Ter·mo·lo·gic

49 Robert St. W., Penetanguishene, Ontario Canada L9M1M5 705-527-3602
www.termologic.com / kara@termologic.com

Term Extraction

Kara Warburton, PhD

Termologic provides a glossary creation service that can significantly improve a company's global communications. Termologic recognizes the need for effective term extraction to support companies' requirements for purpose-built glossaries of terminology and other important language expressions.

Term extraction requires a sophisticated software program and a terminologist to run the program and refine the output. This is why many companies are hesitant to try term extraction, or they are discouraged by the initial results. Termologic addresses this challenge by offering a complete service where our terminologist produces a validated list of terms on your behalf using the best technology available. By using this service, your company no longer has to worry about how to use a term extraction tool and how to "clean up" the output. We deliver high-quality terms in large volumes, specifically tailored to your company's needs.

What is term extraction?

Term extraction refers to the process of identifying the *key terms* in a set of documents. What is considered to be a *key term* depends on the ultimate use to which the list of extracted terms will be put (described in the next section). Generally speaking, key terms are words that express important concepts, i.e. they reflect the topic area of the text. For instance, in an automobile user manual, the names of the car's parts, functions, and features, but also general driving and operating expressions, are important terms. On the other hand, on the car's Web site, the colorful language crafted to influence potential buyers will contain other interesting but possibly less technical terms.

Terms can be extracted manually, by reading the document and highlighting the important terms. Often, though, the text is too large for manual extraction to be feasible. (The information set for most products comprises hundreds if not thousands of individual files.) In this case, we need to use a special software, called a term extraction tool.

Why would we want to extract terms?

Term extraction is useful for a number of situations. The most common is when the text from which the terms are extracted is going to be translated into other languages. The list of extracted terms can be translated in a controlled way where quality is guaranteed, such as by an experienced translator who is already familiar with the topic area. Then, the list of terms, which is now a **bilingual glossary**, can be provided to the translators who will be translating the text. This process guarantees that the translated versions of a text will meet expected standards even when translators with different qualifications or backgrounds are involved.

This approach is particularly useful when multiple translators are employed to translate the same text or collection of texts. Providing translators with pre-determined translations of the key terms ensures that the translations of these terms in the text will be consistent from one translator to the next. Failing to do so means that translators have no guidance on how to translate the important terms, and their choices will vary, leading to inconsistencies in the translated text.

Companies that operate on a multinational or a global scale translate the information about their products and services into multiple languages. Usually the information is spread in multiple files. The information needs to be translated quickly, so that the products can be released to market as soon as possible. Typically a translation company is engaged, and to meet tight deadlines the job is divided into parts and given to different translators. Extracting and pre-translating the key terms are effective strategies to ensure that the quality of the final translation will be acceptable. The process also helps to shorten revision time and reduces the number of errors that need to be corrected.

If the translators use a computer-assisted translation (CAT) tool, the bilingual glossary can be loaded into the CAT tool. In this technical environment, the terms are automatically shown to the translators when they need them. This avoids having to rely on translators looking up terms.

Another use of term extraction is to identify the key words in a document for building an index or a search engine lexicon. The key words can also be used to tag a document as to its main content, such as for a content management system or an automatic content categorization tool. These more advanced information technologies are gaining momentum in commercial settings, to help deal with the exploding volumes of information in electronic form.

Finally, a term extraction process is essential for developing a terminology database (termbase). Many companies are developing their own termbase as a way to store their terminology and other important lexical expressions in a stable environment so that it can be reused when needed, such as to translate additional products or services. Once a company has a termbase, it can begin to leverage the terminology in different ways.

Controlled authoring is one application that is gaining recognition as an effective way to improve content quality and reduce production and translation costs. Controlled authoring refers to measures that are taken to improve the quality of texts in the source language, before translation even takes place, such as adhering to rules of grammar and style, and using company-approved terminology and vocabulary. Software applications are now available to help writers adhere to these rules as they are producing text. These applications require lists of terms that have been approved for use. The company's termbase can provide these lists of terms. Once again, term extraction is a key process to help build up these terminology resources.

What needs to be extracted?

As stated earlier, the lexical items that you would want to extract from a text depends on how you intend to use them. We have just used the word “lexical item” instead of “term” because some things we need to extract are not “terms” in the conventional sense. The conventional notion is that a term is a designation of a concept from a special subject field or domain, such as *gross domestic product* and its acronym *GDP* in economics, or *convection oven* in the field of domestic appliances.

If you are building a list of lexical items for translation purposes, as described in the previous section, you will want to not only identify the key designations of concepts from subject fields (i.e. “terms”), but also, anything that occurs frequently in the text and may be either challenging for translators or likely to be translated inconsistently if multiple translators are involved. These latter expressions may or may not designate domain-specific concepts. For example, certain multi-word expressions are challenging for translators, such as:

1. dynamic signal analyzer
2. direct network measurement

These two terms are ambiguous due to the lack of explicit markers in English to identify the relationships between the modifiers and the head word. There are two interpretations of each term, and each interpretation will be translated differently.

1. analyzer of dynamic signals OR dynamic analyzer of signals
2. measurement of direct networks OR direct measurement of networks

Such ambiguities are common in English and are even more prevalent in multi-word terms that comprise more than three words.

Another challenge for translators is when two different terms occur in the text and they might be synonyms. (Unless terminology is controlled in the source language, as described earlier, inconsistencies and synonyms do occur in a given text.) The translator has to decide whether to translate the two terms the same way (if they are synonyms) or differently (if they are not synonyms). For example, in one company's product, the term *store administrator* and *store manager* co-occurred. On the surface, these two terms look like synonyms. Translators, believing that they were synonyms, chose only one term in their language to translate both English terms, i.e. *gestionnaire du magasin* in French in both cases. In fact, the terms had different meanings in the product, and so different translations were required. This error delayed the product's release to market, because many occurrences of the term had to be changed, and not just for French but for other languages too. Frequency is an important factor to consider. A lexical item that occurs frequently in an information set may need to be pre-translated just to ensure consistency in the target languages.

If you are building a terminology list for a controlled authoring application, it is essential to identify synonyms and other variant expressions of the same concept, such as acronyms and abbreviations. Then, the preferred term among a set of possible terms can be identified and this information then provided to writers through the controlled authoring software.

It is not possible within the limits of this paper to describe all the possible types of lexical expressions that one would want to extract for different purposes. Each use case needs to be considered separately and then some term selection criteria defined.

A brief introduction to term extraction tools

A term extraction tool is a software program that scans a text and outputs a list of terms that were found in that text.

There are a number of commercial term extraction tools on the market. There are also a number of tools that have been developed in research settings such as universities, but these tools are generally not meant for production purposes (support services may be unreliable or unavailable altogether, there may be disclaimers and no guarantees, and so forth). Furthermore, due to IT security controls, many companies do not allow the use of experimental or non-commercial software.

Term extraction tools adopt a range of techniques, from statistical to rules-based or a combination of both. Statistical tools are the most limited; some only extract single words. Most important terms are comprised of more than one word, such as the examples given in the previous section. Tools that use a rules-based (grammatical) approach tend to produce better output than statistical ones. With this approach, the part of speech (noun, verb, etc.) of the words in the text is considered. This allows terms that follow certain patterns to be given priority, such as:

- noun
- noun + noun
- adjective + noun
- adjective + noun + noun

Such an approach therefore requires more sophisticated algorithms involving part-of-speech tagging and syntactic parsing. Some tools also perform morphological stemming to convert inflected forms to their base form. The hybrid approach combines statistical and grammatical calculations to further improve the output.

The output of a term extraction tool is a list of “term candidates”, called so because some of the items in the list are not terms. After a skilled linguist has gone through the list and removed unwanted items, what is left are the real terms.

The unwanted items are called “noise”. Term extraction tools generally produce a lot of noise, often more than 60% of the output. If the effort to remove the noise exceeds the effort of identifying terms manually, then the tool is not useful at all. Many term extraction tools fall into this category, and this

explains why many people decide not to use them.

Aside from the problem of excessive noise, there is also “silence” to be concerned about. Silence refers to the important terms that were NOT extracted by the tool. All term extraction tools fail to identify some important terms.

A term extraction tool that produces the least amount of noise and silence is the most effective. However, there will always be some noise and silence since term extraction tools are not perfect and never will be. A qualified linguist who understands the term extraction process, the limitations of the tools, and the intended purpose of the output, needs to modify and enhance the output before it can be used. This process, called “cleaning,” involves not only removing unwanted terms, but also consolidating families of terms into their key members, and adding new terns through the process of resetting the boundaries of some multi-word terms.

At Termologic, we have our own in-house tool that outperforms tools that are commercially-available, such as those offered with CAT tools. We don't sell it as a product, because we firmly believe that the solution needs to be provided as a service due to the skills required. If you choose Termologic as a service provider, you can be assured that you are now benefiting from the best term extraction tool and accompanying data cleaning service available.

Term Extraction as a SaaS

Purchasing a term extraction tool outright may not be straightforward. The tool might be embedded in a software suite, such as a computer-assisted translation or a controlled authoring product, meaning that you have to purchase the whole suite even if you have no use for the other components. The software might not comply with the company's IT regulations and therefore it may not even be possible to acquire it. Running the tool effectively, that is, knowing how to set the parameters and to develop and use lexical resources such as exclusion lists to get the desired results, presents a challenging learning curve for most. Cleaning the output is necessary yet that requires special knowledge and skills.

Faced with these obstacles, many companies and organizations that could benefit from term extraction don't attempt it. Furthermore, most potential users need to extract terms on a fairly irregular basis, such as for certain translation projects, to provide terminology for a new application such as controlled authoring, or as a means to quickly populate a termbase. The time investment as well as

the capital outlay costs for the software and hardware to set up a process that isn't carried out regularly may not be justified.

A complete term extraction service on a fee-per-use basis is an appealing alternative. Termologic provides such a service. We run the term extraction tool, clean up the output, and provide a final term list that satisfies the customer's requirements. Additional information can be added if desired, such as subject field values, and definitions. If the customer requires a bilingual or multilingual glossary, target language equivalents can be added either by Termologic's linguists or by the customer's linguists, as desired. Existing resources such as translation memories can be leveraged in order to determine the most fitting translations. The service is tailored to the customer's requirements.

Additional services such as to develop a termbase and to import terminology into a termbase are also available. Termologic can connect to the customer's IT systems to integrate the terminology into the globalization workflow, oversee the translation of the terminology, and collect and process feedback. Indeed, Termologic provides expert terminology management services of any kind, from term extraction to end-to-end terminology process development and database engineering.