Contrastive Collocation Analysis – a Comparison of Association Measures across Three Different Languages Using Dependency-Parsed Corpora

Stefan Evert[1], Thomas Proisl[1], Peter Uhrig[1], Maria Khokhlova[2]
[1] University of Erlangen-Nuremberg
[2] St. Petersburg State University
stefan.evert@fau.de, thomas.proisl@fau.de, peter.uhrig@fau.de, m.khokhlova@spbu.ru

Our analysis focusses on association measures for noun-adjective combinations of dependency-related co-occurrences. In the study we limit ourselves to binary collocations, i.e. pairs of two words, for English, German, and Russian. All three languages belong to the same language family (Indo-European); English and German share a longer common ancestry, being both Germanic languages. Typically different levels of attention are paid to them (most often English data tend to be more analyzed) The languages also differ both in their syntactic and their morphological nature. The Russian language is often quoted as a highly inflecting language demonstrating both synthetic and analytical features. Thus to the best of our knowledge the present work is the first study of this kind.

In our study we would like to answer the following question: what differences do we find between the languages concerning noun-adjective collocations extracted from syntactic dependencies? For this task we evaluate lists of relational collocation candidates extracted with the help of association measures and analyze whether the association measures perform the same across the different languages with respect to precision and recall. We also calculate the statistical correlations between the measures to investigate whether they are the same in the different languages.

From a historical point of view, German and English are more similar to each other than each of them is to Russian, so one could expect a similar behaviour when it comes to their collocational patterning. On the other hand, German and Russian are typologically somewhat similar in that they both have a (more or less) elaborate case system and tend to form compounds as single orthographic words (for Russian this is only true to a certain degree). We would thus expect to see bigger differences between German/Russian on the one hand and English on the other hand with regard to noun-adjective collocations.

We aim to use comparable gold standards that include examples extracted from lexicographic works for all three languages. The information about collocations can be presented in different resources (for examples, explanatory or specialized dictionaries, thesauri, wordnets, databases etc.). For example, we use Oxford Collocations Dictionary for Students of English (OCD2) as a gold standard for English. We confine ourselves to noun-adjective collocations as this type of phrases is well-defined in various lexicographic resources. We decided to use large web corpora that comprise billions of tokens and provide a high coverage of the gold standards, in particular DECOW16A, ENCOW16A (Schäfer & Bildhauer, 2012; Schäfer, 2015) and Araneum Russicum II Maximum (Benko, Zakharov, 2016). The corpora were annotated with state-of-the-art dependency parsers that enabled us to extract specific syntactic relations (such as verb+subject and verb+object). Following the approach of Evert & Krenn (2005), we rank the candidates using a wide range of statistical association measures (those evaluated by (Evert et al., 2017) on a smaller English gold standard) and evaluate collocation identification quality in terms of precision-recall graphs; average precision up to 50% recall (AP50) is used

to make quantitative comparisons. We evaluate both a global ranking of all candidates as in (Bartsch & Evert, 2014) and a separate ranking for each node as in (Uhrig & Proisl, 2012).

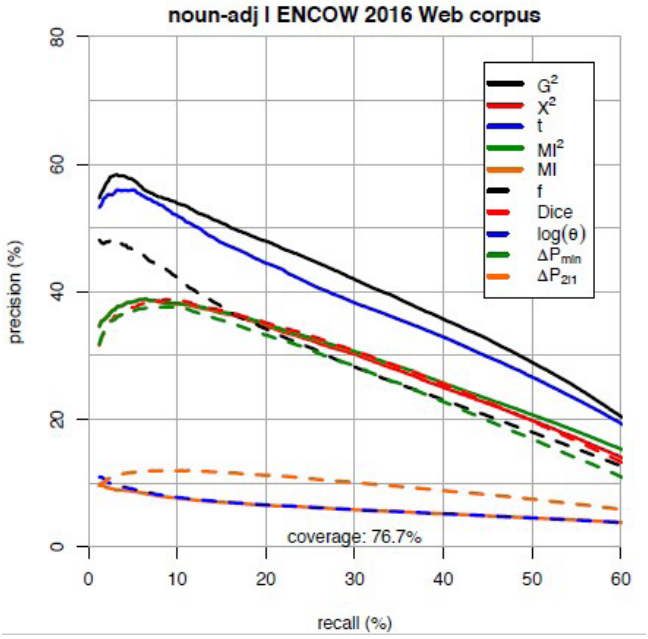

Fig. 1: Precision-recall graphs based on global ranking of all candidates for English verb/object collocations

The experiments on English noun-adjective collocations showed the following results (figure 1). We made a global ranking of all candidates with frequency threshold $f \geq 5$ comparing them to the gold standard from OCD2. The best overall measure is log-likelihood that outperforms the second best measure t-score. The coverage of 76.7% shows that the majority of all noun-adjective collocations from the OCD2 gold standard occur at least 5 times in the large Web corpus. The further experiments suggest that for English collocations log-likelihood proves to be the best measure for all types except adjective-verb collocations. For adjective-verb collocations MI4 gives better results, while subject-verb collocations appear to be particularly hard to identify. Our earlier experiments on Russian data have shown that t-score and Dice proved to extract the largest number of collocations that overlap with the data found in dictionaries (Khokhlova, 2017).

Keywords: collocation, evaluation, contrastive linguistics

References

OCD2 = *Oxford Collocations Dictionary for Students of English*, 2nd edition (2009). Edited by Colin MacIntosh. Oxford: Oxford University Press.

Bartsch, S., Evert, S. (2014). "Towards a Firthian notion of collocation." *OPAL – Online publizierte Arbeiten zur Linguistik* 2/2014: 48–61. Accessed at: http://pub.ids-mannheim.de/laufend/opal/pdf/opal2014-2.pdf [30/03/3018].

Benko, V., Zakharov, V. (2016). "Very Large Russian Corpora: New Opportunities and New Challenges". *Komp-juternaja lingvistika i intellektual'nyje technologii: Po materialam mezdunarodnoj konferencii «Dialog», 15 (22)*. Moskva: Rossijskij gosudarstvennyj gumanitarnyj universitet, pp. 79-93. Accessed at: http://www.dialog-21.ru/media/3383/benkovzakharovvp.pdf [30/03/3018].

Evert, S., Krenn, B.. (2005). "Using small random samples for the manual evaluation of statistical association measures". *Computer Speech and Language*, 19(4), pp. 45–466. https://doi.org/10.1016/j.csl.2005.02.005

Evert, S., Uhrig P., Bartsch S., Proisl, T. (2017). "E-VIEW-alation – a large-scale evaluation study of association measures for collocation identification." In: *Proceedings of eLex 2017 – Electronic lexicography in the 21st century: Lexicography from Scratch*. Leiden: Lexical Computing, pp. 531–549. Accessed at: https://elex.link/elex2017/wp-content/uploads/2017/09/paper32.pdf [30/03/3018].

Khokhlova, M. (2017). On The Differences between Association Measures for Automatic Collocation Extraction: Evaluation against Dictionaries. In *SGEM International Multidisciplinary Scientific Conference on Social Sciences and Arts 2017*. Sofia. V. 2, pp. 887–892.

Schäfer, R. (2015). Processing and Querying Large Web Corpora with the COW14 Architecture. In: Bański, Piotr, Hanno Biber, Evelyn Breiteneder, Marc Kupietz, Harald Lüngen, Andreas Witt (eds.) *Proceedings of the 3rd Workshop on Challenges in the Management of Large Corpora (CMLC-3)*, Mannheim: IDS Publication Server, 28–34. Accessed at: https://ids-pub.bsz-bw.de/files/3826/Schaefer_Processing_and_querying_large_web_corpora_2015.pdf [30/03/3018].

Schäfer, R., Bildhauer, F. (2012). "Building Large Corpora from the Web Using a New Efficient Tool Chain." In: Calzolari, N., Choukri, K., Declerck, T., Uğur Doğan, M., Maegaard B., Mariani J., Moreno A., Odijk, J., Piperidis, S. (eds.). (2012). *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul: European Language Resources Association, 486–493. Accessed at: http://www.lrec-conf.org/proceedings/lrec2012/pdf/834_Paper.pdf [30/03/3018].

Uhrig, P., Proisl, T. (2012). "Less hay, more needles – using dependency-annotated corpora to provide lexicographers with more accurate lists of collocation candidates." *Lexicographica* 28, pp. 141–180. https://doi.org/10.1515/lexi.2012-0009.