The impact of translation competence on error recognition of neural MT

Moritz J. Schaeffer mschaeffer@uni-mainz.de
TRA&CO, University of Mainz, Germersheim, 76726, Germany

Abstract

Schaeffer et al. (2019) studied whether translation student's error recognition processes differed from those in professional translators. The stimuli consisted of complete texts, which contained errors of five kinds, following Mertin's (2006) error typology. Translation students and professionals saw translations which contained errors produced by human translators and which had to be revised. Vardaro et al (2019) followed the same logic, but first determined the frequency of error types produced by the EU commission's NMT system and then presented single sentences containing errors based on the MQM typology. Participants in Vardaro et al (2019) were professional translators employed by the EU. For the current purpose, we present the results from a comparison between those 30 professionals in Vardaro et al (2019) and a group of 30 translation students. We presented the same materials as in Vardaro et al (2019) and tracked participants' eye movements and keystrokes. Results show that translation competence interacts with how errors are recognized and corrected during post-editing. We discuss the results of this study in relation to current models of the translation process by contrasting the predictions these make with the evidence from our study.

1. Introduction

Translation competence has long been a more or less central issue in Translation Studies (e.g., Campbell, 1991; PACTE, 2003; Göpferich, 2009; Malmkjaer, 2009; Kiraly, 2013). In order to draw conclusions regarding what constitutes expert behaviour during translation and in order to eventually be in a position to model translation competence a number of studies have compared behaviour during translation by recruiting participants with different degrees of competence or expertise (e.g., Jakobsen, 2002; Rothe-Neves, 2003; Jensen & Pavlović, 2009; Dragsted, 2010; Carl et al, 2016; Daems et al, 2017). However, participant groups are typically formed in a binary fashion (e.g., students versus professionals), are created adhoc or in a qualitative manner. Few validated instruments which would make it possible to systematically compare different groups of participants beyond adhoc or qualitative categorization. The tool advanced by the PACTE group (Orozco & Hurtado Albir, 2002), e.g., offers hardly any numerical items, which makes quantitative analyses impossible or difficult, and the multiple-choice questions used to differentiate groups include very few response options, thus offering a rather limited coverage of what is to be modelled, i.e., translation competence. It is, in addition, difficult to generalize any findings in relation to this tool, given that two large parts consist in a translation and problem/error analysis task confined to an English text. Finally, PACTE provide scant statistical details about its external validation protocol.

The Translation and Interpreting Competence Questionaire (TICQ) presented by Schaeffer et al (2020) addresses a number of these issues. The TICQ establishes a gold-standard instrument for the systematic assessment of translation and interpreting competence and has

been statistically demonstrated to robustly discriminate among participants with null, incipient, and professional experience. The predictive power of the questionnaire was tested with a discriminant function and results showed (Schaeffer et al 2020: 99) that this function could differentiate between innocent bilinguals, translation students and professional translators with a high degree of accuracy (70-84%).

2. Predictive power of the TICQ

While it has been shown that the TICQ successfully distinguishes between groups with different degrees of training and/or experience in the trade (Schaeffer et al 2020), the discriminant function used to do so models these differences in a continuous two-dimensional space. It therefore does justice to the fact that competence is highly unlikely to be categorical and much more likely to be better modelled on a continuous scale. While the ability to disciminate between groups of participants is useful and important, the purpose of the present paper is to test to what extent the coefficients within the discriminant functions used in Schaeffer et al (2020) are predictive of behaviour during translation.

The current study investigates how errors in translations produced by the neural machine translation (NMT) system employed by the Directorate General of Translation (DGT) of the European Comission are recognized and corrected by two groups of participants: professional translators working in-house at the DGT and translation students studyng at the University of Mainz. Both groups of participants filled in the TICQ, coeffcients for each participant were calculated on the basis of the discriminant function as described in Schaeffer et al (2020) and used to predict error recognition processes during post-editing.

3. Modelling translation competence

Schaeffer and Carl (2017) show that phrase based statistical machine translation systems (PBSMT) and human translators deal with translation ambiguity in a similar manner. Training of such a system involves creating expectations about possible a target texts given a source text. Schaeffer and Carl (2017) show that uncertainty as modeled in PBSMT systems is not unlike the uncertainty as moddeled by human translators – as measured by how the degree of uncertainty about possible target texts affects behaviour during translation: the greater the uncertainty of either human or machine, the longer the production durations. Carl (2021) shows that semantic vector space-based models encode small semantic discrepancies across languages such that they are predictive of behaviour during translation. Broadly, the larger the distances in vector space, the longer it takes human translators to process a translation. It is well known that the predictability of upcoming text has a large and very reliable effect on processing during reading (e.g., Smith and Levy, 2013). Whether phrase based statistical or vector space models of translation are more representative of how humans predict, produce and evaluate translations is an interesting question in itself, however, it is beyond doubt that there are large individual differences as to what kind of and how these expectations interact with how text is processed during translation – age, age of acquisition, expertise in a certain area, geographical factors and many others are likely to all affect how a much more fundamental statistical property of words, i.e. word form frequency, is predictive of reading behaviour (e.g., Chen et al, 2018). In other words, a bilingual person's expectations and associated uncertainty regarding translation is likely to differ substantially from a professional translator's expectations and associated uncertainty. Errors in existing text contravene expectations and how and when errors are recognized as contravening expectations and how and when errors are corrected is indicative of, well, the

nature of those expectations, i.e., of the model of translation operating inside a particular (group of) human bilingual(s).

3.1. The current study: participants

Two groups of 30 participants each took part in the study. The group of professional translators were employed by the DGT and the students of translation were inscribed at the University of Mainz. Table 1 below shows the biographic data for student and professional participants. About half of the student participants were early bilinguals (age of acquisition of L2 before the age of 7), while this was the case for only 12% of the professionals. Language use was relatively balanced in both participant groups and professionals rated their own competence in L1, L2 and L3 higher (on a scale of 0 to 100).

	Students		Professionals	
	mean	sd	mean	sd
Age	21.6	(3.0)	46.0	(9.7)
Age at which L2 learning started	7.2	(2.6)	10.3	(3.0)
Number of years learning L2	13.5	(4.4)	12.0	(8.0)
Hours per week reading in L1	7.6	(5.1)	14.8	(11.3)
Hours per week reading in L2	5.9	(4.5)	16.0	(14.4)
Hours per week consuption of L1 Audio	5.0	(6.2)	2.6	(3.7)
Hours per week consuption of L2 Audio	2.7	(3.6)	2.5	(4.1)
Hours per week consuption of L1 AV material	6.1	(7.9)	5.3	(5.2)
Hours per week consuption of L2 AV material	6.0	(5.4)	3.8	(5.7)
Age at which L3 learning started	11.6	(4.0)	13.9	(5.4)
Number of years learning L3	8.3	(5.9)	9.7	(7.6)
Subjective Competence L1 (scale 0 – 100)	89.7	(8.2)	96.6	(3.4)
Subjective Competence L2 (scale 0 – 100)	76.8	(9.6)	82.6	(10.9)
Subjective Active Competence L3 (scale 0 – 100)	57.7	(19.8)	75.6	(12.4)
Subjective Passive Competence L3 (scale 0 – 100)	68.2	(17.8)	86.1	(7.8)
Early bilinguals (age of acquisition < 7 years of age) %	47%		12%	

Tabel 1: Biographic data for participants

Figure 1 visualises the scoring of participants according to the discriminant functions (F1z and F2z) as proposed by Schaeffer et al (2020). A reference line at the median for F1z has been introduced.

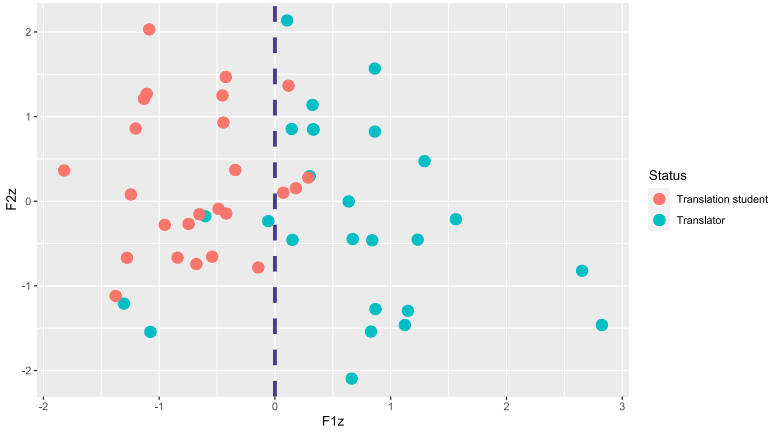

Figure 1: Visualization of scoring according to discriminant functions (Schaeffer et al 2020)

The reference line at the median for function F1z neatly separates participants into the two respective groups with a small number of outliers on each side of the divide.

3.2. Materials and task

The materials are identical to the ones used in Vardaro et al (2019). For a detailed description please consult Vardaro et al (2019). Suffice to say that participants saw single sentences which had been translated by the NMT engine as used at the DGT in 2019 and which had been postedited by in-house translators at the DGT. On the basis of a comparison between the raw NMT and the postedited texts, errors were identified. The sentences which participants saw either contained only one error or none. Each participant always only saw one version of each sentence (with or without error). Participants were asked to correct any mistakes they found nd were told that these had been produced by the in-house NMT of the DGT. In total, participants saw 81 sentences. No time restrictions were given.

3.3. Data gathering method

The sessions were recorded using the non-invasive eye-tracking device SMI RED250Mobile (250 Hz) and the eye-tracking and key-logging tool Translog-II (Carl 2012).

3.4. Data analysis

The statistical analysis with linear mixed-effect regression models (LMER) was carried out in R (R Core Team 2022), using the package lme4 (Bates et al, 2015). The package lmerTest (Kuznetsova et al, 2017) was used to calculate standard errors, effect sizes, and significance values. The effects of the models were visualized in plots for a better interpretation of each model by applying the effects package (Fox and Weisberg, 2019). Residuals outliers (> |2.5| SD) were excluded from the final model. To test for skewness and kurtosis, the package moments (Komsta and Novomestky, 2015) was used. After exclusion of residual outliers, skewness was below |2| and kurtosis below 7, meeting assumpptions of normality (Kim 2013).

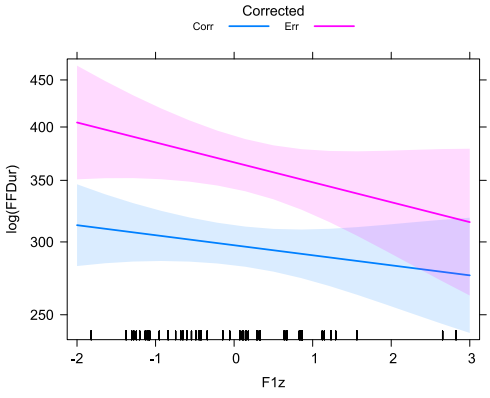

Figure 2: Effect of the presence of an error (Err) on first fixation durations. Errors are recognized during a first fixation, difference between Err and Corr (correct token). However, this effect does not interact with the competence score F1z (on the basis of the TICQ)

3.5. Results

Here, we report two models: We assume that if a token which we considered to be an error was corrected, it must have been recognized as such. The presence of an error did have an effect on first fixation durations ($p < .001$), but this early error recognition effect did not interact significantly ($p > .05$) with the F1z score derived from the TICQ (see above). In other words, irrespective of the participants' degree of translation competence, the time needed to recognize an error remained constant.

The second model traces the interaction between the early error recognition processes and the later stages of the postediting process, i.e., the eye-key span (Dragsted, 2010). The eye-key span (EKS) measures the time between a first visual contact with an error token and the timestamp of the first keystroke which contributed to the correction of this error token. It is reasonable to interpret the duration of the EKS in the following way: The longer the EKS, the more uncertain is the translator regarding the correction of an error that was recognized during a first fixation duration. In other words, the recognition processes taking place during a first fixation duration are likely to recruit largely automatic processes which pitch actual textual material against expectations regarding upcoming text. However, the processes which lead to a correction of the error token are likely to involve deliberations and monitoring processes which are less likely to be automatic. The model we report here, tested a twoway interaction between log-transformed first fixtion duration and the F1z score reported above, the dependent variable being the EKS. This twoway interaction was significant ($p < .01$).

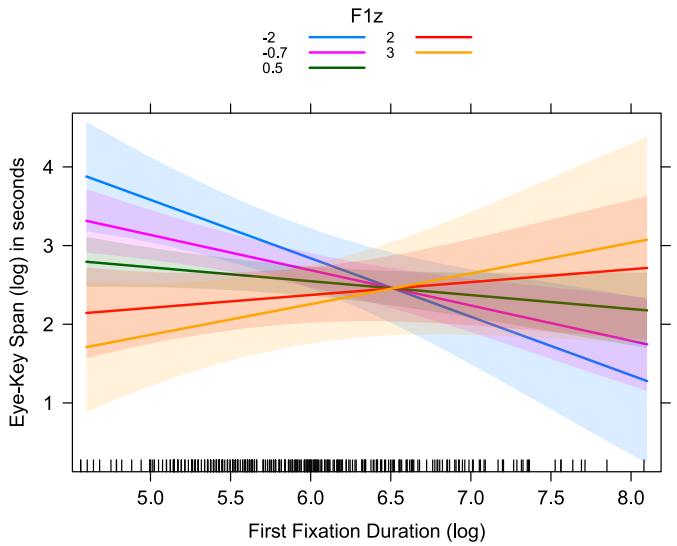

Figure 3: Interaction between log-transformed first fixation duration and F1z score for the Eye-key span.

The interaction was such that for participants with a higher F1z score the log-transformed first fixation duration had a positive effect on the EKS, which for those with a lower F1z score, the opposite was the case. If the error signal in the largely automatic processes was weak (short first fixation duration), for the more competent participants, this resulted in a short EKS. If, however, the error signal in the largely automatic processes was strong (long first

fixation duration), for the more competent participants, this resulted in a long EKS. In other words, participants with a higher F1z score blindly trusted a weak error signal: they corrected the error quickly, while a srong error signal resulted in effortful revision of the output from the early processes before a correction could be carried out. For participants with a lower F1z score, the opposite was the case. Those with less translation competence trusted the largely automatic early error recognition processes blindly only if the error signal was strong (long first fixation durations). The longer the first fixation duration, the shorter the EKS. However, if the error signal from the early processes was weak, they required effortful and lengthy revision of the output from the early processes.

4. Discussion

The present paper shows that a score based on the TICQ (Schaeffer et al 2020) is predictive of error recognition processes in a group of participants with differing degrees of translation competence. The score presented here is on a continuous scale, derived irrespective of a particular language (combination), it can discriminate between differing degrees of translation competence. It does so, in particular, for the interaction between early error recognition and late error correction processes. As such, it is in line with e.g., the model proposed by Schaeffer and Carl (2013), which proposed that output from early, automatic processes is evaluated by later processes. It is the interaction between the early and late stages of error recognition and correction which is carried out differently by participants with differing degrees of translation competence. The scores based on the TICQ are promising not only because they may serve to directly compare participants with different biographies and stages of professional development, but also because they can be predictive of complex behavioural patterns which are relevant to aspiring practicing professional translators, on the one hand, and on the other hand, they may be used to further refine models of the translation process.

References

Bates, D. *et al.* (2015) 'Fitting Linear Mixed-Effects Models Using lme4', *Journal of Statistical Software*, 67(1), pp. 1–48. Available at: https://doi.org/10.18637/jss.v067.i01.

Campbell, S.J. (1991) 'Towards a Model of Translation Competence', *Meta: Journal des traducteurs*, 36(2–3), p. 329. Available at: https://doi.org/10.7202/002190ar.

Carl, M. (2012) 'Translog-II : a Program for Recording User Activity Data for Empirical Reading and Writing Research', in *The Eighth International Conference on Language Resources and Evaluation. 21-27 May 2012, Istanbul, Tyrkiet*. Department of International Language Studies and Computational Linguistics, pp. 2–6.

Carl, M., Aizawa, A. and Yamada, M. (2016) 'English-to-Japanese Translation vs . Dictation vs . Post-editing : Comparing Translation Modes in a Multilingual Setting', in N. Calzolari et al. (eds) *The LREC 2016 Proceedings: Tenth International Conference on Language Resources and Evaluation*. Portorož: ELRA, pp. 4024–4031.

Carl, M. and Schaeffer, M.J. (2017) 'Why Translation Is Difficult : A Corpus-Based Study of Non-Literality in Post-Editing and From-Scratch Translation', *Hermes - Journal of Language and Communication Studies*, (56), pp. 43–57.

Chen, X., Dong, Y. and Yu, X. (2018) 'On the predictive validity of various corpus-based frequency norms in L2 English lexical processing', *Behavior Research Methods*, 50(1), pp. 1–25. Available at: https://doi.org/10.3758/s13428-017-1001-8.

Daems, J. *et al.* (2017) 'Translation Methods and Experience: A Comparative Analysis of Human Translation and Post-editing with Students and Professional Translators', *Meta*, 62(2), pp. 245–270. Available at: https://doi.org/10.7202/1041023ar.

Dragsted, B. (2010) 'Coordination of Reading and Writing Processes in Translation: An Eye on Uncharted Territory', in G.M. Shreve and E. Angelone (eds) *Translation and Cognition*. Amsterdam and Philadelphia: John Benjamins.

Fox, J. and Weisberg, S. (2019) *An R Companion to Applied Regression, 3rd Edition.* Thousand Oaks, CA. Available at: https://socialsciences.mcmaster.ca/jfox/Books/Companion/index.html.

Göpferich, S. (2009) 'Towards a Model of Translation Competence and its Acquisition: the Longitudinal Study TransComp', in S. Göpferich, A.L. Jakobsen, and I.M. Mees (eds) *Behind the Mind: Methods, Models and Results in Translation Process Research*. Copenhagen: Samfundslitteratur (Copenhagen Studies in Language 37), pp. 11–37.

Jakobsen, A.L. (2002) 'Translation drafting by professional translators and by translation students', in G. Hansen (ed.) *Empirical Translation Studies: process and product*. Copenhagen: Samfundslitteratur, pp. 191–204.

Jensen, K.T.H. and Pavlović, N. (2009) 'Eye tracking translation directionality', in A. Pym and A. Perekrestenko (eds) *Translation research projects 2*, pp. 93–109.

Kim, H.-Y. (2013) 'Statistical notes for clinical researchers: assessing normal distribution (2) using skewness and kurtosis', *Restorative Dentistry & Endodontics*, 38(1), p. 52. Available at: https://doi.org/10.5395/rde.2013.38.1.52.

Kiraly, D. (2013) 'Towards a View of Translators Competence as a Emergent Phenomenon: Thinking Outside the Box(es) in Translation Education', *New Prospects and Perspectives for Educating Language Mediators*, (January 2013), p. 197.

Komsta, L. and Novomestky, F. (2015) *moments: Moments, cumulants, skewness, kurtosis and related tests*. Available at: https://CRAN.R-project.org/package=moments.

Kuznetsova, A., Brockhoff, P.B. and Christensen, R.H.B. (2017) 'lmerTest Package: Tests in Linear Mixed Effects Models', *Journal of Statistical Software*, 82(13), pp. 1–26. Available at: https://doi.org/10.18637/jss.v082.i13.

Malmkjær, K. (2009) 'What is translational competence?', *Revue française de linguistique appliquée*, 14(1).

Orozco, M. and Hurtado Albir, A. (2002) 'Measuring Translation Competence Acquisition', *Meta*, 47(3), pp. 375–402. Available at: https://doi.org/10.7202/008022ar.

PACTE (2003) 'Building a Translation Competence Model', in F. Alves (ed.) *Triangulating Translation: Perspectives in Process Oriented Research*. Amsterdam and Philadelphia: John Benjamins, pp. 43–66. Available at: https://doi.org/10.1075/btl.45.06pac.

R Core Team (2022) *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at: https://www.R-project.org/.

Rothe-Neves, R. (2003) 'The Influence of Working Memory Features on Some Formal Aspects of Translation Performance', in *Triangulating Translation: Perspectives in Process Oriented Research*. Amsterdam and Philadelphia: John Benjamins, pp. 97–119.

Schaeffer, M. *et al.* (2020) 'The Translation and Interpreting Competence Questionnaire: an online tool for research on translators and interpreters', *Perspectives*, 28(1), pp. 90–108. Available at: https://doi.org/10.1080/0907676X.2019.1629468.

Smith, N.J. and Levy, R. (2013) 'The effect of word predictability on reading time is logarithmic', *Cognition*, 128(3), pp. 302–319. Available at: https://doi.org/10.1016/j.cognition.2013.02.013.