A Tentative Proposal for Machine Assisted Human Translation (MAHT) – Tool-Specific General Text Typology

Marcin Feder
Department of Translation Studies - School of English – Faculty of Modern Languages – Adam Mickiewicz University, Poznań, Poland

The aim of the present article is to contend the widespread but largely unfounded claim that MAHT tools (popularly known as translator's workbenches) are best suited for translating so-called technical texts. The article calls for establishing a separate MAHT tool-specific text typology by presenting what are – in the author's opinion – the most important features characteristic of an MAHT-suitable text. It transpires that it is difficult to simply relate this particular type of text to any of the existing classifications of translation-related or general text typologies (such as those advocated by, for example, Hatim & Mason 1990; Reiß 1983; Snell-Hornby 1988 or Kussmaul 1997) since the existing methodologies do not take MAHT-specific attributes into account. Therefore, the present article calls for an empirical, corpus-based study that would help establish the relation between the proposed underlying features and actual text types as described elsewhere.

1. Introduction

Before embarking on a discussion concerning the proposed outline of an MAHT-text typology it seems appropriate to define what is meant by Machine Assisted (or Aided) Human Translation. Subject specialists and translation practitioners interested in mechanical aids in translation or translation automation are most frequently confronted with the term Computer Assisted (or Aided) Translation (and less frequently with its variation – Machine Assisted (or Aided) Translation). However, over the years there has been a lot of confusion as to the actual meaning of this designation so that, given the merger of scientific and popular interest in Machine Translation (MT) and CAT, the latter has become somewhat of a catch-all term and its synonymy with MAHT has become questionable. If we treat CAT as a generic term, which it has – in effect – become, then MT and MAHT are its two major instances. Today, MT is generally understood to be the process in which the machine (or rather a computer program) is the pivotal part – i.e. is responsible for preparing the actual translation of a given text whereby the human is entrusted with the task of programming the machine, updating the program, pre- and/or post-editing the text and sometimes interactively conversing with the machine to solve certain translation problems as they appear in the text undergoing translation. At the other end of the CAT spectrum we have the MAHT tools – popularly known as translation workbenches – where the human is the pivotal part of the process. The human translator prepares the target language version of the text in question, and the machine (i.e.

a computer program) assists him in this task by, basically, offering terminological hints (i.e. terms are retrieved automatically, semi-automatically or manually from an appropriate database) and by building and employing a translation database (a Translation Memory - TM) to suggest identical (exact matches) and similar segments (fuzzy matches) that have been used in translations and stored in the database previously. It is this second instance of CAT that is the subject of the present article.

2. MAHT-text typology

The type of text to be translated with the use of a Translation Memory-based tool significantly influences the degree of usefulness of such a tool in the translation process[1]. In other words, if a user or a potential user wants to determine the suitability of the tool he or she uses or is going to use to translate a particular text type, he or she must consider what exactly it is the machine does while assisting him or her in the translation task. Let me stress that the translation outcome is still largely a human product, but to be able to use this label properly, we first have to find out how a MAHT tool actually helps a human translator in his or her otherwise complex task.

The general principle is very simple – the tool looks for strings of text that are similar or the same within a given text or across a number of texts (if a database of previous translations is available and the degree of sameness or similarity is sufficient). There are a number of retrieval techniques that can be employed. These include *linguistic techniques* of syntactic and morphological parsing and analysis, the so-called *traditional techniques* consisting in a mixture of "heuristics applied on syntactic features, morphological reduction, the use of classic (relational) database systems, etc." (Heyn 1995:74) and finally *mathematical* or *statistics-based techniques* used commonly in many information retrieval (IR) applications such as similarity measure, stoplists, successor variety, table lookup, affix removal, n-gram techniques, edit distance and inverted files (cf. Trujillo 1999:61ff.). In the present paper I adopt this general principle of sameness or similarity underlying the operations of a given MAHT tool as the point of departure for developing an outline of an MAHT-specific text typology. I would claim that in order to be compatible with the general principle and therefore suitable for co-operation with a given tool, a text has to display the following, basic characteristics:

1. REPEATABILITY – in the present proposal this is considered the key and overriding feature that a text has to demonstrate in order to be MAHT tool-usable. Repeatability refers to the degree of repetition of textual material within a given text or across a body of texts. If the tool's basic principle of operation is to look for the same or similar sentences or other text elements (e.g. table cell contents, items on a list, etc.) identified as translation units (TU) for the purposes of a given tool, it has to be able to find such identical and similar

parts. There is rarely any point in MAHT-translating a text that has no repetition or similarity within it or across a number of texts at all; the tool would then become only a very expensive word processor. It has already been indicated that repeatability may be measured within one text or across a corpus of texts (of course, not all texts, but original documents and their subsequent versions or updates, documents related to the same subject domain or documents translated for the same client). Bearing that in mind, it is possible to speak of internal and external repeatability (Schüller 1995:13). Such repeatability may be determined at different levels and relate to individual terms and words (terminology management; see also 2.6. below), whole phrases, sentences (segment matching) and paragraphs. However, the repeatability referred to here is not the kind that occurs mainly or exclusively at a sub-sentence or sub-TU level as this would reduce the MAHT tool's success rate and usability (cf. Del Pino 1998:133).

A very important question that arises here is the actual repeatability threshold ratio that would make it useful to employ a MAHT tool to translate a given text. Such quotas are sometimes set at 50% (cf. Spies 1995:3). Corpus studies show that for some so-called technical texts (which, at this point, seems to be a very imprecise appellation) internal recurrence reaches the level of around 40% and external recurrence is sometimes as high as 55% (Merkel 1992). As far as translation practice is concerned, Uniscape, a worldwide provider of Internet-based translation services, has observed typical re-usability rates of previously translated material (equalling external repeatability) to be around 40 to 80% in the case of web site, documentation and software textual content revisions (Khosla & Schwartz 1999:7 and Schwartz & Khosla 1999:9). The Rank Xerox translation department reports that external repeatability ratios across documentation for similar products have been observed to be as high as 70% when exact matches alone are considered (Gaussier et al. 1999:11). Nevertheless, the question is a particularly difficult one. Sometimes a very low degree of internal repeatability would be acceptable if there were a sufficient level of external repeatability (e.g. if there were a large corpus of documents related to the same topic or client; cf. EAGLES 1995:141-2) or, generally, if there were a prospect of translating similar material in the future. On the other hand, some empirical studies (e.g. Brungs 1996) show that the rates of external repeatability across two texts (an original and its revision) do not necessarily achieve 50%[2] (especially in the case of automatic repeatability analysis).

Note that both internal and external repeatability ratios are also greatly influenced by text structure in terms of its segmental composition – i.e. if the segments (or TUs) that a text is made up of are full sentences, the repeatability ratio is likely to be lower

than in the case of documents containing more individual phrasal expressions or, in an extreme case, just lists of items (e.g. listings, rankings, tabular representations, enumerations, etc.; cf. Kenny 1999:77).

The remaining MAHT-text features, surveyed below, are – so I would claim – secondary to repeatability.

2. DOCUMENT LENGTH – An MAHT tool learns as it translates. Its learning-capacity is very limited, of course, since the only thing the system does, is remember what it (or rather the human translator in charge of the tool) has translated before. Thus, quite naturally, the longer the document, the greater the likelihood for improved performance of the tool as the chance of repetition or appearance of similar segments is greater. This, naturally, provided the document is sufficiently repeatable in the first place, which demonstrates the supremacy of criterion 2.1. above. A long document, by practitioners' standards, would be anywhere from 30 to 50 pages. However, better results are achieved with even longer texts of, for example, 100 to 300 pages. This is not to say that documents under 30 pages are not suitable for MAHT purposes.

3. STYLE – in order to guarantee greater repeatability of a text or a number of texts it is necessary to ensure that the style employed by the original is as straightforward as possible, that no or only limited stylistic variance is employed, that the use of metaphor or idiomatic language is limited or eliminated altogether and that no advanced forms of analogy (e.g. anaphoric, cataphoric or exophoric references) are employed. The style of an MAHT-text has to be simple and consistent (also in the sense of terminology, grammatical structure and layout – see 2.5. below) throughout a document or a series of documents. In other words, the language of MAHT-texts has to be *controlled* – i.e. specially trimmed (usually simplified and devoid of ingenuity as displayed by uncontrolled languages) to suit the purposes of specialised communications (cf., for example, Lee's description of the so-called BULL Controlled English; Lee 1993:36 or Newton 1992b).

4. SENTENCE STRUCTURE – not only style, but also sentence structure must be simple and consistent. An MAHT tool generally favours sentences that are short[3] and whose structure (e.g. word order) does not change throughout a text or across a body of textual material (cf., for example, "Press ENTER to exit" vs. "To exit press ENTER"). This will greatly facilitate TU match retrieval (cf. also Trujillo 1999:61ff.). Naturally, the longer the sentence, the greater the likelihood of alterations even if the "content" (i.e. the sense) is similar or identical to that of a previously translated segment (cf. Benis 1998:5). In such cases the tool will "find" it increasingly difficult to cope with structural variation.

5. SOURCE LANGUAGE (SL) TEXT QUALITY – The grammatical quality of a text in a strict sense should be ensured under 2.3., 2.4. and 2.6.; what is meant here is text quality related to such features as correct spelling, punctuation and even formatting. – i.e. typography understood in the broadest sense. An MAHT tool looks for similarities or identicalness, thus any change of spelling (whether incorrect or variant) or inconsistency in punctuation, application of formatting or layout conventions[4] may result in decreasing the tool's chances of finding a proper match and thus introduce the need for some degree of pre-editing to enhance tool performance.

6. PHRASEOLOGICAL CONSISTENCY – Apart from terminological consistency, i.e. terminological and phraseological repeatability and uniform use of the same appellations to denote the same concepts[5], the vocabulary used throughout a text or a collection of texts has to be sufficiently specialised (to be widely supported by adequate terminological resources via a terminology management module) and rich (i.e. a text should preferably contain a large amount of terminology and other specialised phrases as such, provided the above criteria are met). Moreover, repeatability, in the case of the present criterion, will not be achieved if synonyms are employed throughout the text and the tool will stumble on homonymy – i.e. certain terms may be erroneously employed where a different translation is needed.

The above listing clearly indicates that it is difficult to relate the type of text required for MAHT to any of the existing and popularly employed classifications of translations and general text types. Indeed, the text type under discussion is neither dependent on field of discourse (i.e. subject domain) nor on function (e.g. literary, poetic, didactic), nor can it be classified in terms of communicative intentions or rhetorical purpose. In addition, it does not fit within Hatim & Mason's hybrid, multifunctional text classification (1990: 138-64) and it is just as impossible to situate it in Snell-Hornby's spectrum of text types and criteria relevant for translation (1988:32).

Theoretically one might argue that the classification presented above bears some, be it a very distant, semblance to Hatim & Mason's *argumentative text* (1990:153-4) or Reiß's *operative text* (1983), still, these similarities are only partial if not superficial[6]. From the above proposal it clearly transpires that in the case of an MAHT-suitable text it is the **surface features** (i.e. text form rather than its content and function) that are to be taken into consideration[7] in the first place. Hence, it is argued here that an MAHT-suitable text is a text that meets all or most of the criteria given above and must be defined primarily as a **surface text** (i.e. a text distinguished primarily on the basis of its form) without much regard to its function, content and communicative purpose[8] since we have shown that, as understood within the confines of the present argument, it is **largely devoid of typical functional features**.

Some researchers and practitioners (e.g. Brungs 1996; Benis 1998:4; Ray & Ray 1999:280) claim that "technical" or "specialised" texts are best suited for MAHT purposes. However, even though technical specifications, lists of spare parts, manuals, handbooks, instructions[9], product descriptions, etc., match the definition of an MAHT-suitable text as presented above, this does not mean that only "technical" texts are MAHT-suitable to the exclusion of any others, and that an MAHT-text must be based only on criteria related to the field of discourse[10]. Instead, one should bear in mind that the most important factor in determining the MAHT tool suitability of a given text from the point of view of the tool's subsequent linguistic performance (i.e. retrieval success ratio) appears to be the surface structure of such a text.

3. Implications and conclusions

From the above discussion it transpires that the existing function or content-based translation-oriented text typologies are insufficient to fully account for the novel text type that has become the focus of MAHT studies, i.e. an MAHT-suitable text, as I propose to call it. This results chiefly from the fact that such a text should not be considered from a purely linguistic point of view within the framework of the existing typology methodologies but rather, first and foremost, from the tool's extra-linguistic point of view. In such an approach it is the text's surface characteristics that come to the fore and ultimately determine its nature and suitability for the intended purpose. An analysis of MAHT-suitability therefore leads to the creation of a new text typology. This is, however, not to say that traditional approaches to typology should not play any role in determining the characteristics of such texts. Although I have explicitly distanced myself from subject domain typologies (or text type distinctions as suggested by the German name *Textsorte*, cf. Kussmaul 1997:69), i.e. the selection of texts in question with regard to a specific subject area (e.g. computer manuals) or a specific audience (which, usually, also implies concrete field of discourse specifications), it is quite likely that there will be some correlation between these and the actual surface text features that make up the characteristics of an MAHT-suitable text, as suggested in the present article. Since, however, to my knowledge, no major corpus-based study has been carried out so far to actually associate such MAHT-text surface features with the existing text types, mainly because obtaining real-life corpora of MAHT-translated texts might be very difficult due to the often confidential character of the translated material, I would like to stress there is a pressing need for such research. The discrepancy between traditional text typologies and the MAHT-suitable text characteristics outlined above indicates the need to establish which of the traditional text types are best suited for MAHT purposes, given the specific characteristics of MAHT text processing. This would allow the elimination of unsubstantiated rule-of-thumb or common-knowledge claims and also lead to ameliorating the existing typologies. Furthermore, such research could also contribute to the determination of MAHT-related controlled language

principles for a specific source language as opposed to the controlled Englishes developed so far primarily for MT purposes. The aim of such a project would be (employing the results of previous research into text type determination) to use a representative corpus of source texts, translations and SL text revisions or updates[11] to establish those linguistic, and technical, phenomena to which (a) given MAHT tool(s) is/are most sensitive, and to use them as a basis for the development of a catalogue of CL rules.

By way of conclusion, I would like to say that I have decided to write about MAHT-suitability as a text typology in its own right for two reasons. Many claims are made concerning the so-called recyclability of certain methodologies. Such claims are well founded and extensively exploited in the area of natural language processing (NLP), to which MAHT undeniably belongs (cf., for example, Rowley 1992:124). However, as demonstrated above, the postulate of recyclability does not appear to work in the case of function and content-based text typologies applied to MAHT. MT-based research into tool specific text typology would not yield any positive results either. First of all, MT text typology, if applied, will not influence the choice of an MAHT tool significantly and, therefore, cannot be recycled for MAHT. This is due to the fact that MT text typology itself is largely (although not entirely – viz. controlled languages) restricted to field of discourse (content) and text function typology. Such a restriction results principally from the offer of lexicons and text- or domain-specific grammatical rules attached to a given MT system and/or from the user's willingness to customise or produce dictionaries and sets of rules that would suit a particular translation assignment. This appears not to be the case with MAHT applications where a given text suitability depends, first and foremost, on its **surface structure.**

Bibliography

Arnold, Douglas (1990). "Text typology and machine translation: an overview." Mayorcas (1990), 73-79.

Ball, Sylvia (1997). "In the Beginning Was the Glossary: the Development of Integrated Language Support Services at the European Parliament." *Terminologie et Traduction* 1997(2), 74-79.

Benis, Michael (1998). "Review of Atril's Déjà Vu 2. The Happy Hoarder." *Translation Journal* 2(1) On line at: http://accurapid.com/journal/03TM1.htm (consulted 28.05.2002)

Bowker, Lynne et al. (eds) (1998). *Unity in Diversity? Current Trends in Translation Studies.* Manchester: St. Jerome.

Brungs, Bettina (1996). *Translation Memories als Komponenete Integrierter Übersetzungssysteme.* Saarbrücker Studien zu Sprachdatenverarbeitung und Übersetzen. Band 7. Saarbrücken: Universität des Saarlandes.

Del Pino, Santiago (1998). "Using Translation Memory Software (TMS): An Organisational Checklist." *Terminologie et Traduction* 1998(1), 132-139.

EAGLES (1995). *Evaluation of Natural Language Processing Systems, EAGLES document EAG-EWG-PR.2. Version of September 1995.* On line at: http://www.issco.unige.ch/ewg95/ewg95.html (consulted 28.03.2002).

Gaussier, Éric, Hull, David A. & Salah Aït-Mokhtar (1999). *Term Alignment in Use: Machine-Aided Human Translation.* Meylan: Xerox Research Centre Europe. On line at: http://www.xrce.xerox.com/publis/mltt/mlttart.html (consulted 29.05.2002).

Hatim, Basil & Ian Mason (1990). *Discourse and the translator.* London/New York: Longman.

Heyn, Matthias (1995). "Key Technologies. Impacts of Neural Network Technology on Computer Aided Translation." *TAMA'94 Proceedings* (1995), 71-84.

Kenny, Dorothy (1999). "CAT Tools in an Academic Environment: What Are They Good For?" *Target* 11(1), 65-82.

Khosla, Ashok & Howard Schwartz (1999). "Worldwide Enterprise Publishing: A Case Study." http://www.uniscape.com (consulted 25.04.2000).

Kussmaul, Paul (1997). "Text-Type Conventions and Translating: Some Methodological Issues." Trosborg (1997), 67-83.

Lee, Arthur (1993). "Controlled English with and without Machine Translation." *Translating and the Computer* 15 (1993), 35-39.

Mayorcas, Pamela (ed.) (1990). *Translating and the Computer 10. The Translation Environment Ten Years On.* London: Aslib.

Merkel, Magnus (1992). *Recurrent Patterns in Technical Documentation.* Research Report, Department of Computer and Information Science. Linköping: Linköping University.

Newton, John (ed.) (1992a). *Computers in Translation. A Practical Appraisal.* London/New York: Routledge.

Newton, John (1992b). "The Perkins Experience." Newton (1992a), 46-57.

O'Brien, Sharon (1998). "Practical Experience of Computer-Aided Translation Tools in the Software Localization Industry." Bowker et al. (1998), 115-122.

Ray, Deborah S. & Eric J. Ray (1999). "Good, fast, cheap: Translation Memory Systems offer the potential for all three." *Technical Communication* 46(2), 280-5.

Reiß, Katharina (1983). *Texttyp und Übersetsungsmethode. Der operative Text.* Heidelberg: Julius Groos.

Rowley, Jennifer E. (1992). "Evaluation of Software." *Translating and the Computer* 14 (1992), 117-126.

Schüller, Thilo (1995). *Integrierte Übersetzungssysteme.* Saarbrücker Studien zur Sprachdatenverarbeitung und Übersetzen. Band 1. Saarbrücken: Universität des Saarlandes.

Schwartz, Howard & Ashok Khosla (1999). "The Mandate to Translate: An Internet-based Model for Translation Management in the Global Enterprise." http://www.uniscape.com (consulted 25.04.2000).

Snell-Hornby, Mary (1988). *Translation Studies. An Integrated Approach.* Amsterdam/Philadelphia: John Benjamins.

Spies, Christina (1995). *Vergleichende Untersuchung von Integrierten Übersetzungssystemen mit Translation-Memory-Komponente.* Saarbrücker Studien zu Sprachdatenverarbeitung und Übersetzen. Band 3. Saarbrücken: Universität des Saarlandes.

TAMA'94 Proceedings. Third TermNet Symposium. Terminology in Advanced Microcomputer Applications. Recent Advances and User Reports. (1995). Vienna: TermNet.

Translating and the Computer 15. Machine Translation Today. (1993). London: Aslib.

Trosborg, Anna (ed.) (1997). *Text Typology and Translation.* Amsterdam/Philadelphia: John Benjamins.

Trujillo, Arturo (1999). *Translation Engines: Techniques for Machine Translation.* London: Springer.

Turner, R. (ed.) (1992). *Translating and the Computer 14. Quality Standards and the Implementation of Technology in Translation*. London: Aslib.

---

[1] The translation process is understood here as a combination of the creative act of rendering a source language text available in the target language, involving the supreme human skill of translating (human-generated) texts, and the pre-, peri- and post-translation activities such as checking the content for accuracy, formatting, publishing, printing, adjusting the layout, etc.

[2] Brungs examined three original texts (that had been identified as suitable for MAHT purposes, though her text typology was different from the one proposed here, cf. Brungs 1996:17) and their revisions for external repeatability (i.e. exact and fuzzy matches) first manually and then automatically (using Trados Translator's Workbench Analyse tool). The respective values for the manual and automatic analysis of the three texts were: 55.44% vs. 34.78%; 66.15% vs. 30.74% and 14.85% vs. 1.98% (cf. ibid. 44, 55, 65, 82, 92 and 98). Her analysis also shows how important the control of the other factors mentioned here (i.e. points 2.2. to 2.6.) is (ibid. 102-3). Moreover, according to some practical estimates the typical minimal level of acceptable repeatability (for a MAHT tool to be usable in the case of a given project) is about 20% (Schwartz & Khosla 1999:13) though values between 10 and 70% have also been observed by localisation industry practitioners (O'Brien 1998:119). The European Parliament translation services report on their experience of frequently working with texts whose level of repetitiveness reaches almost 100% (Ball 1997:77).

[3] Some industry standards (e.g. controlled language or CL rules) state that sentence length should not exceed 25 words (cf. Lee 1993:36).

[4] The more so that, reportedly, the majority of MAHT tools do employ certain formatting features such as heading, sub-heading, font style, footnote, table of contents and other layout markers and references as segment identification anchors.

[5] This would also facilitate retrieval at the segmental level by contributing to segmental consistency and eliminating variation.

[6] It is true that the MAHT-text type, as proposed here, might be held to comply with Reiß's principle of *Verständlichkeit* or 'comprehensibility' (i.e. short sentences, simple syntax; Reiß 1983:65-6) but Reiß's postulate of *Erinnerungswert* or 'memorability' (Reiß 1983:66) or Hatim & Mason's recurrence/repetition (Hatim & Mason 1990:154) clearly refers to rhetorical repetition at, mainly, sub-sentence level.

[7] Arnold (1990:74-5) suggested a similar typology with regard to discourse structure as one of the factors that could be controlled for the purposes of Machine Translation.

[8] The use of the terms **text type** and **text typology** in the present article is therefore different from Hatim & Mason's definition of text type as "a conceptual framework which enables to classify texts in terms of communicative intentions serving an overall rhetorical purpose" (Hatim & Mason 1990:140). This results from the fact that, in my proposal, the treatment of a text requires a complete shift of focus, i.e. a departure from the function (and content/sense) towards the form (as demonstrated by criteria 2.1. to 2.6.). This shift is naturally imposed by the nature of the tool with which a text is manipulated.

[9] This indicates a possible link between the MAHT-text type and the *instructional text* type (Hatim & Mason 1990:156-8). Since, however, the statement concerning field of discourse classification seems to be based on common knowledge grounds it still remains to be scientifically validated (see 3. Implications and conclusions).

[10] MAHT-text typology, as presented in this article, may have its linguistic limitations and it certainly does not comply with Hatim & Mason's postulate for multifunction-

ality-based text typology (although it might turn out to be true that texts translated with MAHT tools are multifunctional texts and can be classified according to Hatim & Mason's typology - this, however, remains to be proven by an appropriate corpus-based study).

[11] In CAT the terms **update** and **revision** are clearly distinguished. The former is used to denote changes introduced to a document while its translation is still in progress, while the latter stands for a new version of the whole document released at a later date (i.e. after the translation has been completed).