If an atom is a letter, then a molecule is a word: applying corpus linguistic methods to chemistry

Maciej Eder, Michał Woźniak, Urszula Modrzyk, Rafal L. Górski (Institute of Polish Language, Poland) and Bartosz Grzybowski (Ulsan National Institute of Science and Technology, Ulsan, South Korea)

The buzzword quoted in the title of this paper, even if popular in chemistry, sounds very naïve for anyone who has some expertise in linguistics. Nonetheless, despite a rather shallow similarity between linguistics and organic chemistry on theoretical level, one cannot deny the usefulness of (some of) corpus linguistics methods to analyse internal molecule structures. This potential applicability makes such investigations feasible and desirable. The present paper demonstrates a procedure of identifying word forms – originally developed to assess Chinese script with no clear word boundaries – to divide complex chemical molecules into "meaningful" substructures. In this context, "meaningful" means groups of atoms that are local centres of reactions.

Let us assume that a molecule is a sentence (with some obvious caveats in mind, non-linearity of molecules being the most important one). If so, then a list of known molecules can be treated as a corpus. Quite striking is the fact that a commonly used convention of describing chemical structures (referred to as SMILES) uses sequences of characters, what makes any comparisons to corpora even more feasible. E.g., caffeine is coded as follows: CN1C=NC2=C1C(=O)N(C(=O)N2C)C.
Being one of the most crucial issues in organic chemistry, the question why certain groups atoms tend to keep together, while repelling others, has been approached using different methods, which are aimed at finding repetitive fragments of molecules. It can be assumed that methods derived from text mining can be adopted to (partially) solve the task.

If we name the "meaningful" groups of atoms as "words", we need a device for finding them in our corpus since there are no explicit word boundaries. Grzybowski (2015) compared (in pairs) thousands of molecules, in order to extract their maximum common substructures, with the belief that they represent chemical "words"; this step was followed by a term frequency–inverse document frequency (tf/idf) heuristic. The elements which are obtained by this procedure behave as words in respect to Zipf's and Heaps' laws. Still, the linguistic motivation of this approach is rather weak, moreover it is computationally complex. Therefore we adopt a method for establishing word boundaries in Chinese, as proposed by Maosong et al. (1998).

The method uses the concept of sliding window, which divides a string of characters into chunks, and then computes an association measures inside the window as it moves. The association measure is a combination of two classical indexes: Mutual Information and t-tests.

Our preliminary tests, performed on a rather small corpus of 100,000 chemical

molecules, show that one can identify a considerably stable set of repetitive molecule parts that have well-defined "word-boundaries". Certainly, the obtained results cannot be validated directly (there is no a priori definition of a chemical word). However, one can verify the results indirectly, e.g. via linguistic characteristics revealed by our "words". In Fig. 1, an example of such a measure is shown, i.e. the classical frequency/rank dependence, followed flawlessly by our "corpus".

References

Cadeddu, A., Wylie, E.K., Jurczak, J., Wampler-Doty, M. and Grzybowski, B.A. (2014). Organic chemistry as a language and the implications of chemical linguistics for structural and retrosynthetic analyses. *Angewandte Chemie*, 126(31): 8246- 8250.

Maosong, S., Dayang, S. and Tsou, B.K. (1998). Chinese word segmentation without using lexicon and hand-crafted training data. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, Vol. 2 (pp. 1265-1271).