Computer-aided translation

Lynne Bowker & Des Fisher
University of Ottawa

Computer-aided translation (CAT) is the use of computer software to assist a human translator in the translation process. The term applies to translation that remains primarily the responsibility of a person, but involves software that can facilitate certain aspects of it. This contrasts with machine translation (MT), which refers to translation that is carried out principally by computer but may involve some human intervention, such as pre- or post-editing. Indeed, it is helpful to conceive of CAT as part of a continuum of translation possibilities, where various degrees of machine or human assistance are possible.

After recognizing that fully automatic MT was a formidable challenge, researchers gradually turned their attention to CAT in the 1960s, starting with the creation of term banks, which used computers to store large volumes of structured information (see Terminology and translation*). Advances in computing and computational linguistics in the late 1970s and early 1980s spurred the development of modern CAT tools, which rely on computers not only for storing information, but also for actively searching and retrieving it. Visionaries such as Martin Kay (1980), among others, conceived of tools that became the backbone of CAT. However, it was not until the mid-1990s that these tools became widely commercially available. Since then, the rapid evolution of technology has made CAT accessible, affordable, popular and even necessary – to help translators in this globalized information age tackle the enormous volumes of text to translate in ever shorter turnaround times (Esselink 2000; Lagoudaki 2006; see also Globalization and translation*; Localization and translation*).

1. CAT tools

Since many translators are avid technology users, a wide range of tools could fall under the heading of CAT. However, this term is typically reserved for software designed specifically with the translation task proper in mind, rather than tools intended for general applications (e.g., word processors, spelling checkers, e-mail).

The most popular and widely marketed CAT tool in use today is the Translation Environment Tool (TEnT) – sometimes referred to as a translator’s workstation or workbench – which is in fact an integrated suite of tools. Individual components differ from product to product; however, the main module around which a TEnT is generally constructed is a translation memory (TM), which, in turn, most often functions in close association with a terminology management system (TMS).
1.1 Translation memory tools

A TM is a tool that allows users to store previously translated texts and then easily consult them for potential reuse. To permit this, the source and target texts are stored in a TM database as bitexts. An aligned bitext is created by first dividing the texts into segments – which are usually sentences – and then linking each segment from the source text to its corresponding segment in the translation.

When a translator has a new text to translate, the TM system first divides this new text into segments and then compares each with the contents of the TM database. Using pattern-matching, the TM system tries to identify whether any portion of the new text has been previously translated as part of a text stored in the TM database. When the TM system finds matches for a given segment, it presents the translator with them (see Table 1). The translator is never forced to accept the displayed matches; these are offered only for consideration and can be accepted, modified or rejected as the translator sees fit.

Table 1. Types of matches commonly displayed in TMs

<table>
<thead>
<tr>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exact match</td>
<td>A segment from the new text is identical in every way to one in the TM database.</td>
</tr>
<tr>
<td>Full match</td>
<td>A segment from the new text is identical to one in the TM database save for proper nouns, dates, figures, etc.</td>
</tr>
<tr>
<td>Fuzzy match</td>
<td>A segment from the new text has some degree of similarity to a segment stored in the TM database. Fuzzy matches can range from 1% to 99%, and the threshold can be set by the user. Typically, the higher the match percentage, the more useful the match; many systems have default thresholds between 60% and 70%.</td>
</tr>
<tr>
<td>Sub-segment match</td>
<td>A contiguous chunk of text within a segment of the new text is identical to a chunk stored in the TM database.</td>
</tr>
<tr>
<td>Term match</td>
<td>A term found in the new text corresponds to a termbase entry in the TM system's integrated TMS.</td>
</tr>
<tr>
<td>No match</td>
<td>No part of a segment from the new text matches the contents of the TM database or termbase. The translator must start from scratch; however, the new translation can itself go into the TM for future reuse.</td>
</tr>
</tbody>
</table>

1.2 Terminology tools

While TM systems are at the heart of TEnTs, they are typically integrated with terminology tools, which can greatly enhance their functionality. A terminology management system (TMS) is a tool used to store terminological information in and retrieve it from a termbase. Translators can customize term records with various fields (e.g., term, equivalent, definition, context, source), and they can fill these in and consult them at will in a standalone fashion. Retrieval of terms is possible through various types of searches (e.g., exact, fuzzy, wildcard, context).
However, termbases can also be integrated with TM systems and work in a more automated way. For instance, using a feature known as active terminology recognition, a TMS can scan a new text, compare its contents against a specified termbase, and automatically display records for any matches that are found. If desired, users can further automate this step by having the system automatically replace any terms in the text with their target-language equivalents from the termbase. Consistency and appropriateness are maintained by using termbases specific to certain clients or text types.

1.3 Other TEnT components

TEnTs include more than just TMs and TMSs. Table 2 summarizes the functions of other common TEnT components. For more detailed descriptions, see Bowker (2002), Somers (2003) or Quah (2006), as well as Corpora*, Localization and translation*, Machine translation today* and Terminology and translation*.

Table 2. Some common TEnT components

<table>
<thead>
<tr>
<th>Component</th>
<th>Brief description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concordancer</td>
<td>Searches a (bi)text for all occurrences of a user-specified character string and displays these in context.</td>
</tr>
<tr>
<td>Document analysis module</td>
<td>Compares a new text to translate with the contents of a specified TM database or termbase to determine the number/type of matches, allowing users to make decisions about which TM databases to consult, pricing and deadlines.</td>
</tr>
<tr>
<td>Machine translation system</td>
<td>Generates a machine translation of a segment that has no match in the TM database.</td>
</tr>
<tr>
<td>Project management module</td>
<td>Helps users to track client information, manage deadlines, and maintain project files for each translation job.</td>
</tr>
<tr>
<td>Quality control module</td>
<td>May include spelling, grammar, completeness, or terminology-controlled language-compliance checkers.</td>
</tr>
<tr>
<td>Term extractor</td>
<td>Analyzes (bi)texts and identifies candidate terms.</td>
</tr>
</tbody>
</table>

2. Impact on translation

With much touted benefits of increased productivity and improved quality being hard to resist, the incorporation of CAT into the translation profession has been significant: CAT tools have been adopted by many translation agencies, governments, international organizations, transnational corporations and freelance translators (e.g., Joscelyne 2000; Jaekel 2000; Lagoudaki 2006). However, organizations and individuals must take into account their unique needs and, in light of these, must evaluate the costs and benefits of CAT before adopting it.

Given the way TMs operate, any gain in efficiency depends on the TM’s ability to return matches. Texts that are internally repetitive or that are similar to others already
translated (e.g., revisions, updates and texts from specialized fields) will tend to generate useful matches. Texts that are less “predictable”, such as literary works or marketing material, will not. Once matches are found, simply being able to automatically copy and paste desired items from the TM database or termbase directly into the target text can save translators typing time while reducing the potential for typographic errors. However, significant productivity gains are usually realized over the medium to long term, rather than over the short term, because the introduction of CAT tools entails a learning curve during which productivity could decline (Lagoudaki 2006).

With regard to quality, CAT still depends on human translators; should a client impose its own TM for its work, the translator has no control over its contents. Furthermore, the segment-by-segment processing approach underlying most TM tools means that the notion of “text” is sometimes lost (Bowker 2006). Translators may be tempted to stay close to the structure of the source text, and individual segments may bear the differing styles of their authors, leading to poor readability of the resulting target text.

CAT also affects the professional status of translators, their remuneration and their intellectual property rights. For instance, some clients ascribe less value to the work of translators who use CAT tools. If CAT is faster and easier than human translation, clients may ask to pay less for it. When a TM offers exact matches or fuzzy matches, should the translator offer a discount? Clients may be even more demanding if they use their own TM to pre-translate a text before sending it to a translator. Yet, even exact matches do not equate to zero time spent; a translator must evaluate the suggested sentences and make adjustments depending on the communicative context. Traditional fee structures (billing by the word or page) may no longer be appropriate, and among freelancers in particular, these are sometimes being replaced by hourly charges. In addition, the movement away from desktop TM systems towards Web- and server-based access to TM databases is giving more control to clients, who insist that translators use their TMs only, preventing translators from building up their own linguistic assets (Garcia 2008: 62).

Ethical, financial and legal questions surround the ownership and sharing of CAT data (Topping 2000; Gow 2007). The contents of a TM are source texts and translations, the ownership of which presumably remains with the client. However, when collected as a database, control and ownership are thrown in doubt. Translators may wish to exchange or sell a TM or termbase. However, the client or original owner may demand confidentiality of the information (e.g., for copyright protection). Currently, translators often deal with this problem through contracts and nondisclosure agreements, but it remains a murky issue.

3. Emerging possibilities

CAT is here to stay, and translators will continue to adapt to its presence and its increasing importance. However, CAT will undoubtedly proceed to develop at a rapid pace.
Advances in TMs may include the introduction of linguistic analysis and the ability to recall the surrounding context of matching segments, as well as that of their corresponding translations – already a feature of certain TMs. Moreover, in the case of fuzzy matches, current TM systems readily identify differences between the sentence to be translated and the source text segment retrieved from the TM database; however, future enhancements may also indicate which elements of the corresponding target segment should be preserved or modified. Sharing across different products will become easier as standards such as Translation Memory eXchange (TMX), TermBase eXchange (TBX) and XML Localization Interchange File Format (XLIFF) become more widely adopted (Savourel 2007). In addition, the availability of open-source products will make access to CAT tools easier for more translators, while Web- and server-based access to TM databases is also increasing (Garcia 2008).

The Internet provides several atypical possibilities for CAT. For instance, crowdsourcing translations involve leveraging the knowledge and free labour of the crowd – the general public. Collaborative translation – often undertaken as part of a volunteer effort – is similar in that multiple translators participate, but it can be limited to selected or professional ones. Tools such as wikis and weblogs can serve these purposes (Bey et al. 2008; see also Networking and volunteer translators*; Web and translation*). There will be no shortage of innovation while millions of technologically savvy people, speakers of dozens of languages, contribute daily to the information sphere that is our world. However, none of this will obviate the need for professional translators, equipped with specialized skills and knowledge – and even CAT tools – to continue their indispensable work into the next decades.

References


Further reading


