Workshop: Lost for Words—Maximizing Terminological Quality and Value at an LSP

David J. Calvert
TransForm Gesellschaft für Sprachen- und
Mediendienste mbH
Dürener Str. 177–179
Köln 50931
Germany
d.calvert@transformcologne.de

Abstract

Part 1: Theory. Although the economics of the business preclude large-scale investment in terminology, I believe that an iterative approach to collecting and improving terminological data can pay off. The quality and value of terminology are discussed from an LSP's viewpoint and defined for an LSP. The features of an optimal terminology process and the process' relationship to the ISO17100 translation process are identified. The interests of the other parties in the translation process are reviewed and best practices for terminology work are identified for the different parties involved. The objectives of a terminology process are formulated and discussed. The features of two standard terminology modules are compared and my choice of terminology server is explained. A standard terminological record structure for termbases is introduced. Part 2: Practice. The second part of the workshop will present an implementation of termbases using this term record structure. This will include the ways in which TransForm is dealing with the strengths and weaknesses of the terminology server used and an iterative process for improving the value of terminological records. Different approaches to automatic term matching will be evaluated, with particular attention paid to the problem of false positive results in QA checks.

1 Theory

Terminology work is often written about and discussed. Yet the terminology work discussed in conference papers and academic textbooks is mostly concerned with single-language terminology and starts from a completely different perspective to that of a language services provider (LSP) or translation services provider (TSP).

1.1 Why do we do it?

I am looking at the subject of terminology from the point of view of a small LSP. My company specializes in various forms of communication, mostly concerned with corporate image as presented to customers, employees or more specific target groups. A large proportion of our work comes from corporate publishers and is destined for publication in print, online or on multiple channels. The range of subjects covered is correspondingly broad, so we have to deal with a wide range of specialized areas, many of which have their own specialized terminology.

Even within specific subject areas, different clients follow different external and internal standards, and may use different regional variants of their corporate language for different parts of the company.

So we need to keep track of terminology—to ensure that we use the appropriate term for the language variant, for the customer, for the subject area, and for any applicable standard. This is a quality-based argument. There are also economically based arguments for terminology work. These include lower costs thanks to a reduction in the amount of research

Proceedings of the 38th Conference Translating and the Computer, pages 1–9,
London, UK, November 17-18, 2016. ©2016 AsLing

necessary prior to or during the translation and review phases, fewer complaints and increased customer loyalty.

Although the economics of the business preclude large-scale investment in terminology, I believe that a well-planned iterative approach to collecting and improving terminological data can pay off for an LSP.

In short, we do it because it saves money and makes our lives easier.

1.2 What are we doing?

“Terminology is the study of terms and their use,” writes Wikipedia.[1] That sounds logical, but it doesn’t go very far.

TermNet introduces its website with a quotation from Confucius.

ISO TC 37 defines a terminology as “a set of designations belonging to one language for special purposes” and goes on to define such a language as “a language used in a subject field and characterised by the use of a specific linguistic means of expression.”

Pavel, in her *Handbook of Terminology*[2], offers two definitions: “The first meaning of the word ‘terminology’ is ‘the set of special words belonging to a science, an art, an author, or a social entity,’ for example the terminology of medicine or the terminology of computer specialists.” She then goes on to say, “The same term, in a more restrictive sense, means ‘the language discipline dedicated to the scientific study of the concepts and terms used in specialized languages.’ General language is that used in daily life, while a specialized language is used to facilitate unambiguous communication in a particular area of knowledge, based on a vocabulary and language usage specific to that area.”

So it is clear that one term can have two different meanings, i.e. that there are two different approaches to terminology. For the purposes of an LSP, the first definition—a set of words with specific meanings in a specific context—is what we need. The second is the province of professional terminologists, and of practitioners of computational linguistics. As an LSP, we may sometimes rely on the work of such people, but their skills do not form part of our core expertise. It is also important to note that the subject fields referred to in ISO TC 37 span all areas of human activity including commercial activities within vertical industrial or economic sectors[3], so terminology can also be taken to include terms such as department names and job titles, which can be very important to an LSP.

1.3 What are we not doing?

Source language terminology is the customer’s job. Any work we do here is a by-product unless the customer is specifically paying us to work on their terminology—in which case they are probably paying the wrong people.

There appears to be a disconnect between expressed belief and real-world practice, at least in Germany, where an online survey in 2013[4] found that 2/3 of the 504 respondents believed that consistent terminology made work substantially easier or easier, a slightly smaller proportion believed it saved time, and well over 80% believed it made similar improvements in quality and customer understanding of technical documentation. The same survey found that over 40% of the respondents stated that terminology was of little or very little importance in their company.

[1] https://en.wikipedia.org/wiki/Terminology

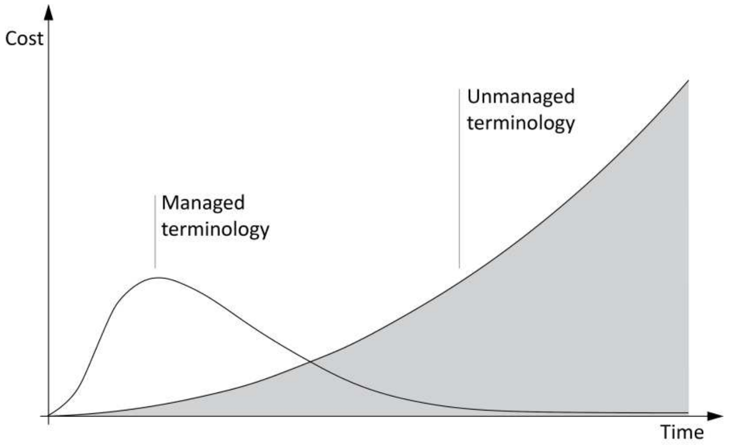
[2] http://itia.ir/farsi/documents/ha.pdf

[3] Warburton, K. after Rondeau, G. Tekom Proceedings tcworld 2013, CHAT 1.

[4] Straub, D and Schmitz, K-D., Tekom Proceedings tcworld 2015, TERM07

1.4 Is it about the money?

The financial return of source-language terminology work can be quantified. Approaches usually attempt to define expenditures and resulting cost savings in order to establish a ROI. These can range from simple “back-of-an-envelope” calculations to more detailed models such as that implemented by ZVEI—the German Electrical and Electronic Manufacturers’ Association—in an Excel spreadsheet. The pain curve illustrating the costs of managed vs. unmanaged terminology provides a conceptual basis for this type of calculation.[5] It is important to note that these models are intended for use by manufacturing companies, and that the primary focus is usually on source language terminology. The cost-benefit considerations for translation are usually a combination of subsets of those for technical documentation and marketing communication, and as such may be substantially different to the overall picture.



Terminology pain curve (after Dunne, 2002)

1.5 Costs and benefits

Cost vs. benefit is different for every customer, and frequently for different jobs for the same customer. The amount of effort an LSP can dedicate to terminology work is extremely limited. Most customers are not prepared to pay for terminology work. They simply expect it to be right, although they may not always notice if it isn’t. So we do terminology work for the benefits it offers us as an LSP. Our investment is determined by the potential returns, so we have a much greater incentive to invest on behalf of a customer offering a relatively high volume or one providing intermittent but well-paid work. For us, terminology tends to be either rushed for a new customer, or slow and steady for an established one.

1.6 How we work

Terminology for translation must be: identified, researched, and recorded. It may be verified and further documentation may be added. It may then be published or fed back to the originating company. These activities give rise to costs. The first cost, that of the system for storing and managing the terminology, may be covered by the purchase of a computer-aided translation system which includes a terminology management application. Some such systems provide only basic terminological functions, and it may be necessary to purchase a separate program to provide adequate functionality.

Directly attributable costs for terminology arise when a customer delivers terminology associated with a particular project, or when the decision is taken to invest in preparing terminology for a job. Costs for maintaining terminology are also directly attributable. I do

[5] R. Herwartz: „Wann macht sich Terminologie bezahlt? Erstellung einer Kosten-Nutzen-Analyse“, in: technische Kommunikation 5/2011, pub. TEKOM

not see scope for a small LSP employing terminologists on anything other than a contract basis related to specific jobs.

The other costs associated with terminology work tend to be difficult to isolate from general overhead.[6] This is also true in an LSP environment, where a significant amount of terminology is identified, researched and (hopefully) documented during the translation and revision processes.

1.7 Types of project and the importance of terminology work

Both the costs and the savings underlying the pain curve follow significantly different patterns for the different types of project dealt with by an LSP. In our case, these types can be classified by volume, frequency, and nature of material.

Recurring regular projects such as employee magazines with a regular publication cycle have characteristic terminological requirements. Such magazines are usually published by corporate communications departments, often with the help of a corporate publisher. Here the target readers are corporate employees, possibly at plants spread throughout the world, and the main purposes of the magazine are promoting a universal corporate culture and a sense of belonging along with conveying essential company information. Articles in such magazines often showcase specific departments, product developments, or management initiatives. Here, vital terminology starts with names—of departments, initiatives or products. Getting the Vice President’s department name or job title wrong is as bad as misspelling his or her name. Advance investment in terminology is strongly advisable for this type of project, as is a good system of documentation for terms such as feature names. At the very least, a copy of the previous issue in the target language will clear up the question of whether the section was entitled “In Focus” or “In the Spotlight”. Careful TM maintenance also helps in this area. There is often a great deal of overlap between documenting terminology and creating a complete style guide.

In-house magazines for industrial companies also frequently require cooperation between internal and external translation providers. End customers frequently require that individual features, special sections or even whole magazines concentrating on research or innovation be translated in house due to the technical nature of the texts. The corporate publisher will then require translation of headlines, captions and general texts such as editorials followed by at least a copy desk process where conformity with English grammar, spelling, and the house style is checked. Here the problem of maintaining consistency is greater, as there is often no access to a source text, and so no way of applying automated checks for terminological consistency.

Company reports, e. g. quarterly or annual financial reports, or reports covering sustainability and/or corporate social responsibility require the use of specialist terminology defined by specific organizations and subject to change. In particular, financial reporting in Europe usually makes use of the International Financial Reporting Standards and International Accounting Standards defined by the International Accounting Standards Board. These standards are subject to change every year. Similarly, sustainability reports are often subject to the guidelines of the GRI Global Reporting Initiative.

Reports generally involve a significant effort to establish source and equivalent target terminology before the translation of the first issue by an LSP. This effort will usually involve previous issues if available, plus general terminology from standards such as IFRS. At this point, problems such as mismatched regional variants may become apparent, e.g. when a company with US English as its corporate language publishes an annual report on the basis of IFRS/IAS, which are written in British English. The process of terminology collection prior to the first issue is usually similar to but more intensive than that required for magazines.

[6] TSS2009_FS_EconomicIssues, Prof Dr Frieda Steurs, Lessius/KULeuven

Terminology for projects belonging to an account with a more or less regular flow of jobs with common topics, e.g. press releases, technical documentation and websites, follows the classic pain curve, with an initial peak subsiding to a low level of effort required for the addition of new terminology and occasional weeding out of obsolete or deprecated terms.

Terminology for projects belonging to an account with an irregular flow of jobs or covering a wide range of (non-repeating) topics is usually less cost-effective, as the initial peak of the pain curve repeats with every new topic. The decision to invest in terminology is on the basis of risks/rewards for the individual job plus a speculative component dependent on the likelihood of the customer sending more regular work.

It is seldom worth carrying out substantial terminology work for apparently one-off jobs with no reasonable expectation of follow-up work, e.g. contracts or static websites. Here, the time covered by the pain curve is the duration of the single project, so the peak of cost due to terminology management may be greater than the costs incurred by not managing the terminology. The decision to invest in terminology is on the basis of risks/rewards for the individual job. One significant exception to this is where the LSP has been brought in to work on a pitch. Providing a limited amount of terminology work as part of a pitch is clearly a gamble, but does demonstrate the team's commitment to quality. This is also a good way to increase customer loyalty.

Terminology linked to a specific account is not available for general use, as it will contain company-specific material and material subject to copyright and confidentiality.

Terminology that is not linked to a specific account is available for general use but is strongly constrained by subject area.

1.8 Risks associated with terminology

The risks associated with incorrect use of terminology are a subset of those associated with incorrect translation. The consequences range from causing amusement among colleagues to bearing responsibility for death or injury due to incorrect operating instructions or service documentation. By drawing up matrices of likelihood of specific consequences occurring vs. the consequences themselves for specific types of terminology error we can determine the level of risk posed by incorrect translation of terminology. Possible immediate consequences can be graded in order of increasing severity, from e.g. Internal communication impaired, no material consequences, to Danger to life or limb. This approach makes clear that while getting the job title of an executive or the name of a department wrong will lead to embarrassment and may lead to a loss of trust, the overall risk is less than that incurred when a product name or description is wrong, as there is a significant risk of expensive corrections at a late stage in prepress work, or worse if the presses have already started to roll.

1.9 Cooperation

The key factor in enabling effective cooperation is making it easy by removing barriers. Translators will not provide services free of charge if they do not see an immediate and direct benefit from doing so. The same applies to convincing in-house staff to willingly identify and record terminology.

1.10 Relationship to ISO 17100

The translation workflow as specified by ISO 17100 Annex A only mentions terminology once, under Section 4, Pre-production processes and activities. It is specified as an optional step in point 4.6.3.2, which states that "…the client and the TSP can agree that the TSP shall ensure that the appropriate terminology is available…". Point 5.3.1 a) of Section 5, Production process, specifies compliance with domain and client terminology and maintenance of terminological consistency. A significant part of the challenge for LSPs is to obtain and validate the terminology in the first place, and this is an area where the tool vendors are a long way from supplying optimal solutions. Although the ability to capture terminology on the fly has been around for some years, there is no simple way of returning such captured terminology as part of a job package.

1.11 Interested parties

The interests of the client are best served by delivering a translation which does not expressly contradict the end client’s existing documentation and material, unless such contradiction has expressly been requested, as part of a product relaunch, for example.

Subcontractors usually want to deliver a product which conforms to the customer’s expectations at the lowest possible cost to themselves.

For suppliers, the best practices in the translation process can basically be summed up as consistency, documentation and communication. Consistency, because it makes problems easy to fix; documentation, because it makes it possible to recognize and avoid the problem the next time around; and communication, because it ensures that people are aware of both the problem and its solution. The most constructive practice from the LSP side is to facilitate and encourage feedback of terminological problems and of the proposed solutions from suppliers. Naturally, this requires LSPs to form close, long-term relationships with selected suppliers.

For clients, the picture is more varied. From the LSP’s viewpoint, the most important best practice is the use of professionally prepared source language terminology in source language documents. The second most important one is to have their target language terminology reviewed by someone who is both familiar with the domain and a native speaker of the target language. Generally, however, the LSP’s role here is limited to asking what, if anything, exists and is available.

It is also important for LSPs to distinguish between different types of client. Agency and publisher customers rarely have the expertise or the need to receive terminology in any form other than a glossary supplied as a PDF. In-house translation departments, on the other hand, are more interested in receiving terminology in a form compatible with their system. End customers will often have their own specific input format specific to their implementation of a terminology database.

1.12 Subject-specific challenges

Different customer accounts present different challenges. Linguistic problems are always present. For example, translating German financial reports into English involves problems such as the German word *Rechnungsabgrenzungsposten*, which translates as *prepaid expenses* when it appears on the assets side of the balance sheet and as *deferred income* when it appears on the liabilities side. Or the German word *Umsatz*, which is variously translated as *sales*, *revenue*, *revenues* or *turnover* for different German companies. If the original accounts have been worked out according to the German HGB standard, then many of the terms used will be conceptually different from English accounting terminology and the text will require a degree of localization. Researching specific subject areas can be problematic; for example, “older” areas of industry such as railway technology are not as well documented online as IT and telecommunications. Technical issues such as the nature and format of available terminology also arise and call for different approaches.

1.13 Starting points

The most common starting point for terminology work for a new account is probably one or more PDF documents. These may be exports from an in-house system, or (possibly protected) PDFs of last year’s annual report. Excel spreadsheets are also popular among users. Possible challenges here include problems caused by the fact that Excel’s default text delimiter varies according to the regional settings of the version of Windows under which it is running. For example, the straight double quote used by Excel in English is also the symbol for inches and can cause problems in Excel glossaries.

End customers’ terminology is usually in a form suitable for the customer’s own use, i.e. arranged as a dictionary or glossary. It usually has not been lemmatized or edited for automatic term recognition.

1.14 What is quality?

The idea of “fit for purpose” is a fundamental tenet of quality assurance. There is no point in wasting effort on producing something that exceeds the required specification. The primary purpose of terminology work at an LSP is satisfying the customer. So it follows that the main considerations on the quality side are:

- Customer acceptance
- Consistency
- Correctness

Correctness here is taken to mean that the term is intelligible to the rest of the world—the Humpty Dumpty problem—and that it does not contradict other established uses. Trade-offs between customer acceptance and correctness are almost always decided in favour of customer acceptance—at least initially.

However, from an LSP’s point of view, we also want to maximize returns and minimize costs. These objectives are achieved by optimizing the content of our termbases and the automatic term recognition settings to maximize the hit rate of terms recognized, while minimizing the rate of incorrect recognitions and false positives generated during automatic quality control. From the LSP’s point of view, the cost-effectiveness of a termbase is its main quality criterion.

1.15 The ideal and the real world

The ideal customer has a well-defined collection of source-language terminology put together by a professional terminologist and coupled with target-language terms approved by in-country reviewers with the relevant expertise. And this terminology is available in TBX or some other form of XML.

In practice, one or all of these features will be missing. Even where the target language terminology exists, it may well have been prepared by interns or students, hopefully working under the direction of a terminologist. It may have been obtained from the development department, and be heavily influenced by the source language, or from an overseas subsidiary, and have little relationship with the source language concepts. Or it may have been crowdsourced.

The LSP’s task here is to convert any existing terminology into a cost-effective termbase with the minimum of effort.

1.16 So where do we want to go?

We want to abolish duplication of effort.

We want to be able to benefit from our efforts by reusing their results.

We know that the journey never ends.

1.17 And how do we intend to get there?

We have to establish a working system and ensure that it minimizes effort and maximises returns. The situation represented by the terminology pain curve is, however, an idealized representation of the cost of terminology for an end customer. It does not take into account the effects of such events as new product launches or version releases, let alone corporate reorganizations or changing documentation standards. There is also little point in an LSP implementing monthly updates to a termbase if the termbase is only used for translating a customer magazine twice a year. This is even more so when the updates need significant effort to port them into the TLS’s system. So the real picture of terminological cost is characterized by occasional peaks either immediately prior to the translation of an issue or immediately after, when feedback has been received after the customer’s review of the translation.

One effective mechanism for improving existing terminology is by iterating through feedback loops. In addition to documenting new terms, reviewers can note problems with existing terminology. The logs from any quality control tool used provide valuable indications of which terms are causing false recognition results and how the results from the termbase can be improved.

1.18 The journey to date

TransForm’s first termbase system was MultiTerm in its file-based incarnation as a part of Trados Translator’s Workbench for Windows. As the technology vendors moved from file-based to server-based systems, software costs for operations of our size increased dramatically. File-based systems were effectively removed from the market and the capabilities of non-server systems were restricted to single users on networks without domain controllers. Server-based systems for around five users were relatively expensive, so we had to find an alternative strategy. This was achieved by making increased use of Wordfast, which was already in use as our backup system and was widely used by our freelancers. We also used an intranet-based system for collecting terminology on a project basis.

Wordfast uses glossaries for terminology. For Wordfast Classic and Wordfast Server, these are simple tag-delimited text files with the first three fields defined as Source Language Term, Target Language Term, and Comment, and three further, user-definable fields which can be used for attributes. The glossaries for Wordfast Pro 3 and 4 can also import TBX, although only a subset of TBX can be accommodated by a glossary structure. Wordfast also enables the use of Blacklists of forbidden terms. Wordfast distinguishes between automatic fuzzy terminology recognition, which does not require editing the source terms in the glossary, and manual fuzzy terminology recognition, which makes use of asterisks in the source terms as wildcards.[7] The asterisks can be placed at the beginning or the end of the term, or in the middle of the term. The trade-off between automatic and manual terminology recognition is less accuracy vs. more initial effort required.

memoQ Translator Pro and memoQ Server use a concept-oriented termbase structure. However, the termbase definition is fixed, and the user is limited to Kilgray’s choice of fields. It has the classic TBX-style three-level structure with concept, language, and term levels. It also contains Kilgray-specific fields and, as previously mentioned, omits some fields which are extremely useful to LSP users. However, Kilgray also supplies a terminology server, known as QTerm. This runs on the Web server integrated into memoQ Server and supports user-defined fields within the three-layer structure. Some of the Kilgray-specific fields from the standard terminology module are included in the termbases to maintain compatibility. Forbidden terms can also be stored in memoQ. However, unlike in Wordfast, they are stored within the termbases, and are distinguished on the term level by a “Forbidden” attribute. This apparently has certain implications for the fragment assembly and predictive typing features of memoQ.[8]

[7] https://www.wordfast.net/wiki/Fuzzy_Terminology_Recognition

[8] Thread “Feature request (or does this already happen?): no forbidden TB entries in predictive typing,” memoQ@yahoogroups.com, March 2016

After trying out memoQ as a TM system I came to the conclusion that it offered a high degree of interoperability and was the best choice for our current operation in terms of capabilities and cost-effectiveness. However, the limitations of the built-in termbase made it necessary to purchase the QTerm terminology server extension for the standard memoQ Server.

This history has led us to define a standard terminological record that offers our ideal balance between the effort put into collecting terminological data and the scope for its current and future utilization.

1.19 The TransForm standard terminological record

The structure of the standard terminological record used at TransForm was originally defined for terminology collection via a form in the translation management database in the company intranet. It was derived from TBX-Basic and automatically associated metadata from the translation management database with source-target term pairs, thus building up account-specific glossaries that could be imported into other systems via TBX-Basic. Our move to QTerm termbases required modifications to the structure to accommodate memoQ-specific fields necessary to maintain compatibility.

The Concept (termEntry) level contains the standard transactional information on creation date and user and last modified date and user. It also contains the memoQ built-in fields for Domain, Subject, Client, and Project (metadata), and Image and Image caption fields. The memoQ termbase field Note and an ID field are also present.

The Language (langSet) level contains the memoQ built-in field Definition. This is directly equivalent to the descrip tag in TBX-Basic.

The term (tig) level contains the term itself and the fields Term source, Usage example, Usage example source, Note, Term type, Validation status and Validated by. It also contains three built-in QTerm fields: Case Sensitivity, Matching, and Forbidden. These are necessary to maintain compatibility with memoQ, in particular with the QA Check feature. The memoQ termbase fields Part of Speech, Number (grammar) and Gender (grammar) have also been included to retain compatibility with memoQ.

The key elements of this structure are the three-level concept-based structure itself and the use of specific fields. In particular, the compatibility with the TM system ensured by the memoQ built-in fields benefits term recording and recognition. However, one of the key factors behind the choice of QTerm instead of simply using the standard memoQ termbase was the need to define a term source field. This is because the source of a term is an extremely useful proxy for the term’s reliability. If a term is used by the customer in the customer’s documentation there is little scope for disagreement about the use of the term. Similarly, the documentation of both a usage example and the source of that example provides a known degree of confidence in the reliability of the term in context.

The Validation and Validated by fields have been brought over from the TransForm intranet terminology record, where they were intended for use as elements in an EN 15038-compatible terminology process.

1.20 To be continued…

We are currently have 20 years’ worth of terminology collected and partly duplicated across four different systems. We are in the process of establishing which parts of this data are worth porting to QTerm and of unifying and porting the data selected, and of optimizing those parts of the data that have already been ported.

The second, practical part of the workshop will look at how some of these ideas and approaches are being implemented at TransForm GmbH.