Term Extraction: 10,000 Term Candidates—Now What?

By Barbara Inge Karsch

Identifying terms and names and researching their concepts should not be the responsibility of just freelance translators. True, they must do it if no one else has, but freelancers are actually not always in the best position to research terminology thoroughly.

In a large translation project, freelancers may be part of a team of translators. If each team member researches the same terms instead of collaborating, it is a duplication of efforts and will not lead to consistent target terminology. Furthermore, freelancers often do not have access to all of the information that resides on the client side. Indeed, they may not recognize an important new concept in a long list of term candidates because they may not have enough product knowledge. Ideally, companies and organizations identify and document terminology before they hand off content for translation.

If you have never performed term mining, term harvesting, or term extraction—all synonyms for the process of identifying terminology in a semi-automated fashion—it could be a daunting task. For one, you need to have a tool, have the files in the right format, and decide on certain extraction parameters. But the most intimidating aspect might be the term mining output: a potentially long list of term candidates that you need to evaluate critically and decide which would be useful in your term base. Let's talk briefly about the initial steps involved in this process and zero in on criteria for identifying good term candidates.

Freelancers may not recognize an important new concept in a long list of term candidates because they may not have enough product knowledge.

Preparation

There is a variety of tools on the market with term extraction capabilities. Many of them are based on statistical extraction engines that look at the number of occurrences of a word or a string of words in order to determine whether or not to classify them as term candidates. Linguistic engines use information, such as part of speech, stemming, etc., for term extraction. Your translation tool may come with such an engine. For instance, memoQ has a term extraction component and Trados MultiTerm has a sibling tool called Extract. Both are statistical extraction engines. Statistical engines generally produce more "noise" (i.e., unwanted term candidates) than linguistic engines. As a result, they require more cleanup efforts.

What is involved in the cleanup?

The tool has provided you with a list of term candidates. Term candidates are words or strings of words that the extraction engine has identified as something that you might want to include in your term base. Good tools also provided you with context, the context source, and the frequency with which the term candidate occurred. You must now go through the list and either pick what you want and transfer it into the term base or delete what you do not want and import the remaining list into the term base.

Making Good and Fast Choices

After you have done this a few times, you might know exactly what you want. But if you have never gone through a list of term candidates, you might need some selection criteria. The following are some of the most obvious ones.

1. Abbreviations and Their Long Forms: We often cannot be sure what an abbreviation stands for. Therefore, acronyms, short forms, initialisms, etc., along with their fully spelt-out forms, are worthy term base entries. For example, in manufacturing software, we would want to document *MRP* along with its full form, *manufacturing resource planning*.

2. Homographs: A term that might have two different meanings is also a good candidate for inclusion in a term base. In an ideal world, we do not have homographs in the same text or subject area, but in reality we often find a term or name that represents different concepts. For example, in a manufacturing context, *MRP* might also stand for *materials requirement planning*, *mid-range planning*, or *maintenance recovery period*. Other examples are standards that are often named exactly like their underlying technology, even down to capitalization (e.g., Bluetooth, USB, VGA). So, since we have two concepts—one for the standard, one for the technology—and target-language designations might not follow the same naming pattern, we should include both in our term base.

> Ideally, companies and organizations should identify and document terminology before they hand off content for translation.

These two categories help us determine a good number of candidates. The following are more refined criteria that have come out of many large term mining efforts. The list is not exhaustive, but it might give you a good idea of what to look for. Based on your specific subject field or type of material (e.g., software strings versus free-flowing text), you may want to refine or extend the list.

3. Novelty: New texts deal with new concepts via their new designations—new terms or names—otherwise they would not need to be written. New terms and names obviously make good term candidates. You might want to be clear about whether new means new to the content supply chain (e.g., new to the translation memory, new to the translation team), or whether it is new to the terminology database.

4. Confusability: The main reason a term or concept can be confused with something else is homography, which we already covered. But you might find other reasons why a term or concept could cause confusion, which is why you should add it to your term base with all appropriate terminological data.

5. Terminologization or Transdisciplinary Borrowing: Another cause for homography, and therefore the potential for confusion, is when common, everyday words become specialized terms and adopt a very specific meaning in a subject field. For example, the word *cloud* has undergone what is called terminologization; it now has a specific meaning in the information technology field. Similarly, an existing term in one field may represent a new concept in another field. *Virus* is a medical term that was borrowed by information technology experts to take on a similar meaning for computers, one with which we are all familiar today. It is tricky to catch these terms in a term mining project when we are not familiar with them.

6. Degree of Specialization: Technical language cannot do without technical terms. In fact, it must contain technically precise language so as not to hinder communication. And some of these technical terms represent highly specialized concepts: ideas with which freelance translators who do not deal with documents in this field on a daily basis might be unfamiliar. For example, if you were to identify complex concepts in this article, *terminologization* might be a term that you would include in your term base, since it represents a pretty specialized concept even in the field of terminology management.

7. Frequency and Distribution: Our term mining output file should give us an indication of how often a particular term or name comes up in the original files. It would also be good to know whether it comes up in one or more files. While we would usually want to include designations that occur, for instance, in the software, in the help text, and on the packaging material, high frequency might mean that we do not include it. This is because terms that come up often might be so well known and so clear that we do not need to include them. Of course, if they have a very awk-

ward spelling you would still include them in a term base. The same goes for names that do not come up very often. This is something we definitely have to keep in mind, but in large projects, we might have to limit our focus on the medium- to high-frequency terms for lack of time.

8. Visibility: A new concept may not be mentioned very often, but it may be so important and prominently mentioned (e.g., only in the title) that it is worth adding the term or name as well. For example, translators working on a project would need to know and understand the slogan of the latest marketing campaign in order to do it justice.

9. System: This warrants a bit more explanation. Terminology research and documentation is ideally always done with a systematic approach in mind. That means that we do not focus on one concept in isolation. Rather, we address related concepts together. For example, if a medical device, such as an external defibrillator, has many different types of cables, we look at all of the cables together. Doing so helps in two ways. First, it helps us understand how one cable is different from another. Second, it is much faster to research all of the cables at one time rather than researching them separately. Even if a particular concept does not come up in the mined material, it might still be good to include it. Doing so will make the system more complete and help those doing the research in the target language understand the concepts.

10. Standardization: Finally, we document terms and names because we always want to use the same spelling, or perhaps never want to use any of the available synonyms. This is one of the main aspects of doing terminology work for authoring in one language. Furthermore, it is impossible for one person to identify all of the terms or names that might need to be included to have harmonized terminology in all target languages. It is perfectly legitimate for a translator to request that additions be made to the database to serve standardization purposes better. This way a team of, say, 10 Japanese translators will use the same term for a new concept instead of leaving it in English, transliterating, or making a choice between the options the different syllabaries might inspire.

Provide a Footprint for Future Productivity

The criteria above have different weight for different terms or names. After you have gone through a few term mining projects, identifying terms and names from the suggested list of candidates becomes second nature. The time you spend on research will grow shorter for subsequent projects, since you will now deal only with terms and concepts that are new instead of starting your research from scratch each time. For example, for Windows Vista over 2000, new concepts and terms were documented in the Microsoft terminology management system. For Windows 8 and 8.1, the new additions were in the hundreds. Also, once you have your source and target terminology entered into a system, you can automate part of the quality assurance process.

Extracting and managing your terminology has many benefits. Building the term base is not an easy process. But the above criteria can give you a good idea of what your database users might like to see. ■

Related References

Warburton, Kara. "Glossary Creation Service Leveraging State-of-the-Art Term Extraction," http://termologic.com/wp-content/uploads/2014/06/termextraction.pdf.

International Standard 1087-1: Terminology Work–Vocabulary–Part 1: Theory and Application (Geneva: International Organization for Standardization, 2000).

Karsch, Barbara Inge. "Term Selection." Blog series by BIK Terminology, http://bikterminology.com/category/process-2/selecting-terms/:2011.