*María Fernández-Parra**

The Workflow of Computer-Assisted Translation Tools in Specialised Translation

1. Introduction

Since the wide availability of computers, the work profile of the professional translator has radically changed. Translators no longer work in isolation relying on a typewriter and a pile of books as their only aids. However, the goal of a totally independent translating machine producing high quality output has not been achieved either, and may never be achieved. It would be practically unthinkable for translators nowadays to work without a computer, but integrating computers into the translation workplace does not mean replacing human translators altogether. The term **computer-assisted translation** (henceforth **CAT**) refers to the integration of computers into the workplace, whereas the term **machine translation** (**MT**) refers to fully automating the translation process. The workflow described here is that of CAT, but some differences and similarities with MT are also pointed out as appropriate.

1.1. Aims

The first aim of this paper is to explain in simple terms what is meant by the term computer-assisted translation and how this type of translation differs from other types of computer-based translation. This is broadly discussed in section 2 below. The second aim of this paper is to raise awareness of the new working methods of translators of specialised texts by describing a simplified but typical workflow in the specialised translation scenario using CAT tools, in section 3 below. In other words, I aim to describe what computers already can do at present. It is hoped that the description of such a workflow will highlight some of the differences between a traditional translation process without computers and the translation process with CAT tools. Finally, the third aim of this paper is to determine how computers could do more for us. In particular, in section 4, I explore the possibility of increasing the functionality of CAT tools, which is the topic of my PhD.

1.2. CAT tools and specialised texts

CAT tools are particularly suited to assist in the translation of scientific and technical texts, which for the sake of convenience I refer to as specialised texts here. These texts typically contain numerous terms and a high degree of repetition. Instruction manuals for printers are a case in hand. They may contain technical terms such as *inking roller train, back unit printing blanket, coldset dry*, or *gum Arabic*. They may also contain identical sentences, or even paragraphs, where the only difference is a model number, e.g. *Replace the colour cartridge with a CX C6578 Series* or *Replace the colour cartridge with a CX No.78 cartridge*. This is why most CAT tools contain two types of database, one for terms (termbase, or terminological database) and one for sentences (translation memory or TM) for their efficient storage and reuse in subsequent translations.

Terms may occur both as single terms and as multi-word units. Since not all multi-word units are terms, it would not be too far-fetched to consider that CAT tools may be used to deal with dif-

* *María Fernández-Parra*
Swansea University
UK
116435@swansea.ac.uk

ferent types of multi-word units, in a similar way to that of terms. In particular, my PhD focuses on a typical pre-translation task of the workflow, namely the terminology extraction task, as explained in section 4 below. This project intends to establish whether CAT tools can be used to extract multi-word units from specialised texts as effectively as terms.

1.3. The specialised translation industry

Scientific and technical translation is an ever-growing industry whose purpose is to disseminate information at an international level. It is somewhat paradoxical that, despite the dominance of English as a lingua franca in settings requiring communication at an international level, the demand for translation has been steadily increasing (Gotti & Šarčević 2006: 9).

Although "technical translation has traditionally been regarded as the poor cousin of "real translation" (Byrne 2006: ix), it has been estimated that technical translation accounts for some 90% of the world's total translation output each year (Kingscott 2002: 247 in Byrne 2006: ix). This is hardly surprising "given the importance attached to the availability of technical information in a variety of languages (…) motivated partly by the increasingly international focus of many companies and partly as a result of legislation", which requires "the provision of comprehensive, accurate and effective technical documentation in a variety of languages" (Byrne 2006: 2).

Byrne (2006: 2) concludes that "it is clear to see why technical translation is one of the most significant employers of translators". It would seem, then, that the increase in the demand for technical translation can only partially be met by an equally increasing automation in the process of translating technical documentation.

2. Approaches to computer-based translation

In characterising the role of human translators and the role of computers in the translation process, we can distinguish between human translating activity, assisted by computers (CAT), and machine translating activity, assisted by humans (MT). In CAT the translation task as such is performed by the human translator with assistance from the computer in tasks such as formatting or terminology lookup. In CAT, the computer does not perform a syntactic and semantic analysis of the source text to any extent, nor does it generate a target text. These are precisely the tasks that the computer performs in MT. The assistance from human translators in MT is typically in pre- or post-editing tasks. In the next sections, I give a brief overview of CAT and MT in order to illustrate the niche that each takes in the translation market at present.

2.1. Computer-assisted translation (CAT)

As mentioned in the previous section, in CAT the machine typically does not translate text but instead it supports the translator in a range of tasks, such as formatting, grammar checking, etc. Some of the tasks the machine performs can be rather sophisticated, e.g. automatic terminology lookup, source and target text alignment, reuse of previously translated material, verification of terminology consistency, etc. The level sophistication in the tasks performed allows us to establish a hierarchy of CAT tools.

On one hand, grammar checkers, spell-checkers and others can be considered to be at the most basic level of use (Somers 2003: 14). On the other hand, on a more sophisticated level of use, there are the so-called **workstations**, defined as "a single integrated system that is made up of a number of translation tools and resources such as translation memory, an alignment tool, a tag filter, electronic dictionaries, a terminology management system and spell and grammar-checkers" (Quah 2006: 93). Examples of workstations include the SDL Trados 2007 suite, the current leader in the CAT tools market (see, e.g., Hutchins 2005: 13; Hatim & Munday 2004: 114) and its competitors, such as Atril's Déjà-Vu, Star's Transit, as well as independent CAT tools such as Wordfast.

Localisation tools can also be included under the more sophisticated group of CAT tools. **Localisation** is understood in broad terms as "the process of changing the documentation of a product, a product itself or the delivery of services so that they are appropriate and acceptable to the target society and culture" (Quah 2006: 19). Examples of localisation tools are Corel Catalyst, Passolo, AppLocalize, etc. For a more detailed definition of localisation and discussion of its state in the industry, see Esselink (2000 and 2003) and Pym (2004).

CAT tools have proven to be extremely useful for the professional (particularly technical) translator, who can increase productivity and quality as well as speed of delivery to the client. However, translators must invest time, effort and sometimes money in training themselves to use such tools. Further, CAT tools liberate the translator from many tedious tasks such as formatting, document layout, etc, so that the translator can concentrate on the translation per se. This is how high quality translations can be achieved at an increasingly shorter time.

2.2. Machine translation (MT)

In MT, the task of analysing and decoding the source texts corresponds to the machine, but the human operator may carry out pre-editing or post-editing tasks. The translations produced in this way are also very fast, but typically of low quality. As Hutchins and Somers (1992: 149) state, "we can have either fully automatic translation or high quality translation but we cannot have both".

Quality in many MT systems, however, is not the main goal. Instead, these systems can provide a 'rough' or 'indicative' translation, sometimes referred to as **gist translation**. This is especially useful in the case of long source texts, where MT programs can be used to determine which portion of the source text would warrant professional translation, which takes us back to CAT tools. This is how CAT and MT have been happily co-existing in the translation industry. Examples of MT systems include SYSTRAN (general MT), SPANAM and ENGSPAN (specialised MT), etc.

However, some MT systems can produce very high quality translations. This is possible in specialised contexts with controlled input, such as MÉTÉO in Canada, which translates from English into French and French into English in the domain of meteorology. In the specific domain of health in Latin America, SPANAM and ENGSPAN translate from English into Spanish and Spanish into English respectively (see Loffler-Laurian 1996: 21). These two examples illustrate that, instead of reaching for high-quality general MT, the way forward may lie in high-quality specialised MT. By taking the **skopos**, or purpose, of the translation into account, we can achieve high-quality output in the particular context where it is needed (see ten Hacken 2001).

3. Typical workflow with CAT tools

In this section, I describe in simple terms the typical use of CAT tools in a specialised translation scenario. The workflow described here is generic to most CAT tools, rather than focused on a particular CAT tool, and it can be equally applied to freelance translators working individually and to translators working as part of a team. Because CAT tools are usually software packages consisting of a number of programs that can interact between them or work as standalone programs, describing the workflow of the specialised translation process with a visual image (see figure 1 below) is useful in that it allows us to see at a glance how the results of certain parallel tasks are the basis for the successful performance of subsequent tasks.

3.1. Visualising the workflow

A visual image of the workflow is given in figure 1 below, where we can trace the source text from the moment the client commissions the translation until the finished product is handed to the client. Figure 1 is broadly divided into three sections, namely Before translation, 'During translation' and 'After translation', but in practice the boundaries between each section may not be as clear-cut. Because the workflow in the translation process is centred around the source text, the

thick arrows in figure 1 indicate where the source text stands at each of the three stages. These arrows indicate how some of the tasks performed affect the source text directly, such as 'Term extraction', whereas the benefits of other tasks, such as 'Text alignment', are more indirect.

In addition to exemplifying the workflow in chronological order, this division into three stages is also intended to highlight the particular importance of the 'Before translation' stage. The quality and quantity of the tasks performed during this stage have a direct bearing on how effective the translation with CAT tools is in the 'During translation' stage. This is further explored in the following section and the 'During translation' and 'After translation' stages are discussed in sections 3.2 and 3.3 respectively.

3.2. Preparing for translation

When undertaking any translation, preparation work is always essential, but in the context of working with CAT tools, the tasks performed prior to translation take on special significance because the higher the quality and quantity of the tasks performed during this stage, the more efficient the translation of the source text will be afterwards. This is because the two main databases common to most CAT tools, the translation memory and the terminological database (or termbase), as explained in section 1.2 above, are initially empty when starting to work with CAT tools. It is up to the CAT tool user to **populate** both the translation memory and the termbase with relevant material prior to translation, whenever this is possible.

The main tasks that can be performed prior to translation are **text alignment** and **terminology extraction**, but some CAT tools also offer the possibility of **converting** previous or third-party databases into standardised formats in the translation industry such as .tbx, which allows the exchange of data between CAT tools. The results of text alignment (typically sentences) will be fed to the **translation memory** and the results of terminology extraction (typically terms) will be fed into the **termbase** (or terminological database).

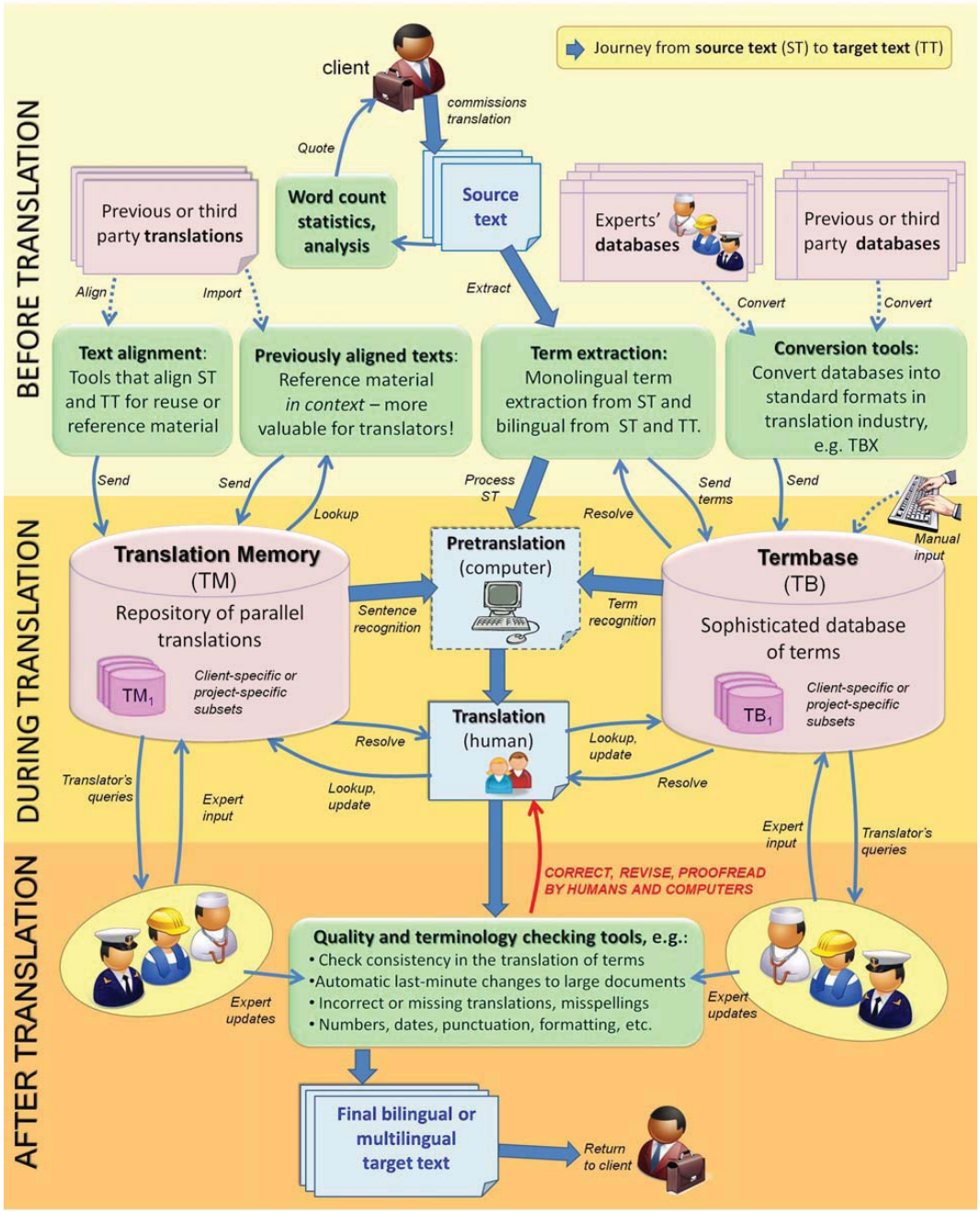

Figure 1. A simplified but typical workflow with CAT tools

As explained in section 1.2 above, the difference between translation memory and termbase is that the translation memory contains source-target segments at sentence level (sentences, paragraphs, titles), whereas the termbase contains source-target segments at term level (single words or multiword items). If there is no previous or relevant terminological material available, CAT tool users can of course create blank translation memories and termbases which can be populated manually when such material becomes available. Additional tasks that CAT tools can perform prior to translation include **word count**, **statistics** and **analysis**.

3.2.1. Word count, statistics and analysis of the source text

By performing word count, statistics and analysis tasks, CAT tools produce the figures on which translators can base the quote for the client, that is, how much the translation will cost. The word count feature can count the words in each source file separately and in the project as a whole. Statistics and analysis tasks compare the source text or texts to the existing translation memories in order to provide information to the translator such as the percentage of repetition in the source text, including full matches, fuzzy (partial) matches (see section 3.3.1) and segments with no match (0% match). This information will allow translators to calculate how much new material there is to translate in the source text and how much material has been previously translated.

As well as the cost of the translation, these figures allow translators to calculate the amount of time needed for the completion of the project, on the basis that 0% matches require most work and 100% matches may require no work at all. However, as any translator who has worked with translation memory software knows, this is simply not true. Even in the case of 100% matches, there is always post-editing work to do.

A common practice in the translation industry is for the client to pay different rates depending on the matching percentage range of each segment (Giammarresi 2008). In other words, 0% matches may be paid full price, mid-level matches may be paid half price and full matches are sometimes not paid at all (Giammarresi 2008), which can be construed as a penalty for using the software. Although Giammarresi admits that "translators have their share of responsibility for this state of affairs", this practice has practically been imposed on them. Giammarresi concludes that there is no easy solution, at least in the short term.

3.2.2. Text alignment

Alignment is the process of breaking down a source text and its target text into segments and matching each source text segment to its corresponding target language segment. Segments are also referred to as **translation units** and are typically sentence-long, although shorter segments are also possible, such as titles, list items or table cells and larger segments such as whole paragraphs. Examples of alignment tools are WinAlign by SDL Trados and Translation Customizer by LanguageWeaver.

The algorithms that CAT tools use for the segmentation of the texts are typically based on punctuation, formatting, or dates (Quah 2006: 100) which should, at least theoretically, be identical in both source and target texts. Such common points between source and target text function as anchors (Simard et al. 1992) which the machine relies on in order to establish the similarities between both texts and thus align each translation unit. Therefore, alignment is most successful when the source and target text are very similar, if not identical, in structure.

The rationale behind the inclusion of an alignment tool in a CAT package is that the user could take advantage of any translations carried out outside the CAT package by making a translation memory with them or integrating them to an existing translation memory, provided that both source and target texts are in electronic form. The purpose of alignment is two-fold; the results of alignment can either be used to create a new translation memory of previous translations for future reference or reuse in subsequent translations; or the results can be added as new translation units to an existing translation memory. Once the segments are added to a translation memory, they are available for immediate use; this is why it is probably most useful to perform text alignment before starting to translate the current source text; this is also the reason why text alignment features in the 'Before translation' stage in figure 1.

Although alignment can be performed manually by the user (Austermühl 2001: 177), most CAT tools packages provide their own automatic alignment tool. However, errors in the automatic alignment of source and target segments are practically inevitable. This is often because the basic unit for the segmentation of the texts is the sentence but "not all text is written in sentence form" (Bowker 2002: 94) but also in tables, lists, etc. There could be errors in automatic alignment even

when the texts consist of straightforward sentences because, for example, the target language may use more word order inversion than the source language, or a cultural reference has been dropped in the target text (Bowker 2002: 111), which could confuse the alignment tool. For best results, the user should check the results of alignment and manually post-edit any misaligned segments. Depending on the texts to be aligned, it may also be worth considering the pre-editing of the source and target texts so that their structure is as similar as possible before alignment, therefore minimising the number of errors in the output.

As mentioned above, any previously aligned material can be inserted (whether edited or not) from a translation memory into a translation, but it can be also used as reference material only. In the use of aligned material as reference, some CAT tools, such as PerfectMatch™ by SDL Trados, make a distinction between reference material and reference material *in context*. This distinction is illustrated in the 'Previously aligned texts' box in the 'Before translation' stage in figure 1. The usefulness of reference material *in context* is enhanced from that of reference material alone because the source text and the reference material share the same domain (medical, legal, etc.) or even sub-domain (microbiology, criminology, etc.). Translators can then be assured of the correct use of specific terminology in the right domain.

3.2.3. Term extraction

Term extraction is one of the newest types of tasks performed by CAT tools (Bowker 2003: 51) and it has been defined as the operation of identifying term candidates in a given text (see Thurmair 2003, Zielinski et al. 2005), with the purpose of creating, enlarging and updating termbases, which will enhance translator productivity by speeding up the process of locating, collecting and storing terminology. Although most CAT tools provide terminology management software, usually in the form of a terminological database, or equivalent, not all CAT tools integrate a terminology extraction program as part of the package. Examples of term extraction software are MultiTerm Extract by SDL Trados, and Xerox Terminology Suite.

Term extraction, or **term identification**, should not be confused with **term recognition**, the former illustrated in the 'Before translation' phase in figure 1 and the latter illustrated in the 'During translation' phase. Term extraction refers to finding new terms to add to a termbase, whereas term recognition refers to spotting occurrences in the source text of terms already in the termbase. Typically, term extraction takes place prior to translation because entering the extracted terms and their translations into the termbase before starting to translate helps to ensure terminological consistency throughout the target text. By contrast, term recognition typically takes place during translation. With automatic term recognition settings enabled, the CAT tool usually opens a new window automatically in which any terms in the source text matching an entry in the termbase are displayed, together with any information about the term contained in the entry. It is always up to the human user to decide to use or ignore the proposed term.

According to Cabré et al. (2001), there are three types of term extraction tools: (a) those that use **linguistic methods**, (b) those that use **statistical methods** and (c) **hybrid systems**, that is, using both statistical-based and linguistic-based term extraction methods. Statistical systems extract potential terms using frequency algorithms of lexical units, based on the premise that, in specialised domains, terms have higher frequencies than in general language texts (Ha et al. 2008). Therefore, to put it simply, the strings with the highest relative frequencies are proposed by the system as potential terms.

Although statistical methods focus on frequencies of lexical items, there is a small linguistic component in the form of **stopwords**, which are words such as *the, on, by*, etc., to be excluded from the extraction process. Linguistics-based systems try to "identify terms by their linguistic (morphological and syntactical) structure" (Ha et al. 2008: 108), whereas hybrid systems extract terms based on a combination of syntactical patterns and statistical regularities (see Dias et al. 2000).

The advantage of statistical methods over linguistic ones is that statistical ones are language-independent, that is, they can be applied to any language because they do not require the integration of dictionaries into the software or the creation of a set of language-specific rules that linguistic systems ones do. By contrast, linguistic methods are designed to produce a higher quality output than statistic ones, that is, a higher number of correctly selected terms and fewer irrelevant strings. More recent research shows a move towards combining both linguistic and statistical techniques in hybrid systems (e.g. Feldman et al. 2006, Koeva 2007).

A term extraction can be **monolingual** or **bilingual**. Monolingual term extractions are usually performed on the source text to be translated. The terms extracted in this way are stored in the termbase without their translations which have to be found elsewhere by the translator and added manually to the termbase prior to translation, or they may be imported from other termbases. A bilingual extraction is usually performed on previously completed translations. The extracted terms from the source text together with their translations from the target text are exported and stored in the termbase and will be looked up and used in future translations as required.

In the typical use of term extraction software, the machine produces a list of so-called **candidate terms**, that is, a list of single words or multi-word items which are likely to be terms in the given text, though it has been claimed that extracting single words is seldom useful (Thurmair 2003) because the majority of terms, at least in technical texts written in English, is made up of multi-word items. Once the candidate term list is produced, it is then up to the user to search through this list in order to seek out the 'true' terms for inclusion in a termbase and manually discard the non-terms.

Once the user has discerned terms from non terms, term extraction software often supports a **validation** facility. Validating a term stands for marking a term, often by ticking a checkbox or by use of colour-coding, in order to confirm that it is a valid term. After validation, all non-validated terms can usually be deleted as a batch. The validated terms are then exported to a termbase, or a new termbase can be created with them.

Before export, term extraction software allows the user to fill in a **terminological record** for every term. In this record the user can enter additional information about the term, such as part of speech, gender, acronym, etc. The user can then export the validated terms together with any additional information to a termbase. The terms and any additional information about them can be directly accessed during translation through the termbase, as shown in figure 1.

3.3. Translating the source text

In order to translate the source text, CAT tools generally provide two types of database, as mentioned in section 1.2 above, where reference material is stored, the **translation memory** and the **termbase**. They contain the reference material which the CAT tool will use to attempt a **pretranslation** of the source text by finding any matches in the source texts of segments in the translation memory or in the termbase.

3.3.1. Translation memory

The **translation memory** is a bilingual or multilingual repository of parallel translations, whether previously completed by the users themselves or collected from other sources. The translation memory contains sentences or sentential items (e.g. headings, titles), called **translation units**. It can also contain whole paragraphs. It makes sense in specialised translation to have such a collection of sentential items because of the repetitive nature of many technical and scientific texts. The translation memory is usually populated with the results of text alignment (see section 3.2.2), in the case of translations produced outside the CAT tool, and also by exporting completed translations. The user can also create project-specific or client-specific translation memories. Programs such as SDL Trados, Lingotek and Wordfast integrate translation memories in their software.

Once a translation memory has been created and populated with material, there are two main ways in which it can be used. One method to retrieve the material in the translation memory is that, when a match occurs between a segment in the source text and the translation units in the translation memory, the CAT tool can automatically insert the corresponding equivalents directly into the target text or propose a number of similar but not exact equivalents by opening a separate window from which the user will select a proposal as appropriate or ignore them altogether. The human user will also check the automatically inserted segments and translate any segments of the source text for which no match was found. Another way to use the translation memory is to manually search it for portions of sentences or whole sentences, when a specific query comes up.

In both methods of use of the translation memory system, a **full match** (or **100% match**) means that the source segment and the segment in the translation memory are identical, both linguistically and in terms of formatting (Bowker 2002: 96). When the source segment and the translation memory segment are similar but not identical, the term **fuzzy match** is used. Thus, a fuzzy match could mean, for example, that the only difference between source segment and the translation memory segment is the use of bold formatting in one of them. A good translation memory system will "highlight the differences between the segments to draw the translator's attention to areas that may need to be edited before the proposed translation can be integrated into the new target text" (Bowker 2002: 99).

The degree of similarity in fuzzy matches may vary from 1% to 99%, a 99% fuzzy match being the most similar to a source segment, although some CAT tools only allow percentages from 30% to 99%, for example. The user can adjust the threshold of similarity for the task in hand, bearing in mind that, if the selected percentage is too high, the translation memory system will produce **silence**, that is, some of the similar segments may not be identified. Similarly, if the selected percentage is too low, the system will produce **noise**, that is, too many irrelevant matches may be returned. Generally speaking, in order to obtain the best results, match values between 65% and 75% are applied, 70% being often the default value. Fuzzy matching can save time and effort, but it almost always requires a certain amount of post-editing by the user.

3.3.2. Termbase

The **termbase** is a bilingual or multilingual terminological database that contains terms, either as single words (e.g. *aneurism*) or as multi-word items (e.g. *soft contact lens*). It makes sense to keep a database of terms because terms are crucial elements in specialised translation. Examples of termbases include MultiTerm by SDL Trados and Terminotix.

A termbase, however, is not a mere bilingual or multilingual collection of terms; a termbase produced with a CAT tool differs from other terminological databases in that entries can be organised by *concept*, rather than by terms. For example, the terms *railroad* (USA) and *railway* (UK) would be found in the same entry. This denotes that the concept 'where trains circulate' can be expressed as *railroad* or as *railway*. However, both *railroad* and *railway* would be indexed separately in the termbase, which in practical terms means that the user only needs to search the termbase for *railway*, for example, in order to find both *railway* and *railroad*.

Entries can contain many types of terminological information about terms, such as part of speech, gender, terminological status, etc. The user can also specify specific client preferences and geographical coverage, and create project-specific termbases. Further, CAT tools support the inclusion of multimedia information about a term, such as video, audio or sound files.

The termbase can be populated with the results of term extraction (see section 3.2.3) and through the conversion of previous or third-party terminological databases. Where previous or third-party databases are not available, it is also possible to populate the termbase manually. Though it may also possible to populate a translation memory with translation units manually, this is probably not so frequently the case. The termbase can be looked up during translation or during the post-

editing stage if specific terms need to be reconfirmed. After translation, the termbase is usually updated with the final version of the target terms for possible use in future translations.

3.3.3. Pretranslation

Pretranslation is the term I use here to refer to a task performed by the CAT tool, usually prior to human translation, in which the CAT tool attempts to automatically translate any portions of the source text that have come up before by comparing the source text to all the material contained in the translation memory and the termbase and proposing it to the user or inserting it in the target text directly. Pretranslation is therefore most relevant when the source text to translate contains a large proportion of identical or similar material to that of previous translations, as in the example of printer instruction manuals mentioned above (see section 1.2).

The pretranslation task contributes to the consistency of the target text by ensuring that terms are translated in the same way throughout the target text and by ensuring that any client- or project-specific preferences are taken into account. This is why a successful translation with CAT tools will largely depend on the work carried out in the 'Before translation' stage in figure 1.

Depending on the user's selected preferences, when the CAT tool finds a **full match** or a **fuzzy match** (see section 3.3.1), it can either list all the options in a separate window for the user's consideration or automatically insert any matches in the target text, which the user will then edit or delete as appropriate. It is worth noting that it is always up to the user to decide how the options proposed by the CAT tool should be used in the current translation. The user can choose to ignore any proposal by the CAT tool when the contents of the source text so require. Regardless of the amount of source text translated by the CAT tool, the user should always post-edit the results of pretranslation.

3.3.4. (Human) translation

In the workflow described here, the translation per se of the source text usually takes place after text alignment, term extraction and pretranslation (when applicable). Of course, alignment and extraction can also take place any time that suitable material becomes available to the user. The degree of similarity (**fuzziness**) between the source text (or portions of it) and the material stored in the termbase and the translation memory will determine to what extent the source text can be 'pretranslated' with the pretranslation task, which is performed by the computer. Then the user can proceed to edit the results of the pretranslation task and fill in anything left blank by the computer, and this is the task I refer to as 'human translation' here.

During human translation, the user can look up and, more crucially, update the material stored in the translation memory and the termbase. If any term or translation unit has been modified or updated after the user has started to translate the target text, CAT tools usually have a facility to update the whole target text automatically with the new translations or modifications, thus avoiding any terminological inconsistencies potentially caused by the human user in this way. Such a facility would also allow the user to consult experts at any time, knowing that the target text can be consistently updated or edited afterwards.

It is important to note that, for the CAT tool to be able to automatically update the whole target text, it is more efficient to make any update or modification of terms or translation units through the termbase or translation memory respectively as appropriate, rather than making the changes in the target text directly. Alternatively, if the changes have only been made in the target text, CAT tools usually support the export of the newly translated text to the translation memory and termbase once the translation is completed. After updating termbase and translation memory, the whole target text may be updated thoroughly once again. Post-editing tasks are further explored in the next section.

3.4. Quality assurance tasks

A number of tasks can be performed with CAT tools to ensure the quality of the translated text, as briefly illustrated in the 'After translation' stage in figure 1. The name for each task and their organisation (in different menus or under different tabs) within each CAT package varies, but quality assurance tasks typically include the following:

- spell checks
- number and date checks
- omissions
- terminological consistency in the translation of terms
- adherence to project glossaries
- untranslatables (e.g. formatting codes)
- capitalisation check
- formatting check
- punctuation check
- incorrect or missing translations
- repetition check

The quality assurance tasks performed by CAT tools are considerably more sophisticated than those performed by spell-checking tools in word processing programs, though this may not be readily obvious from the list above. Users will almost certainly need to invest time to familiarise themselves with the names of each task and what each task does, for example *Automatic change – preserve case* as opposed to *Conditional change – preserve case*.

Quality assurance tasks can be performed manually or automatically, during or after translation, on a portion of text or on the whole text. Each task may have a number of settings, preferences or options for the user to choose from. In addition to the tasks mentioned above, quality assurance software can also be used to alert the user to the problems in a specific MT output (DeCamp 2009).

Quality assurance software is almost always required (Quah 2006: 47) but it becomes particularly useful when it is not possible to provide thorough human post-editing (see DeCamp 2009). This type of software can then provide users with feedback that will allow them to correct their own problems. Examples of quality assurance software can be found in Déjà-Vu X, SDLX QA Check, Star Transit, SDL Trados QA Checker, ErrorSpy, QA Distiller, XBench, etc. For a comparison of some QA tools, see Makoushina (2008).

4. Additional functionality of CAT tools

The previous sections give an overview of the tasks that translators of specialised texts typically carry out in a professional setting and describe the workflow in which such tasks are usually organised. In this section, I explore the possibility of increasing the functionality of CAT tools beyond that already explained. Such exploration is the core of my PhD, which is still in progress, as already introduced in section 1.1 above.

4.1. Topic of the PhD

Because of the centrality of terminology in specialised translation projects, CAT tools are strongly geared, as we have seen, towards the treatment of terminology, in tasks such as the identification/recognition, storage, retrieval and update of terms. In my PhD I argue that, because terms are not the only multi-word items in a language, CAT tools may be equally productive in the treatment of other multi-word units as they already are in the treatment of terms. If this is the case, CAT tools would be able to assist in the specialised translation process in further tasks than those described

in the workflow in figure 1 above, without having to change the specifications of the software. In other words, I analyse the performance of CAT tools in tasks they were not actually designed to but which, at least in theory, they should be able to perform. In order to fit the project into the scope of a PhD, I focus on one CAT tool, SDL Trados, one type of multi-word unit, **formulaic** expressions, and one task, the terminology extraction task. The trends in the performance of CAT tools in a second task, the terminology recognition task (see section 3.2.3) are also briefly analysed and the trends in the performance of other CAT tools with formulaic expressions will also be part of the PhD. It is hoped that, if the functionality of CAT tools cannot be extended in the desired way, the results of this research will nevertheless point towards possible amendments or enhancements for future versions of CAT tools.

4.2. Formulaic expressions

A **formulaic expression** can be defined as a "sequence, continuous or discontinuous, of words or other elements, which is or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar" (Wray 2002: 9). Wray's definition is used here as a working definition. Formulaic expressions may include a rather wide variety of expressions, such as idioms (*shoot the breeze*), complex prepositions (*in spite of*), institutionalised expressions (*fish and chips*), routine formulae (*How do you do?*) etc. Their unifying criterion is that, although written as separate words, they form a holistic unit in the mental lexicon. For a discussion of other possible defining criteria, see Fernández-Parra 2008).

Formulaic expressions are pervasive in all types of language use (e.g. Nattinger & DeCarrico 1992, Jackendoff 1997) and this includes specialised texts (e.g. Goźdź-Roszkowski 2006). Formulaic expressions, like terms, may be single words (*Sorry*) or multi-word units (*free of charge*), but in my PhD I focus on multi-word formulaic expressions, as they are more interesting from the computational point of view. Terminology and formulaicity are not usually linked "because the concepts are based in very different domains of linguistics" (ten Hacken & Fernández-Parra 2008: 1). However, in their treatment with CAT tools, both formulaic expressions and terms pose similar problems.

4.3. Using a CAT tool to deal with formulaic expressions

In this section I describe the methodology followed and the initial results obtained in the first stages of my PhD. The only CAT tool that I have analysed to date is SDL Trados, in particular its two term extraction programs, MultiTerm Extract and PhraseFinder. These two programs differ mainly in the file formats supported and in the method of term extraction used: MultiTerm Extract can be considered a statistical term extraction tool, and PhraseFinder a hybrid system (see section 3.2.3).

The differences in the method of extraction between MultiTerm Extract and PhraseFinder mean that there is no reason to expect a significant overlap in the results returned with each of these functions, at least theoretically. Far from becoming a duplication of efforts, or from competing with each other, each program has its place within SDL Trados. The choice of one over another will depend on external factors, at least in part, such as the file format or size of the source text, rather than on the inherent properties of each one.

4.3.1. Methodology

The research is divided into three main stages. Firstly, a corpus of texts belonging to specialised domains was collected, totalling about 420,000 words. The source texts are in English (200,000 words) and the target texts in Spanish (220,000 words).

Secondly, a subset of this corpus of about 10,000 words was selected in order to try out in depth all the settings and combinations of settings available both in MultiTerm Extract and PhraseFind-

er. This sub-corpus was processed manually first and 90 formulaic expressions were extracted. This figure of 90 formulaic expressions is used as a baseline to compare the number of formulaic expressions extracted with different settings and combinations of settings in each program.

Finally, in the last part of the research, which is currently in progress, the settings with which the best results are achieved are used to process the remainder of the corpus, in order to try to establish more accurately the settings that produce the best results in the automatic extraction of formulaic expressions from a source text.

In order to evaluate the performance of the software, I use measures of **precision** and **recall**, which are common in disciplines such as natural language processing and information retrieval. Precision can be described as "purity of retrieval" (Buckland/Gey 1999: 12), that is, how many of the 'candidate' expressions extracted by the system are correct. Recall can be described as "completeness of retrieval" (Buckland/Gey 1999: 12), that is, how many expressions the system extracted or failed to extract.

Because it seems that a system may produce high recall or high precision but not both at the same time (Buckland/Gey 1999: 12), I also use the so-called **F-measure,** as described by Manning and Schütze (1999: 269), which combines measures of precision and recall into a single measure of overall performance.

4.3.2. Initial results

In this section, I give a summary of the results obtained in the second stage of the research, as described in the previous section. Before presenting any results, I would like to emphasize a limitation in the scope of this work, which is that the two term extraction programs analysed here, MultiTerm Extract and PhraseFinder were not designed for the purpose I used them for, so that any conclusion should not be taken as an evaluation of the software.

Because of the different settings and combinations of settings available in both programs, I carried out and evaluated a total of 246 extractions with MultiTerm Extract and 142 extractions with PhraseFinder. In each case, half the extractions were monolingual and the other half bilingual. Monolingual and bilingual extractions can be compared as to the results they achieve for English. In PhraseFinder, both monolingual and bilingual extractions scored identical measures of precision and recall. In MultiTerm Extract, however, the trend observed was that monolingual extractions gave slightly better results than their bilingual counterparts, but it was not possible to establish when bilingual extractions might score better.

Another trend observed in both programs is that very often formulaic expressions are identified as part of a longer string, rather than as expressions in their own right. For example, the expression *learn a lesson* was identified as 'embedded' in the string *Lessons learnt while trying to meet that challenge have led*. Therefore, this string would have to be counted as a correct identification of *learn a lesson*. It was not possible with either program to reduce the string to *Lessons learnt* without drastically increasing the amount of noise (see section 3.3.1). And, of course, the larger amount of noise, the longer the post-editing tasks become for the human user.

Precision was very low throughout all of the extractions with both programs. The highest precision recorded with MultiTerm Extract was 16% and 1.8% with PhraseFinder. Recall was much higher with MultiTerm Extract (85%) than with PhraseFinder (10%).

In the case of MultiTerm Extract, these figures show that there is more recall to be gained than precision to be lost, as precision was so low throughout the experiments. Thus, on the basis of these figures, the better settings were those with which 85% recall was achieved, although the precision measure achieved with these settings was very low, 5%. In practical terms, this means that the human user would have to search through the 1,675 strings returned by the system in this case in order to find 77 formulaic expressions.

In order to search through the large number of strings returned by MultiTerm Extract, the user can refer to **scores**. Scores (or **ranks** in PhraseFinder) are assigned to every candidate term and

they range from 1 to 99 in MultiTerm Extract. The higher the score the more confident the system is that a given candidate term is an actual term, although here I used scores as indicators of formulaic expressions, rather than terms. In this stage of the research I found that most formulaic expressions are assigned scores between 60 and 77. This means that the retrieval of the 77 expressions from 1,675 strings could be could be speeded up by only searching through those which were assigned a score of 60 to 77. However, this range of scores was determined after the analysis of the full set of candidate terms. This raises the question whether it is possible to determine the optimum range of scores before such an analysis.

In the case of PhraseFinder, it was surprising to find that 93% of all the 142 extractions performed also returned the same set of nine formulaic expressions (10% recall), regardless of the settings applied. The other 7% of extractions returned 8 formulaic expressions, which does not allow for many meaningful comparisons. Because precision was also extremely low throughout (1.8%), the F-measure was also very low throughout, the highest being 3%. These figures suggest that, as a general rule, we cannot expect to obtain higher precision and recall measures when extracting formulaic expressions with PhraseFinder without making some changes to its in-built linguistic rules.

5. Conclusion

This paper attempts to raise awareness of the new working methods of translators of specialised texts by describing in simple terms a typical workflow with CAT tools. The term CAT (computer-assisted translation) refers to the integration of computers into the translation workplace, rather than to replacing human translators altogether, as it seems that the increase in the demand for technical translation can (at least partially) be met by equally increasing the automation in the process of translating specialised texts. CAT should not be confused with MT (machine translation) which refers to fully automating the translation process.

The wide range and sophistication of the tasks described in the workflow illustrates how the translating profession has changed since the wide availability of computers. Translators no longer work in isolation and CAT tools allow translators to access external databases and glossaries and to consult experts remotely. Also, the workflow highlights the fact that, when working with CAT tools, the tasks performed prior to translation are essential as they prepare the ground for an efficient translation process. During translation, CAT tools can speed up lookup, retrieval and updating tasks whilst maintaining consistency in the target text. After translation, CAT tools can perform a vast array of quality assurance tasks, thus minimising human oversights and allowing for automatic, last-minute changes or updates to large documents. To sum up, CAT tools can be "the best tools to assist in the management of [specialised] translations from all parties involved" [a]nd "until a computer can think like a human, it seems to be the best choice for starting to automate the process" (Iverson 2003).

6. References

Austermühl, Frank 2001: *Electronic Tools for Translators*. Manchester: St. Jerome Publishing.

Bowker, Lynne 2003: Terminology Tools for Translators. In Somers, H. (ed.), *Computers and Translation: A Translator's Guide*. Amsterdam: John Benjamin.

Bowker, Lynne 2002: *Computer-Aided Translation Technology*: *A Practical Introduction*. Ottawa: University of Ottawa Press.

Buckland, Michael/Gey, Fredric 1999: The Relationship between Recall and Precision. In *Journal of the American Society for Information Science* 45, 12-19.

Byrne, Jodie 2006: *Technical Translation*. Dordrecht: Springer.

Cabré, Maria Teresa et al. 2001: Automatic Term Detection: A Review of the Current Systems. In Bourigault, Didier/Jacquemin, Christian/L'Homme, Marie-Claude (eds.), *Recent Advances in Computational Terminology*. Amsterdam: John Benjamin.

DeCamp, Jennifer 2009: Translation Tools: Status, Practice and Gaps [workshop]. In *Proceedings of the Twelfth Machine Translation Summit*, Canada.

Dias, Gaël et al. 2000: Combining Linguistics with Statistics for Multiword Term Extraction: A Fruitful Association? In *Proceedings of Recherche d'Informations Assistée par Ordinateur, RIAO 2000*, Paris, France.

Esselink, Bert 2003: Localization and Translation. In Somers, Harold (ed.), *Computers and Translation: A Translator's Guide*. Amsterdam: John Benjamin.

Esselink, Bert 2000: *A Practical Guide to Localization*. Amsterdam: John Benjamin.

Feldman, Ronen et al. 2006: TEG: A Hybrid Approach to Information Extraction. In *Knowledge and Information Systems* 9 (1), 1-18.

Fernández-Parra, María 2008: Selecting Criteria to Define Formulaic Language. In Jenset, Gard et al. (eds.), *Linguistics in the Making*. Oslo: Novus Press, 77-97.

Giammarresi, Salvatore 2008: Second Generation Translation Memory Systems and Formulaic Sequences. In *Lodz Studies in Language* 17, 419-430.

Gotti, Maurizio/Šarčević, Susan 2006: *Insights into Specialized Translation*. Bern: Peter Lang.

Goźdź-Roszkowski, Stanisław 2006: Frequent Phraseology in Contractual Instruments: A Corpus-Based Study. In Gotti, Maurizio/Giannoni, Davide (eds.), *New Trends in Specialized Discourse Analysis*. Bern: Peter Lang.

Ha, Le An et al. 2008: Mutual Terminology Extraction Using a Statistical Framework. In *Procesamiento del Lenguaje Natural* 41, 107-112.

ten Hacken, Pius/Fernández-Parra, María 2008: Terminology and Formulaic Language in Computer-Assisted Translation. In *SKASE Journal of Translation and Interpretation* 3, 1-16.

ten Hacken, Pius: 2001: Has There Been a Revolution in Machine Translation? In *Machine Translation* 16, 1-19.

Hatim, Basil/Munday, Jeremy 2004: *Translation: An Advance Resource Book*. London: Routledge.

Hutchins, John 2005: Current Commercial Machine Translation Systems and Computer-Based Translation Tools: System Types and their Uses. In *International Journal of Translation* 17 (1-2), 5-38.

Hutchins, John/Somers, Harold 1992: *Introduction to Machine Translation*. London: Academic.

Iverson, Steve 2003: Working with Translation Memory. In *Multilingual Computing and Technology* 59, vol. 14 (7).

Jackendoff, Ray 1997: *The Architecture of the Language Faculty*. Cambridge, MA: Newbury House.

Kingscott, Geoffrey 2002: The Impact of Technology and the Implications for Teaching. In Dollerup, Cay/Appel, Vibeke. (eds.), *Teaching Translation and Interpreting 3: New Horizons*. Amsterdam: John Benjamin.

Koeva, Svetla 2007: Multi-Word Term Extraction for Bulgarian. In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing: Information Extraction and Enabling Technologies*, Prague, Czech Republic.

Loffler-Laurian, Anne-Marie 1996: *La Traduction Automatique*. Villeneuve d'Ascq, France: Presses Universitaires du Septentrion.

Makoushina, Julia 2008: A Comparison of Eight Quality Assurance Tools. In *Multilingual*, June 2008.

Manning, Christopher/Schütze, Hinrich 1999: *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Nattinger, James/DeCarrico, Jeannette 1992: *Lexical Phrases and Language Teaching*. Oxford: Oxford University Press.

Pym, Anthony 2004: *The Moving Text: Localization, Translation and Distribution*. Amsterdam: John Benjamin.

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

Simard, Michel et al. 1992: Using Cognates to Align Sentences in Bilingual Corpora. In *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation*, Canada.

Somers, Harold 2003: *Computers and Translation: A Translator's Guide*. Amsterdam: John Benjamin.

Thurmair, Gregor 2003: Making Term Extraction Tools Usable. In *Proceedings of EAMT-CLAW `03*, Dublin.

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

Zielinski, Daniel/Ramírez-Safar, Yamile 2005: *Research Meets Practice: T-Survey 2005: An Online Survey on Terminology Extraction and Terminology Management* [online]. http://fr46.uni-saarland.de/download/publs/sdv/t-survey_aslib2005_zielinski.htm#_ftn1 (accessed 23 September 2008).