# Simplifying terminology extraction: OneClick Terms

Vít Baisa (Masaryk University, Czech Republic), Jan Michelfeit (Lexical Computing Ltd, Czech Republic) and Ondřej Matuška (Lexical Computing Ltd, Czech Republic)

Despite the fact that methods for terminology extraction represent a heavily studied topic, many industrial applications use rather simplistic solutions to tackle this problem (see Steurs, 2015, for an overview). They generally work with word forms (not lemmas) and word n-grams and their occurrence counts in the analyzed document. This leads to results which are often heavily polluted by items that do not qualify for terms and a lengthy manual cleaning process is needed to produce a good quality list of terminology.

The state-of-the art approaches can, however, extract terminology lists which require little or no manual cleaning at all. This requires that the source text be lemmatized and tagged for parts of speech which will create the base for applying two complementary principles: unithood and termhood.

Unithood is the quality of a lexical item to qualify for a term in a language. A combination of *preposition + verb + preposition* will not be considered a valid term structure in most languages while *adjective + (optional) adjective + noun* will. Language dependent rules referred to as term grammar defining valid word combinations are applied during terminology processing.

Termhood is the quality of a lexical unit to be specific to the domain. Termhood is established by comparing the frequency of the whole (possibly multi-word) lexical unit in the analysed text to the frequency of the same (multi-word) lexical unit in a reference corpus. Best results are achieved with a large general language reference corpora of billions of words.

The procedure outlined above is already fully functional and available in 20 languages through the Sketch Engine (Kilgarriff, 2014). However, the typical user looking for term extraction is a translator or a terminologist without the ambition of using a complete suite of Sketch Engine features and without the time or will to learn a complex system for completing a task which might seem so basic on the superficial level.

With this in mind, a decision was taken to design a single-purpose user-friendly interface to accomplish this task in as few steps as possible. In Sketch Engine, the user has to visit 5 screens and click about 10 times before the list of terminology is produced. A rather unrealistic target of reducing the numbers to 1 click and 1 screen was set.

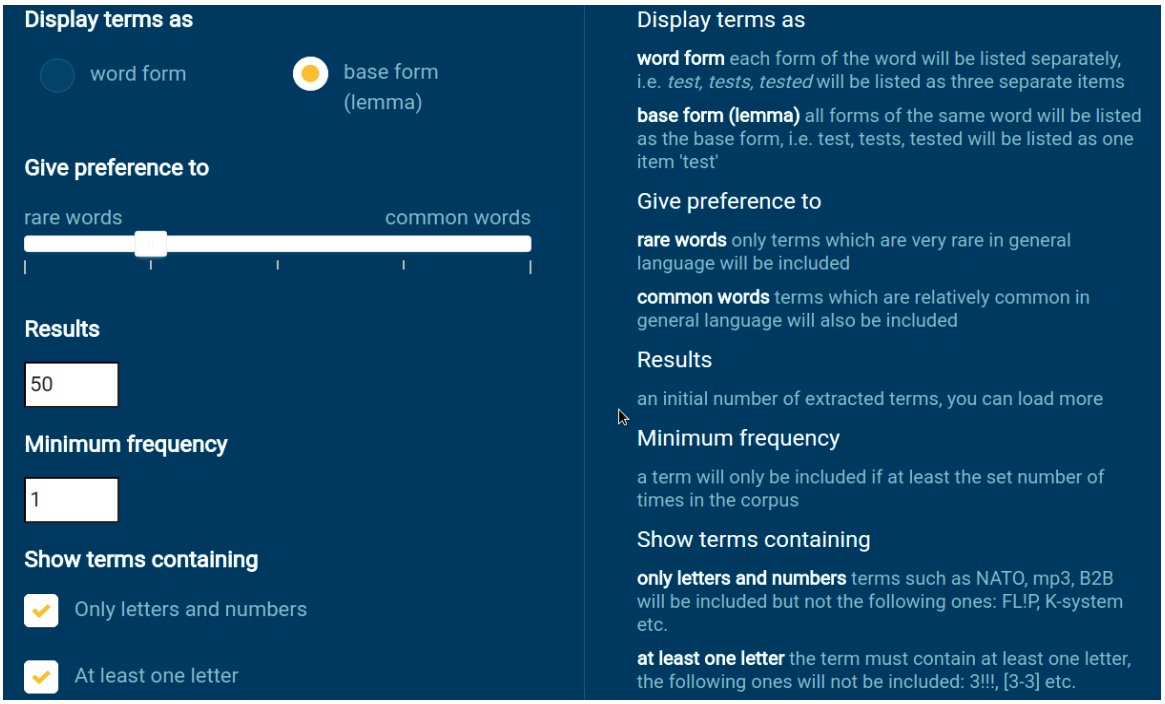
To extract terminology, the source text has to be converted into a corpus. A series of complex pipelines was meticulously developed to shield the user from the need to control this process, from taking decisions about the best settings and from launching each step of the corpus creating procedure individually. The corpus creation now happens with a simple drag&drop interface.

The supported formats are: TMX, XLIFF (from version 2.0), PDF, DOC, DOCX, HTML and TXT and the ever growing number of supported languages currently includes Chinese (simplified, traditional), Czech, Dutch, English, French, German, Italian, Japanese, Korean, Polish, Portuguese, Russian, Slovenian and Spanish. Support for more languages is continuously developed.

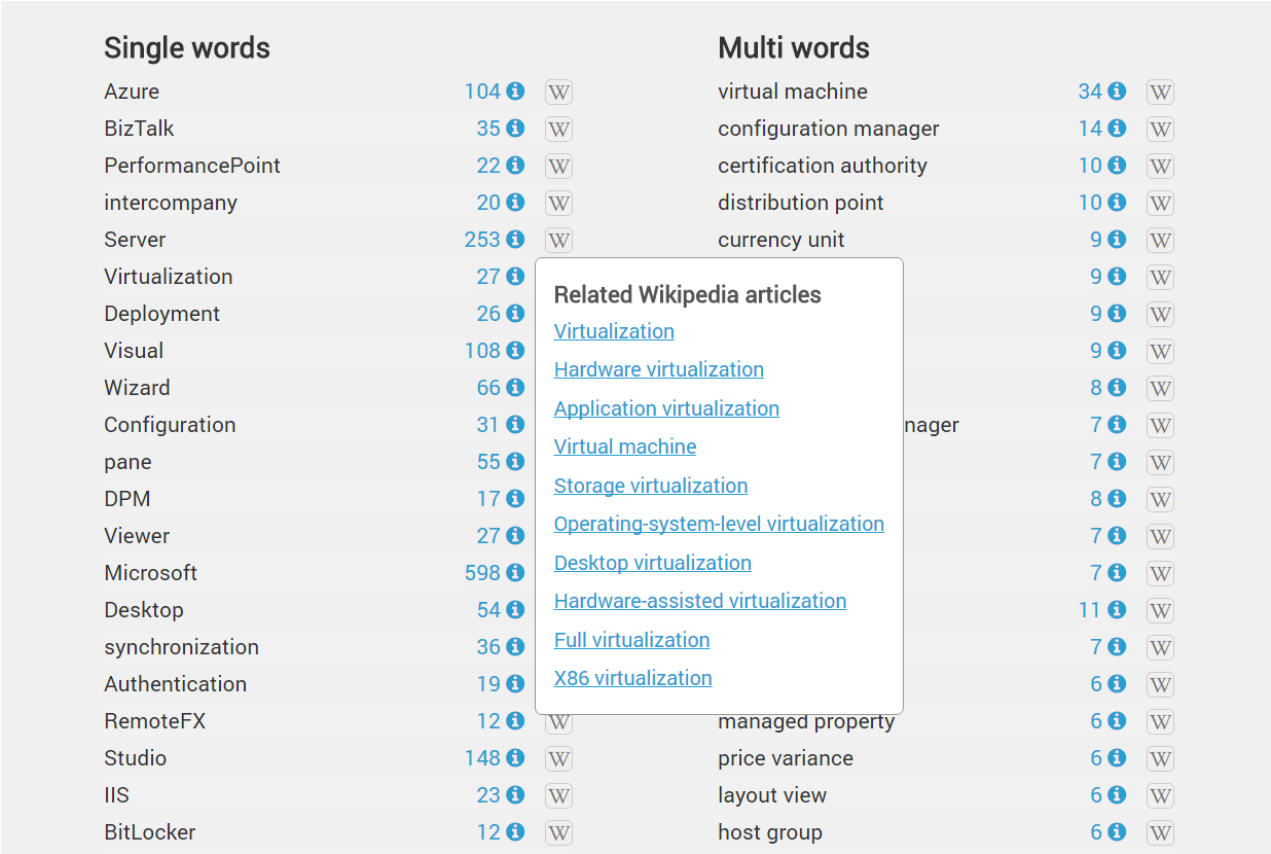

After dropping the file, the upload starts and the user can decide to launch the term extraction immediately or to interfere with the default settings should they wish to do so. If the user prefers the former, then this is the only click required from the user to produce a clean term list.

Since it is reasonable to expect some users to wisth to have a certain level of control over the process, an intermediary settings page was introduced. The settings include easy to understand options with glossary, any originally numerical settings have been converted to visual controls. The user can control the following options:

- lemmatized or non-lemmatized list,
- a slider with 5 stops was introduced to control to which extent the algorithm should prefer rare words, i.e. words specific to the focus text, or common words, i.e. words common relatively frequent in general language,
- the number of items to extract, more terms can be loaded from the result screen,
- a minimum frequency of the term in the focus text (default is 1) can be changed to help filter out certain unwanted items and
- two more options (ON by default) can be switched OFF: the term has to contain at least one letter and the term must not contain non-alphanumeric characters. The latter might be useful with the OFF setting to include various product names or model numbers, e.g. CN-9030b, should this be important to the user.



The result screen lists terminology in two columns: single-word and multi-word items. The user can download the lists as a CSV file supported by a vast selection of software. The result screen also gives direct access to 10 most relevant Wikipedia articles for each term and the user can display the term used in context as it appears in the focus text.




What seemed rather unrealistic at the beginning was achieved. The user can, without any knowledge of NLP, text corpora, tagging or corpus management software, drop a file into the interface, click once and produce a clean list of terminology. The length of the whole process is dependent on the size of the source file and lasts only a couple of seconds for average-sized files. The most significant achievement is the sheer simplicity. The user can operate the system instantly, without the need to take any decisions and without any introductory training. This led to the decision to give the system a name that describes this achievement best: *OneClick Terms*.

## References


Steurs, F., De Wachter, K. & De Malsche, E. (2015). Terminology tools. In H. J. Kockaert & F. Steurs (Eds.), *Handbook of Terminology* (222–249). John Benjamins Publishing Company: Amsterdam.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. & Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography*, *1*(1), pp. 7–36.