	3.729	92	77	414.2255
7	and	2,616,708	945.263	<u>1564</u>	890	347-4752
8	force	15,497	5.598	<u>85</u>	74	304.1962
9	lobby	1,118	0.404	39	36	280.6709
10	<u>rich</u>	6,691	2.417	<u>61</u>	51	277.2885
11	weapon	1,938	0.700	42	39	262.2437
12	<u>than</u>	144,567	52.224	<u>207</u>	172	261.3607
13	enough	31,123	11.243	<u>96</u>	89	242.6597
14	<u>is</u>	990,281	357.730	<u>675</u>	454	225.2373
15	immensely	642	0.232	27	24	204.5034

Example: powerful noun collocates

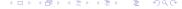
Collocation parameters:			
Information:	collocations ‡	Statistics:	Log-likelihood :
Collocation window span:	1 Left : 1 Right :	Basis:	whole BNC :
Freq(node, collocate) at least:	5 ;	Freq(collocate) at least:	5 :
Filter results by:	Specific collocate:	and/or tag: any noun :	Submit changed parameters : Go!

There are 9846 different types in your collocation database for "[word="powerful"%c]". (Your query "powerful" returned 7062 hits in 1919 different texts)

No.	Word	Total No. in whole BNC	Expected collocate frequency	Observed collocate frequency	In No. of texts	Log-likelihood value
1	influence	7,831	0.981	72	61	477.6129
2	tool	2,177	0.273	49	44	412.5753
3	<u>weapon</u>	1,938	0.243	35	34	279.1873
4	force	13,308	1.667	<u>48</u>	45	230.2471
5	man	58,595	7.338	<u>64</u>	55	164.1834
6	voice	16,127	2.020	38	35	151.2464
7	body	24,588	3.079	42	27	141.8188
8	lobby	944	0.118	<u>16</u>	16	125.5725
9	header	713	0.089	<u>15</u>	10	124.2254
10	effect	22,545	2.823	37	33	122.1823

Min frequency may matter

k = 3, medicine



References I

Hoffmann, S., S. Evert, N. Smith, D. Lee, and B. P. Prytz. 2008. Corpus Linguistics with BNCweb: a Practical Guide. New York: Peter Lang.