

A New Approach To Computer-Assisted Human Translation

Balázs Kis, István Lengyel

MorphoLogic
{kis,lengyel}@morphologic.hu

The article presents methods to fight off major limitations of CAT tools currently available on the market. These limitations include language-independence and CAT tools delivered as framework products without databases.

Three types of tools or operations are concerned; only the first two are considered language technology tools in the narrow sense:

(1) Translation memory: recycles previous translations of the translator or the translation community;

(2) Terminology management system: efficiently looks up and handles terminology specific to the translation domain;

(3) Workflow management tool: chops up long texts, manages translation „threads” (texts or parts of long texts where one translator and optionally one proofreader is involved), monitors and archives translations, measures efficiency, etc.

A translation memory is meant to find equivalent or similar matches to the segments of the current source text in the database of previous translations. In most cases, only similarities are found; it is crucial to have high quality hits constituting real linguistic similarity. The authors point out the need to implement the parsing of morphology and a limited syntax of the source language in CAT tools. Moreover, it would be a breakthrough in translation memories if they did not only show the exact hit in the database, the „canned translation”, but also attempted to adapt the translation to the context according to the nature of similarity between the source and the stored segments. Translation memories currently available are all delivered with an empty database. However, the subject of source texts, respective vocations and sciences – and their use of language – exist independently: it would be very useful if the CAT tool had access to a general corpus of the given context, preventing the translation memory from appearing to be empty at the time of installation.

Current terminology management systems treat terminology as a rule, considering technical terms as unambiguous. This approach has many limitations. The authors' experience in translation and terminology definition shows that the major feature of terms is not disambiguity but the topic-specific formation of the context. Also, words and expressions present in everyday language are often attributed a different meaning (in another context, perhaps with a different syntax); these also have to be treated as terms. Therefore, terminology is not necessary nominal, nor unambiguous.

Also, it is very important to provide a means to spot terms: translators and translation team managers cannot be expected to know all topics they have to deal with in great detail. Recognising terminological situations – situations where a given word or collocation should be considered technical term – is not trivial and cannot be performed independently of general corpora and glossaries of the domain. The authors present the development of a terminology spotting network service.