A Collocation Analysis of 'energy' in Brown Family Corpus

Prihantoro*

Universitas Diponegoro

Abstract. I here argue that Corpus Linguistic (CL) investigations can show evidence that *renewable energy* has become increasingly important in the last 20 years as shown in the Brown Family corpus, a linguistic database of both British and American English whose diachronic data span from 1930s to 2000s. I use *collocation analysis,* a well-known CL technique, to discover collocates (accompanying words) that significantly associate with *energy*. The significance is statistically calculated using Log Likelihood (LL). No content word is found up to 1930s data. Some content words related to the categorization of *energy* are found in 1960s data. In 1990s data *renewable* is within top-3. In 2006 data, *renewable* is found to rank first, showing a very strong significance with *energy*.

1 Why linguistics?

Linguistics investigations are useful approaches to finding answers to some research questions within language learning or other disciplines. For instance, in medical studies, as shown in [1], people who suffer from *aphasia* (a disease caused by damage in a particular area of the brain) may further be categorized on the basis of their speech fluency and accuracy. Speech fluency and accuracy tests are also parts of the diagnostic procedure for patients suspected to suffer from speech delay. These identifications may lead to important decisions on administering medical aids (drugs, therapies, among many others).

In the field of discourse analysis, linguistic findings in numerous studies have proven to be essential. For instance, this study [2] shows that a number of linguistic units, when presented in contexts, help show how *muslim* is characterized (in news published by the British Press from 1998 to 2009) as, for example, a group instead of an individual. While this seems trivial, it has a tremendous effect on society, particularly the consumers of the media who digest the news they publish. An unjustified act of an individual may be labeled as a group act if such framing is strictly followed.

A very simple experiment can also be carried out on COHA (Corpus of Historical American English https://www.english-corpora.org/coha/), a diachronic linguistic database of American English. Finding from queries using words related to *terror* such as *terror, terrors, terrorist, terrorism,* etc shows a dramatic increase in 2000s data, as shown by the figure below (reproduced from [3], We can correspond this to WTC Bombing in 2001, as

* E-mail : prihantoro@live.undip.ac.id, Orcid ID: https://orcid.org/0000-0001-7708-9785, Scopus ID: 57094247500, SINTA ID: 6079217

from that point of time, the United States (US) of America has shown a high level of awareness so that the frequency of these words are significantly high in the news articles posted by the US media during that period.

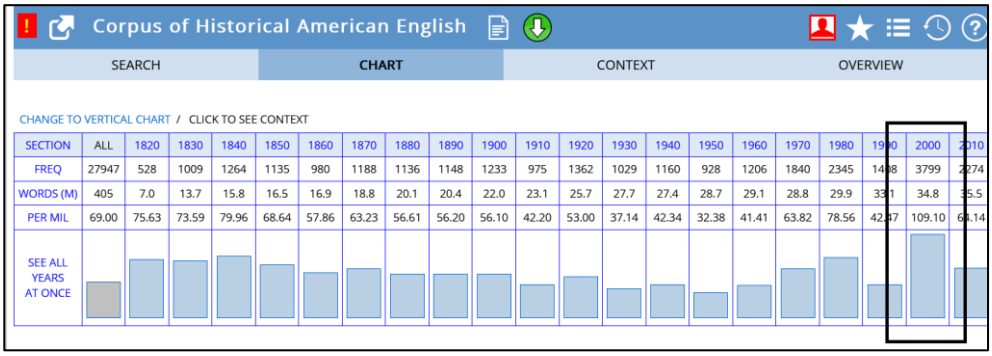

Fig. 1. A dramatic increase of the frequency of words related to *terror* (2000-2010 data) in COHA

In this study, I analyse how certain words associated with *energy* progress over different periods of time. For this purpose, I study Brown Family (BF) corpus, a 6,897,517 words linguistic database whose periods of data collection diachronically span over different periods of time, from 1931 to 2006. BF corpus incorporates linguistic database from a number of different English corpora (American and British) developed using Brown Corpus architecture [4].

Brown Family (extended): *powered by CQPweb*	
Metadata for Brown Family (extended)	
Corpus ID number	**22** (base 36 code: **00000m**)
Date of installation on system	(unrecorded)
Corpus title	Brown Family (extended)
CQPweb's short handles for this corpus	familyx / FAMILYX
Total number of texts in corpus	3,000
Total word tokens in all corpus texts	6,897,517
Word types in the corpus	142,034
Standardised type:token ratio (1,000-token basis)	0.4142 types per token
Non-standardised type:token ratio	0.0206 types per token

Fig. 2. Metadata for Brown Family Corpus CQPweb
(https://cqpweb.lancs.ac.uk/familyx/index.php?ui=corpusMetadata)

To this corpus, I apply *collocation analysis,* a corpus linguistic analytical technique that aims at finding words with statistically strong association to one or more target words. I here adopt the collocation analysis procedure suggested in [5] .

The prominence of collocation analysis in combination with other corpus linguistic techniques and other analyses has been attested in studies across disciplines. For instance, in the study of fictions [6] [7] and [8], health science [9] or news media [10], among many others. These various studies show that collocation analysis may help reveal how a social phenomenon is represented in a collection of texts, i.e. a corpus. Thus, I argue that collocation analysis is one proper technique to analyze my data in the context of my study here.

2 Collocation parameters

The version of Brown Family (BF) Corpus I study here is the one indexed in CQPweb [11], which can be accessed via CQPweb Lancaster https://cqpweb.lancs.ac.uk/familyx/. CQPweb (CQP stands for Corpus Query Program) is a powerful web-based corpus query tool, using which, users can implement a variety of Corpus Linguistics (CL) analytical techniques. As it is a web-based tool, it can be used across platforms without requiring users to install or index corpora from their own local PC even though such functionality is also present in CQPweb Lancaster.

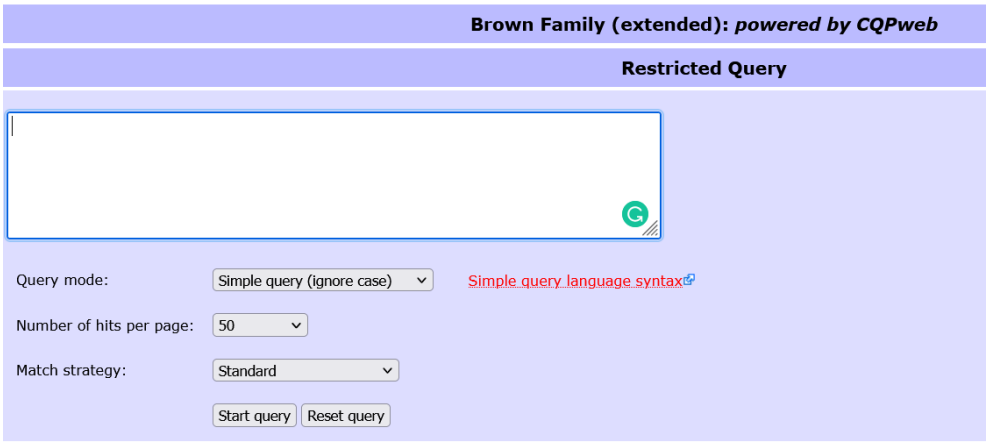

Fig. 3. Brown Family corpus CQPweb query box (https://cqpweb.lancs.ac.uk/familyx/)

The query I supplied to CQPweb is *energy.* The collocation *parameters* [12] can be briefly summarised as follows. The collocation span is 3LR. This means CQPweb will consider 3 words to the left and right sides of *energy,* as the collocate candidates. Next, I set a filter of minimum *observed* collocate of 5 words. This means words which occur less than 5 times within the collocation span will not be considered. This filter is used to exclude words with low frequency. Finally, the statistical measure used here is Log Likelihood (LL), a measure of significance. The higher the LL score, the less the probability of collocates occurring by chance. This procedure is applied to all periods of data collection namely 1931, 1961, 1990, and 2006, using *restricted query* mode. This search mode allows users to limit searches only on one or more particular categories of interests, in this case, time periods.

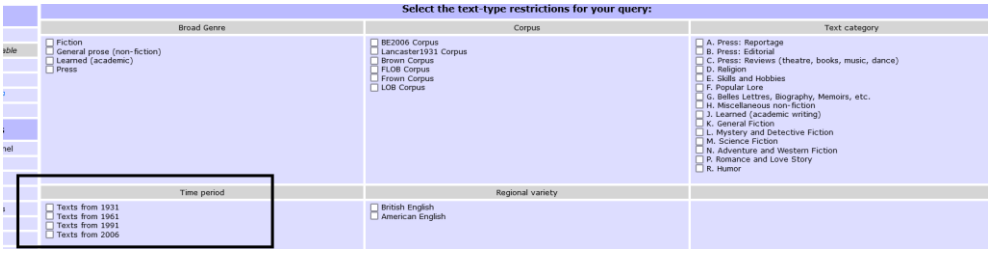

Fig. 2. Brown Family corpus CQPweb in restricted query mode (https://cqpweb.lancs.ac.uk/familyx/index.php?ui=restrict)

3 Renewable Energy is getting prominent!

Using the search procedure described in the previous section, the following findings are obtained, as shown by the table below. We can see that the top-three collocates from 1931 data are: *with, or,* and *of.* These three words are called *function words* or *grammatical words* or *closed class words.* They are words that create grammatical relationship. We here do not see any nouns verbs, adjectives or adverbs, which in linguistics are categorised as *content words.* Unlike function words, the meanings of content words are salient. The distinction between these two kinds of words are further described in [1].

Table 1. Findings from collocation analyses over different time periods in Brown Family corpus CQPweb.

Period	Top three collocates (1st,2nd and 3rd)
1931	*of, is,* and *for*
1961	*atomic, kinetic,* and *potential*
1991	*task, force,* and *renewable.*
2006	*renewable, source,* and *strain*

Let us now consider finding from 1961 data. The table offers us three content words namely *atomic, kinetic,* and *potential.* Grammatically, all these three words are adjectives (content words) which function to describe the noun that they refer to, which is very likely to be *energy. Atomic* and *kinetic* indicates the types of energy. *Potential* is another classification of energy. Consider the concordance lines below which show how the three collocates are used in contexts. While we don't see any use of *renewable* as a strong collocate, we can at least see how the high ranked collocates have shifted from function to content words.

```
it also contains nearly all the mass , and the atomic energy. You may ask what else
Heat of condensation ( work function ) plus kinetic energy of the electrons impinging
replace the momentum p of a free particle ) and the potential energy V(r) being derived
```

Now, let's move on to a more recent period, 1990. Data from this period shows the presence of *renewable* as the third rank collocate based on the LL scores. The first and second most significant collocates are *task* and *force.* How they are used in contexts is shown by the following concordance lines.

```
Senate committees , the Energy Task Force 's jurisdiction was broadened
directed an interim legislative committee , the Energy Task Force , to study
Contrary to Taylor 's assertions , renewable forms of energy now provide for
```

We can see here that a task force called *The Energy Task Force,* is specifically mentioned, therefore, both *task* and *force* are captured as strong collocates. Unlike previous concordance lines where collocates are immediately present, here, the coverage of collocation span is attested as *force* is captured, even though it is not an immediate collocate as shown in the above concordance. Likewise, the collocate *renewable* is also not an immediate collocate, but still can be captured by CQPweb. Let us consider an extended context where the above concordance is presented.

The route towards a greener source of energy is not to promote uneconomic and unsafe nuclear reactors , nor fossil fuels . A comprehensive programme of energy efficiency is needed , together with an increase in the use of renewables . Contrary to Taylor 's assertions , **renewable** forms of **energy** now provide for about 20 per cent of the world 's primary supply ; not just from wind , but also from bio-mass , hydro power and solar energy .

From the extended context, we can learn that the author of the text emphasises on the urgency of a greener source of energy other than nuclear and fossil based energy. This is accurate as many studies have shown concerns over nuclear and fossil based energies. See [13]. Finally, in the most recent data (2006) we can see that *renewable* ranks 1st as the strongest collocate, followed by *source* and *strain.*

barriers are described by the **Renewable Energy** Association (REA)
renewable **source** of **energy** should be undertaken. The Biosciences Federation
As a structure is damaged the internal **strain energy** increases until the

The above concordance lines show collective efforts on renewable energy. We can observe this from the presence of organisations such as *Renewable Energy Association* and *Bioscience Federation* (also Royal Society of Chemistry, but removed from the concordance line due to the word limit) present around the node word. This can be considered as an improvement, as in 1961 data, a concern was shown at task force level. That the concerns have shifted to a larger organisation, in the form of an association, means that renewable energy is getting more prominent over years. We can also see from the extended context below that the organisation specifically concerns about the uptake of biomass for producing heat and electricity.

There are several barriers to increasing the uptake of biomass for producing heat and electricity , despite its apparent superiority over biofuels in terms of potential carbon savings . These barriers are described by the **Renewable Energy** Association (REA) : the significance of biomass in contributing to our carbon abatement targets , our climate change targets and also , increasingly , to the question of fuel security has simply failed to be recognised and given the significance that it probably deserves there is an inflexibility when it comes to biomass in that it does not recognise some of the other benefits in being able to present base load capacity at the end of transmission lines .

4 More ‘renewable energy’ in corpora data?

Using Brown Family (BF) corpus, I managed to show that the association of *energy* to *renewable* has positively progressed over different periods of time. Observing this, we can at least ask one question. How long this trend will last? Only time can answer this question. It really depends on the degree of our concerns about this issue, and to what extent the discussions are going on, then recorded as corpora data.

In US, the contribution of renewable energy sources was measured at 10%, very similar to nuclear, as reported in [14]. The author discussed his view on the future of renewable energies. Other studies, such as [15], [16] or [17] are only a few among numerous studies discussing the urgency of renewable energies.

However, we need to note that those are scientific reports, whose audience are very much specific. Bringing the narrative of renewable energy to public discourse is the required move as 1) it will reach a wider audience, 2) more chances for the data to be incorporated as corpus data, but more importantly 3) bringing more applications so that renewable energy can contribute more.

References

1. V. Fromkin, R. Rodman and N. Hymes, An Introduction to Language (9th Edition), New York: Wadsworth Cengage Learning (2011)
2. P. Baker, C. Gabrielatos and T. McEnery, "Sketching Muslims: A corpus driven analysis of representations around the word 'Muslim'in the British press 1998–2009," *Applied linguistics,* **34(3)**, 255-278 (2013)
3. Prihantoro, Buku referensi pengantar linguistik korpus: lensa digital data bahasa, Semarang: Undip Press (2022)
4. W. Francis and H. Kucera, Brown corpus manual (1972)
5. P. Baker, "Corpus methods in linguistics," in *Research methods in linguistics*, London & New York, Continuum, 93-113 (2010)
6. S. Reichelt dan M. Durham, "Adjective intensification as a means of characterization: Portraying in-group membership and Britishness in Buffy the Vampire Slayer," *Journal of English Linguistics,* **45(1)**, 60-87 (2017)
7. M. Bednarek, ""Get us the hell out of here": Key words and trigrams in fictional television series," *International Journal of Corpus Linguistics,* **17(1)**, 35-63 (2012)
8. M. Mahlberg, P. Stockwell, J. Joode, C. Smith and M. O'Donnell, "CLiC Dickens: Novel uses of concordances for the integration of corpus stylistics and cognitive poetics," *Corpora,* **11(3)**, 433-463 (2016)
9. L. Collins, E. Semino, Z. Demjén, A. Hardie, P. Moseley, A. Woods and B. Alderson-Day, "A linguistic approach to the psychosis continuum:(dis) similarities and (dis) continuities in how clinical and non-clinical voice-hearers talk about the in how clinical and non-clinical voice-hearers talk about their voices.," *Cognitive neuropsychiatry,* **25 (6)**, 447-465 (2020)
10. A. Potts, M. Bednarek and H. Caple, "How can computer-based methods help researchers to investigate news values in large datasets? A corpus linguistic study of the construction of newsworthiness in the reporting on Hurricane Katrina.," *Discourse & Communication,* **9(2)**, 149-172 (2015)
11. A. Hardie, "CQPweb—combining power, flexibility and usability in a corpus analysis tool," *International journal of corpus linguistics,* **17(3)**, 380-409 (2012)
12. V. Brezina, Statistics in corpus linguistics: a practical introduction, Cambridge: Cambridge university press (2018)
13. B. Schlamadinger, M. Apps, F. Bohlin, L. Gustavsson, G. Jungmeier, G. Marland and I. Savolainen, "Towards a standard methodology for greenhouse gas balances of bioenergy systems in comparison with fossil energy systems," *Biomass and bioenergy,* **13(6)**, 359-375 (1997)
14. S. Bull, "Renewable energy today and tomorrow," in *Proceedings of the IEEE,* **89(8)**.
15. I. Dincer, "Renewable energy and sustainable development: a crucial review," *Renewable and sustainable energy reviews,* **4(2)**, 157-175 (2012)
16. R. Gross, M. Leach and A. Bauen, "Progress in renewable energy," *Environment international,* **29(1)**, 105-122 (2003)
17. P. Moriarty and D. Honnery, "What is the global potential for renewable energy?," *Renewable and Sustainable Energy Reviews,* **16(1)**, 244-252 (2012)