# Terminology and Formulaic Language in Computer-Assisted Translation

Pius ten Hacken & María Fernández Parra

Terminology is the study of technical vocabulary, whereas formulaic language is based on the study of the mental lexicon. In translation, both require a holistic approach. Therefore, it is not so far-fetched to consider whether the tools for terminology in Computer-Assisted Translation software can also be used to improve the translation of formulaic language. In order to explore this possibility we first consider the theoretical background of the relevant concepts and then study a number of individual cases in detail. The result is the formulation of some general conditions on the felicity of this approach.

Terminology and formulaic language are not usually linked, because the concepts are based in very different domains of linguistics. In translation, however, both concepts are relevant. Moreover, their translation turns out to pose strikingly similar problems. Therefore we will here first address terminology and formulaic language in the domain they originate from (section 1). Then we turn to the problems they cause in translation (section 2). After that, we will briefly describe the relevant tools available in Computer-Assisted Translation (CAT) packages (section 3). On the basis of this background, we will then analyse a number of expressions in section 4 and draw some tentative conclusions about the optimal treatment of formulaic expression in relation to terminology in section 5.

## 1. Formulaic Language and Terminology in Language

In order to explain the different backgrounds of formulaic language and terminology, it is useful to start by considering the nature of language. Arguably, one of the most important contributions of Chomskyan linguistics to the study of language is the distinction of a number of different concepts, each of which has sometimes been understood as the meaning of *language*. Ten Hacken (2007: 41-53) discusses these concepts and the context in which they were introduced in more detail.

A first pair of concepts is competence and performance. Chomsky (1965: 4) calls *competence* "the speaker-hearer's knowledge of his language" and *performance* "the actual use of language in concrete situations". Both competence and performance are empirical phenomena in the sense that they exist independently of the linguist observing them. Competence is realized in the speaker's brain whereas performance is realized as sound waves, ink on paper, digital characters, etc. Competence underlies performance in the sense that the former is a necessary component in the production and comprehension of the latter.

A second pair of concepts is I-language and E-language. Chomsky introduces I-language as a "notion of structure" that is an "element of the mind of the person who knows the language" (1986: 22). There is no reason to consider *I-language* as something else than a synonym of *competence*. *E-language*, however, is "a collection of actions, or utterances, or linguistic forms (words, sentences) paired with meanings" (1986: 19). It is therefore an entirely different type of concept from performance. Whereas performance is an empirical

concept, based on competence, E-language is an abstract, non-empirical concept, "understood independently of the properties of the mind/brain" (1986: 20).

The term *formulaic language* stems from the study of lexical retrieval. The question here is what are the units in the mental lexicon. It is introduced by Wray (2002: 9) to refer to expressions that consist of more than one word or other element, but are stored and retrieved as a single unit. Some examples of formulaic language are given in (1).

(1) a. Good morning.
b. Good night.
c. Nice to meet you.
d. Nice meeting you.

Although the examples in (1) can be understood compositionally and could be constructed by applying normal syntactic rules to the individual words, it is unlikely that they are constructed each time they are used. Apart from the relative frequency of these expressions, also the rules for their proper use argue against such a view. An example of these rules is the contrast between (1a) and (1b). Whereas (1a) is used only in greeting, (1b) is used only on leaving. This information cannot be included in the lexical entries for *morning* or *night*. Another case is the contrast between (1c) and (1d). Whereas (1c) is commonly used when being introduced to someone, (1d) is more likely to be used when saying goodbye. Of course this information cannot be stored as parts of the meaning of the words (which are the same) or the construction. The only place where it can be stored is in the entry for the full expressions in the mental lexicon. The perspective of language that is central in the study of formulaic language is therefore that of competence/I-language.

The phenomenon we refer to by *formulaic language* is often discussed under different names. Jackendoff (2002: 167-182), for instance, uses *idiom* in his discussion of lexical storage versus on-line construction. However, as Tschichold's (2000: 11-24) overview shows, this term has been used in a variety of more specialized meanings, so that we tend to avoid it in a technical sense. As a practical guide for the recognition of formulaic expressions we adopt Fernández Parra's (2007) working definition in (2).

(2) A formulaic expression is an expression of at least two words which
a. is prefabricated,
b. shows frozenness in its word order,
c. allows limited substitutability of its component words by synonyms or quasi-synonyms,
d. shows conventionalization, and
e. has a non-compositional meaning.

The essential condition is (2a). This is also the central condition Wray (2002:9) gives. It is a well-known fact that competence/I-language is not immediately available for inspection. Therefore, we cannot observe (2a) directly. The properties (2b-e) are used as more readily accessible criteria to determine (2a).

When we turn to terminology, we enter a field with a rather different character. Terminology can be seen as a part of specialist communication. As outlined by Wright (1997), there are two main strands in terminology, the descriptive and the prescriptive

approach. They can be illustrated on the basis of (3), an example of a statement which includes terms.

(3) It is decidable for an arbitrary context-free grammars whether it generates any terminal strings.

(3) is a statement in mathematical linguistics which uses the terms listed in (4).

(4) a. decidable
b. context-free grammar
c. generate
d. terminal string

For each of the expressions in (4), there exists a well-defined correct use. Where the expression exists in general language, as in (4a), the terminological definition is more specific. In the case of *decidable*, it will specify, for instance, the range of procedures by which a decision can be reached. Where the expression exists in other fields, as for (4c) in electrical engineering, there will be different, independent definitions. The descriptive strand of terminology aims to describe the meaning and use of such terms.

A central issue in the prescriptive strand of terminology is standardization. As Wright (2006: 19-20) mentions, the idea of standardization is often misunderstood. It is not a matter of crushing diversity by imposing a standard using economic and political power, but of ensuring optimal communication in a field. As ten Hacken (2006: 10-11) suggests, the prescriptive strand of terminology, i.e. the process of finding an appropriate standard in the form of a set of concepts and names for them, might actually be seen as a type of applied science.

A standard is not an empirical phenomenon in the same way as competence and performance. It is created consciously by an authority. Therefore, in the Chomskyan characterization of language, it belongs to E-language. The procedure of composing such a standard is strongly based on actual use, i.e. performance. In fact, Strehlow (1997: 206) sees this procedure as "closer to what most people think of as comprising terminology management", i.e. descriptive terminology. The standard has to be as close as possible to actual use in order to maximize the chances of it being accepted in the relevant community. The role of competence in terminology is that of a general mediator: observed use is based on competence; the creation of a standard requires the use of competence; and the standard obtained should inform the relevant speakers' competence so that it will constrain their performance.

## 2. Formulaic Language and Terminology in Translation

The nature of formulaic language and of terminology imposes special constraints on their translation. In view of the differences between formulaic language and terminology considered above, they will at first be considered separately here.

In (5), we give a compositional and an idiomatic translation of (1a) into French. A literal back translation is given in brackets.

(5) a. ?bon matin ('good morning')
b. bonjour ('good day')

The literal translation in (5a) can be used as a noun phrase to refer to a *morning* that is in some way *good*, but it cannot be used as a formulaic expression corresponding to (1a). Instead, (5b) must be used. This example shows, therefore, that formulaic expressions cannot be relied on to be translated compositionally but have to be considered holistically. The literal English translation of (5b) is common in Australia but not in Britain. This illustrates the fact that *English* is not in all cases the correct level at which to state formulaic expressions.

The translation of a term such as (4b) is slightly more complex. In (6), five versions of a French translation are given.

(6) a. *contexte-libre grammaire ('context-free grammar')
b. ?grammaire libre de contexte ('grammar free of context')
c. grammaire hors-contexte ('grammar out_of context')
d. grammaire indépendante de contexte ('grammar independent of context')
e. grammaire de type 2 ('grammar of type 2')

The translation in (6a) concatenates the translations of the three components of the English term. It is ungrammatical, because of general word order constraints in French. In (6b), the elements of (6a) are reordered to make the expression grammatical. However, this is not a form that is in common use. A Google search produced only 25 hits (4 Sept. 2007).

In order to understand the other translations, it is necessary to look at the nature of the concept in more detail. Context-free grammars are formal grammars of a particular type. In general, a formal grammar is a system that generates strings and assigns structure to them. It characterizes the language consisting of the strings it generates. A grammar consists of a set of terminal symbols (the symbols making up the strings), a set of non-terminal symbols (auxiliary symbols that cannot appear in strings of the language), a designated start symbol (conventionally *S*), and a set of rewrite rules. Chomsky (1959a: 142-3) defines a number of different types of grammar by restrictions on rewrite rules which can be illustrated with the help of (7).

(7) a. $\alpha \rightarrow \beta$
b. $A \rightarrow BC$
c. $AC \rightarrow BC$

The general form of a rewrite rule is (7a). Here $\alpha$ and $\beta$ can be any string of terminal or non-terminal symbols. Context-free grammars have rules of the type illustrated in (7b). Every rule in a context-free grammar has $\alpha$ instantiated to a single symbol. A grammar containing a rule such as (7c) is not context-free.

On the basis of (7) we can understand the forms (6c) and (6d). In (7b), *A* is rewritten as *BC*, independently of the context of *A*. Whereas (6c) sounds slightly awkward, (6d) is very clear but somewhat long. In fact, (6c) is used relatively frequently, e.g. in the *Wikipedia* (http://fr.wikipedia.org/wiki/Grammaire_hors-contexte, 31 July 2007). (6d) was suggested to us by Eric Wehrli, but it does not seem to be in regular use (no hits on Google, 31 July 2007).

As all of (6a-d) have certain disadvantages, it is interesting to consider (6e). This is the translation proposed by the *Office de la langue française* in Canada (http://w3.granddictionnaire.com, 20 July 2007). It is not based on the same components as the original English term, but on a classification known as the *Chomsky hierarchy*. This hierarchy is defined by Chomsky's (1959a:142) series of restrictions on rewrite rules. The term *context-free grammar* is only introduced by Chomsky (1959b: 393). In English, *type 2 grammar* is normally only used when presenting context-free grammars in the context of the Chomsky hierarchy. In a Google search, *context-free grammar* yielded 346,000 hits, as against 592 for *type 2 grammar* (4 Sept. 2007). In French, (6e) is also less frequent than (6c), but with 218 hits for (6e) as against 528 for (6c), the proportions are clearly of a different order. The figures suggest that the concept is used significantly more frequently in English than in French.

The discussion of (6) gives a good impression of the complexity of the issues arising in the translation of terms. In the field of mathematical linguistics, much of the research has been done in English-speaking countries and published in English. Terminological decisions were therefore based in part on idiosyncratic properties of English, which makes it hard to translate the terms into other languages. It is interesting to note in this context that Maegaard et al. (1975) use a form parallel to (6e) in Danish, *type 2 grammatik*, although they mention *kontekstfri* as an equivalent of *type 2* (1975:167).

A comparison of the problems in translating formulaic language and terminology illustrated in (5) and (6) respectively shows one major similarity and one major difference. The similarity is that both involve expressions that have to be translated holistically. Compositional translations as in (5b) or (6a-b) are inappropriate. If they are correct in other cases this is accidental in the same way that cognates may be correct translations or false friends. The difference in translation strategy, at least in the examples discussed so far, is that for formulaic language the correct translation can be found by direct recourse to performance or native-speaker judgements, whereas for terminology a more elaborate analysis of the concept referred to is necessary.

The obvious question here is to what extent the observed similarity and difference can be extrapolated to formulaic language and terminology in general. Given the nature of formulaic language and of terminology as stored units of a form and meaning, the former in the mental lexicon, the latter in a standardized termbase, it would be highly surprising to find more than accidental counterexamples to the generalization that both classes have to be translated holistically. The observed difference can also be related directly to the characterization of the two classes. A formulaic expression is stored in the mental lexicon. A key issue in terminology is standardization, which involves a conscious operation to facilitate technical or scientific communication, based on the study of the concepts in the relevant field. It is therefore *a priori* plausible that the translation of a formulaic expression can in general be retrieved from the mental lexicon, whereas the translation of terminology generally requires an analysis of the concepts referred to.

## 3. Termbases in Computer-Assisted Translation

Computer-Assisted Translation (CAT) is the name of a type of software package that helps the translator while leaving the translator in control of the translation process. It is important to distinguish CAT and Machine Translation (MT). Whereas in the case of MT, it

is the computer that structures the translation process, in the case of CAT the user determines the workflow. Quah (2006) gives a useful overview of the field of CAT and MT.

CAT packages typically provide two databases, a Translation Memory (TM) and a termbase (TDB). The TM stores previously translated segments. Typically, a segment is a sentence or an expression occurring independently (e.g. as a heading or in a table). A TM is particularly useful in repetitive texts or if a new version of a previously translated text is translated. The TDB stores terms in order to recognize them in the source text and give access to the information stored for them.

The user of a CAT package remains in control of the translation process. This means that TM and TDB can be consulted and can propose translations, but the human translator determines when they are consulted and how the information they come up with is used. Typically, a CAT package will make the TDB accessible in three different ways. First it is available for browsing. The translator can open the TDB, look up a term, and use the information. Secondly, it can be used in automatic term recognition. The user asks the CAT tool to match the source text with the TDB and the CAT tool signals for each segment which terms have been found. Typically, they are displayed in a list and the user can select an entry to consult the information in the database. Thirdly, the user can ask the CAT tool to translate the terms automatically. In this case, the source language terms recognized by the CAT tool are replaced by the corresponding target language terms in the TDB. The translator can then adapt the form to match the morphosyntactic constraints imposed by the context (e.g. number, gender, case).

It is interesting to compare termbases as used in terminology with TDBs in CAT tools. For a proper comparison, we also have to distinguish two types of use of CAT tools. One type is the kind of translation project that is too large for a single translator. The reason can be the volume of text, the number of target languages, or a combination of both. The other type is the individual translator using the databases in the CAT tool to activate previous experience more efficiently. Although it is not always possible to draw a clear boundary between the two scenarios in the case of TMs, the distinction is reasonably clear and highly useful when we concentrate on TDBs.

In a termbase used for the documentation and standardization of terminology, we can expect to find at least the information types listed in (8). An example of a more extensive template for term records is given by Cabré (1999: 125).

(8)
a. Term identification: form, abbreviation (if applicable)
b. Syntactic properties: syntactic category, subcategorization
c. Subject field
d. Definition (with source)
e. Examples of use (with sources)
f. Semantic relations to other terms (hyperlinks)
g. Administrative information: status, author, date

Most of the information types in (8) are straightforward. By *subcategorization* in (8b) we mean any further syntactic constraints on usage, e.g. gender and count/mass for nouns, as well as complementation. As an example for semantic relations in (8f), *context-free grammar* might have links to *type 2 grammar* as a synonym, *grammar* as a hyperonym, as well as a syntagmatic relationship to *context-free language* (‘any language that can be generated by a context-free grammar’). The status in (8g) refers to the standardization process. It is meant to

record which decisions have been taken by which standardization bodies or authoritative handbooks in the field.

Term records with information types as in (8) can also be specified in TDBs in CAT tools. Many CAT tools provide a number of basic templates for term specification and all allow the user to determine exactly which fields are available in the TDB.

The main difference between the types of termbase appears when we consider the treatment of translations. At first sight, the simplest approach seems to be to add the translation into other languages as a further information type. This implies that (8b-g) are applicable to all languages. Of course this is not correct. There is no reason to suppose, for instance, that if a noun is feminine in French, its translation will be feminine in German. Also the status in one language does not depend on that in another one, but on individual decisions by different standardization bodies or other authorities. Examples of use are clearly language-specific. Even in an ideal world, only (8c-d) and (8f) can be treated as applicable to all languages.

However, our real world is not ideal in the relevant respect. In mathematical linguistics, the subject field is formal and general enough to expect that concepts are the same in all languages, but if we consider subject fields in general there are two sources for divergence in this respect. One source is independent legislative decisions. The consequences are illustrated, for instance, in a field such as traffic law. Although *motorway* can be translated into German as *Autobahn* and into French as *autoroute*, they do not refer to exactly the same concept. In fact, *Autobahn* is likely to refer to slightly different concepts in Germany, Austria, and Switzerland.

Another source of divergence is the different structure of the general vocabulary in languages. Thus in describing washing machines, *easy-care* is an adjective denoting a certain kind of textiles and the detergents and washing programmes appropriate for them. In German translations, *Feinwäsche* can be used, but it is not a precise equivalent. It is a compound consisting of the adjective *fein* (‘fine’) and the noun *Wäsche* (‘textile’). Whereas the adjective alone translates less than the English term, the full compound translates more. The reason for this mismatch is that German has a very productive word formation process combining adjectives and nouns into a compound, whereas English does not have an equally productive equivalent. Naming and ultimately term formation are influenced by the availability of this process.

The most principled way to address such issues is to describe each term in each language separately. Cabré (1999: 127) proposes to use *correspondence records* to record translations. In a correspondence record, the equivalence between terms in different languages is recorded in a way that ensures, first, that the conceptual system of each language is maintained, and second, that any divergence can be adequately expressed.

Let us now return to TDBs as a component of CAT tools. In the case of a large translation project, the central concern is terminological consistency. The maintenance manual of a complex device such as a nuclear power station or a fighter plane is of a length that requires a large team of translators to work on it. It is essential, however, that terms are translated consistently, so that, for instance, a part of which the use is described in one section and the procedure for replacing it in another is referred to in both sections by the same name. In a properly organized translation project, at least one person is designated to maintain the terminology database.

In most CAT tools, the distinction between properties of the concept and properties of the expression in a particular language can be maintained when defining terms. In SDL

MultiTerm (version 2006), for instance, information can be specified at *Entry level* or at *Term level*. The Entry level is for information pertaining to all languages. It includes the subject field and the definition (8c-d). Information at the Term level is specified independently for each language. It includes the form and syntactic properties (8a-b) as well as examples and hyperlinks (8e-f). This organization structure is less flexible than the one involving correspondence records, because a single definition is assumed. It can be motivated by the nature of the TDB. When the TDB is compiled for a particular translation project, the concepts are the ones referred to in the source language. The preferred way to use the TDB is to ask the CAT tool to recognize the terms in the source language and make a link available to the relevant TDB entry. From the entry, the meaning of the source language term can be retrieved, its translation, and also grammatical information and target language examples that support its correct embedding in the target text produced.

The outlook of individual translators, producing a regular output of often rather small documents with short deadlines, is rather different. Their main reason for using a TDB in a CAT tool is the gain in efficiency. When confronted with the model of a term record along the lines of (8), their main concern is therefore to specify (only) the information that is actually needed. A translator who regularly translates texts in the domain of mathematical linguistics will not need to look up the syntactic category of the terms in (4). A common simplification is also to give either a definition or an example, rather than both. The main function of a TDB for the individual translator is to record the results of terminological research. Thus for (4b), they might want to record not only the translation chosen, but also the main considerations in the choice among the alternatives considered in (6). In this way, if doubt arises or their translation is criticized, they can easily retrieve the reasoning behind their choice.

In the context of an individual translator, the TDB tends to get the character of well-organized personal notes rather than a full, systematic description of the terminology of a particular field. Efficiency considerations are an important part of the explanation for this phenomenon. If the TDB is used in the mode that replaces source language terms by their translations automatically, any further information beyond the form of the source and target language terms remains unused. Arguably, searching for such information is therefore only efficient if the translator suspects there will be a need to consult it later.

The three profiles in the specification of TDBs can therefore be characterized as follows. The terminologist working in a multilingual terminological project is concerned with avoiding a bias towards one of the languages. This can be achieved by specifying term records for each language independently and record translational equivalence in correspondence records. The terminologist working in a multilingual translation project is concerned first of all with the overall terminological consistency of the translation. A bias towards the source language of the project is natural. Full term records with a general and a language-specific level of specification are adequate for this context. The individual translator doing occasional terminological research as part of their translation work is concerned first of all with efficiency. Therefore, the main considerations in selecting information to be recorded in the termbase are the ease of finding it and the likelihood that it will be consulted in future.

## 4. Formulaic Language in Computer-Assisted Translation

Formulaic expressions such as (1) do not refer to a concept. Their meaning is almost entirely determined by the conditions on the situations in which they can be appropriately used. This makes it unattractive to treat these items in the same way as terms. Although it is in principle possible to specify the appropriate situations as a kind of definition, the properties of such a definition violate basic assumptions underlying a TDB. The reason why a TDB as a component of a CAT tool can legitimately simplify the correspondence of terms to a system that attributes different names to the same concept is that in terminology this is the usual situation. In terminology, we can assume that concepts are largely determined by outside reality, independently of the naming process. Exceptions discussed above are indeed relatively exceptional. This is not the case for formulaic expressions, as can be seen by comparing (9), (10), and (11).

(9) a. good morning
b. good afternoon
c. good evening

(10) a. bonjour ('good day')
b. bonsoir ('good evening')

(11) a. buon giorno ('good day')
b. buona sera ('good evening')

English has the three expressions in (9) used at different times of day. The boundary between the times when (9a) and (9b) are appropriate is conventionally fixed at noon. The one between (9b) and (9c) is somewhat less strict. Conventionally, (9c) is used between the end of the working day and bedtime. French only has the two expressions in (10) and Italian the two in (11). On the basis of the glosses, we might expect (10a) and (11a) to correspond to (9a-b) combined and (10b) and (11b) to (9c). In fact, however, the boundary between (10a) and (10b) falls somewhere in the period covered by (9c), whereas the one between (11a) and (11b) falls somewhere in the period covered by (9b). In Italian the logic of the division is to wish someone well for a period that is about to start rather than one that is almost over.

From this description of the situation, it is clear why there is no point in entering expressions such as (9) in a TDB. It is not only impossible to define them as terms, in the absence of a genuine concept, but their translation is dependent on contextual information to such an extent that automatic replacement is not efficient.

Not all formulaic expressions behave as the ones in (9). In a sense, the expressions in (9) are extreme cases because their meaning is (almost) completely situational. In order to explore the behaviour of more referentially or conceptually oriented formulaic expressions, we analysed an English-Spanish translation corpus of technical texts. This corpus is described in more detail by Fernández Parra (2007). Here we will concentrate on the expressions listed in (12). We used the corpus as a source of examples, but in their discussion we will not limit ourselves to the occurrences in our corpus.

(12) a. special needs
b. raw material
c. code of practice
d. in progress

### 4.1. *Special needs*

The expression *special needs* in (12a) is used in a variety of contexts. Its meaning is transparent and compositional, but there are a number of indications that it is prefabricated. This can be seen in the examples in (13).

(13) a. This is a student with special needs.
b. This is a special needs student.
c. #The needs of this student are special.
d. #This student has only one special need.

The contexts of *special needs* in (13a-b) are typical. In (13b) it is what ten Hacken (2003) calls a *phrase word*, occurring in the non-head position of a compound. As indicated by #, in (13c-d) the expression does not occur in its usual sense. In an educational context, (13c) is not an appropriate paraphrase of (13a-b), indicating that it violates the definition of formulaic language in (2). Similarly, (13d) cannot be used as a parallel to (13a).

As suggested by the examples in (13), a common field in which the expression *special needs* is used is education. However, it also occurs in other fields, e.g. when referring to passengers in public transport. We also found (14) in our corpus in a text on chiropractic healthcare.

(14) Modified adjusting approaches can help children, the elderly or those with special needs.

In the translation to Spanish, two main alternatives should be considered. They are listed in (15).

(15) a. necesidades especiales ('special needs')
b. necesidades educativas especiales (NEE) ('special educational needs')

In cases such as (14), the literal translation in (15a) is correct. In the educational field, (15b) has been established as a standardized term. In the translation of (13a-b), (15b) should be used.

From this discussion, the following treatment in a CAT tool can be deduced. In the field of education, *special needs* is a term with (15b) as its translation. It is important that whenever this term is activated, the translator checks whether the context is appropriate, because (15b) is more restricted than the English (12a). For the occurrence of (12a) in other fields, it remains to be established separately whether there is any terminological motivation for a TDB entry. Even if in (14) no standardized definition is presupposed, it may still be worthwhile to specify an entry with (15a) as its translation. Of course, such an entry will not match contexts with variations of the type of (13c-d). The added value of such an entry is determined only by the gain in efficiency.

### 4.2. *Raw material*

The expression *raw material* in (12b) originates from the description of manufacturing processes. A representative example from our corpus is (16a).

(16) a. Through its major derivative, sulfuric acid, sulfur ranks as one of the more-important elements used as an industrial raw material.
b. #Sulphur is raw.
c. #Sulphur is a material which is raw.

The non-compositionality of the expression can be observed in the semantic deviance of (16b-c). The last example also indicates that the word order is fixed. The only possible variation is that the expression may appear in the plural.

At first sight, we may conclude that *raw material* is a term in the field of manufacturing. This is surely correct, but it is not the full story. The expression shows a particular tendency to be used metaphorically. Two examples are given in (17).

(17) a. Innovation is inherent in software, since it is the only way to compete because "manufacturing" (producing copies), distribution and raw materials are equally cheap for all players.
b. raw material for the army

Example (17a) is from our corpus. It shows an explicit metaphorical extension of manufacturing with its accompanying terminology to the field of software engineering. Example (17b) is from the *Collins English Dictionary* (5th edition, sub *raw material*, sense 2). Here we have an implicit metaphorical extension in which military training is conceptualized as a manufacturing process.

The translation of *raw material* into Spanish cannot be the compositional #*materia cruda*. In the literal sense in (16a) and in mild extensions such as (17a), the correct translation is *materia prima* ('basic material'). In (17b), however, a different metaphor has established itself in Spanish. An idiomatic translation is *madera para el ejército* ('wood for the army').

Let us now return to the question whether or not (12b) is a term. In the context of manufacturing, it can definitely be treated as a term. It is possible to come up with a rigid, terminological definition listing the conditions for something to be a *raw material*. Such a definition will normally not cover the uses in (17). However, if we have a TDB entry, it will also recognize the cases in (17) unless we take special precautions. We might for instance deactivate the entries not pertaining to the subject field of the source text. There may be various procedures for doing so, depending on the specific CAT tool used. If no such restriction is specified, the translator will have to override the TDB suggestion in the case of (17b).

It is interesting to speculate under what circumstances it would be appropriate to specify a TDB entry for *raw material* with a translation as *madera*. In the relevant sense, (12b) is clearly not a term. The only reason to include such a TDB entry would be to speed up translation and avoid errors. The entry is only efficient if the expression occurs frequently. However, when a marked image such as (17b) is used often in a text, this is usually regarded as bad style. While opinions on the correct approach to bad style in the source text may diverge, there is certainly no obligation on the part of the translator to be *consistent* in the

translation of overused images (cf. Newmark (1988: 147)). Therefore, it is arguably better to avoid TDB entries of this type on all occasions.

### 4.3. *Code of practice*

The case of *code of practice* in (12c) is interesting because the translation into Spanish highlights an ambiguity that is not directly apparent in English. In our corpus, the expression occurs in (18).

(18) The laws and regulations adopted in pursuance of Article 4 above may provide for their practical application through technical standards or codes of practice, or by other appropriate methods consistent with national conditions and practice.

In Spanish, the two translations in (19) are the most likely to be chosen for *code of practice*.

(19) a. código profesional ('professional code')
b. repertorio de recomendaciones ('catalogue of recommendations')

The translation of (18) in our corpus contains (19b). In the context of (18), however, only (19a) is correct. The difference between (19a) and (19b) is that the former has a much more institutionalized authority. An example of a definition of this sense of the concept is (20)

(20) Rules established by regulatory bodies or trade associations, which are intended as a guide to acceptable behaviour. As such they do not have the force of law behind them.

Taken from the website of EDP Health and Safety Consultants (http://www.edp-uk.com/glossaries/terms.htm, 10 Sept. 2007), the definition in (20) indicates the status of a code of practice quite explicitly. However, not all instances of *code of practice* have the official character implied by (20). An example of a less formal use is (21).

(21) We aim to put our customers at the heart of everything we do, so we have written a code of practice that we hope is clear and useful. This code intends to:
- outline the main services we offer;
- tell you how to contact us;

[…]

(21) is taken from the website of British Telecom (www.btplc.com/Thegroup/Regulatory information/Codeofpractice/Consumercodeofpractice/ConsumerCodeofPractice.htm, 6 Sept. 2007). It uses a much less formalized concept of *code of practice* than (20). For the translation of (21) into Spanish, (19b) is to be preferred over (19a).

The problem of translating (12c) highlights a general limitation of TDBs. As we concluded in the discussion of the translation alternatives in (6), the translation of a term involves the analysis of the concept it refers to. The English term *code of practice* refers to a catalogue of regulations that may have different degrees of authority. The form itself does not specify the degree of authority. Arguably, only the higher end of the spectrum, exemplified by (20), is a genuine term. In Spanish, two expressions are available corresponding to different degrees of authority.

If the TDB is used in a large translation project, the primary concern is consistency. One approach to achieve this is for the terminologist to determine the nature of the concept referred to before the text is translated. If all occurrences refer to one type of *code of practice*, only the relevant entry is made available. Given that this scenario assumes a single, large text as input, it is not unlikely that this is indeed the case. By centrally selecting the right translation, terminological consistency can be achieved.

In the case of a TDB used in the single translator scenario, the variety of different, typically short source texts makes it much more likely that both readings occur in the domain of application of the TDB. The possibility for encoding different, 'conflicting' entries and the strategies for resolving such conflicts will differ between CAT tools. A possibility offered by any CAT tool is to make a note of the problem in the TDB entry. However, if the TDB is used for batch translation of terms, this information will not be automatically displayed. Only if the translator remembers that there was an issue and consults the TDB manually on translating the relevant segment will the note with the relevant information appear.

### 4.4. *In progress*

The expression *in progress* in (12d) differs from the ones in (12a-c) because, in the terminology of Jackendoff's (1983) Conceptual Structure, it refers to a PROPERTY rather than a THING. The property is typically attributed to a process. Of the examples below, (22a) is taken from our corpus and (22b) from the British National Corpus (BNC).

(22) a. Additional studies using more sequences and more taxa are in progress.
b. The papers presented to the seminars were essentially 'work in progress' reports and should not be read as completed pieces of research.

There are strong arguments for treating *in progress* as a formulaic expression. They are based in particular on the relative frequency of the expression, suggesting that it is preferred to alternatives, e.g. *in course*, *in process*, *in development*. Of the 551 occurrences of *in progress* in the BNC, 113 are for *work in progress*. This suggests that *work in progress* is a formulaic expression as well. This is not an argument against treating *in progress* as a formulaic expression as well. There is no reason to assume that the mental lexicon has to be storage-efficient in the sense of only storing information that cannot be compositionally derived. If that were the case, there would not be formulaic language.

From the monolingual perspective, it seems tempting to treat *in progress* by means of a TDB entry, because it refers to a concept and occurs frequently. When we consider translations into Spanish, however, the problems of such an approach are evident. The translation of (22a) in our corpus is (23a). (23b) gives a word for word back translation into English.

(23) a. Se están realizando actualmente estudios adicionales usando más secuencias y taxones.
b. 'Are being carried_out currently studies additional using more sequences and taxa'

No direct translation of *in progress* occurs in (23a). Instead, *is in progress* is translated by the progressive form of *realizar* and the adverb *actualmente*. The selection of *realizar* is

determined by the object *estudio*. From the Spanish perspective, the only formulaic expression involved is *realizar un estudio* (‘carry out a study’).

This picture is confirmed by the entry for *in progress* in the *Collins Spanish-English Dictionary* (6th edition). In all of the examples, a direct translation of *in progress* is avoided. The two examples in (24) illustrate this.

(24) a. Negotiations are still in progress.
Aún se están manteniendo las negociaciones.
(lit. ‘Still are being maintained the negotiations’).
b. “Silence: exam in progress”.
“Silencio: examen”.
(lit. ‘Silence: exam’).

In (24a), the same translation strategy is used as in (23). The expression *in progress* is combined with the verb and translated by a progressive form of a verb that fits the object. In this case, the adverb *still* is also integrated in the selection of the verb *mantener*. In (24b), where this strategy is impossible because there is no verb, the entire expression *in progress* is left untranslated.

From this discussion we can draw the conclusion that it is not possible to translate *in progress* idiomatically into Spanish without taking into account the larger context. In the absence of a uniform translation, no proper TDB entry can be specified.

## 5. Conclusion

The question we studied in this paper was whether the tools available in CAT for the treatment of terminology could also be used fruitfully for the translation of formulaic expressions. A central difference between terminology and formulaic language is that the former refers to a standard whereas the latter refers to the mental lexicon. As a consequence, although both require a holistic approach in translation, the translation of a term always passes through an analysis of the concept it refers to, whereas a formulaic expression can be translated on the basis of the situations in which it is used and general language competence.

CAT tools make available TDBs for the translation of terminology. These TDBs differ in their approach to translation from the preferred terminological approach, which assumes that a term is described independently in each language. TDBs do not offer an equivalent to the correspondence records that are central in such an approach. The purpose of a TDB in a CAT tool is not standardization without bias to any of the languages involved, but improvement of consistency and efficiency in translation.

In the use of a TDB by translators, priorities may vary. In a scenario in which a large translation project is carried out by a group of translators, consistency is the most important concern. In a scenario in which a single translator uses a TDB to record previous terminology work, efficiency is the first concern. This difference in priorities leads to a slightly different set of preferences for the organization of a TDB. In the former scenario, more explicit, formal specification of terminology is desirable, whereas in the latter the selection of information can be influenced to a larger degree by the expected use of individual pieces of information.

For the optimal treatment of formulaic expressions by means of a TDB, two conditions apply. First, the expression has to refer to a concept that exists in more or less the same form in both languages. Second, it must be possible to identify the concept on the basis

of the form as it appears in the source text. In the discussion of various formulaic expressions in section 4, we found a number of cases in which these two conditions were violated to different degrees.

A clear violation of the first condition is in the translation of *good morning*, cf. (9), (10) and (11). The meaning in this case is a set of appropriate situations that differ from one language to another. The case of *in progress* in English-Spanish translation is similar, because although it can be seen as a concept in English, there is no directly corresponding concept in Spanish. In the case of metaphors that are still felt to be alive, as in *raw material* in (17b), good style favours an attitude in which each instance is considered separately. This makes the establishment of a concept in a TDB less desirable.

The translation of *good morning* also presents a good example of a violation of the second condition. The comparison of the use of (9), (10) and (11) shows that conditions on appropriateness depend on subtle indications that are not readily deduced from the form alone. For a number of expressions discussed, we found that they are ambiguous and that the two readings require different treatments in translation. The translation problems they present are slightly different, however. In the case of *raw material*, we have a term-like expression and its extended metaphorical use. In the case of *special needs*, we find a term and a more compositional expression. In the case of *code of practice*, what seems to be vagueness in English corresponds to two translations in Spanish.

It is clear from the outset that in a scenario with a terminologist maintaining a TDB for a large translation project, a larger degree of support for the translators can be expected, because the context of the translation can be more narrowly constrained so as to exclude some of the readings of ambiguous expressions. In a scenario with a single translator, more of the information about the ambiguity must be stated in the TDB. The retrieval of this information then depends on the way the database is used. If terms are translated in batch mode, there is no visible reminder of the problem unless the translator puts a code character in the target language string as a personal reminder that potentially crucial information for the translation of the expression is available in the TDB.

To what extent the methods and considerations discussed here are in fact generalizable is a question that requires further research. Another question that we have left for future research is how the availability of other tools in CAT packages, such as the Lexicon in Atril's DéjàVu, affects the considerations formulated here.

## References


CABRÉ, M. Teresa. 1999. *Terminology: Theory, methods and applications* [DeCesaris, Janet Ann, transl.; Sager, Juan C., ed.]. Amsterdam: Benjamins.

CHOMSKY, Noam. 1959a. 'On Certain Formal Properties of Grammars'. *Information and Control* 2:137-167.

CHOMSKY, Noam. 1959b. 'A Note on Phrase Structure Grammars'. *Information and Control* 2:393-395.

CHOMSKY, Noam. 1965. *Aspects of the Theory of Syntax*. Cambridge (Mass.): MIT Press.

CHOMSKY, Noam. 1986. *Knowledge of Language: Its Nature, Origin, and Use*. Westport (Conn.): Praeger.

FERNÁNDEZ PARRA, María. 2007. 'Towards a definition and classification of formulaic language for its translation in specialized texts'. In NENONEN, Marja & NIEMI, Sinikka (eds.). *Collocations and Idioms 1: Papers from the First Nordic Conference on Syntactic Freezes, Joensuu, May 19-20*, 2006. Joensuu: University of Joensuu, pp. 113-127.

TEN HACKEN, Pius. 2003. 'Phrases in Words'. In TSCHICHOLD, Cornelia (ed.). *English Core Linguistics*. Bern: Lang, pp. 185-203.

TEN HACKEN, Pius. 2006. 'Introduction'. In TEN HACKEN (ed.), pp. 7-15.

TEN HACKEN, Pius (ed.). 2006. *Terminology, Computing and Translation*. Tübingen: Narr.

TEN HACKEN, Pius. 2007. *Chomskyan Linguistics and its Competitors*. London: Equinox.

JACKENDOFF, Ray. 1983. *Semantics and Cognition*. Cambridge (Mass.): MIT Press.

JACKENDOFF, Ray. 2002. *Foundations of Language: Brain, Meaning, Grammar, Evolution*. Oxford: Oxford University Press.

MAEGAARD, Bente; PREBENSEN, Henrik & VIKNER, Carl. 1975. *Matematik og Lingvistik*. Odense: Odense Universitetsforlag.

NEWMARK, Peter. 1988. *Approaches to Translation*. New York: Prentice Hall.

QUAH, Chiew Kin. 2006. *Translation and Technology*. Basingstoke: Palgrave Macmillan.

STREHLOW, Richard A. 1997. 'ISO10241: Preparation and Layout of Terminology Standards'. In WRIGHT & BUDIN (eds.), pp. 203-208.

TSCHICHOLD, Cornelia. 2000. *Multi-Word Units in Natural Language Processing*. Hildesheim: Olms.

WRAY, Alison. 2002. *Formulaic Language and the Lexicon*. Cambridge: Cambridge University Press.

WRIGHT, Sue Ellen. 1997. 'Term Selection'. In WRIGHT & BUDIN (eds.), pp. 13-23.

WRIGHT, Sue Ellen & BUDIN, Gerhard (eds.). 1997. *Handbook of Terminology Management*. Amsterdam: Benjamins.

WRIGHT, Sue Ellen. 2006. 'Terminology Standards for the Language Industry'. In TEN HACKEN (ed.), pp. 19-39.

*Authors' address:*

*Swansea University*
*Department of Modern Languages*
*Singleton Park*
*Swansea, SA2 8PP*
*United Kingdom*
*p.ten-hacken@swansea.ac.uk*
*116435@swansea.ac.uk*

In *SKASE Journal of Translation and Interpretation* [online]. 2008, vol. 3, no. 1 [cit. 2008-04-21]. Available on web page <http://www.skase.sk/Volumes/JTI03/pdf_doc/1.pdf>. ISSN 1336-7811.