Terminology in the age of multilingual corpora

Alan K. Melby, Brigham Young University[1]

ABSTRACT

Terminology management has long played an important role in translation and localisation. It has been asserted, however, that the need for terminology management is declining with the rise of widely accessible aligned multilingual corpora, such as bi-texts. In this view, translators will be able to identify terms and their translations by using previous translations to automatically identify the best translation for a term. This article, however, argues that while bi-text resources will assist in human-oriented terminology management, they will actually increase the need for skilled terminology work and termbases. Furthermore, because more tools will generate terminological data, the need for exchange between tools will increase. After discussing the case for terminology management and terminology exchange in the age of aligned multilingual corpora, the paper describes the role of the TermBase eXchange (TBX) standard in terminology exchange, including typical scenarios for its use, and some of the challenges faced in using it.

KEYWORDS

Concept-oriented terminology, data exchange, glossary, interoperability, aligned multilingual corpus, multilingual corpus, bi-text, TBX, termbase, TermBase eXchange, terminology, terminology management.

> "There is unfortunately no cure for terminology; you can only hope to manage it."
> — Kelly Washbourne (personal communication)

1. Introduction

In recent years corpus linguistics has seen tremendous success in translation-related fields. For the purposes of this article the term *multilingual corpus* refers specifically to aligned corpora, collections of texts and their translations that have been segmented (i.e. divided into linguistic units, such as sentences) and aligned so that the segments in one language correspond to those in the other. The most common multilingual corpus is bilingual and consists of a collection of *bi-texts*. Statistical machine translation based on such corpora has become widely used (while still remaining the subject of considerable debate), as evidenced by tools, such as Google Translate, Bing Translate, and the open-source Moses project. The TAUS Data Association has been successful in encouraging erstwhile competitors to share translation memory data with each other in an attempt to improve translation memory results (and to seed machine translation efforts).

Despite these successes, however, the impact of corpus resources on terminology management is still to be seen. If terminology management is understood broadly to include all terminology-related activities (as

discussed in section 2 below) then multilingual corpora have the potential to transform and simplify many aspects of terminology management. Corpus-based resources, however, will likely not eliminate the need for terminology management, for reasons to be discussed in this article.

2. Managing Terminology

Before addressing how multilingual corpora will impact terminology management, it is important to establish what is meant by a *termbase* in this article and whether termbases are still useful to translators, who can now access terminology in ways other than by using termbases. For example, if termbases could be entirely replaced by searching on-line collections of texts and their translations — a possibility explored by Bowker (2011) — then termbases would be generally irrelevant to today's translator. The argument concerning the irrelevance of termbases, however, combines different sorts of terminology resources and may not apply equally to them all; it is important, for example, to distinguish in particular between glossaries and termbases. This article will show that although one type of two-column, bilingual glossary may, in some limited circumstances, no longer be necessary to the everyday work of a translator, both termbases — as understood here — and the exchange of information between termbases are important to the multilingual document production chain.

A *termbase* is a computer database consisting primarily of information about *domain-specific* concepts and the terms that designate them. Specialised translation deals with *domains* of knowledge, and every domain is organised through *concepts* that are linked to objects or ideas relevant to that domain. Termbases may be monolingual, bilingual, or multilingual[2].

A termbase is therefore organised differently from an electronic, general-purpose lexicographical dictionary where an entry consists of all the varied senses of a headword used in general language. For example, a large general dictionary entry for the headword "pig" will probably list an animal sense (a swine), a metallurgical sense (a crude casting of metal)[3], a slang sense (a boorish and uncultured person), and so forth. In contrast, a termbase is organised by concept rather than headword and represents domain-specific knowledge that has been divided up into concepts by consensus among experts in a particular domain, such as "lung cancer," "agriculture," "integrated circuits," or "metallurgy" and so on[4]. If a term has more than one meaning, each meaning would be contained in a separate concept entry. In general, a termbase would not even contain uses of a term from outside of its particular domain; e.g. a metallurgical termbase would contain the metallurgy-specific term and meaning of *pig*, but not the agricultural term.

Concept orientation does not imply that concepts are universal across all cultures and time periods. Terminological concepts are the creation of the experts who work in a particular domain. They evolve over time, but hopefully, for any well-established domain of knowledge, there is a well-defined set of concepts understood by most members of the discourse community.

2.1. Termbases vs. glossaries

There is an uneasy relationship between termbases and glossaries, as explored in this section.

Although glossaries and termbases share certain similarities and the term 'glossary' is often used as a generic referent to any terminological resource, a glossary may or may not be equivalent to a termbase. There are two primary types of glossaries. Depending on which type is meant, a glossary may be equivalent to a termbase or it may have more in common with general-language lexicographical resources, such as bilingual dictionaries.

The first type of glossary is a monolingual collection of terms and definitions that are relevant to a particular domain. This type of glossary is often created as an aid for authors that lists company — or project — specific terms and their definitions to help ensure that authors follow a particular language style. This type of glossary could be represented as a monolingual termbase.

The second type of glossary is a two-column list of terms, created from scratch (and not exported from a termbase), in which the first column displays terms in one language and the second column displays corresponding terms in another language, so that each term in one language matches up with exactly one term in the other[5]. Lists of this sort are frequently created by translators in tools such as Microsoft Excel. Lists may serve as mnemonic aids to help ensure that terminology is consistent, by recording decisions they have made about how to translate particular terms in one document. If the terms in such a glossary are not all from the same domain, then not even the minimal requirement for a termbase (concept orientation) is satisfied. Thus, a two-column, bilingual glossary, consisting of a list of various terms in two languages, without further information such as the domain to which they apply, would *not* qualify as the content of a termbase.

2.2. Structured terminology

Leaving two-column bilingual glossaries aside, the concept entries in a termbase are typically sub-divided into sections: a section describing the concept as a whole, a section for each language, and a section for each term. The following information is often found in these sections. Using a bottom-up approach, these sections are focused on terms, languages, and the concept they all designate[6]:

- Term sections each consist of one term and information about the term, such as:
  - term type (full form, acronym, abbreviation, etc.)
  - part of speech (noun, verb, etc.)
  - contextual example of the term in a sentence or paragraph
  - customer code (if this term is specific to a particular customer)
  - project code (if this term is specific to a particular project)
  - responsibility (if more than one person works on this termbase)
  - cross-reference to another term (if applicable)
  - usage note (usage notes can, for example, indicate regional variation).
- Language sections each consist of one or more term sections, and information about the language section as a whole:
  - language of the terms in this language section (required)
  - definition of the concept (optional).
- Information about the concept as a whole, as designated by all the terms across all the language sections of the entry, includes:
  - domain to which the concept belongs (required)
  - link to an image illustrating the concept (optional)
  - Note: Definitions can be at the concept rather than language level.

The information in a termbase is divided into discrete units of information often called *elements* or *fields*, which can be searched, modified, and manipulated individually. Each element is associated with what is called a *data category*. So far, this article has listed only the most common data categories found in corporate termbases (LISA Terminology SIG 2008: 4). This list includes only a small portion of the data categories used in termbases; many more are available. The International Organization for Standardization (ISO) has published an international standard for the data categories used in termbases: ISO 12620. The 1999 version of ISO 12620 was a traditional standard published on paper, but the 2009 version of ISO 12620 is tied to an on-line database called ISOcat (http://www.isocat.org) in which there are several hundred data categories found in termbases around the world. While no termbase uses all of these data categories, every terminological data category in ISOcat is used in some termbase.

A useful set of data categories not mentioned so far consists of concept relations, such as generic-specific and part-whole (see Marshman *et al*. 2012 in this issue for a discussion of concept relations, which are also called terminological relations).

3. The impact of multilingual corpora and translation technology

With this background on termbases and glossaries it is worth examining whether increased use of multilingual corpora has impacted the need for termbases. Translation memories, statistical machine translations, and even lists of possible terms can be automatically derived from the same bi-texts. Non-digital bi-texts have been around for a very long time, at least since the Rosetta Stone was engraved about 2,200 years ago, but the term 'bi-text' was coined only a little over 20 years ago by Brian Harris (1988). A segment of text paired with its translation has always been important to descriptive terminology work. That importance is increasing as Translation Environment Tools (TEnTs) are becoming more common and are increasingly required in commercial translation activities. New techniques for utilising multilingual corpora are bringing machine translation (MT) and translation memory features in TEnTs together. Some translation memory systems not only retrieve segments of already translated text, they also match a segment in a source text with a similar segment in a translation memory database of previous translations and then adjust details of the target-language segment according to the source-text segment in question. Some translation memory systems also perform lookups at the sub-segment level and even suggest target-language terms. This is not far from the way statistical machine translation (SMT) systems automatically build tables of words and suggest translations of a piece of text from previous translations. This blurring of the distinction between fully automatic machine translation, accessed from within a TEnT, and other TEnT features may impact termbases.

In a recent presentation at the 2011 American Translators Association conference in Boston, Jost Zetzsche, a well-known figure among freelance translators for his translation technology newsletter (*http://www.internationalwriters.com/toolkit/*), claimed that "translator-created glossaries" are no longer needed, even though there will be an on-going need for client-supplied glossaries (Zetzsche 2011). This claim must be understood in the context of three other claims that Zetzsche made in the same presentation:

- The quality of translation memories is becoming much more important.
- Very large and diverse translation memories are not useful as replacements for glossaries and termbases; a translation memory should be derived from a set of closely related documents.
- High quality termbases will become more important as fewer two-column glossaries are created manually by translators (2011).

3.1. The importance of 'clean' data

Supposing that a corpus of bi-texts from a single domain exists, and further supposing that every term is translated consistently throughout the corpus, then a simple termbase could be automatically compiled from the bi-text corpus, using techniques from statistical machine translation and sophisticated translation memory software.

Multilingual corpora are, however, seldom as 'clean' as is required in the scenario described above. Automatically generated glossaries would likely contain mistakes (inaccurate pairings of terms)[7] and irrelevant terms or even non-terms, but after suitable checking and correcting by a human, such a termbase would be usable. In that case, a two-column, bilingual glossary created manually from that corpus by a translator would not be needed. There is much discussion currently in the translation industry of the need to clean up translation memories (and thus the bi-text corpus that underlies them). The following statement is from a company that specialises in 'cleaning' translation memories:

> We also understand that the quality of your translation resources may deteriorate [over] time, especially if many translators work on the same project and update the translation memory [TM] with their translations with average [or] no quality control. This is why many translation memories can sometimes get filled with inconsistent translations and errors. If you are also using a machine translation system along with your TM system, there are higher chances that your translation memory repository becomes contaminated with lesser quality translations (Linguaspot n.d.: online).

The company website goes on to offer services to spot erroneous translations, duplicate segments and untranslated segments, and to correct these errors. All this work may make a translation memory cleaner but still does not guarantee that the use of terminology will be consistent, which is a crucial aspect of 'cleanliness' implicit in the vision of Zetzsche. Note that the issue of clean and consistent terminology typically arises in discussion of translation, but problems frequently stem from inconsistencies in the source text that are undetected until the text is translated, and translators, striving to be faithful to the source text, may magnify them. In the words of Alison Toon of Hewlett-Packard, translators are "the garbage collectors of the documentation world" (as related by Arle Lommel, personal communication).

Uwe Muegge suggests that termbases remain vital in the age of translation memories and, by implication, multilingual corpora:

> Many language service providers use a translation memory system for storing and reusing translations. While it is true that a translation memory makes it possible to retrieve not only translated sentences but also sub-sentential elements such as terminology, this so-called concordance feature is no substitute for creating a

> termbase. Here is why. In the absence of a termbase, translation memories typically contain synonyms, i.e. multiple translations, abbreviated forms and variants of the same term, making it very difficult, if not impossible, for teams of translation professionals to consistently pick the same translated term. Also, using the concordance function every time a term occurs in a text to be translated is very time consuming and results in low productivity. And that's the best-case scenario where the term has actually been translated before: For new terminology, the translation memory system is no help at all (2011: online).

Muegge makes the important point that even if needed terminological information is embedded in a translation memory database (or the underlying bi-text corpus from which the translation memory was derived), a direct search of the data may retrieve either too much information or incorrect information or both. This may slow down and impede the translation process.

Zetzsche counters Muegge (personal communication) by pointing out that recent developments in translator tools, especially a feature sometimes called *subsegmenting*, overcomes the slowness of concordance searching by automatically inserting in the target text the most statistically common target language term or phrase. In subsegmenting, no distinction is made between classic terms and phraseological units. While this approach could lead to greater efficiency, it raises two important questions:

- How much work is required to clean up a translation memory or bi-text corpus to the point where subsegmenting provides appropriate terminology? Might it be more work than validating a termbase derived from a less clean bi-text corpus?
- Is the automatic insertion of target terms — resulting in total consistency — always the best approach in specialised translation?

Perhaps ensuring consistency of terminology in a bi-text corpus should be viewed as a type of terminology work.

3.2. Is absolute consistency desirable?

In fact, consistency might not be an absolute good. The question of whether the same target-language terms should always be used for a given source-language term is raised by Margaret Rogers, who points out that absolute, mechanical consistency of terminology may not always be the best choice for a translator:

> A re-orientation has therefore been suggested towards a view of term selection in technical translation which focuses on motivations rather than on a one-dimensional notion of consistency. Such an approach to translation decisions can clearly be seen as a part of translator competence and contrasts with what has been seen as an advantage of machine translation and computer-assisted translation systems over human translation, namely the automatic substitution of equivalents (2008: 112).

Rogers suggests that rather than mechanically substituting target terms for source terms, especially when the source text itself may be inconsistent in its use of terminology, a human translator should consider textuality, that is, context at a higher level than the current segment being translated. In other words, context beyond the words immediately surrounding a term is also relevant. For a discussion of types of context in translation, see Melby and Foster (2010).

3.3. The human role in terminology work

The reader is requested to remain open to the possibility that just as raw MT will likely never replace polished translation by highly competent translators who are asked to produce accurate and fluent documents for a demanding audience, automatically generated termbases will likely never replace 'post-edited' termbases corrected and enriched by highly competent terminologists. Inaccuracies in automatically generated termbases may result from errors in the underlying bi-text corpus or from mechanical processing of data (that is, from the software not understanding human language). Enrichment of termbases consists of adding metadata that cannot be derived automatically from the corpus.

A high-end termbase is one in which concept entries have been validated to ensure that all the terms in an entry do indeed designate the same or nearly the same concept and that all associated information is accurate. Information is included that helps a translator choose appropriately among terms for the same concept, such as terms specific to a product or company, obsolete terms, acronyms, and short forms. Linguistic metadata about each term, such as its part of speech, a definition, or at least an example of its use in a sentence, is included. The subject field i.e. domain, to which the concept applies is indicated in each entry, and, ideally, semantic relationships among concept entries are made explicit. In addition, various items of administrative information, such as the party responsible for the term and the date the information was updated, are included.

The information in a high-end termbase (a) allows a translator to dynamically select a subset of the termbase that is relevant to a particular translation project (e.g. only those terms related to a particular product or company), (b) guides a translator to choose appropriate terms, and (c) enables automated processing in translation tools.

Without a high-end termbase, a translator consulting a bi-text corpus directly is vulnerable to two perils: (1) if the corpus is too small, then the terms the translator needs may not be in it; (2) if the corpus is too large or varied, then a search for a term as only a word or combination of words is in danger of finding multiple target-language terms for a given

source-language term and making suggestions that conflict with the project specifications or with other translators working on the same project.

New domains or conceptual changes to existing domains result in new source-language terms for which there is not yet a target-language equivalent and in those cases, an entry in a termbase may precede any use in a bi-text corpus. Such a situation is not uncommon in new and emerging fields of knowledge where translators may be called upon to create neologisms.

3.4. Clouds on the horizon?

Garcia (2009) paints a rather bleak picture of the future of translation as an 'independent' profession. As defined by Garcia, translation is an independent profession when the primary training of translators is in translation theory and practice, rather than in domain-specific knowledge (2009: 200). He predicts that after the year 2010, the bulk of professional translators will be pushed into new roles based on the "utility center" and "hive" models of translation:

- Working in low-paid jobs in translation "utility centers" where their activity is limited to simple post-editing, at a segment-by-segment level, of output from software systems that combine translation memory and machine translation; or
- Working in a "hive" environment where translation is done by non-translator subject-matter specialists who volunteer and a few professional translators who deal with quality assurance (Garcia 2009: 211).

There is, however, an alternate view, based on the assumption that translation never has been an independent profession. According to this expanded view, a professional translator is one who combines subject-matter expertise, usually at the masters or doctoral level, with training in translation theory and practice, including current technology. This pair of knowledge and skills, one domain-specific and the other translation-specific, can be obtained through a combination of formal education and on-the-job training. The independent profession described by Garcia may well be in trouble.

To sum up, translators should not use technology as a substitute for understanding. In relation to the current topic, this means that termbases are not a substitute for a deep understanding of a domain.

3.5. Optimistic views

Jaap van der Meer, former CEO of AlpNet and current president of the Translation Automation User Society (TAUS), who was perhaps the first to predict that translation will become a "utility" (that is, a service as universally available as electricity and running water, c.f. Garcia's utility centers), now sees a bright future for professional translators. In his invited address to the world congress of translators in San Francisco, he suggested that:

> ...non-perfect MT output will stimulate the need for high-quality translation in a broad range of communication situations. The challenge we face as an industry is to agree on the criteria and the measurements for the level of quality that is needed for each situation (van der Meer 2011: online).

Chris Durban, former president of the French translators association (http://www.sft.fr), also paints a positive picture of the profession in her 2010 book, *The Prosperous Translator*.

It is possible to summarise optimistic views of the future of human translation using an analogy. Humans will never replace calculators (they are far too slow at doing arithmetic), but computers will never replace certified accountants who use calculators and other tools to make informed recommendations. Likewise, humans will never replace computers to search for words in a bi-text corpus (they are far too slow at skimming large collections of documents for particular words), but the only human translators who will be replaced by computers are those who translate like computers, that is, mechanically.

3.6. The need for termbases

Assuming that human translators will be around for the foreseeable future, will they need access to termbases?

Kara Warburton, chair of the ISO technical committee responsible for standards on terminology practice, has worked for IBM as their head terminologist and has contact with various other large organisations, such as the World Bank, SAS, SAP, Medtronic, and others. She asks:

> Why are all these organizations spending time and money to develop termbases if they already have access to rich repositories of bilingual corpora in which, with existing technology, it is quick and easy to search for terminology? From what I am hearing at these organizations, the main reason to justify a termbase is that specific contextual environments of an individual term provide insufficient information for controlling the use of that term across the organization. Additional metadata — requiring a dedicated, structured repository — are required to indicate when and where a term can be used, such as for specific product lines or projects, in specific grammatical constructions and not others, and in specific forms such as capitalized or hyphenated. Another reason is that structured terminological data

> can serve a wide range of natural language processing applications that are becoming increasingly popular in commercial environments to improve content management applications (personal communication).

3.7. The bottom line for termbases

What does this discussion mean for termbases? The generation of preliminary termbases from multilingual corpora using sophisticated software tools will become more and more automated, just as the generation of statistical machine translation systems using such software tools as Moses has become commonplace (see, for example, Let's MT! 2011). This shift is to be expected and welcomed for its ability to simplify manual tasks, along with the improved generation of machine translations from the same mutlilingual corpora. However, as automation of MT and termbases evolves, the need for professional translators and professional terminologists who can enhance these automatically-generated resources will also increase.

The author envisions a future in which translation tools are able to access large multilingual corpora (ideally classified by domain) to find term candidates. However, rather than adding these terms to local, two-column glossaries, as is currently done, these tools will simplify the submission of these terms as starter entries in structured termbases (which may, themselves, be shared with others).

A termbase includes information that has been validated by a human terminologist. While some of this information can be inferred, with varying degrees of success, from a bi-text corpus (e.g. domain may be inferred if the bi-text is classified as belonging to a particular domain), terminology work adds value to a bi-text corpus. Termbases will thus be around for the foreseeable future and, as suggested by Zetzsche (2011) and Warburton, the value of high-end termbases will increase. The author anticipates that termbases will be used not only reactively to fix problems in a bi-text corpus but also proactively to avoid introducing problems into translations (see Schmitz and Straub 2010: 292) and for various additional tasks, including automatic processing of text and human processing of text with the assistance of termbases. All this will result in an increased need for skilled terminology work.

4. TermBase eXchange (TBX): addressing the challenges of interoperability

Having argued that termbases remain relevant despite improvements in bi-text processing and that they are a valuable resource in the production of multilingual documents, the rest of this article will focus on the role of TermBase eXchange (TBX) (ISO 30042 2008), a family of formats for representing the information in a high-end termbase in a neutral

intermediate format in a manner compliant with the Terminological Markup Framework (TMF) (ISO 11642 2003).

In addition to a brief introduction to TBX and TMF, this article provides use cases — a kind of usage scenario for attaining a particular goal — for TBX (section 6) and suggestions for its implementation (see Appendix) that complement the information available in the standards themselves. The reader is invited to consult ISO 30042 and 16642 for more detailed technical information.

TBX may not be needed to represent a two-column bilingual glossary consisting of terms in two languages relevant to a particular translation project, but it is flexible enough to represent the information in such simple glossaries if desired, as well as the information in complex termbases.

TBX does not specify how to create and maintain a termbase. One can choose to create the termbase manually, or by using a semi-automated process whereby terms are extracted from a variety of internally consistent bi-texts. In the latter case, as discussed earlier, human validation and enhancement are necessary.

TBX is simultaneously an international standard (ISO 30042) and an industry standard. The industry standard version, which differs from the ISO standard only by having different title pages, is available at http://www.ttt.org/oscarstandards[8]. The host organisation (LISA) for OSCAR, the industry standards body that developed TBX, was dissolved in February 2011, but in September 2011, ETSI took over maintenance of the OSCAR standards. ETSI has established an interest group for translation/localisation standards and a liaison relationship with ISO so that TBX can continue to be published as both an ISO standard and an industry standard.

There are many types of termbases in use, ranging from huge termbases (usually called 'term banks') operated by governments[9], to medium-size termbases maintained by corporations and NGOs, to smaller termbases maintained by translation service providers and individual translators. The problem addressed by the designers of TBX was that existing termbases are generally not interoperable. They are based on different data models that use a variety of data categories. And even if the same data category is used for a particular piece of information, the name of the data category and the values allowed for the data category may be different.

For example, one termbase may use the 'part-of-speech' as the data category name while another may use 'pos' for the same data category. And one termbase may allow 'noun' and 'verb' as values for this data

category while another may only allow 'n' and 'v'. Even more problematically, a termbase may use the non-standard data category 'grammar' and combine both part of speech and grammatical gender in one element. In addition, the overall structure may differ.

4.1. Terminological markup framework

At a very high level, any termbase that is to be represented in TBX must conform to the abstract data model called Terminological Markup Framework (TMF), as defined in ISO 16642 (2003), the basic structure of which is shown in Figure 1 (below). Only TMF-compliant termbases can be fully represented in TBX.

Note that the TMF metamodel does not specify which data categories, other than 'term' and 'language' must appear in a termbase. This flexibility is a strength of TMF in that it allows many diverse termbases to comply with the TMF metamodel. On the other hand, this flexibility is a weakness in that two TMF-compliant termbases may not be compatible because they use vastly different data categories. This section will describe the fundamental structure of TMF, and section 5 will discuss how TBX deals with differences in data categories.

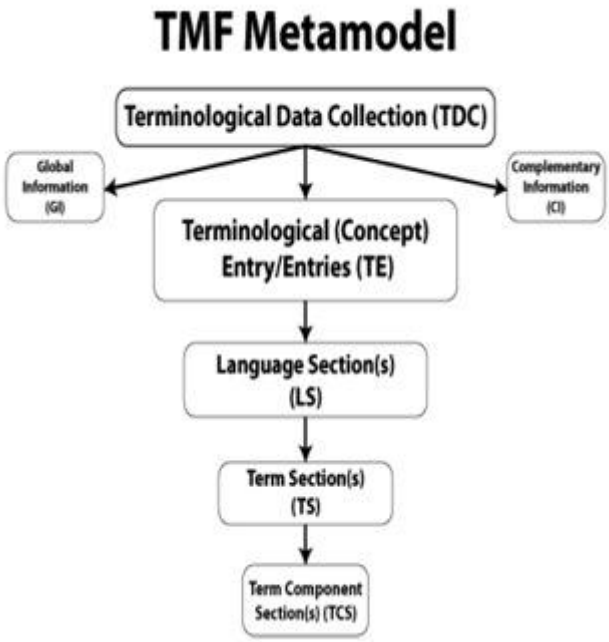

Figure 1. High-level structure of the TMF Metamodel (from ISO 30042 2008: 8)

TMF includes minimal requirements on the design of termbases. The one fundamental requirement is that a compliant termbase be concept-oriented rather than organised by headword with all the senses of a headword grouped together.

The TMF model includes the following levels within the concept entry:

- **Language Sections**. All the terms in a particular language for a given concept must be grouped together into the language section of that concept entry. Most current termbases satisfy this requirement.
- **Term Sections**. Within a language section, each term has its own section. The same types of metadata can be stored for each term, a principle referred to as term autonomy. Some termbases do not satisfy this principle. For example, they may list synonyms of a term without allowing a full set of metadata to be stored for each synonym.
- **Term Components**. Sometimes multiword terms, such as "uninterruptible power supply" are broken up into individual words and the words are stored separately in a termbase, allowing linguistic information, such as part of speech and lexical gender, to be stored for each word. Indicating the gender of a noun inside a term is probably the most common use of components. Other information that could be stored about the components of a term includes how words are hyphenated, inflected and pronounced. Linguistic information can be useful in automated processing of a text. For example, in highly inflected languages, the same term may appear in texts in many forms. Linguistic information is needed by a translation tool in order to automatically look up terms in a text and display filtered information from the concept entry for the benefit of a translator or reviser who is comparing the source text and the target text. According to TMF, linguistic information about the components of a multiword term is stored in term component sections associated with a particular term section. The term component level is optional.

In addition to the hierarchical structure in the graphic shown earlier (i.e. a concept entry consists of language sections, which consist of term sections, and each term can potentially be split into components and annotated), there can be two additional sections in a termbase that conforms to the TMF model: *global information* and *complementary information*:

- **Global information** applies to the entire termbase. It includes the name, origin, and ownership of the termbase.
- **Complementary information** typically consists of entries for people and entries for references. When several people are involved in the maintenance of a termbase, it is against basic principles of information management to repeat their contact information in many places. Instead, a short, unique identifier for a person is stored at some level in each applicable concept entry. Then, the detailed contact information for that person is stored in one place, in the complementary information section of the termbase. Likewise,

information about books or other references that are the source of several terms in a termbase are stored in the complementary information section, rather than being stored multiple times in concept entries. Only a short, unique identifier for the reference is stored in the entry.

4.1.1. Links

Some concept entries may, of course, refer to external information on the Internet. It makes a termbase extremely fragile to include links that point directly to an external resource that might change at any time. The ideal solution is to link only to persistent identifiers. The topic of persistent identifiers is beyond the scope of this article. However, an intermediate solution to the problem of the rapidly changing nature of URLs and other Internet identifiers is to store the actual URLs in the complementary information section of a termbase and refer to them indirectly from concept entries through links to entries in the complementary information section. With this approach, a broken link can be fixed in one place, and all external links are in one place for easy periodic verification.

In addition to links between concept entries and external resources, entries and elements of an entry can be linked internally (that is, within a termbase) in two ways: term-to-term and concept-to-concept. For example, a term can be linked to its antonym in another concept entry and concept entries can be linked using concept relations typically found in ontologies. Alternatively, a concept entry can be linked to a node in an external concept system.

4.1.2. Concept relations

One advantage that concept-oriented termbases have over lexicographically oriented glossaries is that they are more suited to managing conceptual relations between entries. Conceptual relations include hierarchical relations, such as broader (superordinate), narrower (subordinate) and sibling (coordinate) relations, as well as partitive relations (for instance, the relation between "pupil" and "eye"), and associated relations (for instance, the relation between "pitcher" and "baseball"). Concept relations have traditionally been intended primarily as a benefit to authors, translators, and others who want to navigate through a domain. The potential of such relations in terminology resources is becoming increasingly recognised for also enabling certain extended applications of terminology, such as for search engine optimisation and content management. Concept relations are basic to the Semantic Web. TBX provides a solid foundation for representing such relations in termbases, and there should be more interaction between TBX in terminology work and in OWL (from the Semantic Web community).

The information in this section is a basic explanation of TMF. For more information on TMF, please consult the ISO specification.

5. TBX and data exchange

As mentioned earlier, TMF provides a very flexible framework for defining termbase structure, but this flexibility can complicate data exchange. TBX approaches the many differences among termbases by requiring the developer of termbase software to map every data category in their termbase to a standard data category, defined by ISOcat, and to map the structure of the termbase to a core structure that corresponds to the TMF metamodel. This mapping may be done by an export routine, i.e. the internal structure and data categories of the termbase need not exactly match those prescribed by TBX; it is, however, necessary to be able to convert automatically between the termbase's internal representation and TBX's data categories.

As for data categories, not only does the name of a data category in a proprietary termbase need to be mapped to its standard name to comply with TBX, but the usage of the data category also needs to correspond to its usage as prescribed in ISOcat. For instance, the TBX/ISOcat data category **context** refers *only* to a sentence or other segment in which the term occurs, and not to some other kind of contextual information, such as subject field or product usage.

Once a TMS has been mapped to TBX and represented in the form of a TBX file, then it is possible for another TMS to import the terminology, provided that both software systems use the same 'dialect' of TBX. A dialect of TBX is defined primarily by the list of data categories from ISOcat that are allowed and the levels in the TMF metamodel at which they are allowed. There are also some technical aspects of defining TBX dialects that are relevant to software engineers implementing TBX but not to translators.

Given the large inventory of terminological data categories in ISOcat, there are potentially thousands of dialects of TBX. Of these potential dialects, so far three have been given privileged status.

(1) **TBX-Default.** TBX-Default is a large dialect of TBX defined in the TBX standard. In TBX-Default, there are over one hundred data categories.
(2) **TBX-Basic.** A team of terminologists who work for large organisations identified a subset of TBX-Default that is minimally sufficient for representing the most important information in a typical corporate termbase (as opposed to a national termbase

maintained by a government agency or a research-oriented termbase). This subset, called TBX-Basic, consists of a subset of the data categories in TBX-Default and a subset of the structural options in the TBX-Default core structure. TBX-Basic, which includes only about 20 data categories, of which only a few are mandatory, is, as expected, much smaller than TBX-Default. The TBX-Basic documentation is available at http://www.ttt.org/oscarstandards/.

(3) **TBX-Glossary.** In order to facilitate migration of glossaries that are represented in simple spreadsheets, with one row per concept entry, a third, even smaller dialect of TBX called TBX-Glossary has been proposed. Documentation about TBX-Glossary is available at http://www.ttt.org/tbxg; this site also links to software that can convert terminology from TBX-Glossary format to other formats.

While it is beyond the scope of this article to describe exactly how to design and implement a TBX export routine for an existing termbase and a particular dialect of TBX, the appendix to this article provides an overview of what is involved. For more details about implementing TBX, the actual TBX standard should be studied and, if needed, the services of a translation technology consultant familiar with TBX can be engaged.

It is important to note that, on the one hand, if TBX had only one dialect and only termbases that conform fully to that dialect could use TBX, then termbase exchange would be easy for those termbases and impossible (at least using TBX) for all others. On the other hand, if many TBX dialects are used, then most termbases can export at least one of those dialects, but exchange is complicated by the clash between different dialects on the sending and receiving ends. This complexity is not a problem resulting from the design of TBX itself, but rather a statement of the complexity and variety of existing termbases.

Although TBX offers considerable potential for sharing resources, its benefits can be realised only when it is implemented properly. While TBX has indeed been widely implemented, most implementations have not yet been subjected to third-party evaluation.

Several years ago, there was some question as to whether the translation industry would support more than one viable translation tool vendor. Since then, the number of translation tools on the market has increased, not decreased, and even if some of them do not prosper in the long term, there will likely always be a variety of tools on the market and thus a need for interoperability. At the translator tool forum held by the American Translators Association in 2011, a majority of the approximately fifteen vendors present indicated that their terminology management component now supports TBX. That is a significant increase over previous

years. In addition to commercial translation tools, TBX is being implemented in Termium and in the upcoming NATO terminology management system, and Microsoft glossaries are now available in TBX format.

This increase in implementation also means that these tools now need to be independently tested to verify their ability to import and export TBX files. There will likely be issues around TBX compliance and interoperability, but this is normal for any standard during its initial implementation stage. In particular, questions of converting between different dialects of TBX will likely take a long time to work out.

6. TBX Use Cases

In anticipation of more widespread support for TBX import and export features, the following are a few likely scenarios in which TBX will find application:

- **Sharing terminology within the supply chain**. A language service provider (LSP) receives work from several clients that each maintain corporate termbases, and passes the work on to various freelance translators. A number of different termbases and translation tools are involved, and it is desirable to make it feasible for translators to import terminological data into their own translation tools, so that terms can be automatically looked up during translation and quality assurance and consistency with the relevant corporate termbase can be guaranteed. The ideal solution is for everyone to use the same dialect of TBX to represent terminological data.
- **Asset protection**. An organisation implements a central **termbase** but does not want to be locked into using the same terminology management software forever. The organisation should require of the terminology management system the ability to export the content of the termbase to a dialect of TBX that can fully represent the information in the termbase. This protects the investment made by the organisation in developing a central termbase. This investment can be substantial. In addition to protecting a valuable asset (the terminological data), design with TBX in mind assures the use of best practices, since TBX is a direct reflection of a set of terminology best practices embodied in the set of ISO standards on which TBX is based.
- **Integration with authoring processes**. An organisation wants to improve its document production chain. The obvious place to start is with authoring. In the 2010 SDL Global Authoring Survey, over four hundred technical communicators were asked what concerned them most. The top answer was quality and clarity of content, something which is also of significant concern to translators. In the 2011 Term

Net survey, 74% of the participants observed discrepancies of terminology between documents. TBX is equally valuable as a best practice for monolingual authoring-oriented terminology resources as it is for multilingual translation-oriented terminology resources, and ideally both kinds of terminology should be managed in the same central termbase.

7. Conclusion

The increased availability and use of multilingual corpora have been shown to be relevant to terminology management, but not in the way that some have supposed. The author predicts that rather than going away, termbases will see greater applicability to the entire multilingual document production chain. This in turn will demand more exchange of terminological data between different termbases.

However, termbases are not, generally speaking, directly compatible. This lack of compatibility is why TBX is important.

Bibliography

- **Bowker, Lynne** (2011). "Off the record and on the fly." Alet Kruger, Kim Wallmach and Jeremy Munday (eds). (2011) *Corpus-based Translation Studies: Research and Applications*. London: Continuum, 211-236.
- **Durban, Chris** (2011). *The Prosperous Translator.* (self-published). http://prosperoustranslator.com (consulted 2 May 2012).
- **Garcia, Ignacio** (2009). "Beyond Translation Memory: Computers and the Professional Translator." *The Journal of Specialised Translation* 12, 199–214.
- **Harris, Brian** (1988). "Bi-text, a new concept in translation theory." *Language Monthly* 54, 8–10.
- **IBM** (2011). "Terminology Management: Use a Terminology Database." http://www-01.ibm.com/software/globalization/topics/terminology/database.html (consulted 2 May 2012).
- **ISO 704:2009** (2009). *Terminology Work – Principles and Methods*. Geneva: International Organization for Standardization.
- **ISO 12620:1999** (1999). *Computer applications in terminology – Data categories*. Geneva: International Organization for Standardization.
- **ISO 16642:2003** (2003). *Computer applications in terminology – Terminological markup framework*. Geneva: International Organization for Standardization.
- **ISO 30042:2008** (2008). *Terminology and other language and content resources — Computer applications in terminology — TermBase eXchange Format Specification (TBX).* Geneva: International Organization for Standardization.

- **Let's MT!** (2011). "Build your own Machine Translation System." http://www.letsmt.eu (consulted 2 May 2012).

- **Linguaspot** (n.d.) "Data Compilation and Cleaning." http://www.linguaspot.com (consulted 2 May 2012).

- **Localization Industry Standards Association** (2008). *TermBase eXchange (TBX)*. http://ttt.org/oscarstandards/tbx/tbx_oscar.pdf (consulted 2 May 2012).

- **LISA Terminology SIG** (2008). *TBX-Basic.* Romainmôtier: Localization Industry Standards Association, 1-91. http://ttt.org/oscarstandards/tbx/TBXBasic.zip (consulted 2 May 2012).

- **Marshman, Elizabeth, Julie L. Gariépy and Charissa Harms** (2012). "Helping Language Professionals relate to terms: Terminological Relations and Termbases." *The Journal of Specialised Translation* 18, 45-71.

- **Melby, Alan** (2008). "TBX-Basic Translation-Oriented Terminology Made Simple." *Revista Tradumàtica* 6. http://www.fti.uab.es/tradumatica/revista/num6/articles/02/02central.htm (consulted 2 May 2012).

- **Melby, Alan and Christopher Foster** (2010). "Context in Translation: definition, access, and teamwork." *The International Journal of Translation and Interpreting* 2(2), 1–15. http://www.trans-int.org/index.php/transint/article/viewFile/87/70 (consulted 2 May 2012).

- **Muegge, Uwe** (2011). "Why Terminology Management Plays a Critical Role in International Launches." http://www.csoftintl.com/shownews.php?id=161 (consulted 2 May 2012).

- **Rogers, Margaret** (2008). "Consistency in terminological choice: Holy Grail or false Prophet?" *SYNAPS* 21, 107–113.

- **Schmitz, Klaus-Dirk and Daniela Straub** (2010). *Successful Terminology Management in Companies*. Stuttgart: TC and More.

- **Translation Bureau** (2011). "The Pavel, Terminology Tutorial." http://www.termiumplus.gc.ca/didacticiel-tutorial/lecon-lesson-1/index-eng.html (consulted 2 May 2012).

- **van der Meer, Jaap** (2011). "The future for translators looks bright, but they will have to reinvent the profession first." http://www.translationautomation.com/perspectives/the-future-for-translators-looks-bright-but-they-will-have-to-reinvent-the-profession-first.html (consulted 2 May 2012).

- **Wright, Sue Ellen and Gerhard Budin** (eds) (1999). *Handbook of Terminology Management*. Vol. 1. Amsterdam/Philadelphia: John Benjamins.

- **Zetzsche, Jost** (2011). "Conference Fatigue." *The Tool Box Newsletter* 11-11-202. http://archive.constantcontact.com/fs090/1101859302759/archive/1108553384354.html (consulted 02.07.2012).

Websites

- "**ETSI**." http://www.etsi.org/ (consulted 4 July 2012).
- "**IATE**." http://iate.europa.eu/ (consulted 4 July 2012).
- "**ISOcat**." http://www.isocat.org/ (consulted 4 July 2012).
- "**Moses**." http://www.statmt.org/moses/ (consulted 4 July 2012).
- "**OWL**." http://www.w3.org/TR/owl-features/ (consulted 4 July 2012).
- "**TAUS Data Association**." http://www.tausdata.org/ (consulted 4 July 2012).
- "**Terminum**." http://www.termiumplus.gc.ca/ (consulted 4 July 2012).

Appendix: Implementing TBX

This appendix presents a set of questions and issues for a terminologist and a software engineer to read together as they plan an implementation of TBX.

Export issues

(1) Does the termbase's structure comply with the TMF metamodel (including whether the structure can be 'mapped' to the TMF metamodel)?
   a. **Yes**: The termbase is suitable for TBX-based import or export.
   b. **No**: The termbase is not suitable for TBX export/import or any other sophisticated use and should be re-designed.

(2) Does the termbase allow users to define their own data categories? If so, there are two approaches to deal with variation in data categories in termbases created by the same terminology management system.
   a. Particular database definitions (templates) can be provided that correspond directly to particular TBX dialects, to allow the system's export feature to output TBX compliant with a particular TBX dialect.
   b. A mapping table can be created between a particular user-defined termbase and a dialect of TBX, and then a software application can be developed that post-processes the raw output of the termbase into TBX compliant with a particular TBX dialect, according to the mapping table.

(3) Regardless of how a TBX file is generated from an existing termbase, it should be checked for compliance to the TBX standard. One way to do this is to use the open-source TBX Checker (available from SourceForge through http://www.ttt.org/oscarstandards). This tool assumes familiarity with XML and the TBX standard.

Import issues

(1) Does the termbase's structure comply with the TMF metamodel? If it does not, it is still possible to design and implement an import routine, but some information will likely be lost during the import process. This data loss may not be a problem if the exchange of information is mono-directional (i.e. only into the target termbase and not back out) and, depending on the type of information that is lost, the consequences may be insignificant.

(2) Which data categories will be allowed in an incoming TBX file? One option is to allow all the data categories in TBX-Default and map each one to a data

category supported by the target termbase. Of course, there may be some sets of data categories that are conflated to a single data category because the target termbase does not support the fine-grained distinctions in TBX-Default. The worst case is that an incoming data category must be mapped to a general-purpose note element in the target termbase.

(3) Is the termbase expected to accept incoming TBX files from unknown parties? If so, it must be made very clear which dialect or dialects of TBX are supported for import, and care must be taken to deal with fields in the termbase whose allowed values are more restricted than the allowed values of the corresponding item in a supported TBX dialect. If the terminology management system allows user-defined termbase data models, then import becomes much more involved and TBX support must be expressed in terms of particular termbase data models.

Biography

Alan K. Melby is professor of linguistics at the Provo campus of Brigham Young University. He has been interested in terminology management and terminology exchange since the mid-1980s. He has been involved in various efforts to develop terminology exchange formats, including the terminological data chapter of version P3 of the Text Encoding Initiative guidelines and various terminology-related standards within ISO Technical Committee 37. E-mail: akmtrg@byu.edu.

Notes

[1] Comments from a number of colleagues, including Barbara Inge Karsch and Jost Zetzsche on early drafts were very helpful. Particular thanks are given to Kara Warburton and Arle Lommel for their extensive comments and suggestions.

[2] When working with termbases, it is useful to be acquainted with the principles of concept-oriented terminology work. There is an international standard for terminology work (ISO 704:2009). Another resource is the comprehensive *Handbook of Terminology Management* (Wright and Budin 1999). There are also freely available introductions to concept-oriented terminology work on the Web (see, for example, Translation Bureau 2011 and IBM 2011).

[3] Metallurgy, in addition to animal husbandry, is one purported source domain of the idiom "sweating like a pig," since farmyard pigs do not in fact sweat noticeably but cooling pig iron does.

[4] The issue of how to define domains and their granularity is beyond the scope of this article.
[5] In some instances more columns are used to accommodate more languages, but the fundamental structure is unchanged.
[6] For a gentle introduction to manually constructing a termbase with the structure and content described in this article, see Melby (2008).
[7] An example of such an inaccurate pairing was encountered by one of the author's colleagues who reported using an English-to-German machine translation system that stated with high confidence that the appropriate translation of "1865" in English was *deutsches-ungarisches* ('German-Hungarian') in German. No other years appeared to generate such bizarre results. Such a result shows that statistical results can state incorrect results with a high degree of certainty.
[8] With the dissolution of LISA, the standards were made available under a Creative Commons licence and are now available through the author's TTT.org domain, among other places.
[9] Such term banks include the Canadian government's Termium, and the European Union's IATE.