

Terminology and Translation Quality Assurance

Joachim Van den Bogaert

Resum: Tot i que avui dia tothom està al cas de la gestió terminològica, encara hi ha una gran distància entre una gestió diligent de la terminologia i la pràctica habitual. Aquest article descriu com aquestes practiques poden afectar la traducció i el procés de control de qualitat. També presenta una metodologia que permet processar dades terminològiques gestionades amb carències i utilitzar la informació recuperada durant el processament i la traducció per tal de millorar la organització i la consistència d'una base de dades.

Paraules clau: Traducció, control de qualitat, gestió terminològica, Kaizen, optimització del flux ce treball.

Resumen: A pesar de que hoy día todo el mundo presta atención a la gestión terminológica, todavía hay una gran distancia entre una gestión diligente de la terminología y la práctica habitual. Este artículo describe cómo estas prácticas pueden afectar la traducción y el proceso de control de calidad. También presenta una metodología que permite procesar datos terminológicos gestionados con carencias y utilizar la información recuperada durante el procesamiento y la traducción para mejorar la organización y la consistencia de una base de datos.

Palabras clave: Traducción, control de calidad, gestión terminológica, Kaizen, optimización del flujo de trabajo.

Abstract: Although terminology management is on everyone's agenda nowadays, there is still a major gap between best practices and real life implementations/old habits. In this article we describe how this affects translation and quality assurance of translation. We also present a methodology to process poorly managed terminology data and use the information retrieved during processing and translation to improve the organization and consistency of a term base.

Key words: Translation, Quality Assurance, Terminology Management, Kaizen, Workflow Optimization

1. FAQ2859

There is an old story on the Internet¹ about users getting confused by the message "Press any key to continue ...". A well-known computer manufacturer got so many customer inquiries about a missing "any" key on the keyboard, that they decided to devote a FAQ-topic to it:

"This is not a key. When you are instructed to press any key, this means you can press any of the keys on the keyboard (such as the Enter key, the R key, or the space bar)."²

This may sound funny, but what if *you* were the person running the technical support division? Would it not be great to get rid of this support question and stop losing precious time? Would it not - for example - be great to replace the message by "Press the space bar to continue ..." to avoid confusion? And would it not also be great if all occurrences in all applications and manuals, from now on, would contain the correct replacement sentence?

¹ http://en.wikipedia.org/wiki/Any_key

² <http://web.archive.org/web/20010317221900/web14.compaq.com/falco/detail.asp?FAQnum=FAQ2859>

Of course it would. But how would you do this?

2. Terminology Management

Enter Terminology Management. Terminology Management is a collection of tools and methodologies that allow an organization to organize its use of terminology. It is not a specific software application in

itself, but a strategy to keep track of the storage and use of terminology. Terminology Management is a fundamental part of a company's Globalization Information Management (GIM).

It helps you to use the right words at the right time. When authoring text, it helps you to write correct and consistent copy. When translating documents, it provides you with the correct wording in your target language.

3. Implementing Terminology Management

We can define the aim of terminology management as "assuring that corporate terminology is used correctly and consequently in all business communication". This sounds simpler than it is. With many parties and sources for terminology involved (technical writers, translators, sales and accounting departments, ...), maintaining consistency becomes a complex task.

To achieve the aim, a professional design and implementation strategy are required, especially when terminology needs to be translated.

Too often, terminology management focuses on identifying and just storing terminology data. *How* terms are stored seems to be less important than *the fact that they are being stored*. And although recommendations and standards for storing and exchanging terminology data are widely available, many people still tend to collect terms as undocumented translation pairs in simple spreadsheet files. This results in poor data sets that are difficult to use and manage, or as an SDL whitepaper on terminology management (SDL 2007) states:

"[...] many organizations rely on spreadsheets and word lists to track and manage terminology. These flat files are not linguistically powerful, and cannot handle the breadth and depth of functionality required to truly manage the complexity of terminology."

For a language service provider this is a challenging situation and a perfect opportunity to excel in customer service (some experiences we had at Yamagata Europe will be presented below). From a purely practical and economic point of view, compiling just a bunch of files is bad practice and a waste of valuable resources.

Harvested terminology needs revision by humans before it can be used. To unlock the full potential of terminology, terms need to be categorized, reviewed, validated and so on. This is an intensive process: there are examples of companies processing only 20 terms a day (Massion 2007), which may seem rather poor, but the results are rewarding. Recent studies have estimated the ROI on terminology management as high as 900% (Childress 2007).

4. Integrated terminology management

Collecting data properly is an important task when implementing terminology management, but it is not the only one. We have seen companies setting up impressive database systems while still struggling with their terminology. We can illustrate this with an example: sometimes we see database exports contain mistakes that have been corrected by translators and proofreaders in previous projects. How is this possible? Apparently, nobody is in charge of updating the central database from which the defect terms originate. What is the use of storing data if no one changes incorrect entries?

This brings us to a more general issue of terminology management: integration. Terminology management should be implemented on a global company level (Warburton 2006a):

“Terminology [...] has traditionally been viewed as a translation issue. [...] Terminology management tools are usually buried in translation software. And most terminology databases, online websites, and so forth, were developed to respond to the needs of translators. With this historical baggage, it is hardly surprising that the authoring community

has not assumed its equal responsibility for managing its terms, and that business executives still fail to see the need to manage terminology at all.”

The focus should be on how terminology is being used in corporate processes. Simply storing data is not enough. It is important to know what will happen to your terms once they have left the database. Protocols for obtaining, adding, updating, and deleting terms need to be defined in order to keep the central term base healthy.

Also, an attitude change towards dealing with terminology in everyday situations should be achieved. Ideally, checking the use of correct terminology should become as natural as spell checking internal documents. This, of course, requires the development and use of support systems. Good support systems enable you to direct and monitor the use of terminology. Designing such support systems is not so difficult, given today's technologies. The important thing is that the design of such subsystems should be included in the design of a terminology management system. This can save a lot of money.

For example, a web service attached to a central term base can feed a word processor plug-in to look up fuzzy matches and at the same time publish the latest changes to terms on the corporate intranet. Analysis software can provide figures on the use of your auxiliary and core terminology systems. This is only easily achieved if the design of the web service and the analysis capabilities has been included in the design of the core architecture. Note that a web service interface offers open possibilities for extension.

5. Terminology use in documents

The last couple of years, there has been an interesting evolution in the development and use of tools to help organizing terminology. Nevertheless, we see that there is still a lot of ground to be covered when it comes to using their full potential.

We argued earlier in this paper that the usefulness of terminology is determined by how thoroughly the terminology data has been defined. If we look at how terminology is used in documents, we see that it is equally important to mark where terminology is being used.

For example, a translator needs to know how to handle terms in a document when translating. For instance, early versions of technical equipment not always have localized versions of on-screen displays and speech commands. In order not to confuse the user, the documentation should reflect this. Consequently, in order to process a term in translation, a translator needs instructions about when to, and when not to translate a term. Sometimes it is possible to derive this from the context, but when no linguistic cues are provided, time and money are saved when the translator can make the correct choice based on tagging.

There is another argument in favor of tagged terminology in documents: it helps companies take full benefit of having their terminology being used, revised and pushed through a whole workflow. After a translation has been round up, terminology assets can be evaluated with metadata stored inside tags during the process. Tagged terminology and tagged translations can be extracted from translated

documents. Updates can be committed, new suggestions for terms can be made, and noise terms can be removed from the database. Without tagging, valuable information gathered during a translation project is discarded and no feedback data is available to improve terminology quality in the term base.

6. Challenges for a language service provider

As a language service provider we make it our business to organize large translation projects. We are highly specialized in converting standardized formats into proprietary formats and the other way around. We input and output data from and to a wide range of systems – worldwide – for our translators, who are not always equipped with the latest technologies. Handling terminology data plays

an important role in all this. It is obvious that quite some processing is required to overcome the impedance mismatch between our clients' output of terminology data and our own systems. There is an interesting choice to make here though: we can choose to process, translate and just do our job, or we can standardize, measure, optimize, deliver feedback and eventually try to improve our clients' understanding and handling of terminology.

The last choice is more or less a description of the Kaizen approach we use at Yamagata Europe (De Sutter 2005, p. 22):

“go to the shop floor, observe, measure and take immediate action [...] the immediate action is only a temporary countermeasure. It is more important to find the root cause of the problem and subsequently standardise so as to create the necessary procedures and tools to prevent a reoccurrence.”

In what follows, we will outline how we deal with the difficulties described in the previous paragraphs. We will first have a look at how we cope with unorganized terminology. Then we will discuss how we address the lack of integration. Eventually, we will show how a solution to the problems described, can be offered to clients as a value-added service.

7. The project

In a project we ran last year, the challenge was to translate a user manual with complex translation rules into 24 languages. Our client provided us with:

- a set of spreadsheets containing
 - untranslatables
 - terminology entries with complete translation
 - terminology entries with incomplete translations
 - terminology entries without translation
- a set of rules to identify new terminology
- a set of rules to identify text in user interface elements
- a very sharp deadline

8. The standard workflow

In an everyday simple workflow, clients deliver terminology lists along with documents that need to be translated. The quality of these lists is very unstable. Sometimes we receive clean database exports. Other times, we get a bunch of disorderly compiled spreadsheets with a lot of noise.

If a project has been prepared well by our client's localization specialist, we may receive reference documentation. Usually this comes as a pdf document with terminology marked according to our client's instructions. For example, red is used for untranslatables, blue for compiled terminology, and green for new terminology.

When a project starts, translators receive bilingual files for TagEditor (our preferred CAT-tool), a translation memory, clean Excel files and a pdf reference document. This way we are sure that a minimally equipped translator can do the job.

Note that at this moment, we present a simple system consisting of non-standardized compiled lists and a document with translation rules compiled by a human.

9. Quality Assurance

Yamagata Europe has been specializing in the quality assurance of translation projects. Let us have a look at how we implement quality assurance.

Traditionally, quality assurance in translation focuses on the three P's of quality assessment (Stejskal 2006, p. 41):

"[...] provider, process and product. The provider is a translator or a translation company, that is, a physical or legal person. The process is a sequence of steps used to produce a target text (translation) that corresponds to the source text (the original document). Finally, the product is the translation itself. The quality assessment method will be very different for each of these."

At our company, all three P's have been covered by an integrated project management system that keeps track of translator selection and performance, our translation processes, and the final product quality.

Our procedures to check consistent and correct translation of terminology are a part of the second P.

To monitor the quality of terminology during translation, we devised a methodology that takes software unit tests as a model. In software unit testing, any testing procedure must be atomic and isolated, reveal the intention of the test, be easy to set up, and fast to carry out³. From this point of view, the proposed standard workflow can inherently not be tested.

We will explain why, by first summarizing our intentions briefly:

- We want to know if the non-standardized collection of terms can be used by translation tools.
- We want to investigate if the set of instructions is communicated adequately, i.e. if the coloring of the pdf file is carried out well.
- We want to make sure that the translation of each term is carried out according to the instructions.

Unless a logging system is added to this workflow, and objective checking parameters (such as an allowed error margin on a given total of terms) have been defined, we cannot effectively check the quality of each job. A second person will be needed to check the first person's work. But this does not check what we want to know. It only checks whether the first person has done his job well. We do not even have an objective source to tell us how many terms need to be checked. Moreover, no new specific information about the terminology is added and we do not know what exactly happens to terms during each single stage. Finally, the process is not easy to set up and certainly not fast.

³ <http://codebetter.com/blogs/jeremy.miller/archive/2005/07/20/qualities-of-a-good-unit-test.aspx>

10. A first optimization: garbage in, quality assured terminology out

As a first step, we introduced a terminology cleaning and conversion tool. It helps us to clean out the terminology spreadsheets (remove doubles, detect inconsistencies, remove unwanted characters such as redundant spaces and hard returns) and to standardize and convert the lists in the most common formats translators use. A log is also created to keep track of removed and inconsistent terms. By using the tool we can assure fast and easily that all translators use the same terminology and that corrupted terminology is discarded (see first intention).

Before the tool was created, we cleaned out files manually and for each target format, we used a different third-party conversion tool. Although we ended up with standardized files, the process was labor-intensive and error-prone, and more importantly, no logs were created.

There were also some specific problems with the non-Unicode MultiTerm 5 format, which is still widely used. Sometimes the conversion did not succeed immediately, because the operator was not aware of

technical problems that could arise and time was lost looking for someone who could solve the problem. Using this process, we managed to provide translators with standardized documents, but it usually took us four to six man-hours and sometimes the terminology files were sent to the translators after the translation had been initiated, which is far from ideal.

With the conversion and cleaning tool, we can now scan the received terminology for noise, process it with the tool, get an output of clean and standardized terminology databases in Excel, MultiTerm 5, MultiTerm 7, and if required, TMX⁴ and TBX⁵ format. We also get a log file that reports what happened to each single term. Processing time is reduced with at least 4 man-hours.

To be able to evaluate the translation of terminology, we generate a .dict dictionary file for QA Distiller™. This is an in-house developed, automated quality assurance application that helps us to detect incorrectly translated terminology. It generates an error log which can be used to correct terminology mistakes immediately within TagEditor. The tool is indispensable, not only in managing the correct translation of terminology, but also in managing the translation quality of our projects as a whole. QA Distiller™ makes it possible to automatically detect omissions, inconsistencies, and formatting problems, and is able to batch process files and supports language dependent settings.

If we look at the third intention we formulated, we see that we still need to consult the instructions to verify if a term was translated correctly, but we do have the advantage that all incorrect translations will be flagged for correction automatically.

11. A second optimization: objective and clear translation guidelines

We already briefly mentioned the discrepancy between terminology occurrence and terminology translation. Not every occurrence of a term needs to be translated and sometimes a complex set of cascading rules determines whether a term needs to be translated or not (for example: always translate this term, but when it occurs in a title, do not translate it, but add a translation between brackets instead).

⁴ TMX is a LISA standard for the exchange of Translation Memory (TM) data, but apparently there are people who use it as a terminology data format. You can find the TMX home page here:
<http://www.lisa.org/Translation-Memory-e.34.0.html>

⁵ TBX is the LISA standard for the exchange of structured terminology data, but the first three formats are far more popular (!). TBX homepage: <http://www.lisa.org/Term-Base-eXchange.32.0.html>

The traditional way of coping with this problem is to add a reference pdf file with highlighted instructions. This is a very inefficient way. During preparation, a localization specialist needs a lot of time to highlight the text correctly. To prevent mistakes, a second person needs to check the highlighting. With terminology lists consisting of up to 10.000 terms, this is undoable. More time is lost during translation, when the translator needs to verify from the document whether an occurrence needs to be translated or not. Note that a real life project often contains translations that have been reused from an earlier project. Sometimes these translations are locked and cannot be edited by the translator. This makes it not always possible to “follow” the highlighted reference text as you translate a document.

In the quality assurance phase, exactly the same amount of time is lost, multiplied by the amount of languages. QA-checkers need to consult the pdf file in order to find out whether a term occurrence has been translated correctly.

To deal with this problem, we developed an application, called TermTracker. It automatically detects and colors terminology and sections that need a translator’s attention, such as untranslatable chapter titles and navigation items. We also added support for regular expressions to define rules to color

terminology that had not been compiled yet. By using TermTracker we could adequately communicate translation instructions to translators and QA-operators (see second intention).

The application consists of an editor to correct wrong coloring, caused by the inconsistent requirements of the project, the poor quality of the database and the nature of language. To give some examples: sometimes the term “All” needed to be translated as an on-screen display, sometimes as a menu-item with a different translation as the one for on-screen display, sometimes it was to be left untranslated and in most cases, it was just part of speech and not terminology. Editing was done with a very easy click-and-change-color interface. All edit information was stored inside the document for later use.

The application engine generated .ttx files for TagEditor with colored terminology for the translators. We had a color code to signal how terms needed to be processed. Sometimes they had to be translated, other times they were marked as untranslatables, or as terms that needed a suitable translation after the original term between brackets, and so on.

Doing so, we realized the following optimizations:

- Instead of coloring, we had an operator “uncolor” the documents, which is a lot more objective, faster and less error-prone.
- We succeeded in conveying objective translation rules to translators. They no longer needed to decide for themselves how to process a term. The colors told them what to do. Subjective interpretation of the complex rule set was ruled out.
- A lot of time was saved because all reference information was inside the document on the correct location. No one needed to consult an external pdf document, losing precious time scrolling around. The translators could tell by the color what action was required and QA-operators could easily check this.
- We implemented full traceability of all terminology in the project (see next paragraph).

Eventually, we managed to make a difficult and error-prone process a lot easier for everyone. We reduced lookup and rule interpretation time with a factor 2N, with N being the amount of target languages of the project. Indeed, the interpretation of the translation guidelines/rules had to be carried out only once (with a check afterwards of course).

12. A third optimization: feedback on terminology quality

A final optimization, and a first step into a better integration of our client's terminology management, was achieved by tracking all terms. During the preparation phase of the translation project we colored terminology and we edited wrong coloring. All edit-information was preserved inside tags. These tags made it possible to track every occurrence of a term and its translation in the bilingual documents.

For example, occurrences that were not regarded as terminology were tagged as noise terms. Occurrences with multiple translations could be identified and marked for review. TermTracker was also able to collect statistical data about each single term and give an indication about the quality of term entries in the term base.

Finally, by giving each occurrence a unique identifier, TermTracker enabled us to extract new translations from the translated bilingual documents. Using the extracted translation pairs, we could update and add entries in the term base. Incomplete terminology could be completed with the assurance that the provided translation was correct for the given context.

13. Final thoughts

The project was a success, not only in terms (no pun intended) of methodology, standardization and optimization, but also in terms of integration. We were able to help our client look at terminology from a

totally new perspective. The result was a stronger relationship. We now build together an improved version of their term base, according to their needs *and* compliant with existing standards. The first time right, the Kaizen way.

References

- Childress, Mark D. (2006) "Terminologiphobia", *MultiLingual*, June 2006, p. 86
- Childress, Mark D. (2007) "Terminology work saves more than it costs", *MultiLingual*, April/May 2007, p. 43-46
- De Sutter, Nathalie. (2005) "Automated translation quality control", *Communicator*, Summer 2005, p. 22-25
- Dunne, Keiran J. (2007) "Terminology: ignore it at your peril", *MultiLingual*, April/May 2007, p. 32-38
- Fidura, Christie. (2007) "The benefits of managing terminology with tools", *MultiLingual*, April/May 2007, p. 39-41
- LISA. (2008) "SIG Mission and Charter." Terminology Special Interest Group, <http://www.lisa.org/Terminology-Special.102.0.html>
- Massion, François. (2007) "Terminology management a luxury or a necessity?", *MultiLingual*, April/May 2007, p. 47-50
- Muegge, Uwe. (2007) "Why Manage Terminology? Ten Quick Answers", *The Globalization Insider*, http://www.lisa.org/globalizationinsider/2007/07/uwes_article.htm
- Ray, Rebecca. (2007) "My 5 Tips for Intelligent Terminology Management", *The Globalization Insider*, http://www.lisa.org/globalizationinsider/2007/08/my_5_tips_for_i.html
- Rirdance, Signe. (2007) "IP vs. Customer Satisfaction: EuroTermBank and the Business Case for Terminology Sharing", *The Globalization Insider*, http://www.lisa.org/globalizationinsider/2007/08/ip_vs_customer.html
- SDL International. (2007) "Terminology matters", SDL International, http://www.lisa.org/index.php?eID=tx_nawsecuredl&u=0&file=fileadmin/filestore/wp/SDL_Terminology_Matters.pdf&t=1213695815&hash=ae6829537caeb62ac1175893f3848b48
- Stejskal, Jiri. (2006) "Quality assessment in translation", *MultiLingual*, June 2006, p. 41-44
- Warburton, Kara. (2003) "The Terms of Business: Saving Money Through Terminology Management", *The Globalization Insider*, November 2003, http://www.lisa.org/globalizationinsider/2003/11/the_terms_of_bu.html
- Warburton, Kara. (2005) "Terminology: Getting Down to Business", *The Globalization Insider*, July 2005, http://www.lisa.org/globalizationinsider/2005/07/terminology_get.html
- Warburton, Kara. (2006a) "Terminology as a Key Driver in Business Communications, Bridging the Divide Between Authoring and Translation", *The Globalization Insider*, April 2006, http://www.lisa.org/globalizationinsider/2006/04/terminology_as.html
- Warburton, Kara. (2006b) "LISA Promotes Terminology Standards in Beijing", *The Globalization Insider*, September 2006, http://www.lisa.org/globalizationinsider/2006/09/lisa_promotes_t.html

Warburton, Kara. (2008) "Introduction to terminology management", IBM Corporation, <http://www-306.ibm.com/software/globalization/topics/terminology/introduction.jsp>

Wittner, Janaina. (2007) "Unexpected ROI from terminology", *MultiLingual*, April/May 2007, p. 51-54

Zerfaß, Angelika. (2007) "Terminologieprüfung", *eDITion*, Ausgabe 2 2007, p. 18-20, http://www.iim.fh-koeln.de/dtt/Dokumente/edition_2007_2_webartikel/edition_2007_2_zerfass.pdf