Chemical Linguistics Hot Paper

DOI: 10.1002/anie.201403708

Organic Chemistry as a Language and the Implications of Chemical Linguistics for Structural and Retrosynthetic Analyses**

*Andrea Cadeddu, Elizabeth K. Wylie, Janusz Jurczak, Matthew Wampler-Doty, and Bartosz A. Grzybowski**

Dedicated to Professor George M. Whitesides on the occasion of his 75th birthday

Abstract: *Methods of computational linguistics are used to demonstrate that a natural language such as English and organic chemistry have the same structure in terms of the frequency of, respectively, text fragments and molecular fragments. This quantitative correspondence suggests that it is possible to extend the methods of computational corpus linguistics to the analysis of organic molecules. It is shown that within organic molecules bonds that have highest information content are the ones that 1) define repeat/symmetry subunits and 2) in asymmetric molecules, define the loci of potential retrosynthetic disconnections. Linguistics-based analysis appears well-suited to the analysis of complex structural and reactivity patterns within organic molecules.*

Informally, there is a clear analogy between a chemist's understanding of a compound and an English speaker's understanding of a sentence. In either case, the trained mind effortlessly identifies part-whole relationships, often despite a lack of demarcations. The intuitive analogy between the spoken language and the language of chemistry has been suggested in Jean-Marie Lehn's famous quote stating that "Atoms are letters, molecules are the words, supramolecular entities are the sentences and the chapters".[1] No matter how imaginative such analogies are, however, they lack a formal verification; that is, they become valid only when the structures of the language and of chemistry are both quantified and then compared with one another. With natural languages, such structural analyses have been largely enabled by the development of modern computers and today, natural language processing[2] is perhaps the most vibrant subfield of linguistics research. NLP has already been proven capable of useful applications, such as internet search engines, automatic translators, automatic summarization, natural language generation and understanding, parsing, segmentation, and information retrieval and extraction. In chemistry, however, the linguistic terms of a corpus, formal grammar, collocations, or n-grams, are largely unknown and not even considered as relevant to the practice of our discipline. As we show herein, this is a rather premature omission. First, we show that organic molecules contain fragments whose rank distribution is essentially identical to that of sentence fragments in English; that is, the "dictionaries" of organic chemistry and of English follow very similar laws. What is chemically surprising is that the "words" of chemistry are not necessarily the functional groups we are accustomed to (methyl, ethyl, hydroxy, etc.) but which do not follow rank distribution of English words; instead, the language-like "words" of chemistry are the common sub-fragments of diverse sets of molecules. Second, extending the concepts used in text search engines to organic chemistry, we show how the dictionary composed of these sub-fragments/"words" can be used to identify in organic molecules high-information-content bonds defining molecular symmetry and/or locations most suitable for retrosynthetic cuts. While we have a clear understanding that our current work is only an opening chapter, we suggest that application of computational linguistics to chemistry can have immense impact on the use of computers to analyze organic molecules and reactions to find structural and reactivity regularities that are hidden in the big-data ouvre of chemistry.

As early as 1935, George Zipf noted[3] that in a corpus (that is, large and structured set of texts) of a natural language, the frequencies of words are inversely proportional to their ranks in the frequency table; that is, the most frequent word occurs twice as often as the second most frequent one, three times more often as the third most frequent word, etc. This regularity is nowadays known as the Zipf's law[4] and its basic premise (that is, that language can be described and analyzed by statistical measures), is one of the cornerstones of modern corpus linguistics.[5,6] One of the algorithms popular in quantifying the structure of natural language corpora is as follows.[7] A text comprising n sentences is selected for analysis. All possible pairs of sentences within this corpus are

[*] Dr. A. Cadeddu,[+] E. K. Wylie,[+] M. Wampler-Doty, Prof. Dr. B. A. Grzybowski
Department of Chemical and Biological Engineering
Department of Chemistry, Northwestern University
2145 Sheridan Rd., Evanston, IL 60208 (USA)
E-mail: grzybor@northwestern.edu
Homepage: http://dysa.northwestern.edu
Prof. Dr. J. Jurczak
Institute of Organic Chemistry, Polish Academy of Sciences
ul. Kasprzaka 44/52, Warsaw (Poland)
[+] These authors contributed equally to this work.

[**] This work was supported by the Non-equilibrium Energy Research Center, which is an Energy Frontier Research Center funded by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences under Award Number DE-SC0000989. We thank Prof. Maciej Eder and Prof. Rafal Gorski from the Institute of the Polish Language, Polish Academy of Sciences, for helpful comments.

Supporting information for this article is available on the WWW under http://dx.doi.org/10.1002/anie.201403708.

then analyzed to find maximum common substrings (MCS); for instance, if one sentence is "George Whitesides is a famous chemist" and another is "Versailles is a famous palace," then the MCS of these two sentences is "is a famous". All such maximum common substrings identified are saved in a "dictionary". The numbers of times each MCS is repeated in the dictionary are calculated and the MCS values are then ranked according to the frequency of their occurrence and are plotted in the form of rank versus occurrence distribution. For instance, blue cross markers in Figure 1a correspond to the distribution of MCS derived from the analysis of 100 sentences chosen at random from English Wikipedia. Perhaps not surprisingly, the most common fragment of the sentences is "e", followed by "a" and "o". In corpus linguistics, it is customary to normalize such distributions and plot them on doubly logarithmic scales to yield characteristic truncated power laws such as those given by the solid markers in Figure 1b; here, different distributions correspond to different numbers of sentences in the texts analyzed, $n = 30, 60, 100, 200, 400, 1000, 2000$.

We now apply similar concepts to the analysis of organic chemistry. Our corpus here is a collection of molecules, and each molecule corresponds to a "sentence"; the question is whether there are "words" or fragments in this chemical corpus that would make it akin to English. An approach that probably every chemist would consider first is to look at functional groups (OH, NH_2, COOH, COOR, etc.) as potential "words" of chemistry. This, however is a poor choice; when molecules composing the corpus are divided into functional groups (314 of them, listed in Ref. [8]), and when these groups are ranked according to their frequencies of occurrence, a distribution is obtained that deviates strongly from that of English (red triangular markers in Figure 1a). Emphatically, the functional groups are not the building blocks of a chemical language. Upon reflection, this is perhaps not that surprising, because when we inspect organic molecules we recognize in them patterns that are often more complex than individual groups. Indeed, we often categorize molecules by larger common repeat patterns, as in the common substrings of English sentences. Therefore, building on this analogy, we consider analysis in which for a set of n molecules we inspect all molecule–molecule pairs and extract from each pair the largest common structural fragment (see Figure 1c for an example). We then count the repetitions of identical fragments, rank them according to the frequency of their occurrence, and plot the frequency distribution. The blue circle markers in Figure 1a demonstrate that the distribution of such common molecular fragments follows closely the distribution of maximal common substrings in sentences. Moreover, as shown in Figure 1b, this correspondence holds for different sizes of the diverse chemical and English corpora (that is, n molecules vs. n sentences, where n values shown are 30, 60, 100, 200, 400, 1000, 2000), and the distributions do not change when different sets of molecules are analyzed (Supporting Information, Section S1). Within these distributions, we discover thousands of fragments corresponding to functional groups but also those that are not intuitive at all (Figure 1e). Together, these are the common "utterances" from which

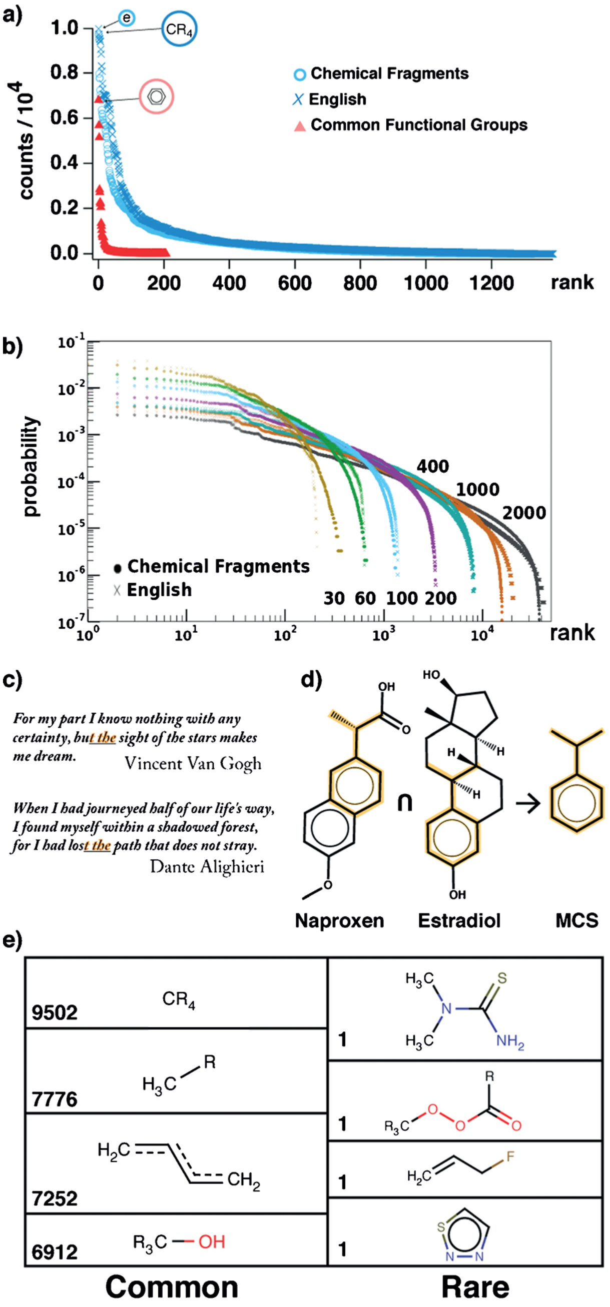

Figure 1. a) Dark blue crosses have the rank-frequency distribution of maximum common substrings in English sentences (here, based on the analysis of $n = 100$ sentences). Distribution of common molecular sub-fragments follows a similar dependence (light-blue circles). In contrast, rank-distribution of common functional groups (red triangles) is significantly different. Note: rank = 1 means the most frequent substring/fragment/group; rank = 2 means second most frequent substring/fragment/group, and so on. Frequency is the normalized frequency of occurrence of substrings/fragments/groups in a given set of sentences or molecules analyzed. b) The probability versus rank distributions plotted on double-logarithmic scales and derived from corpora of $n = 30, 60, 100, 200, 400, 1000, 2000$ sentences or the same number of molecules. Within each pair of nearly overlapping curves, one (×) is for English and one (●) is for chemistry. c) Example of a maximal substring (highlighted) of two English sentences and d) a maximal subfragment of two organic molecules. e) Some most common and less common subfragments constituting the dictionary of organic chemistry.

chemical "language" is made of. Interestingly, the 40000-odd vocabulary of chemistry derived from the analyses of 2000-molecule data sets is similar in size to the vocabulary of a typical English speaker.

We have just shown that there exists a set of molecular fragments with which organic molecules can be described akin to a language. This finding suggests that it should be possible to apply the tools of computational linguistics to the analysis of molecules. Of particular interest to us are the linguistic methods that examine the information content of text documents and could, potentially, be extended to identify the information-rich bonds of organic molecules; as we will see shortly, these information-rich bonds define molecular symmetry and, in general, are also likely candidates for making retrosynthetic disconnections.

The linguistic approach we consider is based on the so-called tf-idf scoring (term frequency–inverse document frequency) which is a numerical statistic used by computer search engines and other information retrieval schemes to determine how relevant certain words are to a given text. The tf part of the score is simply the frequency of how often a given word occurs in a document of interested; tf = word frequency in a document. This measure, however, is rather misleading because some words (for example, "the", "this") generally occur in a language/corpus more often than others. To correct for this bias, tf is multiplied by an idf, which expresses the frequency with which a given word occurs in the entire corpus: it is typically given as a logarithm, idf = −log (documents in which a word is found/total number of documents). Note that if a word like "the" occurs in all documents under study, idf = log(1) = 0 and this word is not "characteristic" of/relevant to any given document. In general, tf-idf scores are highest for the most characteristic words in documents (Figure 2); another way of putting it is that the tf-idf scores identify the most information-rich content of documents.

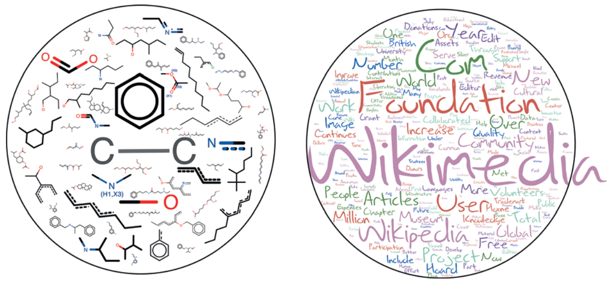

Figure 2. In the so-called "word cloud", the word's size is proportional to its tf-idf score and expresses this word's relative relevance in various documents (see main text). The left panel shows the "cloud" of some of the organic chemistry "keywords" (that is, structural motifs) we identified. The right panel has the word cloud for the Wikimedia Foundation's 2009–2010 annual report.

We extended the above analysis to organic molecules in which we sought the most information-rich bonds. Our hypothesis has been that these bonds would be chemically most meaningful in, for instance, defining molecular symmetry and also defining loci of potential retrosynthetic disconnections. Our algorithm (with the scheme in Figure 3a and the source code included in the Supporting Information) used as dictionary the set of N_{dict} largest common structural fragments discussed earlier (see Figure 1c). First, for any molecule of interest, it was established which structural fragments from the dictionary this molecule contained; we denote the total number of fragments identified in a molecule as F_{mol}. Next, the term frequency was calculated by dividing the number of times a given fragment of type i was found in the molecule by the total number of fragments in this molecule, tf = F_i/F_{mol}. For example, if a molecule contains, say, $F_{mol} = 20$ different fragments from the dictionary and two of these fragments are phenyls ($F_i = 2$), then the term frequency of phenyls in this molecule is 2/20. Then, an idf score was calculated based on the ratio of the number of molecules M containing a given fragment divided by the total number of molecules in the dictionary; idf = $-\log(M_i/M_{dict})$. As in the case of linguistics, the tf·idf product weighted the relevance of fragments from our chemical dictionary to the molecule under analysis; at the same time, it did not yet provide information at the level of individual bonds within the molecule. To achieve this, we considered each bond in the molecule and established to which m structural fragments it belonged (note: since dictionary fragments are not necessarily "orthogonal," each bond can belong to more than one fragment). We then summed up and averaged the tf-idf scores of all $j = 1,..,m$ fragments containing the particular bond:

$$\text{bond score} = \left(\sum_{j=1}^{m} \text{tf}(j)\cdot\text{idf}(j)\right)/m$$

All bonds in the molecule were assigned such scores (Figure 3a).

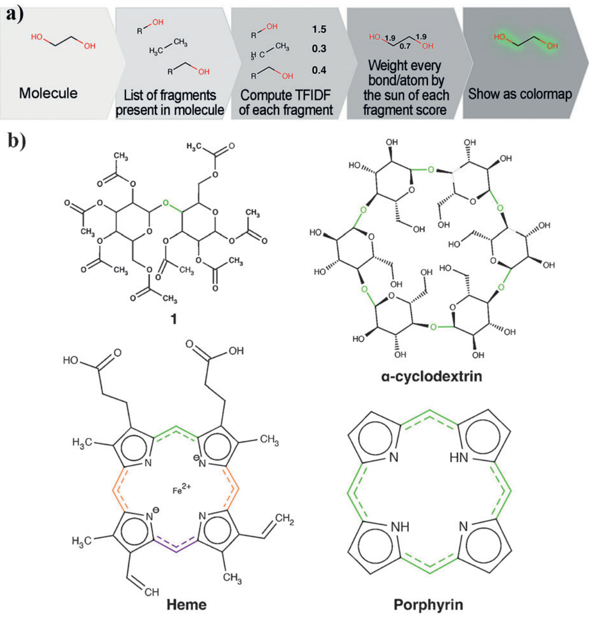
a)
Molecule
List of fragments present in molecule
1.5
0.3
0.4
Compute TFIDF of each fragment
1.9
0.7
1.9
Weight every bond/atom by the sun of each fragment score
Show as colormap
b)
1
α-cyclodextrin
Heme
Porphyrin

Figure 3. a) Scheme illustrating the process of tf-idf bond scoring. b) Four examples highlighting the bonds with highest scores; note that these bonds delineate repeat/symmetry units of the molecules. All analyses were based on a disctionary of 8305 fragments.

Figure 3b shows examples of molecules in which the highest scoring bonds are colored in green; as we see by the examples of dimer **1**, the α-cyclodextrin, and the porphyrin, the linguistic approach identifies the symmetry/repeat units of all these molecules. We emphasize that this is not a small feat given we have not even considered any (x,y,z) coordinates of the atoms making up these molecules and performed no linear-algebra analyses to find symmetries which, incidentally, can be a computationally intensive procedure involving manipulation of matrices. Another point we wish to make is that by matching into molecules entire chemical "words"/patterns (typically spanning several atoms), the tf-idf scoring is capable of distinguishing bonds based on their environments. This is illustrated by the heme example in Figure 3, whereby the "locally" equivalent bonds joining the imidazole rings are given different scores based on the presence of different side chains; the highest-information-content bonds are in green, followed by the violet-colored ones and the orange ones.

This last example also suggests potential uses of linguistic analyses in the context of computer-assisted retrosynthesis. The question we ask here is whether the most characteristic/informative bonds with highest tf-idf scores are the likely candidates for retrosynthetic disconnections in molecules of practically relevant complexity. To explore this possibility, we analyzed 68 molecules taken from a natural product database (CRC Dictionary of Natural Products, http://dnp.chemnetbase.com/tour/). All bonds (between non-H atoms) in these molecules were automatically scored according to the tf-idf scores. Figure 4 shows examples of four such molecules in which the top-three ranking bonds are colored, respectively, light green, dark green, and yellow, whereas the two lowest scoring bonds are colored brown and red. As described in the Supporting Information, Section S2, these and other 64 molecules were also inspected by ten Ph.D.-level organic chemists who, independently, provided their "human" suggestions as to the potential disconnection sites. In about 97% of cases, at least one chemist suggested one of the top-three computer-chosen bonds. A more stringent comparison, however, is against a hypothetical scenario in which computer would be choosing bonds within a molecule at random. In this case, the tf-idf algorithm performed better in 75% of cases, though this number is artificially low because of overabundance of phenyl groups in the dictionary (so that the algorithm overestimates their tf scores and suggests cuts of aromatic rings), and also in the case of molecules comprising less than 18 or 19 bonds such that only few and simple fragments match these molecules rather than larger, more characteristic structural motifs (see the Supporting Information, Section S3 for more discussion). When these two cases are filtered out, the tf-idf scores (for large/diverse-enough dictionaries, see the Supporting Information, Section S4) give chemically reasonable predictions of retrosynthetic cuts in about 90% of cases, which we consider quite remarkable given that tf-idf scores are one of the simplest methods of linguistic analyses (accordingly, we have been actively studying more complex methods such as n-grams or collocations that deal not only with the occurrences of individual words/motifs in the linguistic corpora, but also quantify their associations and connectivities). In the meantime, we have found this linguistics-based approach useful in computer-assisted retrosynthesis where it helps to asses the viability of possible disconnections over millions of molecules that are being considered (as, for example, in our Chematica software).[9–11]

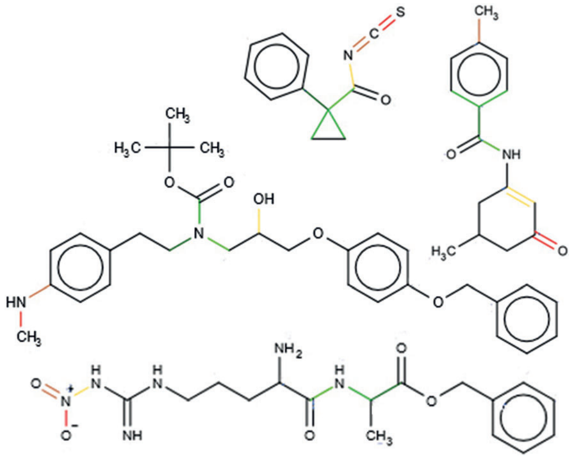

Figure 4. Examples of molecules in whch bonds were tf-idf scored as decribed in the main text. In each molecule, bonds with the highest scores are colored light green, followed by dark green and yellow. Bonds with the lowest scores and not likely to be disconnected are colored brown and red. The disconnections are in general reasonable, though the method is certainly not performing perfectly as illustrated by the sugestion to cut the aromatic ring in the upper-right molecule (which is due to imprecise statistics of substituted phenyl groups in the dictionary). For details of the scoring analyses for the above and other molecules, see the Supporting Information, Section S2.

In summary, the current work aims to 1) document that organic chemistry has a fragment/"word" structure similar to a natural language and 2) introduce the basic concepts and potential impact of linguistics-based analyses to a general chemical audience. We believe that the methods of computational corpus linguistics might be useful in analyzing organic molecules and their reactivity patterns. These statistical methods have been designed and tested to recognize and process characteristic and most information-rich patterns within large and diverse sets of texts; this activity is familiar to chemists whose brains are constantly trained to recognize structural and reactivity patterns in diverse sets of organic molecules. We therefore suggest that the linguistic approach can help us codify and perhaps one day automate the analysis of organic molecules and "the art" of organic synthesis.

Received: March 26, 2014
Revised: July 1, 2014
Published online: July 10, 2014

Keywords: chemical linguistics · graphs · information technology · retrosynthesis · symmetry

[1] J. M. Lehn, *Supramolecular Chemistry: Concepts and Perspectives*, *1 ed.*, Wiley VCH, Weinheim, **1995**.

[2] a) C. D. Manning, H. Schuetze, *Foundations of statistical language processing*, MIT, Cambridge, **1999**; b) K. S. Jones, *The-*

saurus, *Encyclopedia of artificial intelligence*, *2nd ed.* (Ed.: S. C. Shapiro), Wiley, New York, **1992**, pp. 1605–1613; c) H. Saggion, T. Poibeau, P. Piskorski, *Theory and Applications of Natural Language Processing*, *Multi-source*, *Multilingual Information Extraction and Summarization*, Springer, Berlin, **2013**.

[3] a) G. Zipf, *The Psychobiology of Language*: *An Introduction to Dynamic Philology*, MIT, Cambridge, **1935**; b) G. Zipf, *Human Behavior and the Principle of Last Effort*, Addison-Wesley, Cambridge, **1949**.

[4] While the Zipf law is the most universal charactericticic of any natural language, it alone does not describe complexity of languages, relations between grammar and vocabulary, and so on. Zipf's law is also applicable to phenomena outside of linguistics, such as dolphins' sounds, brain waves, and stock markets. For additional literature, see: a) R. Ferrer-i-Cancho, B. McCowan, *Entropy* **2009**, *11*, 688–701; b) J. Baixeries, B. Elvevåg, R. Ferrer-i-Cancho, *PLoS One* **2013**, *8*, e53227; c) J. Kwapien, S. Drozdz, *Phys. Rep.* **2012**, *515*, 115–226; d) X. Gabaix, *Quart. J. Econ.* **1999**, *3*, 739–767.

[5] a) T. McEnery, A. Hardie, *Corpus Linguistics*: *Method*, *Theory and Practice*, Cambridge University Press, Cambridge, **2012**; b) N. S. Dash, *Corpus Linguistics and Language Technology*, Mittal, New Delhi, **2005**.

[6] a) E. L. Thorndike, I. Lorge, *The Teacher's Word Book of 30000 Words*, Teachers' College, Columbia University, **l944**; b) I. S. P. Nation, *Vocabulary size*, *text coverage*, *and word lists*, *in Schmitt*; *McCarthy*, *Vocabulary*: *Description*, *Acquisition and Pedagogy*, Cambridge University Press, Cambridge, **1997**.

[7] We use a modification of a Bag of Words algorithm, based on the Longest Common Substring. BOW algorithms were probably developed by 15th century cryptographers. For a modern mathematical treatment, see: C. E. Shannon, *Bell Syst. Tech. J.* **1948**, *27*, 379–423; C. E. Shannon, *Bell Syst. Tech. J.* **1948**, *27*, 623–656; for the discussion of largest common substrings, see: D. Gusfield, *Algorithms on Strings*, *Trees and Sequences*: *Computer Science and Computational Biology*, Cambridge University Press, Cambridge, **1999**.

[8] S. Soh, Y. Wei, B. Kowalczyk, C. M. Gothard, B. Baytekin, N. Gothard, B. A. Grzybowski, *Chem. Sci.* **2012**, *3*, 1497–1502.

[9] For example, in Chematica's retrosynthetic module, each potential disconnection is checked against a database of several thousand expert-coded reaction mechanisms. The program also scrutinizes stereochemistry and regiochemistry before allowing any disconnection. Still, performing such analyses for all bonds in a molecule is very computationally costly and the linguistic approach helps shorten calculation times by preselecting only certain, most likely bonds, and preventing the exponential "explosion" of possible retrosynthetic "trees".

[10] a) M. Kowalik, C. M. Gothard, A. M. Drews, N. A. Gothard, A. Wieckiewicz, P. E. Fuller, B. A. Grzybowski, K. J. M. Bishop, *Angew. Chem.* **2012**, *124*, 8052–8056; *Angew. Chem. Int. Ed.* **2012**, *51*, 7928–7932; b) P. E. Fuller, C. M. Gothard, N. A. Gothard, A. Wieckiewicz, B. A. Grzybowski, *Angew. Chem.* **2012**, *124*, 8057–8061; *Angew. Chem. Int. Ed.* **2012**, *51*, 7933–7937.

[11] Disclosure: B. A. G. has a financial interest in Chematica, which is distributed by GSI (Grzybowski Scientific Inventions).