

July 31, 2012

## Ten things you should know about automatic terminology extraction (Part II)

Uwe Muegge

## Ten Things You Should Know about Automatic Terminology Extraction (Part Two)

Last week we introduced the [first part](#) of a two-part series on automatic terminology extraction, where translation tools expert and terminologist, [Uwe Muegge](#), broke down terminology extraction and its role in the careful management of terminology. In this post, you'll find suggestions regarding methods and tools for automating terminology extraction. Read on to see what tips Uwe has to give.

1  
tweet

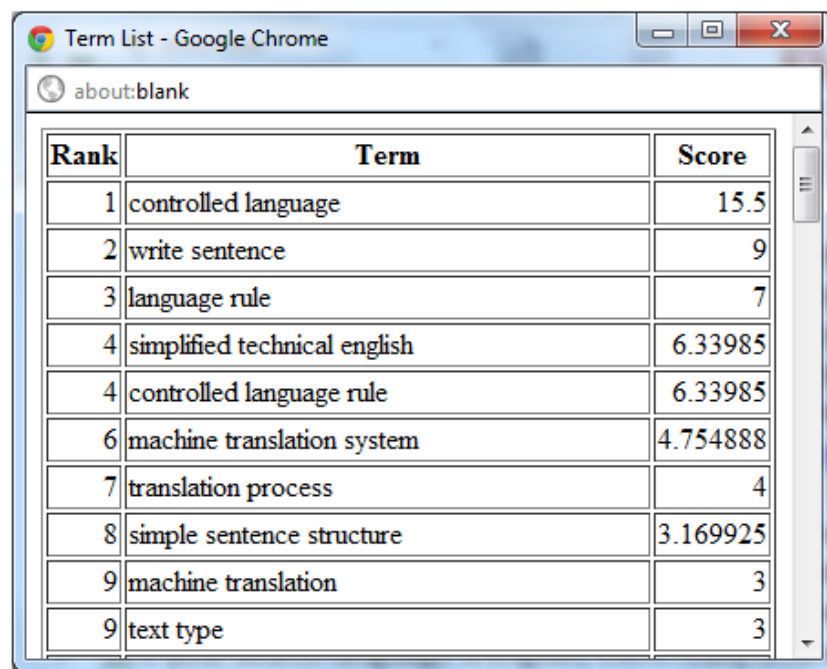
retweet



### 6. Free online tools provide powerful terminology extraction—and there is nothing to install

If, by now, you are still not convinced that automatic terminology extraction is for you, let me introduce to you a set of tools that will only require you to create a term list to specify a source text and then press a button or two. There is no software to install, no manual to read and, of course, no price to pay. With web-based terminology extraction services like [TerMine](#) and [fivefilters Term Extraction](#), automatic terminology extraction really is child's play.

Don't let the simple interface of these sites fool you: Both of these online tools produce professional quality extraction lists that include compound nouns, and, in the case of TerMine, even scored rankings of term candidates.



Rank	Term	Score
1	controlled language	15.5
2	write sentence	9
3	language rule	7
4	simplified technical english	6.33985
4	controlled language rule	6.33985
6	machine translation system	4.754888
7	translation process	4
8	simple sentence structure	3.169925
9	machine translation	3
9	text type	3

TerMine Generated Term List – free online terminology extraction service

### 7. Are you using free MT? Start post-editing with a glossary

As using free machine translation services becomes more and more popular among professional translators, so does the desire to control terminology in the final output that is delivered to clients. [Google Translator Toolkit](#) is a free, full-featured online translation memory system that allows users to post-edit translations generated by Google Translate, Google's proprietary machine translation system. Since Google Translate is a statistical MT system that has been, and continues to be, trained on a wide variety of documents, the same source term might get translated in multiple ways even within the same document, not to mention across documents.

While it is currently not possible to submit user glossaries to Google's machine translation engine, it is possible to upload glossaries to the Translator Toolkit. And using one of the tools mentioned in this article to extract terminology and build a bilingual glossary before translating/post-editing in Google Translator Toolkit may be the best thing linguists can do to improve the efficiency of an already very efficient process.

### 8. Cleaning-up your terminology extraction list to identify most important term types

In my professional experience, term lists generated by automatic terminology extraction tools are never perfect: Even the best term extraction systems introduce 'noise.' For example, in the TerMine term list shown above, I would argue that at least three of the ten term candidates in this list require editing.

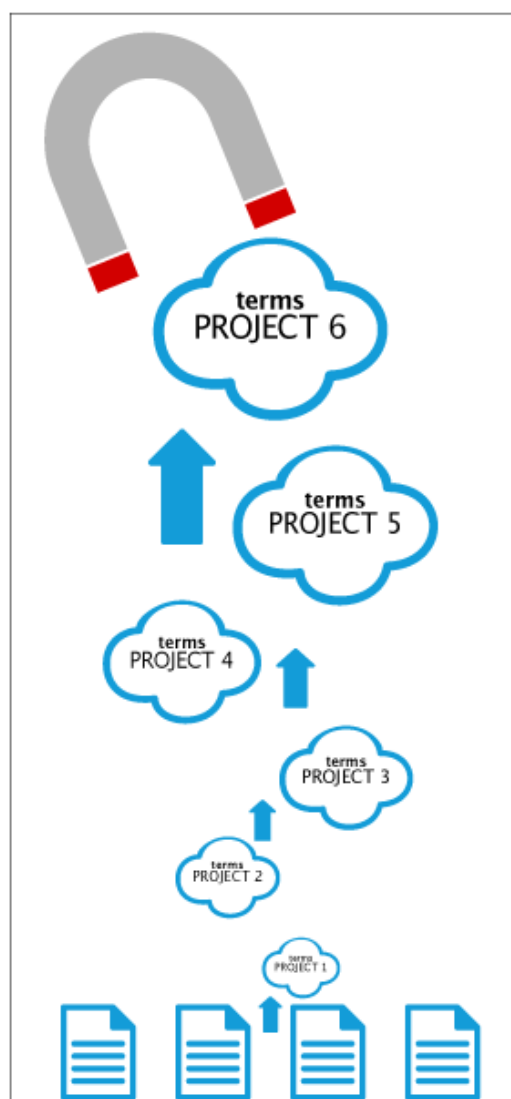
While most illegitimate term candidates are easy to identify, e.g. misspelled, truncated, incorrectly hyphenated words, many linguists have a hard time answering the following question: Which term candidates should users of terminology extraction systems actually develop into multilingual glossaries? There is no simple solution to this problem, as each translation project has its own limiting factors, available time typically being the most important one.

My recommendation for commercial translation projects is to always include the following types of terms in a project glossary, *even if the term occurs only once in a source document*. Yes, I know: This piece of advice runs counter to what many other terminology experts say; namely, if a term occurs only once in a text, there is no risk of inconsistency, and therefore single terms should not be included in glossaries. To that I say: There are terms that are so important that if a linguist gets them wrong *even just once*, it would be a huge embarrassment for all parties involved. Mandatory term types include:

- **Client Business Name**
- **Product Names**
- **Trademarks**

Including every single term of the above-mentioned types is particularly important when working with machine translation, as MT systems are notorious for ‘making-up’ their own terminology in the target language. The following term types should be included in glossaries based on the frequency of their occurrence in the source text (many terminology extraction tools provide frequency information):

- **Feature Names**
- **Function names**
- **Domain-specific Terms**
- **Generic Terms**



Automatic Terminology Extraction leverages your existing TMS

### 9. Recommended data categories when integrating extraction list into a terminology management system

Once the extraction list has been cleaned-up, the next logical step is to develop a multilingual glossary that will add value not only to the translation process but ideally to the entire translation cycle. The most valuable glossaries are those that provide information that goes beyond simple word pairs of ‘source term’ and ‘target term.’

Here is the minimum data model I recommend for commercial projects:

- **Client and/or Business Unit and/or Project Name**
  - **Source Term**
    - **Part of Speech** (e.g. *noun, proper noun, compound noun, verb, adjective, other*)
    - **Context** (e.g. a sample sentence in which the source term occurs)
  - **Target Term**

The big question at this stage is: What software platform to use for developing and managing terminology after extraction? This is an important question as many, if not most, linguists don't have a proper terminology system in place. While it may be tempting to use Microsoft Word or Excel tables to manage terminology—after all, these are programs that most linguists own and know how to use—word processors and spreadsheet applications are not good choices for managing terminology data. The systems I recommend are [TermWiki](#) (if you are willing to share terminology) and [TermWiki Pro](#) (if you need to keep your terminology data private). Full disclosure: I have been, and keep, contributing to the development of TermWiki, which is already changing the way thousands of users around the globe manage linguistic assets.

Here are some of the benefits of using either version of TermWiki:

- Completely web-based (no software to install)
- Platform independent (runs on Windows, Mac, Linux, Android, iOS, etc.)
- Wiki user interface (intuitively familiar, easy-to-use)
- Powerful collaboration features (automatic workflow management, etc.)
- No-cost/low-cost solution (TermWiki is free, TermWiki Pro is 9.95/user/month)



#### **10. A small investment in automatic terminology extraction can yield a big return in efficiency and client satisfaction**

Being able to extract terminology quickly and efficiently is a wonderful thing. With automatic terminology extraction as part of a comprehensive terminology management effort, you are able to:

- create comprehensive multilingual glossaries *before* translation
- have the client authorize project-specific, multilingual glossaries *before* translation
- have translation memory systems automatically suggest authorized translations for every term *during* translation
- have all members of a translation team use the same terminology *during* translation
- eliminate (terminology) review and corrections *after* translation

With so many powerful terminology extraction tools to choose from, as long as the source language is a major language, there really is no excuse for *not* extracting terminology and creating a glossary as part of every translation project.

---

*Uwe Muegge has more than 15 years of experience in the translation and localization industry, having worked in leadership functions on both the vendor and buyer side. He has published numerous articles on translation tools and processes, and taught computer-assisted translation and terminology management courses at the college level in both the United States and Europe. Uwe has been with CSOFT since 2008, and currently serves as Senior Translation Tools Strategist for North America.*