

CORPORA IN LEXICOGRAPHY (PART TWO)

Iztok Kosem

Trojina, Institute for Applied Slovene Studies
Ljubljana, Slovenia

Contact: iztok.kosem@trojina.si

Sketch Engine – key lexicographic features

- Word Sketches
 - ▣ collocations can often be used to distinguish between different meanings of a word (Hoey, 2005)
 - ▣ List of collocates not so helpful
 - ▣ Word Sketches are “summaries of a word’s grammatical and collocational behaviour” (Kilgarriff et al., 2004)
 - ▣ In some dictionary projects, Word Sketches now point of departure (Word Sketch then concordances)

PP_PP_between-i	44	2.5
for	<u>39</u>	28.68

unary_rels		
it_constr	<u>1027</u>	53.6
that_0	<u>17935</u>	49.3
that_0_subj	<u>91</u>	31.2
quote	<u>115</u>	11.7

PP_against-i	405	41.7
possibility	<u>18</u>	20.97
view	<u>20</u>	19.37
interpretation	<u>13</u>	17.23
idea	<u>12</u>	15.25
hypothesis	<u>10</u>	14.52
assumption	<u>8</u>	12.76
notion	<u>7</u>	12.41
mechanism	<u>9</u>	12.16
role	<u>9</u>	10.46
use	<u>9</u>	10.35
analysis	<u>8</u>	7.49
effect	<u>8</u>	6.86
model	<u>6</u>	4.84

pro_subject	3465	12.7
I	<u>1910</u>	75.28
he	<u>697</u>	52.33
we	<u>478</u>	42.25
she	<u>162</u>	37.73
they	<u>135</u>	26.78
it	<u>78</u>	12.15

PP_for-i	1127	4.8
importance	<u>31</u>	22.1
existence	<u>24</u>	21.75
nullification	<u>7</u>	21.3
necessity	<u>12</u>	19.9
view	<u>22</u>	15.75
claim	<u>16</u>	15.17
superiority	<u>6</u>	15.17
thesis	<u>8</u>	14.87
conception	<u>10</u>	14.59
need	<u>16</u>	14.45
presence	<u>15</u>	13.72
interpretation	<u>12</u>	12.66
role	<u>19</u>	12.41
distinction	<u>8</u>	11.54
approach	<u>18</u>	11.5
protection	<u>7</u>	10.58
recognition	<u>7</u>	10.4
link	<u>7</u>	10.3
reading	<u>7</u>	10.16
influence	<u>8</u>	9.32
consideration	<u>6</u>	9.03
possibility	<u>7</u>	8.91
connection	<u>6</u>	8.71
conclusion	<u>6</u>	8.42
use	<u>11</u>	8.21

AVP_mod	2210	4.2
convincingly	<u>64</u>	49.8
also	<u>437</u>	38.4
persuasively	<u>23</u>	38.38
further	<u>130</u>	37.89
recently	<u>78</u>	33.36
famously	<u>21</u>	31.56
as	<u>190</u>	27.93
rightly	<u>18</u>	27.69
cogently	<u>9</u>	27.31
then	<u>80</u>	24.41
plausibly	<u>12</u>	24.34
now	<u>55</u>	24.21
similarly	<u>27</u>	24.21
often	<u>68</u>	24.13
strongly	<u>36</u>	23.51
forcefully	<u>10</u>	23.43
eloquently	<u>7</u>	22.46
reasonably	<u>15</u>	22.13
already	<u>38</u>	21.85
therefore	<u>42</u>	21.49
even	<u>56</u>	20.83
repeatedly	<u>13</u>	19.27
explicitly	<u>19</u>	18.73
long	<u>20</u>	18.13
powerfully	<u>6</u>	17.36

subj_NP	7039	3.7
DOJ	<u>47</u>	40.13
Schmalensee	<u>44</u>	39.89
article	<u>201</u>	33.8
Kant	<u>56</u>	30.19
scholar	<u>89</u>	30.08
paper	<u>165</u>	29.47
author	<u>113</u>	29.11
researcher	<u>94</u>	27.45
other	<u>119</u>	27.1
critic	<u>50</u>	27.04
Fisher	<u>21</u>	24.09
theorist	<u>35</u>	22.85
Hume	<u>20</u>	22.62
Rawls	<u>14</u>	22.26
Cohen	<u>19</u>	22.2
Baudrillard	<u>8</u>	21.82
Adorno	<u>18</u>	21.74
Simmel	<u>10</u>	21.72
commentator	<u>22</u>	21.29
Sturgeon	<u>7</u>	21.27
Mulvey	<u>7</u>	21.1
Kelsen	<u>10</u>	21.08
Jackson	<u>17</u>	20.51
Smith	<u>27</u>	20.51
Eritrea	<u>8</u>	20.3

□ Exercise 2

- In the Sketch Engine, select the CAJA corpus.
- Make a Word Sketch for the verb *argue*.
- Make a list of relations or collocates that you would include in a dictionary. What role would you attribute to each of them – a sense, a phrase, or an example?

argue verb

1. (**transitive**) If someone **argues** a view or an idea in an article or a book, they present the idea and support it with evidence. Note: **argue** is very often followed by a **that**-clause.
2. (**intransitive**) If you **argue for** or you **argue in favour of** an idea or theory, you agree with it and provide evidence that supports it.
3. (**intransitive**) If you **argue against** an idea or theory, you provide evidence that opposes it.
4. (**intransitive**) If you **argue with** someone about/over something, you discuss it because you have different opinions.
5. (**intransitive**) If you **argue with** someone or someone's view, you disagree with it.
6. (**intransitive**) If people **argue**, they talk angrily to each other because they disagree.

WORD FORMS:

argue, argues, arguing, argued

① **[transitive]** If someone **argues** a view or an idea in an article or book, they present the idea and support it with evidence. Note: **argue** is very often followed by a **that**-clause.

- *Mazrui (1999: 1) also argued that Africa developed the West.*
- *The question, I will argue in this essay, cannot be answered without a clearer sense of how Greece relates to Rome in Shelley's work.*
- *Some researchers have argued that a decentralized economy will have difficulty in fully exploiting the growth returns of general purpose technologies.*
- *"Some frames," argues Gamson (1992: 135), "have a natural advantage because their ideas and language resonate with a broader political culture."*

article/paper/essay argues that...

- *This article argues that literacy is an important sociological phenomenon, but one largely under-researched in British sociology.*

it could/can/might/may be argued that...

- *It could be argued that, by suggesting complete openness the liberals initially tried to handle the issue differently.*

one could/can/might/may argue that...

- *In almost simplistic terms, one could argue that, without sound, there would be no music.*

as someone argues somewhere / as argued somewhere

- *As argued elsewhere (Steadman and Palmer 1995; Palmer and Steadman 2004), the key to this distinction appears to be the behaviour of the listeners.*
- *No doubt state leaders in new states often follow the ethnic model of nation-building, but, as I argue below (=later in the article), this is not the only possible solution, and normally not the best one in order to survive and flourish as nations.*

as argued by someone / as someone argues

- *Stravinsky's music was never intended to be complete in itself, but to be made complete by the choreography, as Irene Alm has **persuasively argued**.*
- *The results are consistent with the hypothesis that capital markets are integrated, as argued by Campbell and Hamao (1992) and Harvey (1991), among others.*

② **[intransitive]** If you **argue for** or you **argue in favour of** an idea or theory, you agree with it and provide evidence that supports it.

- *Vygotsky (1978) argued for the importance of language as both a psychological and cultural tool.*
- *This article argues in favour of putting into place a legal framework for feedback intermediaries.*

③ **[intransitive]** If you **argue against** an idea or theory, you provide evidence that opposes it.

- *Ian Sefton is physics educator who has strongly argued against the mistaken view that electrons possess potential energy.*
- *Nevertheless, four pieces of circumstantial evidence argue against this narrower view.*

④ **[intransitive]** If you **argue with** someone about/over something, you discuss it because you have different opinions.

- *Living cells had of course been seen through the microscope before, and a range of their activities had been described and argued over.*
- *The committee also argued about and negotiated the unorthodox multimedia exhibition style.*

⑤ **[intransitive]** If you **argue with** someone or someone's view, you disagree with it.

- *Few would argue with the idea that we should maximise positive value and minimise evil.*
- *According to Derrida, the machine is dangerous because it is the opposite of life; it is 'pure representation' and 'never runs by itself'. So far it is difficult to argue with Derrida.*

⑥ **[intransitive]** If people **argue**, they talk angrily to each other because they disagree.

- *Again, the father went to argue with the doctors. The chief physician went almost wild, according to Hallvard.*

argue about/over something

- *Bernardino Antonio, a resident of Actopan, beat his wife fatally in 1808 after they argued about the punishment of their child.*

WORD ORIGIN

Date: 1300-1400

Language: Old French

Origin: *arguer* (to assert, charge with), from Latin *arguere* (to make clear)

Sketch Engine – key lexicographic features

- TickBox lexicography
 - ▣ Quick selection and export of examples and collocates (and grammatical relations) into dictionary-writing system
- GDEX (Good Dictionary Examples)
 - ▣ Finding good dictionary examples is time-consuming
 - ▣ can it be automated?
 - ▣ GDEX first used by Macmillan (needed examples for learner's dictionary)
 - ▣ Lexicographers *usually* select from top ten examples (Rundell, 2011)

GDEX heuristics (for Macmillan)

- Sentence length: 10-25 words
- Whole sentences only
- Most common words
- Rare words penalized
- Pronouns and anaphora penalized
- Proper names penalized
- Sentences with target collocation in the main clause are preferred
- Sentences with target collocation towards the end preferred

GDEX

- Obtaining a score for each sentence
 - ▣ Adding weights to each heuristic (how relevant is it)
 - ▣ Score on each heuristic
 - ▣ Add together weighted score

They discussed the discrimination against pardos by **local authorities.**

	<i>weight</i>	<i>points</i>	score
Sentence length: 10-25 words	2	0.1	0.2
Whole sentences only	X		✓
Most common words	1	0.1	0.1
Rare words penalized	1	-0.1	-0.1
Pronouns and anaphora penalized	2	-0.1	-0.2
Proper names penalized	2	-0.1	0
Sentences with target collocation in the main clause are preferred	1	0.1	0.1
Sentences with target collocation towards the end preferred	1	0.1	0.1

GDEX

- Obtaining a score for each sentence
 - ▣ Adding weights to each heuristic (how relevant is it)
 - ▣ Score on each heuristic
 - ▣ Add together weighted score
- Further developments (e.g. GDEX for Slovene)
 - ▣ First, lexicographers select best examples manually
 - ▣ Computer learns from human choices
 - ▣ Heuristics devised, tested and refined

Sketch Engine – key lexicographic features

- Thesaurus and Sketch Difference
 - Thesaurus helps with identifying “nearest neighbours”, i.e. potential (near)synonym and, in some cases, antonyms
 - Sketch Difference helps with describing differences or similarities between two words (or word forms)

Thesaurus: *clever*

Lemma	Score	Freq
intelligent	0.296	1844
funny	0.282	4315
lucky	0.249	3951
silly	0.245	2645
brave	0.242	1615
stupid	0.238	3089
ambitious	0.226	1512
pretty	0.22	2617
smart	0.219	1410
brilliant	0.214	3411
subtle	0.209	1773
charming	0.207	1334
handsome	0.197	1578

clever/intelligent

preloaded/bnc freq = 2237/1844

Common patterns

clever	6.0	4.0	2.0	0	-2.0	-4.0	-6.0	intelligent
---------------	-----	-----	-----	---	------	------	------	--------------------

adj_subject	128	119	8.5	7.9
man	<u>3</u>	<u>3</u>	4.6	4.7

adj_comp_of	37	20	6.3	3.4
look	<u>6</u>	<u>4</u>	13.4	11.6
become	<u>3</u>	<u>3</u>	8.2	9.4

and/or	396	742	1.9	3.6
witty	<u>5</u>	<u>10</u>	19.2	25.9
clever	<u>10</u>	<u>4</u>	23.5	12.2
charming	<u>3</u>	<u>10</u>	11.7	22.2
beautiful	<u>7</u>	<u>17</u>	14.7	21.6
bright	<u>4</u>	<u>13</u>	10.7	19.6
amusing	<u>5</u>	<u>4</u>	18.1	14.3
young	<u>11</u>	<u>18</u>	14.3	16.4
ambitious	<u>5</u>	<u>4</u>	16.1	12.5
quick	<u>7</u>	<u>6</u>	16.0	12.8
sexy	<u>3</u>	<u>3</u>	13.3	12.1

modifies	1033	1018	2.8	2.7
girl	<u>70</u>	<u>10</u>	29.5	9.9
boy	<u>62</u>	<u>7</u>	29.1	8.1
man	<u>55</u>	<u>66</u>	18.1	20.1
people	<u>25</u>	<u>58</u>	9.9	17.5
woman	<u>19</u>	<u>33</u>	11.4	16.4
person	<u>4</u>	<u>16</u>	3.8	12.5
use	<u>16</u>	<u>15</u>	11.6	11.1
system	<u>6</u>	<u>24</u>	2.8	11.2
eye	<u>5</u>	<u>14</u>	3.7	9.6
thing	<u>16</u>	<u>5</u>	9.5	3.0
face	<u>3</u>	<u>12</u>	2.3	9.4
one	<u>7</u>	<u>3</u>	9.3	4.5
software	<u>4</u>	<u>5</u>	6.2	7.5
dog	<u>4</u>	<u>5</u>	6.1	7.4
way	<u>15</u>	<u>11</u>	7.3	5.5

modifier	666	474	1.5	1.0
very	<u>272</u>	<u>74</u>	46.6	29.8
too	<u>76</u>	<u>20</u>	33.9	19.4
more	<u>21</u>	<u>90</u>	14.0	32.2
so	<u>69</u>	<u>16</u>	29.5	15.1
as	<u>30</u>	<u>12</u>	22.8	14.9
less	<u>4</u>	<u>21</u>	7.2	22.7
extremely	<u>14</u>	<u>5</u>	21.0	12.1
really	<u>18</u>	<u>8</u>	18.1	12.0
quite	<u>13</u>	<u>15</u>	15.2	17.9
incredibly	<u>5</u>	<u>3</u>	17.1	12.9
also	<u>3</u>	<u>3</u>	2.7	3.4
pp_in-p	12	8	0.3	0.2
way	<u>3</u>	<u>3</u>	9.3	10.2

"clever" only patterns

adj_subject	128	8.5
boxing	<u>3</u>	13.7
George	<u>3</u>	9.1
bit	<u>3</u>	8.3

adj_comp_of	37	6.3
box	<u>3</u>	19.2
get	<u>13</u>	17.8
feel	<u>3</u>	8.9
seem	<u>3</u>	8.8

np_adj_comp_of	14	6.1
think	<u>3</u>	9.6

modifies	1033	2.8
dick	<u>12</u>	28.8
trick	<u>18</u>	25.3
clog	<u>8</u>	24.9
folly	<u>10</u>	20.9
chap	<u>11</u>	19.8
ploy	<u>7</u>	19.8
idea	<u>30</u>	18.3
pun	<u>4</u>	16.0
swine	<u>4</u>	15.5
wheeze	<u>3</u>	14.6
concealment	<u>3</u>	14.1
lass	<u>4</u>	14.1

and/or	396	1.9
cunning	<u>5</u>	19.9
little	<u>19</u>	19.3
brave	<u>6</u>	18.0
subtle	<u>6</u>	17.5
devious	<u>3</u>	14.6
intelligent	<u>4</u>	13.4
bloody	<u>5</u>	13.1
cool	<u>4</u>	12.4
efficient	<u>4</u>	11.7
dangerous	<u>4</u>	11.1
rich	<u>4</u>	10.6
pretty	<u>3</u>	10.5

modifier	666	1.5
fiendishly	<u>5</u>	25.3
that	<u>11</u>	21.5
extraordinarily	<u>6</u>	19.7
rather	<u>13</u>	17.9
pretty	<u>8</u>	16.8
particularly	<u>11</u>	16.5
fucking	<u>5</u>	16.5
awfully	<u>4</u>	16.3
jolly	<u>3</u>	13.6
ever so	<u>3</u>	13.2
terribly	<u>3</u>	11.2
bloody	<u>3</u>	10.8

"intelligent" only patterns

adj_subject	119	7.9
human	<u>3</u>	12.1
child	<u>5</u>	8.7
woman	<u>4</u>	7.6

np_adj_comp_of	12	5.2
make	<u>5</u>	11.2

and/or	742	3.6
sensitive	<u>27</u>	30.4
articulate	<u>11</u>	28.9
thoughtful	<u>8</u>	22.1
lively	<u>10</u>	21.6
educated	<u>7</u>	19.9
well-educated	<u>4</u>	18.7
adaptive	<u>5</u>	18.3
attractive	<u>9</u>	16.3
good-looking	<u>4</u>	15.7
honest	<u>6</u>	15.6
human	<u>13</u>	15.3
informative	<u>4</u>	14.8

modifies	1018	2.7
being	<u>33</u>	27.9
hub	<u>9</u>	22.8
robotic	<u>4</u>	18.7
guess	<u>7</u>	18.5
conversation	<u>11</u>	15.4
read	<u>3</u>	14.0
life	<u>26</u>	12.7
robot	<u>4</u>	12.6
controller	<u>5</u>	12.4
subsystem	<u>3</u>	11.7
creature	<u>6</u>	11.2
electronics	<u>4</u>	10.8

modifier	474	1.0
highly	<u>65</u>	41.6
most	<u>40</u>	25.4
reasonably	<u>6</u>	16.6
obviously	<u>7</u>	16.5
fairly	<u>5</u>	12.5
seemingly	<u>3</u>	12.2
otherwise	<u>3</u>	9.9
equally	<u>3</u>	9.0

□ Exercise 3

- Select the CAJA corpus in the Sketch Engine.
- Go to Thesaurus function and create a list for the verb *argue*.
- Click on the verb *suggest* on the list. This will open a new Sketch difference window for *argue* and *suggest*. What differences/similarities can you observe?

□ Exercise 4

- Go to the Sketch Difference and create a Sketch Diff for the singular (*authority*) and plural form (*authorities*) of the noun *authority*.
- What differences can you observe in the relations and collocates of the two forms of *authority*? Are these differences useful in relation to the meanings of *authority* you found earlier?

Latest developments

□ Labelling

- Currently manual, can be unreliable, unsystematic
- How to automate + make it more systematic?
- Likely candidates:
 - Regional labels (e.g. *American English, British English*)
 - Grammar labels (e.g. *usually plural, usually passive*)
 - Domain labels (e.g. *business, chemistry, physics*)

E.g. find nouns that are *usually plural*

- For each noun in the corpus
 - Count, under condition 1
 - all plural instances
 - Count, under condition 2
 - all instances
 - Compute ratio
 - Sort all words according to ratio
 - Words at top of list are best candidates

In development: automating domain labelling

- Classifying web genres: big research area
- Compare general and specialised corpora to identify 'keywords' in each domain
 - ▣ e.g. keywords for Botany: *rhizome, anther, pistil, stamen, integument, stigma, etc*
- Software applies labels to keywords automatically