Testing Rigidity of MWEs

Radovan Garabík
Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences
Panská 26
Bratislava, Slovakia
garabik@kassiopeia.juls.savba.sk

Relevant UniDive working groups: WG2, WG4

1 Introduction

In this paper, we present a simple web-based tool/application for exploring the "rigidity" of Multi-Word Expressions (MWEs) in large text corpora. The tool has been internally used at the authors' institution for several years to help research Slovak collocations, and now we expanded it to include corpora from the ARANEA family of comparable web corpora (Benko, 2014).

The application is aimed at researchers familiar with text corpora (e.g. via using the (No)Sketch Engine (Kilgarriff et al., 2014)) able to interpret the results and their linguistic meaning; although the basic usage does not require any familiarity with CQL and the visualizations are intended to be directly interpretable. The tool uses the SketchEngine API[1] to access the corpus data and is therefore dependent on a (No)Sketch Engine installation.

We currently cover following languages: Bulgarian, Czech, Dutch, English, Estonian, Finnish, French, Georgian, Hungarian, Italian, Latin, Latvian, Persian, Polish, Romanian, Russian, Slovak (not just the ARANEA corpus, but also the *prim-10* representative corpus[2] and other specialized corpora), Spanish, Swedish, Ukrainian, and Uzbek; although the size and quality of the corpora varies (for our purposes, the accuracy of the lemmatization is rather relevant, although the corpus does not need to be lemmatized, if we are content to query only the inflected word forms).

The tool explores the distribution of distances between two words, typically either a two-word expression or two words from a MWE. The distance between the words is calculated as the difference in their positions within the corpus. Given two words, the tool accesses the corpus to find the dependency between them and plots the number of occurrences of the distance for each occurrence of the first word. This allows the users to explore the dependency of the second word on the first word based on their distance. As such, this is purely a statistic based approach and the results reflect a mixture of idiomaticity, valency and syntax and are best interpreted in conjunction with direct corpus queries.

If the second word is independent of the first word, the number of occurrences of the second word is expected to be constant, and deviations from this expectation are a good indicator of some correlation between the two words, usually being a part of a MWE, and the visualization of the dependency on the occurrences on the distance helps us quickly see the (perhaps unexpectedly interesting) structure of the MWE.

2 Calculations

In the following, we denote these words as w_1 and w_2, and the number of their occurrences in the corpus as n_1 and n_2.

We define the distance between those two words w_1, w_2 in the corpus as the difference between their positions: $d(w_1, w_2) = pos(w_2) - pos(w_1)$; thus the distance of the word to itself is zero, the distance to the following word is 1, to the previous one it is -1.

We take the w_1 as the principal word and look at its occurrences in the corpus, and for each of the distances $d(w_1, w_2)$ (within some reasonably small context), we plot a number of the occurrences of the distance. In other words, we explore the dependency of w_2 on w_1 as reflected by their distance.

For each w_1 concordance, we can model the occurrences of w_2 at a specific distance as a Poisson process which allows us to estimate the confidence interval of the number of occurrences.

If w_2 is completely independent from w_1, we expect the number of occurrences of w_2 within our context window to be roughly constant (independent on $d(w_1, w_2)$ with the (at any position relative to w_1) expected number $c_2 = n_1 \cdot p_2$, where p_2 is

[1] https://www.sketchengine.eu/documentation/api-documentation/

[2] https://korpus.sk/korpusy-a-databazy/korpusy-snk/prim-10-0/

the probability of the occurrence of w_2, which we assume to be $p_2 = \frac{n_2}{N}$, where N is the corpus size.

As for the dependency of the number of occurrences of w_2 in a slot with a given distance d on the distance, we present only a very loose and not rigorous argument. Since two related words depend on each other, and in a natural language we expect locality, it is not too incorrect to assume the w_2 distance from w_1 might follow an exponential distribution (probability distribution of the distance to the second word), and the sum of exponential random variables is the Erlang distribution, which generalizes to the Gamma distribution. We therefore fit the $[d, c_2]$ pairs with the Gamma distribution, which seems to work reasonably well – an idealization where the words are bound and the distance between them is subject to independent random fluctuations caused e.g. by grammar or context. And a deviation from this hints at the inner structure of our MWE.

As a convenience for the users, we plot also the reverse collocation (w_2, w_1) in the left half of the graph (negative values of the distance d), this way it is be apparent if there is some phenomenon that warrants looking into the reverse word order.

2.1 Examples and Output Description

We have noticed that the results are more interesting for synthetic languages – it appears that in an analytic language, the function words "get in the way" and "smear" the distribution (although this is only our unfounded speculation).

As an example we use the query *pull string*[3] (as lemmas) in the English-language corpus *Araneum Anglicum II Maximum*. First we get some basic statistics about the collocation (Output 1). The frequencies of both words in the corpus are self-explanatory; the frequency of tight collocation is the number of occurrences of the exact collocation *pull string* without any intervening words; the mean frequency of *string* is the average frequency of the second word calculated in the right context of the concordance; and the *frequency assuming they are independent* is the expected frequency of the second word in the right context, if it were distributed absolutely randomly according to the Poisson distribution.

The graphical output has been designed to provide most of the information at a glance, using various colors to encode the information. The x-axis represents the distance of the second word from the first one, and the y-axis represents the number of occurrences of the word.

The purple dots represent the absolute frequencies of the occurrences of w_2 at a given distance x from w_1. The error bars mark the 95% confidence interval assuming a Poisson process for the occurrence of the words. The light green curve is the Gamma distribution fit for the frequencies of w_2 at a distance $x > 0$ from w_1, extended to be equal to c_2 for distances $x < 0$. The light blue curve is the Gamma distribution fit for the negative distances (word w_2 being to the left of w_1), also extended to c_2 for distances $x > 0$. Thus, the combined green+blue horizontal line marks the baseline frequency of w_2 if it were independent from w_1. The position of the purple dots above that line shows how much the second word is correlated with the first one at the given distance. The light green and blue curves show the "ideal" distribution according to our model. The two vertical yellow lines mark the arithmetic mean of the distance of the second word from the first one (within our context window). Therefore, they can be used, for example, to compare the "span" or "width" of various MWEs.

We can immediately observe (looking at the right half of the graph at Figure 1) that the independent frequency is significantly below the real occurence of the second word (purple dots) with Output 1 providing the numbers – the independent frequency would be 37.8 and the actual mean frequency in our sample (right context up to 6 tokens) which is 1319. This indicates that these two words are (as expected) strongly correlated. The vertical error bars provide an immediate indication of the statistical reliability, which in this case, due to the large size of the corpus and the relatively frequent nature of the words, is reasonably solid.

The position at $x = 1$ corresponds to the form *pull strings*; the position at $x = 2$ to the phrase *pull the string(s)* and variants (*pull our strings ...*); further at $x = 3$ the most frequent phrases are *pulling all the strings*, *pull a few strings*, *pull on the strings* etc.; and then ever decreasing amount of longer and longer variants (*pull on/at the/your/my heart strings, pull a lot of strings, pull all the right strings* being the most frequent ones).

[3] This is actually a conflation of two frequent forms of this MWE, *pull strings* (no article, plural) and *pull the string(s)* (definite article, singular or plural). We could, of course, query the word forms, to make this explicit.

```
corpus AranAngl_a has 11373661010 tokens
frequency of pull:  901297; ipm=79.0
frequency of string:  475211; ipm=42.0
frequency of tight collocation pull string:
1028; ipm=0.09
mean frequency of string in our sample (right
context of pull):  1319; ipm=0.12
frequency of string in collocation with pull,
assuming they are independent:  37.854±0.186;
ipm=0.0033
```

Output 1: Basic statistics about the query *pull string* in the English language corpus.

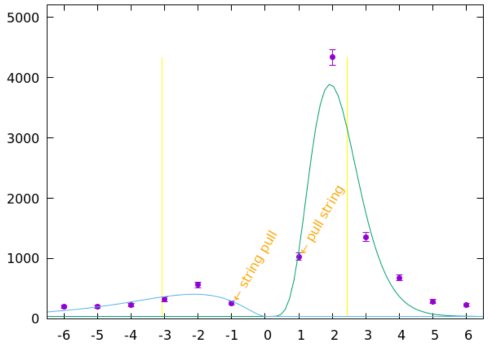

Figure 1: Example of the query *pull string* in the English language corpus.

The second example is the query in the Polish corpus *Araneum Polonicum Maius).* We query the lemma *nauczyciel* (teacher) in collocation with the word *angielskiego* (English-GEN.SG), 'teacher of (the) English (language)'. Since we are interested only in genitive singular of the second word, we will write the second term as `[lemma="angielski" & tag=".*:sg:gen:.*"]`.[4]

The collocation *nauczyciel angielskiego* is significantly above the baseline, the words are strongly correlated. However, the maximum is at the distance $x = 2$, i.e. there is one word between these MWE consituents – predominantly *nauczyciel języka angielskiego* (teacher language-GEN English-ADJ.GEN, 'teacher of (the) English language'), but there are some occurrences of *nauczyciel j. angielskiego* (*j.* as an abbreviation of *języka*) and *nauczyciel od angielskiego* (teacher from English-GEN). Although the size of the concordance is too small to get a true statistical significance, there is a discernible secondary maximum at the distance $x = 4$ corresponding to the related MWE *nauczyciel i lektor angielskiego* (teacher and instructor English-GEN, 'teacher and instructor of English').

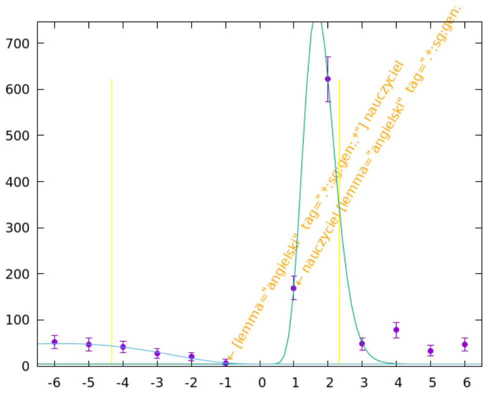

Figure 2: Example of the query *nauczyciel* `[lemma="angielski" & tag=".*:sg:gen:.*"]` in the Polish language corpus.

3 Availability

The application is publicly accessible at `https://www.juls.savba.sk/kolokat_en.html`. The web interface may be rudimentary, but it is fully functional for internal use. The users can query the application using either two word forms or lemmas, or by employing use full CQL syntax, which enables them to search for parts of speech, grammatical categories or even syntactic relations, although the only syntactically annotated (Nivre et al., 2020) corpus connected at the time is the Slovak Corpus of Court Decisions and the Slovak Corpus of Legal Texts (Garabík, 2022).

References

Vladimír Benko. 2014. Aranea: Yet Another Family of (Comparable) Web Corpora. In *Text, Speech and Dialogue: 17th International Conference*, pages 247–254, Brno. Springer.

Radovan Garabík. 2022. Corpus of Slovak legislative documents. *Jazykovedný časopis*, 73(2):175–189.

Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. The Sketch Engine: ten years on. *Lexicography*, 1:7–36.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajic, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

[4] The corpus is MSD-tagged with the IPI PAN tagset: `http://nkjp.pl/poliqarp/help/ense2.html`