Eötvös Lóránd University of Budapest
Faculty of Arts and Humanities
Doctoral School of Linguistics
Translation Studies Doctoral Programme

DOCTORAL THESES

ISTVÁN LENGYEL

The Concept of Translation Error from a Functional Perspective

Supervisor: Dr. Gábor Prószéky, DSc, Professor

Budapest, 2013

1. Object of the research, objective of the dissertation

What is a translation error? What types of translation errors exist? Under what circumstances do we consider a translation phenomenon an error in translation? The category of translation error cannot be interpreted in itself: "The perception of what constitutes a translation 'error' varies according to translation theories and norms" (Hansen 2010). An error is a deviation from the expectations, thus we first need to establish the expectations. Expectations can be theoretical, such as the implementation of a type of equivalence, or practical, like conformance to a translation brief, to the requirements imposed by a text-type, or social, such as compliance with the behavioral norms defined by the community of professional translators (Chesterman 1993).

There are ambiguous and unambiguous errors, though – Pym (1992) calls them non-binary and binary errors. Not all translation errors are of equal importance: while a spelling mistake is binary in many languages, it hardly affects the value of a translation, unlike a mistranslation – unless the typo results in a different meaningful and misleading word.

The objective of my research was to create an automatic, language-independent statistical method that spots a certain type of error. After assessing error typologies and corporate quality management practices, and getting an overview of automated translation quality assurance tools, I chose to investigate omissions and insertions further. It is questionable, though, to what extent we may call these errors binary – as I prove it in the experiment carried out with four corpora, in two language pairs, with three reviewers each.

2. Structure of the dissertation

The dissertation first surveys the possible error types, then the possibilities of human and automated error detection, and finally attempts to propose a method for automatically detecting insertion and omission errors. I start the mapping of possible error types in Chapter 2 by researching the scientific literature, then examine the widely available error typologies in Chapter 3. In Chapter 4 I survey the quality evaluation methods of seven enterprises and a non-profit organisation. In Chapter 5 I give an overview of the possibilities of human and automated error detection by

performing a thorough analysis of the functionality of automated quality assurance tools. In Chapter 6 I present an experiment aimed at evaluating the precision of human review as regards insertions and omissions. In Chapter 7 I attempt to discover omissions and insertions automatically through a method that involves the so-called Muse. This is then followed by a summary of the findings and the bibliography.

3. Research methods

During my research I have applied several complementary methods to answer the different research questions. I have performed traditional research work in libraries and on the internet to gather the relevant research literature. During my literary research I have concentrated on the narrow field of evaluating corporate human (as opposed to machine) translations, and I have not examined the assessment methods of either translation training or machine translation.

I had to gather the publicly available error typologies first which was facilitated by the internet and staff from several language service providers. This is how I got access to the description of the LISA QA and the SAE J2450 models which are not public. For a detailed evaluation of the error typologies I referred mostly to the internet, but I also found valuable material in the ATA Scholarly Monograph Series, published by John Benjamins. The TAUS DQF model is only available for TAUS members. I got access to the restricted area and I did an interview with Rahzeb Choudhury, one of the editors of the model.

In order to familiarise myself with corporate error typologies, I asked my contacts at language service providers to suggest enterprises that perform a continuous evaluation on the quality of their translations. I attempted to get in touch with 25 companies, out of which 8 showed interest in allowing me to evaluate their methods in the dissertation. I have been presented their models through telephone interviews, and many of them have actually sent me the Excel sheet they use for the evaluations.

To survey the functionality of automated quality assurance tools, I relied on literature – Makoushina (2007) gives a great summary –, and downloaded trial versions to the relevant tools, or read the online help files.

In my research I performed two experiments. In both experiments I used the same corpus, but while in one experiment the focus was on quantifying human review

decisions, in the other experiment I created a method to automatically assess a certain error type, and compared it to the manipulated segments in the corpora. During the first experiment I simulated a real review task, involving 3 reviewers per corpus (all together 12 reviewers), and I introduced errors into some of the segments that contained supposedly correct translations (for more details please see below). The reason why I chose to have the text reviewed by three reviewers had to do with the fact that the metric used to compare machine translation against human translations, BLEU, also uses three reference translations (Papineni et al 2002). The reviewers got their tasks as a real commission, ordered by an LSP, paid by the hour, without mentioning that this is ordered to assist in a scientific research project. The results of the experiment came in the same bilingual XLIFF file type (Krenz – Ramlow 2008) that reviewers normally return after the review. This is how I tried to avoid the reviewers spotting a pattern and identifying this as an experiment, which could influence the results. Reviewing a sample of the text is common practice.

The other experiment, implementing an automated error detection, consisted of two main steps: development and evaluation. During development I created an algorithm based on an existing tool, the so-called Muse. Choosing to work with the Muse was a logical step because of the characteristics of the tool. The Muse extracts corresponding source and target subsegments from corpora. The actual development work involved the Muse-based evaluation of insertions and omissions, trying to cover the entirety of the segment with hits coming from the Muse. Evaluation took place using automated measurement. I have compared the results of the algorithm with the list of the manipulated segments in the four corpora. As the correlation based on empirical distribution was not strong enough to be predictive, it made no sense to compare the results of the experiment with the reviewers' judgments and also the use of the corpora that included also the original state of the manipulated segments as training corpora did not cause a major distortion in the results (removing these segments could only have deteriorated the results).

In the next section I will present the corpora I used in the experiments.

4. The composition, size and preparation of the corpora

In the two experiments I used four parallel corpora, two of them in English-German, two of them in English-Hungarian language pairs. In the experiment my goal

was to compare the English-Hungarian and English-German ratings, therefore I was looking for corpora of similar domains and text complexity. Thus I finally chose the following corpora:

1. The 2012 volumes of the DGT English-Hungarian translation memory (Steinberger et al 2012). DGT is the Directorate General of Translation at the European Commission. Since 2007 they publish the translation memory compiled from computer-assisted translation in 23 languages. Translation is performed by professional translators, following many guidelines and pre-defined terminology. The translation memory was not created by alignment. I will also refer to this corpus as EU EN-HU.

2. The 2012 volumes of the DGT English-German translation memory (EU EN-DE). This is almost identical to the English-Hungarian translation memory when it comes to English-language source segments.

3. Part of the IT subset of the Hunglish English-Hungarian corpus (Varga et al 2005). This corpus was created using alignment, rather than from human translation assisted by CAT tools.

4. The corpus from the English-German translations of the memoQ translation environment's online help. This text is also IT, but it was created using computer-assisted translation software.

After selecting the corpora, I removed the alternative translations and the easy-to-spot mistranslations from the corpora. This is how I obtained the corpora used in the experiment. The following table illustrates the main characteristics of the corpora from the perspective of the experiment:

Corpus	**Number of segments**	**Average length of source segments (words)**	**Average length of source segments (characters)**	**Average length of target segments (words)**	**Average length of target segments (characters)**
Hunglish EN-HU	14,518	15.60	92.05	13.04	97.15
memoQ EN-DE	20,195	10.25	62.72	9.58	75.89
EU EN-DE	203,313	19.59	128.51	17.40	136.05
EU EN-HU	203,617	19.57	128.74	15.78	132.20

Assuming that the translations of the segments are correct in the corpora, I have introduced errors into approximately 100 segments of each corpus by removing some words from the translation or inserting some words at one or more locations in the translated text. During the omission and insertion I tried to introduce errors but at the same time retain the grammatical and meaningful nature of the sentence.

I have only used the entire corpora to train the automated error detection method. In order to compare the reviewers' decisions and evaluate the automated method I created subsets of 300 segments from the corpora, including all manipulated segments and twice as many untouched segments. From the manipulated segments I have removed the information about the manipulation, making it impossible to distinguish formally between the erroneous and the untouched segments.

5. Novelty of the research

Translation evaluation standards are current topics again in 2012-2013 because of the acceptance of machine translation and post-editing, as evaluating post-editing and the typical errors of machine translation are slightly different from those of human translations. This research gives the most extensive description and evaluation of these categories, and it also supplies best practices to review and evaluate error categories. Among the reviewed standards there are several new standards (such as ISO WD 14080, TAUS DQF, MQM) that were released in 2012-2013 or are still under development. I did not find a review of these standards in the literature.

As far as I know, there has been no scientific research into the corporate evaluation of quality, these interviews created brand new scientific content.

I have also not found much research into the harmony of human reviewers. Similar research was only carried out by the GREVIS project (Brunette et al 2011) and Péter Iván Horváth (2011). My research measured only bilingual review, the unison of several reviewers, comparing the individual results against the "gold standard", spotting the intentionally introduced errors.

It is a novel aspect for translation studies – although similar industry experiments are taking place – to identify how language technology can assist in finding actual linguistic translation errors. Concepts and methods of such research were unknown to me at the time of writing, and still are, but I am aware that two large LSPs, Lionbridge and Welocalize, are working on developing similar tools.

6. Future research areas triggered by the dissertation

The dissertation paves the way for a number of new research questions: for example it would be interesting to learn how many reviewers you need for "optimal" review. TEP, *translation-editing-proofreading*, is a traditional LSP practice that has been questioned by many industry experts, but no research has been carried out into proving or refuting their points. Do two reviewers find all major errors? What error types can one reviewer, two reviewers, three reviewers, etc. identify, with or without technology? The dissertation gives a methodology for such experiments.

In my dissertation I also wanted to examine one more question, the categorization of translation errors based on error typologies. How many reviewers categorize the same error the same way? How many error types can an error belong to according to reviewers and error typologies? The data on the CD-ROM includes further experimental data waiting to be evaluated, because in the case of three out of four corpora the reviewers could categorize the errors not only as omission/insertion but also as mistranslation (even though we have not intentionally introduced mistranslation errors into the text). Mistranslation is encoded as 2 in the Excel worksheets as opposed to insertions/omissions encoded as 1. As we have only intentionally introduced omissions/insertions into the documents, I have not paid attention to the actual number during the evaluation, and have counted with omission/insertion errors even if the reviewer marked it as mistranslation. I am happy to offer this experimental data set for further research.

It would also be interesting to further analyse insertion/omission error types: I have not attempted to cluster the different types of insertions and omissions, but it may be that reviewers judge for example the omission of text in parenthesis differently from the omission of a word in a multi-word proper name. It is already possible to analyse this with the existing experimental data.

It is also an interesting development that I have defined an English-Hungarian and an English-German reviewer "profile": I have found significant differences between the number of marked segments and – not surprisingly – the precision of the review between the two groups. It would make sense to further elaborate on this research and substantiate or refute the findings with further data.

The automated method is not very successful, though it is able to mark such errors with a higher probability than random sampling, and thus save review time.

However, this is not reliable enough to introduce this check into computer-assisted translation software. If we replaced the Muse with another indexing method, even if that's just the conditional probability method, I believe we would get different results. It may be that by fine-tuning this method we could also arrive at significant improvements, but because the Muse-based evaluation seemed logical to me, and I expected better results than what I got, I am not at all sure about it.

7. Research results

Thesis 1. In my dissertation I am the first to present the quality assurance standards and their error typologies in Hungarian language, and I point out that it is worth choosing the error typology according to the specifics of the text. I claim that the error typology is appropriate if it does not confuse the roles in the workflow and the production technology with spotting the error types, i.e. it looks either for causes or for errors.

I have performed a detailed analysis of the SAE J2450, LISA QA model, MQM, DQF and ISO 14080 WD error categories, evaluating the individual error types. I have described the error categories applied in the ATA and the ELTE exam rating, the error categories of ITS Tag Set 2.0 and the built-in error categories of SDL TMS and memoQ. I have pointed out that some error categories such as grammatical errors, terminology compliance errors, etc. show symptoms, whereas others such as mistranslation due to source text error, or incorrect 100% match, are looking for reasons. I have demonstrated why these create ambiguity in the model. Besides substantiating the claim of the thesis I have also given practical advice to create unambiguous error typologies.

Thesis 2. I examine the translation evaluation practices of enterprises, primarily IT enterprises, what error typologies they use, what errors they identify and what information they collect from the reviewers. I point out that enterprise error typologies take inspiration from standards but hardly use standards in their pure forms.

In my dissertation I was the first to describe the error typologies of so many companies either in the Hungarian or in the English-language literature. I gathered the information through interviews. We could see that only HP and Google claim to follow the LISA QA model, but even they deviate from the standard. Google, for

example, introduces the misinterpretation of the source into the model. The other error typologies – applied at VeriSign, Microsoft – show a significant similarity to the LISA QA model, the two distinctly different models are those of McAfee and Symantec.

Thesis 3. I introduce the forms and use of automated error detection and point out that only simple, rule-based error detection has been automated so far. I examine and summarize what error types can be spotted and what cannot be spotted using computer tools today.

In the dissertation I have presented the – very limited – literature on automated translation quality tools, and I gave a summary of the translation errors that can or cannot be spotted using computer tools.

Thesis 4. I perform an experiment to assess the precision of experienced reviewers in the detection of omission/insertion errors in English-German and English-Hungarian translations. I do this to assess what baseline precision would be needed for automated detection. I prove that depending on the definition of translation error there can be significant differences in the results.

The experiment found that depending on our definition of an error (is it an error if we intentionally introduced omissions/insertions, and how many reviewers have to mark the segment erroneous in order for the segment to qualify as erroneous) there were big differences in the results. Among the four corpora in the best and the worst cases the difference between the lowest and the highest number of segments for correctly detected errors was 2.34 to 5.1 times, for undetected errors it was 4.18 to 8.46 times, for correctly unmarked errors – the largest segment number category – it was between 1.07 and 1.68 times, whereas for false positives it was between 3.8 and 16.6 times. We found that reviewers do not find a significant proportion, 35-45% of omission/insertion errors in the case of a medium permissive error definition. The most thorough reviewer missed 26% of the errors, whereas the least thorough reviewer missed 57% of the errors. We proved that the reviewer is likely to mark those segments that are erroneous, but also marks segments that do not contain an introduced error, and the miss rate increases together with the amount of segments marked: thus the reviewers are likely to first mark the errors that are certain. This suggests a consensus type error definition and proves that the two theoretical error categories of Pym (1992) really exist, but they don't necessarily relate to the error

type: there are characteristics along which there is consensus, and there are such characteristics that divide the reviewers.

Thesis 5. I present an attempt to automatically detect insertions and omissions in a language-independent way and I evaluate its reliability compared to human review.

In my dissertation I have presented two methods, the Muse-based method and the segment length proportion based method that have not come close to the utility of human review. The segment length proportion difference could better predict the occurrence of errors, and gave better results for insertions than for omissions, whereas the Muse-based method could spot omissions better than insertions. The fact that selecting segments with these methods offer erroneous segments with 7-128% higher probability than random sampling offers limited practical use. Using such methods may be useful during sampling if the goal of the sampling is to detect errors with a lower effort or budget.

8. Bibliography

Chesterman, A. 1993. From 'is' to 'ought': Laws, norms and strategies in translation studies. *Target.* Vol 5. No. 1. 1–20.

Hansen, G. 2010. Translation 'errors'. In: Gambier, Y., van Doorslaer, L. (eds). *Handbook of Translation Studies*. Amsterdam: John Benjamins. 385–388.

Horváth Péter Iván 2011. *A szakfordítások lektorálása. Elmélet és gyakorlat. Segédkönyvek a nyelvészet tanulmányozásához 117*. Budapest: Tinta Könyvkiadó.

Krenz, M., Ramlow, M. 2008. *Maschinelle Übersetzung und XML im Übersetzungsprozess*. Berlin: Frank & Timme Verlag.

Makoushina, J. 2007. Translation Quality Assurance Tools: Current State and Future Approaches. In: *Proceedings of the 29th International Conference on Translating and the Computer.* London: ASLIB.

Papineni, K., Roukos, S., Ward, T., Zhu, W.-J. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia: ACL. 311–318.

Pym, A. 1992. Translation Error Analysis and the Interface with Language Teaching. In: Dollerup, C., Loddegaard, A. (eds.). *Teaching Translation and Interpreting*. Amsterdam: John Benjamins. 279–290.

Steinberger, J., Eisele, A., Kloczek, S., Pilos, S., Schlüter, P. 2012. DGT-TM: A freely Available Translation Memory in 22 Languages. In: *Proceedings of the 8th international conference on Language Resources and Evaluation (LREC'2012)*. Istanbul: LREC.

Varga, D., Németh, L., Halácsy, P., Kornai, A., Trón, V., Nagy, V. 2005. Parallel corpora for medium density languages. In: *Proceedings of the RANLP 2005*. Borovets: RANLP. 590–596.

9. Publications

Publications and presentations related to the topic of the dissertation

Lengyel I., Kis B. 2003. Új módszerek az emberi fordítás gépi támogatásában. In: Gyimóthy Tibor (ed.) *Az I. Magyar Számítógépes Nyelvészeti Konferencia gyűjteményes kötete*. Szeged: Szegedi Tudományegyetem. 268–274.

Lengyel I. 2003. Dynamic dictionaries: similarity vs. equivalence. In: Ballard, Michel (ed.) *Qu'est-ce que la traductologie?* Arras: Université d'Artois.

Lengyel I., Kis B., Ugray G. 2004. MemoQ – új megközelítés a fordítástámogatásban. In: Gyimóthy Tibor (ed.) *A II. Magyar Számítógépes Nyelvészeti Konferencia gyűjteményes kötete*. Szeged: Szegedi Tudományegyetem. 268–274.

Lengyel I. 2004. Group translation – exploiting synergy. In: Birgit Strotmann (ed.) *Proceedings of IV Jornadas sobre la Formación y Profesión del Traductor e Intérprete*. Madrid: Universidad Europea de Madrid. Available online: http://www.uem.es/web/fil/invest/publicaciones/web/EN/RED.HTM

Lengyel I., Kis B. 2005. A fordítás számítógépes segédeszközeiről. In: Dobos Csilla et al. (ed.) *Mindent fordítunk, és mindenki fordít. Tanulmánykötet Klaudy Kinga tiszteletére*. Bicske: SZAK Kiadó. 53–60.

Lengyel I. 2004. Az Európai Unió és a túlterminologizálás. In: Cs. Jónás Erzsébet és Székely Gábor (ed.) *Nyelvek és nyelvoktatás Európa és a Kárpát-medence régióiban. 1/2. kötet. A XIV. Magyar Alkalmazott Nyelvészeti Kongresszus Előadásai*. Pécs–Nyíregyháza: MANYE – Bessenyei Kiadó. 213–218

Book review:

Lengyel I. 2005. Ljuba Tarvi: Comparative Translation Assessment: Quantifying Quality. *Fordítástudomány.* Vol. 7. No. 1. 112–115.

Presentations:

Lengyel I. 2006. A fordító, a fordítómemória meg a konzisztencia. Delivered at: XVI. MANYE Kongresszus. Gödöllő: SZIE, 10-12 April 2006

Lengyel I. 2009. (presentation) The Future of TM Management. Delivered at: tekom Jahrestagung. Wiesbaden: Rhein-Main Hallen, 5 November 2009

Lengyel I. 2011. (presentation) Catalyzing Business Model Innovation. Delivered at: TAUS. Enabling better translation. User conference. Santa Clara: Hyatt Regency, 6 October 2011

Lengyel I. 2013. (presentation) Translation Tools Processing Each Other's Formats – Dream or Not? Delivered at: AFIT Romanian Translation Forum. Bucharest: Capital Plaza, 25 October 2013

Lengyel I. 2013. (presentation) Terminology in the Cloud with memoQ and TAAS. Delivered at: Creation, Harmonization and Application of Terminology resources. Wiesbaden: Rhein-Main Hallen, 7 November 2013