This is NOT a proposal about mass surveillance!

Analysing the terminology of the UK's Snooper's Charter

34C3, Leipzig

December 29, 2017, 11.30 pm., Saal Adams

Introduction

What is this talk going to aim at?

Introduction

What is the Investigatory Powers Act (aka Snooper's Charter)?

Corpus Linguistics

What is Corpus Linguistics?

- Corpus Linguistics is the study of language based on examples of real life language use (McEnery and Wilson 1).
- Corpus = collection of machine-readable texts
- Size: several thousand to millions of words

Corpus Linguistics

Why is Corpus Linguistics useful?

- allows for working with large amounts of data
- helps to reduce researcher bias

Corpus Linguistics

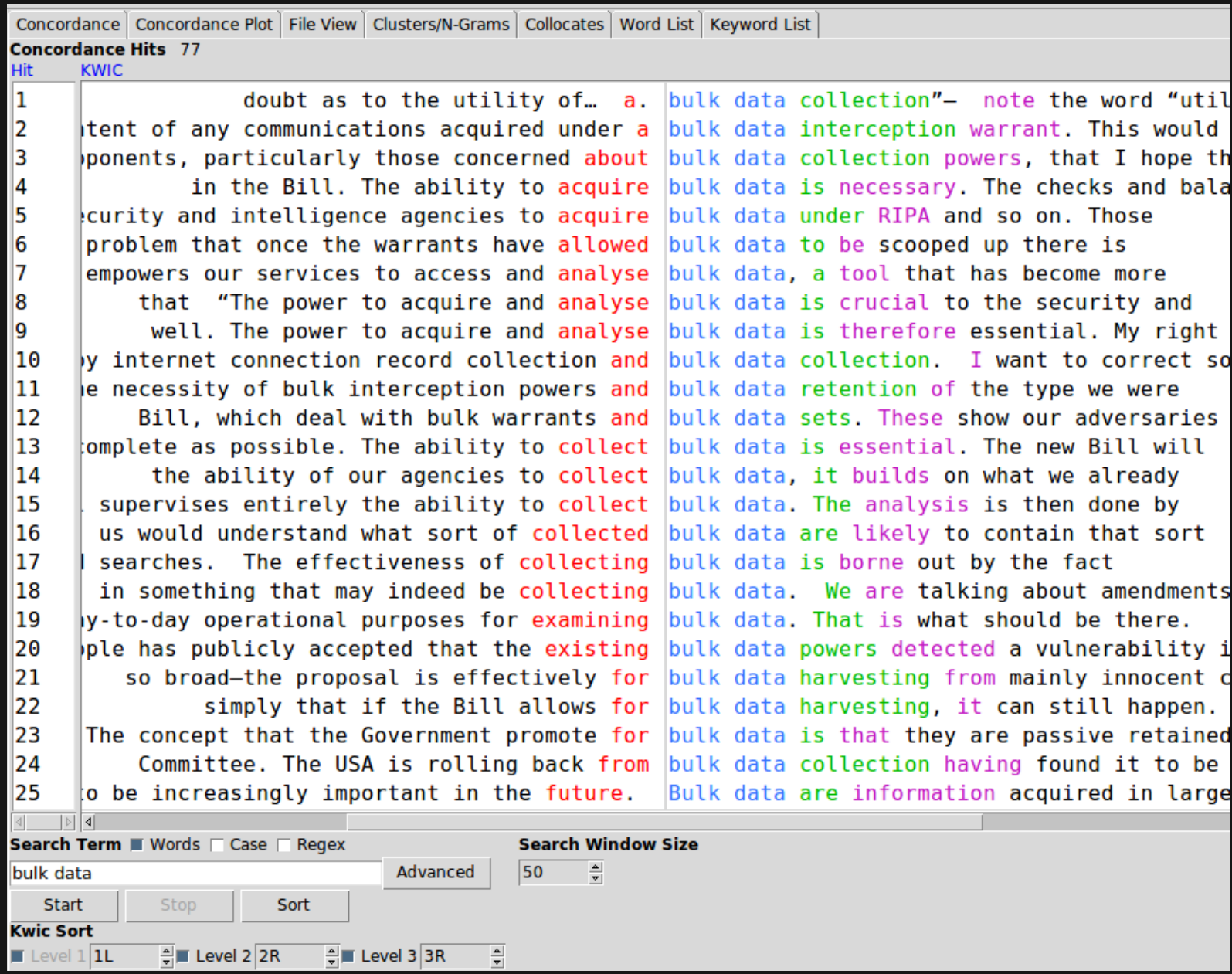
Methods of Corpus Linguistics

- Concordance analysis
- Collocation analysis

Concordance Analysis

Concordance analysis

- lists several words surrounding the keywords
- span of words around the keyword can be adjusted


Concordance
Concordance Plot
File View
Clusters/N-Grams
Collocates
Word List
Keyword List
Concordance Hits 77
Hit KWIC
1 doubt as to the utility of… a. bulk data collection"– note the word "util
2 tent of any communications acquired under a bulk data interception warrant. This would
3 ponents, particularly those concerned about bulk data collection powers, that I hope th
4 in the Bill. The ability to acquire bulk data is necessary. The checks and bala
5 curity and intelligence agencies to acquire bulk data under RIPA and so on. Those
6 problem that once the warrants have allowed bulk data to be scooped up there is
7 empowers our services to access and analyse bulk data, a tool that has become more
8 that "The power to acquire and analyse bulk data is crucial to the security and
9 well. The power to acquire and analyse bulk data is therefore essential. My right
10 y internet connection record collection and bulk data collection. I want to correct so
11 e necessity of bulk interception powers and bulk data retention of the type we were
12 Bill, which deal with bulk warrants and bulk data sets. These show our adversaries
13 omplete as possible. The ability to collect bulk data is essential. The new Bill will
14 the ability of our agencies to collect bulk data, it builds on what we already
15 supervises entirely the ability to collect bulk data. The analysis is then done by
16 us would understand what sort of collected bulk data are likely to contain that sort
17 searches. The effectiveness of collecting bulk data is borne out by the fact
18 in something that may indeed be collecting bulk data. We are talking about amendments
19 y-to-day operational purposes for examining bulk data. That is what should be there.
20 ple has publicly accepted that the existing bulk data powers detected a vulnerability i
21 so broad—the proposal is effectively for bulk data harvesting from mainly innocent c
22 simply that if the Bill allows for bulk data harvesting, it can still happen.
23 The concept that the Government promote for bulk data is that they are passive retained
24 Committee. The USA is rolling back from bulk data collection having found it to be
25 o be increasingly important in the future. Bulk data are information acquired in large
Search Term Words Case Regex
bulk data
Advanced
Search Window Size
50
Start
Stop
Sort
Kwic Sort
Level 1 1L Level 2 2R Level 3 3R

Corpus Linguistics

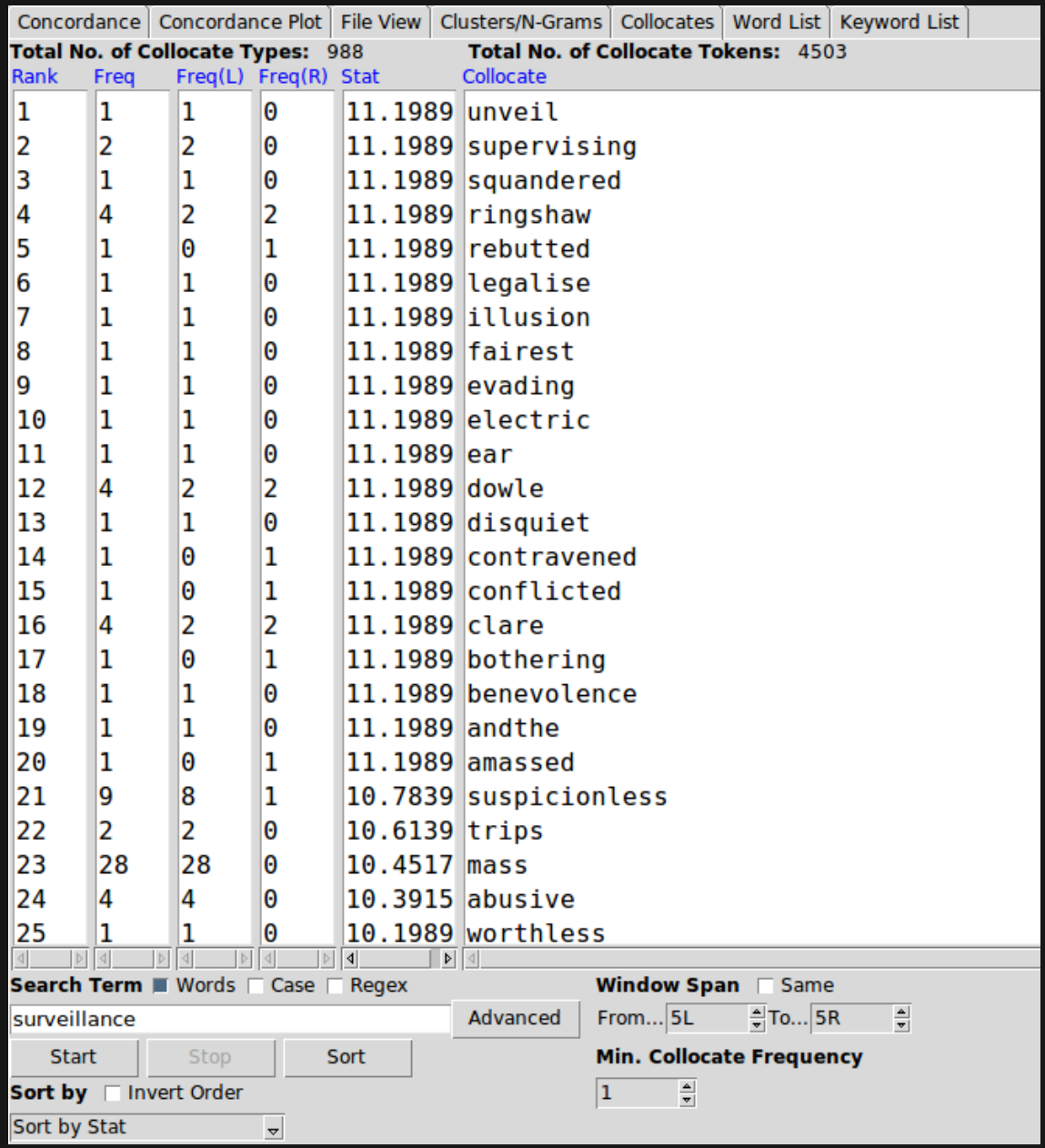
Collocation analysis

- useful to examine the connotations and associations between words
- collocation = above-chance frequent co-occurrence of two words within a pre-determined span (Baker et. al. 278)

Corpus Linguistics

Mutual Information (=MI) score:

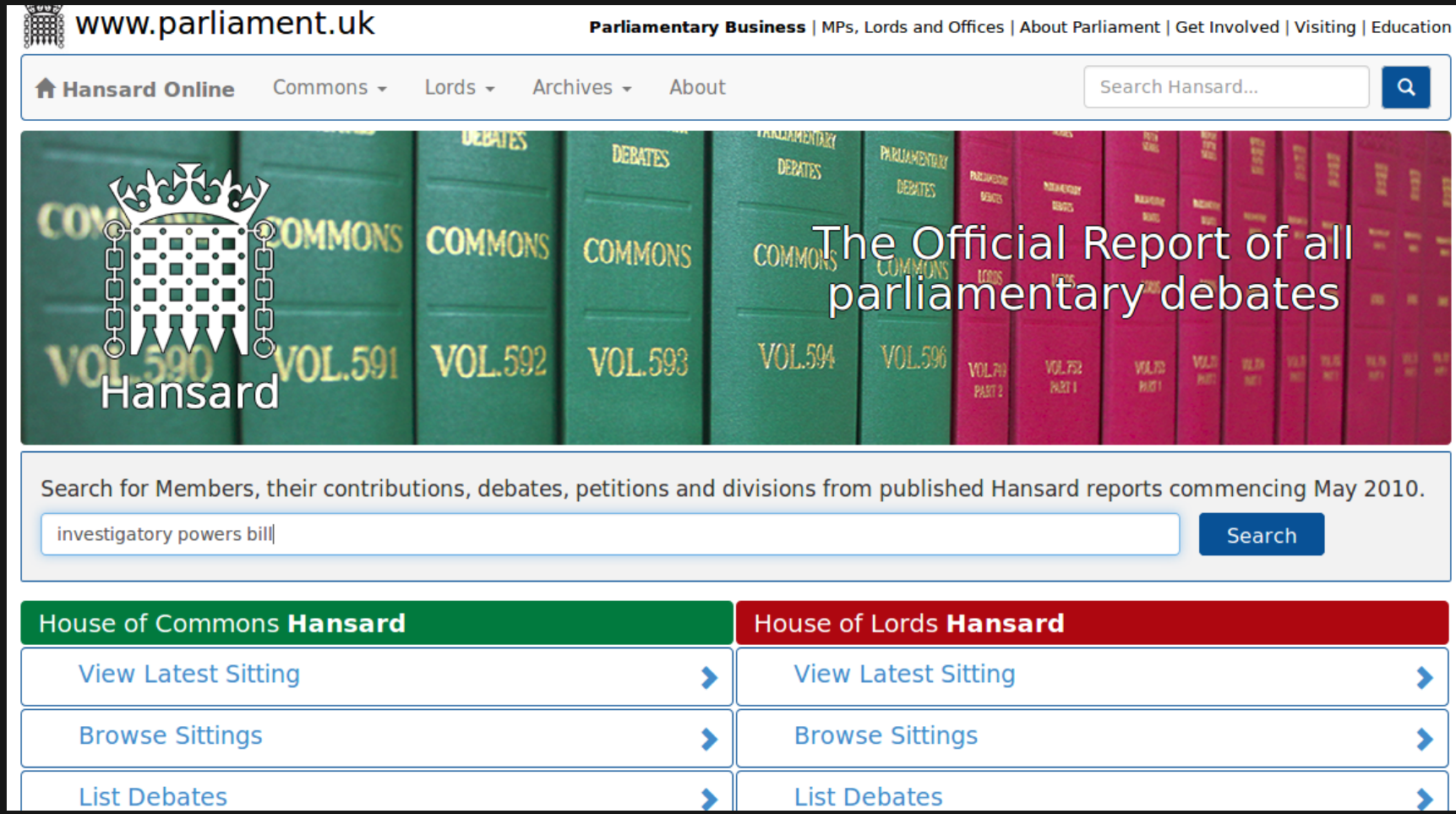
- expected probability of two words appearing near to each other, their relative frequencies and the overall size of the corpus
- comparison between expected probability and observed probability
- the higher the number, the stronger the collocation

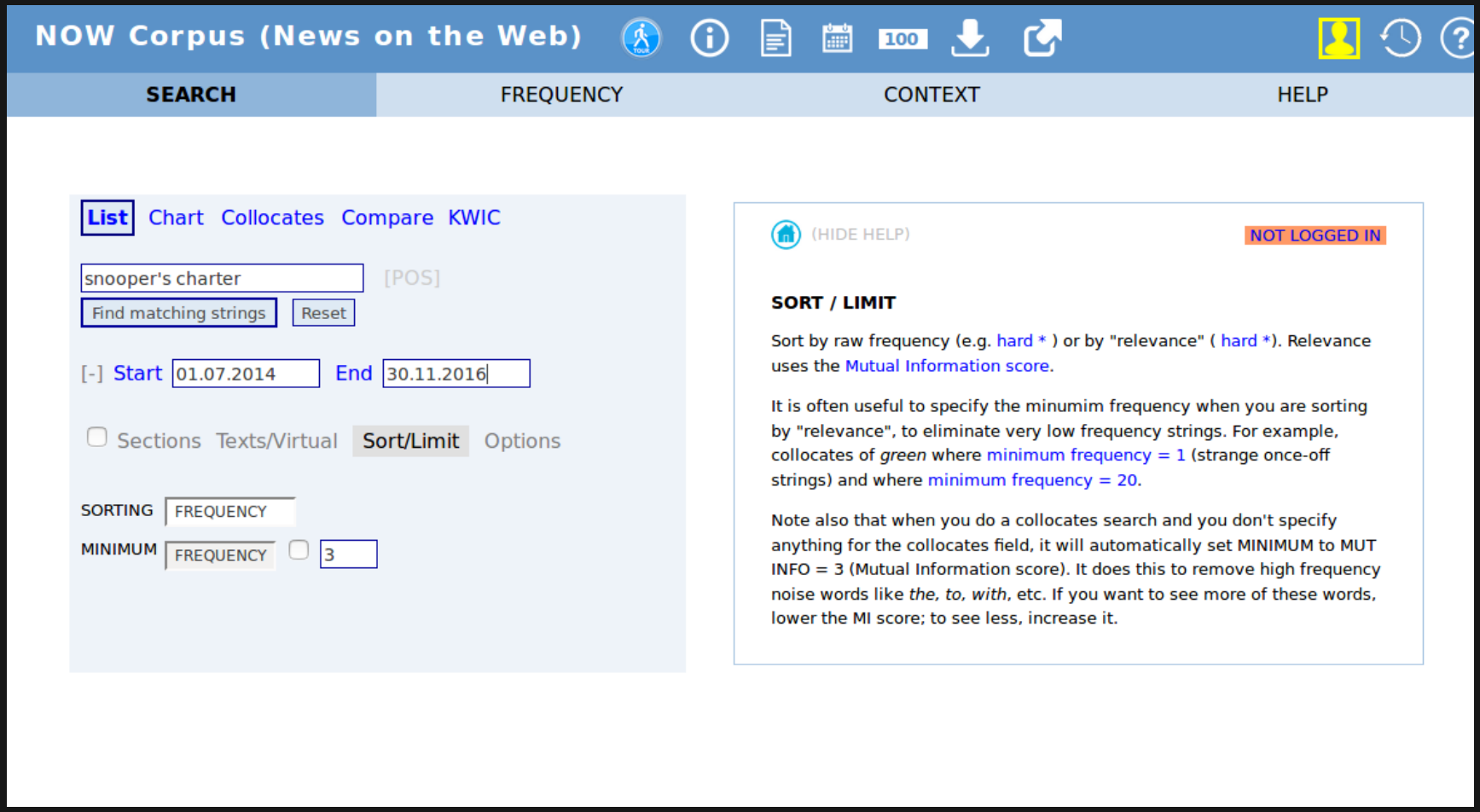

Concordance
Concordance Plot
File View
Clusters/N-Grams
Collocates
Word List
Keyword List
Total No. of Collocate Types: 988
Total No. of Collocate Tokens: 4503
Rank Freq Freq(L) Freq(R) Stat Collocate
1 1 1 0 11.1989 unveil
2 2 2 0 11.1989 supervising
3 1 1 0 11.1989 squandered
4 4 2 2 11.1989 ringshaw
5 1 0 1 11.1989 rebutted
6 1 1 0 11.1989 legalise
7 1 1 0 11.1989 illusion
8 1 1 0 11.1989 fairest
9 1 1 0 11.1989 evading
10 1 1 0 11.1989 electric
11 1 1 0 11.1989 ear
12 4 2 2 11.1989 dowle
13 1 1 0 11.1989 disquiet
14 1 0 1 11.1989 contravened
15 1 0 1 11.1989 conflicted
16 4 2 2 11.1989 clare
17 1 0 1 11.1989 bothering
18 1 1 0 11.1989 benevolence
19 1 1 0 11.1989 andthe
20 1 0 1 11.1989 amassed
21 9 8 1 10.7839 suspicionless
22 2 2 0 10.6139 trips
23 28 28 0 10.4517 mass
24 4 4 0 10.3915 abusive
25 1 1 0 10.1989 worthless
Search Term Words Case Regex
surveillance
Advanced
Start
Stop
Sort
Sort by Invert Order
Sort by Stat
Window Span Same
From... 5L To... 5R
Min. Collocate Frequency
1

Corpora

Two corpora for this analysis:

- Investigatory Powers Act-Corpus (IPAC)
- News on the Web – Corpus (NOWC)


www.parliament.uk
Parliamentary Business | MPs, Lords and Offices | About Parliament | Get Involved | Visiting | Education
Hansard Online
Commons
Lords
Archives
About
Search Hansard...
The Official Report of all parliamentary debates
Hansard
Search for Members, their contributions, debates, petitions and divisions from published Hansard reports commencing May 2010.
investigatory powers bill
Search
House of Commons Hansard
View Latest Sitting
Browse Sittings
List Debates
House of Lords Hansard
View Latest Sitting
Browse Sittings
List Debates


NOW Corpus (News on the Web)
100
SEARCH
FREQUENCY
CONTEXT
HELP
List
Chart
Collocates
Compare
KWIC
snooper's charter
[POS]
Find matching strings
Reset
[-] Start
01.07.2014
End
30.11.2016
Sections
Texts/Virtual
Sort/Limit
Options
SORTING
FREQUENCY
MINIMUM
FREQUENCY
3
(HIDE HELP)
NOT LOGGED IN
SORT / LIMIT
Sort by raw frequency (e.g. hard *) or by "relevance" (hard *). Relevance uses the Mutual Information score.
It is often useful to specify the minumim frequency when you are sorting by "relevance", to eliminate very low frequency strings. For example, collocates of green where minimum frequency = 1 (strange once-off strings) and where minimum frequency = 20.
Note also that when you do a collocates search and you don't specify anything for the collocates field, it will automatically set MINIMUM to MUT INFO = 3 (Mutual Information score). It does this to remove high frequency noise words like the, to, with, etc. If you want to see more of these words, lower the MI score; to see less, increase it.

Snooper's Charter

OED Online: to snoop: "to pry into matters one need not be concerned with"

Table 1: Concordances IPAC

knock out completely that lazy label of	”snooper’s charter”.	That is
I was not going to use the phrase	”snoopers’ charter”	because it is
attack him for the phrase	”snooper’s charter”,	but he
seriously misleading phrase	”snooper’s charter”	has been
snooping, hence the populist phrase	”snooper’s charter”	That view is

Snooper's Charter

In my view, it is lazy to label the Bill as a snoopers charter or a plan for mass surveillance. In fact, it is worse than lazy: it is insulting to people who work in the police and in the security services. *(Burnham, Draft Investigatory Powers Bill: Volume 601, Column 825)*

Snooper's Charter

Table 2: Concordances NOWC

reintroduce a beefed-up version of the	"snooper's charter"	In an
national security. The	snooper's charter	is discredited,
saying that the bill is "neither a	snooper's charter,	nor a plan for
renewed effort to pass a	"snooper's charter"	bill of
sweeten the pill of her revived	snooper's charter	on Wedn

Snooper's Charter

IPAC

- 75% of occurrences negated
- no reference to previous snooper's charter
- criticism for using term snooper's charter

NOWC

- hardly negated
- references to previous snooper's charter
- criticism of content/implications

Mass Surveillance

Table 3: Concordances IPAC

that this is not a proposal for	mass surveillance	and to restate
not accept that the Bill is a plan for	mass surveillance	but we need to
neither a snooper's charter, nor a plan for	mass surveillance	[Hon. Member
a snoopers charter or a plan for	mass surveillance	In fact, it is
neither a snooper's charter, nor a plan for	mass surveillance	"—[Official

Mass Surveillance

Table 4: Concordances NOWC

GCHQ's alleged	mass surveillance	of private communications breaks
the fallout of the NSA's	mass surveillance	programs are also uniting to
under criticism for its	mass surveillance	practices. Notably, the country
reports about the NSA's	mass surveillance	programs. Vincent Yu
of carrying out	mass surveillance	as their critics have claimed.

Mass Surveillance

IPAC

- over 50% of occurrences negated

NOWC

- hardly negated
- full extent of negative connotations

Bulk Data

”bulk”:

OED: ”great or considerable volume, a mass; the collective mass of any object”

Figure 1:
https://commons.wikimedia.org/wiki/File:Spice_Market_Istanbul_Turkey_2007.J

Table 5: Concordances IPAC

The ability to acquire	bulk data	is necessary. The checks
The ability to collect	bulk data	is essential. The new Bill will help
the ability to collect	bulk data.	The analysis is then done by trustec
agencies were able to use	bulk data	to identify that he had recently
services to access and analyse	bulk data	a tool that has become more import

Bulk Data

For example, in 2010, an airline worker in the UK who had access to airline capability was stopped as a result of access to bulk data. We have information on GCHQ in- telligence uncovering networks of extremists who had travelled to Pakistan and then been stopped as a result of the acquisition of bulk data. *(Hanson, Investigatory Powers Bill: Volume 607, Column 867)*

Bulk Data

Figure 2: By Herry Lawford - originally posted to Flickr as Harvest, CC BY 2.0, https://commons.wikimedia.org/w/index.php?curid=11269097

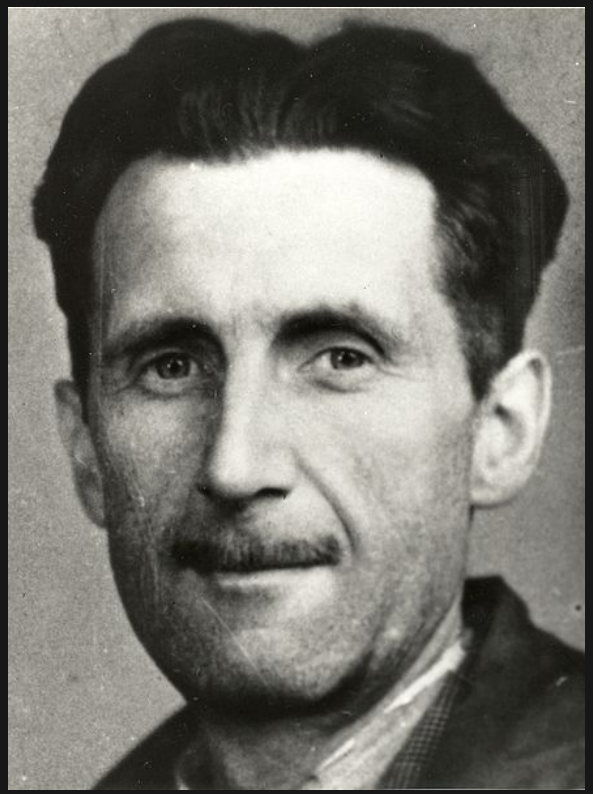
Bulk Data

I appreciate that bulk powers are controversial, but I am absolutely sure that we do not conduct data harvesting in this country. It simply does not happen. The use of bulk powers is not for that purpose, but for the purpose of examining material. Even though that may be done in bulk, it is done in a way that does not amount to the generalised harvesting of data for their examination. It simply is not. *(Grieve, Investigatory Powers Bill: Volume 611, Column 1056)*

Bulk Data

Doublethink:

> *The power of holding two contradictory beliefs in one's mind simultaneously, and accepting both of them (...)* *Orwell, 1984, p. 220.*



Bulk Data

Redefinition of ”mass surveillance” as ”bulk data collection”

Bulk Data

"Mass surveillance" and "bulk data" are not to be equated:

> *Collection of bulk data, most of which are never even read, does not constitute mass surveillance.*
> *(Rooker, Investigatory Powers Bill: Volume 773, Column 1423).*

Table 6: Concordances NOWC

concluded that the	bulk data	collection was "illegal and unproductive"
businesses already face	bulk data	collection of a different kind, as
in the first world war to	bulk data	collection exposed by Ed Snowden,
broader surveillance and	bulk data	gathering. There is as yet
freedom to preserve ours.	bulk data	collection: Neither lawful nor effective

Bulk Data

IPAC

- majority: importance and necessity
- redefinition of ”mass surveillance” as ”bulk data collection”

NOWC

- majority: Snowden Revelations/USA Freedom Act

TOP 3 quotes

I sincerely hope that as the Bill proceeds—we have a way to go yet—we will explain that we do not conduct mass surveillance in the UK. Indeed, it is not done in the USA. Collection of bulk data, most of which are never even read, does not constitute mass surveillance.
(Rooker, Investigatory Powers Bill: Volume 773, Column 1423).

TOP 3 quotes

However uneasy we may feel about internet connection records or thematic warrants, that does not compare to the infinitely greater unease we ought to feel about our intelligence agencies being unable to use those tools to keep us safe. (Warburton, Investigatory Powers Bill: Volume 607, Column 891).

TOP 3 quotes

The power of holding two contradictory beliefs in one's mind simultaneously, and accepting both of them... To tell deliberate lies while genuinely believing in them, to forget any fact that has become inconvenient, and then, when it becomes necessary again, to draw it back from oblivion for just as long as it is needed, to deny the existence of objective reality and all the while to take account of the reality which one denies—all this is indispensably necessary. *Orwell, 1984, p. 220.*

Last But Not Least

I appreciate that bulk powers are controversial, but I am absolutely sure that we do not conduct data harvesting in this country. It simply does not happen. The use of bulk powers is not for that purpose, but for the purpose of examining material. Even though that may be done in bulk, it is done in a way that does not amount to the generalised harvesting of data for their examination. It simply is not. *(Grieve, Investigatory Powers Bill: Volume 611, Column 1056)*

Bibliography

Baker, Paul, et al. “A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press.” Discourse Society 19.3 (2008): 273–306

Harris, Mike. “The politics of surveillance are about politics, not keeping us safe.” The Telegraph

McEnery, Anthony M and Anita Wilson. Corpus linguistics: an introduction. Edinburgh University Press, 2001.

NOW, Corpus. https://corpus.byu.edu/now/

OED Online. http://www.oed.com/.

Any questions?

lilalaser@posteo.de