

# TransType2 - A New Paradigm for Translation Automation

Antonio S. Valderrábanos\*, José Esteban\*\*, Luis Iraola\*\*

Bitext SL\*, SchlumbergerSema SAE\*\*

Madrid, Spain

[asv@bitext.com](mailto:asv@bitext.com); [jfernando.esteban@madrid.sema.slb.com](mailto:jfernando.esteban@madrid.sema.slb.com)

## Abstract

The aim of TransType2 (TT2<sup>1</sup>) is to develop a new kind of Computer-Assisted Translation (CAT) system that will help solve a very pressing social problem: how to meet the growing demand for high-quality translation. To date, translation technology has not been able to keep pace with the demand for high-quality translation. The innovative solution proposed by TT2 is to embed a data driven Machine Translation (MT) engine within an interactive translation environment. In this way, the system combines the best of two paradigms: the CAT paradigm, in which the human translator ensures high-quality output; and the MT paradigm, in which the machine ensures significant productivity gains.

## 1. Objectives

TransType2 (TT2) aims at facilitating the task of producing high-quality translations, and making the translation task more cost-effective for human translators. TT2 accomplishes this goal of speeding up and facilitating the work of translators by automatically suggesting translation completions. These suggestions are provided as follows: initially, TT2 suggests a possible translation for a given sentence. If the translator doesn't approve this translation proposal, he/she will start typing a new translation, and with each new character the translator enters, the system will provide new suggestions that are compatible with the translator's new input. If the system provides the right suggestion, the translator has only to accept it, thereby saving time in producing his or her translation. Otherwise, the translator ignores the system's suggestions and continues to type his or her intended translation. The interaction is thus a natural one for the translator, since its focus is the target language text he is trained to produce.

The proposed translation process will retain two major facts as general requirements:

- the human translator is the key component for producing high quality output
- machines provide the means to speed up the translation task and make it more efficient

Finding a way to reconcile the two requirements is the key challenge of the TT2 project.

The solution proposed by TT2 is to properly combine the positive aspects of the MT and CAT paradigms within one single environment. In order to produce high-quality translations, TT2 will include a data-driven MT engine to assist the translator with suggestions.

Using this kind of translation technology will allow TT2 to:

- quickly and cheaply develop different MT engines: from English into French, Spanish and German, and vice versa. Developing six MT systems in a reasonable period of time and with feasible resources is a goal that can only be tackled by using machine learning techniques.
  - comply with the severe performance requirements imposed by an interactive MT system. The MT engine of TT2 must provide translation suggestions within the time constraints required by a real-time environment. Any delay in this regard would affect the productivity benefits of TT2 and, more important, its level of user acceptance.
- At a more general scientific level, the main objectives of TT2 are two-fold:
- to provide a framework in which leading-edge research can be conducted in the area of data

---

<sup>1</sup> TransType2 (TT2) is an RTD project funded by the European Commission under the Information Society Technologies Programme (IST-2001-32091). For further details, see <http://tt2.sema.es>.

driven methods in natural language processing and, more specifically, translation

- to provide a practical application for that research which will help solve a pressing social problem: how to meet the ever-growing demand for high-quality translation

## 2. TT2 and translation automation

There are basically two kinds of specialized technology that are available to translation services today: fully automatic MT systems, and on the other hand, translator support tools (which are often called CAT tools, e.g. translation memory systems, usually combined with other programs to manage term glossaries or align translations).

In the next sections we will outline the positive aspects and limitations of these two paradigms, MT and CAT. Then, we will describe how TT2 combines the strengths of both paradigms. Finally, we will present TT2 as new kind of tool in the paradigm of IMT (Interactive MT).

### 2.1. TT2 and the CAT paradigm

From an industrial point of view, CAT tools are probably the most impressive success story of the last decade in the market of translation automation. For this reason, CAT tools look like a suitable environment to foster the co-operation between translators and machines for the task of producing high-quality translation. Different products have emerged in the area of CAT tools; some have been in the market for years and have solid plans for expansion.

CAT tools packages normally bundle Translation Memory (TM) technology with other utilities like terminology management or alignment tools. TM technology exploits the fact that machines have bigger and more precise memory than humans. TT2 has been designed so it can be easily integrated with these tools and, thus, be readily accepted by translators.

Users outline two major benefits of CAT tools:

- consistency: in certain translation domains (like software or hardware manuals) being consistent is a synonym of quality, as the translation process is meant to produce a clear and easy to understand text, rather than to be creative
- speed: in environments where the time to market impels the translation cycle to be as

short as possible, shortening this cycle is regarded as a major benefit

Probably, the most important reason why TM technology has been so widely adopted is the fact that the translator is always in control of the translation process, approving, correcting or disregarding the translations proposed by the computer, as it is the case in TT2. CAT tools have succeeded in creating a positive attitude towards translation automation within the translation industry. In view of these facts, CAT tools seem to offer the perfect environment to foster the co-operation of humans and machines for the task of producing high-quality translation.

### 2.2. Limitations of the CAT paradigm

For CAT tools based on TM to be cost-effective, the source text needs to have a minimum level of repetition. Whenever the source text lacks this minimum level, CAT tools provide little, if any, productivity gains to translators. This level of repetition is highly dependent on the domain and type of text. For example, it's normally high for hardware documentation but low for news.

In this context, there is a strong need for tools and techniques that can significantly increase translator productivity without such a heavy dependency on repetitiveness. This is one of the most important facts addressed by TT2: the purpose of the data driven MT engines incorporated in TT2 is to help solve this problem.

### 2.3. TT2 and the MT paradigm

If we look at the other translation paradigm, Machine Translation (MT), the general picture is quite different: expectations have always been high and results have rarely been up to these expectations. MT, understood as technology that allows machines to translate with high (or human-like) quality levels, and at a significantly faster speed, has always been the dream of researchers and cost managers. Bar-Hillel (Bar-Hillel, 1951) first formulated this challenge as FAHQT, or fully automatic high-quality translation of unrestricted texts.

Over the last decade, the MT paradigm has seen its most significant progress come from the area of machine learning and data-driven techniques, particularly the approach known as Statistical MT. Compared to classic, rule-based MT, the most attractive features of this approach are: low

development cost, quick deployment, and easy portability across languages. The MT engines of TT2 are based on this type of data-driven technology, integrating both statistical and finite-state techniques.

#### **2.4. Limitations of the MT paradigm**

In the MT framework the machine aims at replacing, rather than helping, the translator. This is probably the main reason for its low acceptance in the translation industry. As such, MT may have been the dream of researchers, but never the dream (rather the nightmare) of translators. Nowadays, the main application of MT is not for the production of high-quality translations, but rather for the production of "informative translations", i.e. translations that are meant to be read to provide an idea, or a gist, of the contents of the source text. These translations are not meant to be published or saved for future reference, as human translations are, but to be discarded after use.

Given that the output of most MT systems does not usually achieve the necessary quality levels, some sort of quality assurance phase is needed after the translation is automatically produced. This phase is commonly known as post-editing and is generally considered the weak point of the fully automatic MT approach. Post-editing is a time-consuming task, to the point that some translators claim it takes more time than actually producing the translation from scratch. So in terms of productivity, post-editing is one of the major drawbacks for the acceptance of MT technology.

#### **2.5. TT2 - The integration of two paradigms**

Between fully automatic MT, on the one hand, and CAT tools like TM on the other, there is currently a major vacuum which offers a window of opportunity for TT2. In this context, the aim of TT2 is to merge the best aspects of both paradigms and integrate them in a single application. As in the CAT paradigm, translators working with TT2 will be in control of the production of the target text. As in the MT paradigm, data-driven techniques will be used to help the translator produce faster and more consistent translations. The CAT paradigm will provide the environment to interact with translation professionals, while statistical MT will provide a more intelligent way of exploiting (and learning from) previous translations.

TT2 can also be seen as an innovative way of exploiting MT technology while avoiding the post-editing phase. TT2 provides the translator with the right environment to accept or reject the proposals of the MT system at their point of entry, rather than after the fact. Indeed, the translator can even guide the suggestions of the MT system via the partial text he or she types. In this way, the translator retains full control of the target text. As a result, TT2 should allow us to obtain the principal benefits of MT technology while avoiding the inconveniences of post-editing.

#### **3. TT2 - A new kind of IMT**

TT2 can also be viewed as an Interactive Machine Translation (IMT) system, since the final translation is produced by combining contributions both from the machine and the translator. However, our project's approach to interaction is highly original. In past IMT systems, the focus of the interaction between translator and machine has always been the source text: typically the translator is asked to resolve syntactic or lexical ambiguities that the machine cannot solve on its own. In TT2, on the other hand, the object of the interaction is not the source text but the target language translation. Both the human translator and the system contribute to the drafting of this text, with the system's contributions taking the form of suggested completions of portions of target text which the translator has in mind for the current sentence.

#### **4. TT2 and multimodality**

Multimodality is a relevant dimension of TT2. TT2 has two input modalities: text (keyboard- and mouse-based) and speech. Speech will be integrated to allow the translator interact with TT2; for example, it will allow the translator to manage suggestions provided by the MT engine. This way of exploiting speech is different with respect to other projects where MT and speech have been used together: in TT2 the translator uses speech to choose or to change the translations that the MT engine suggests; and speech is used in combination with regular text input. As a result, both input modalities can be used in parallel, to the convenience of the translator, without excluding each other.

Another relevant feature of TT2 is related to the input considered by the speech recognition system.

While in most speech translation systems the only input to the translation engine is the utterance in the source language, in TT2 the input to the system is a combination of text for the source language (typically, the sentence that the translator needs to translate) and speech for the target language (the translation uttered by the translator for the source text).

## **5. Current Status**

The duration of the project is 36 months. At the time of submission, July 2003, the project has just covered its first year. At this point, a prototype is already available and this will be the focus of the presentation.

## **6. Bibliographical References**

Bar-Hillel, Y. 1951. The State of Machine Translation in 1951, *American Documentation*, vol.2, 1951, pp.229-237.