Dynamic pause assessment of keystroke logged data for the detection of complexity in translation and monolingual text production

Arndt Heilmann
RWTH Aachen
University / Aachen, Germany
heilmann@anglistik.rwth-aachen.de

Stella Neumanm
RWTH Aachen
University / Aachen, Germany
neumann@anglistik.rwth-aachen.de

Abstract

Pause analysis of key-stroke logged translations is a hallmark of process based translation studies. However, an exact definition of what a cognitively effortful pause during the translation process is has not been found yet (Saldanha and O'Brien, 2013). This paper investigates the design of a key-stroke and participant- dependent identification system of cognitive effort to track complexity in translation with keystroke logging (cf. also (Dragsted, 2005) (Couto-Vale, in preparation)). It is an elastic measure that takes into account idiosyncratic pause duration of translators as well as further confounds such as bi-gram frequency, letter frequency and some motor tasks involved in writing. The method is compared to a common static threshold of 1,000 ms in an analysis of cognitive effort during the translation of grammatical functions from English to German. Additionally, the results are combined with an eye-tracking analysis for further validation. The findings show, that at least for smaller data sets a dynamic pause assessment may lead to more accurate results than a generic static pause threshold of similar duration.

1 Introduction

Translation studies can be grouped into two major fields of research: Product-based translation studies and process-based translation studies. In the former, corpus based studies are currently being used for example to find out what makes translation different from original texts in a lingua-culture. Process based studies are rather interested in the emergent translation and how it comes into being. The interest in process based translation research has been growing exponentially in the past twenty years due to technological advances like affordable eye-tracking equipment and key-stroke logging programs that allow researchers to analyze the translators' behavior while translating along with the emergent product. Both methods can be used to operationalize cognitive effort during translation. Straightforwardly, reading times of stretches of the source text can for example be operationalize processing difficulties during the process of translation. For stretches of text with higher complexity longer reading times would be expected (Shreve et al., 2010). Complexity in the process of translation does not only encompass syntactic features of the source text but also typological differences between the two languages in question modulating the possibilities of rendering the lexical-semantics of the source to something equivalent in the target language. Also multiple translation possibilities for words add to the complexity of the task of translation(Schaeffer et al., 2016). This is also reflected in longer production times and thus pauses (Dragsted, 2005). However, what exactly a pause is, is subject to debate in translation process research. Typical pause measures that have been applied by scholars of translation process research, range from one (Jakobsen, 1998) to five seconds (cf. Saldanha and O'Brien (2013) for an overview). However, these pause limits were chosen relatively arbitrarily and it is hard to tell if they are too high or too low to capture translation related cognitive effort (ibid., 121). Also, different translators may have different baselines of production speed and an inter-keystroke span of 1,000 ms (Jakobsen, 1998), for example, might be an indicator of increased cognitive effort for one person but not for another. More recent approaches have been modelling translator dependent baselines for identifying a minimal pause length of cognitive effort (cf. Dragsted, 2005 and Couto-Vale, in preparation) thus neutralizing a confounding factor like participant-specific baselines of pausing behavior. Dragsted manually searched for a plausible pause value for one participant. This allowed her to develop a relative measure for pause length across subjects. She divided the randomly selected participant's production speed minus the time spent on revision by the identified pause value (Dragsted, 2005). While superior too a very rigid pause identification measure, Couto-Vale, (in preparation) rightly notes that Dragsted's measure can still be improved upon. Couto-Vale's method involves a further adjustment of a minimal cognitively effortful pause based on the type of character produced. This means he employs a classification system for characters that need one or more action keys like 'shift' to be produced. He classifies characters by means of the action key combinations first and then classifies them into categories of pauses 32 ms, which is a custom threshold . The category with the highest number of pauses is multiplied by two and additional 128 ms (also arbitrarily) are added to the threshold. Inter-key spans of a character of a respective key-combination higher than this threshold are considered to be cognitively effortful. Couto-Vale, (in preparation) suggests to combine participant specific and key combination specific thresholds. This appears reasonable and useful, however, why exactly this formula should yield cognitively effortful pauses does not become clear. There is, thus, still a need for further differentiation between mere typing related inter key spans (which are essentially confounds) and cognitively effortful pauses. Among such confounding factors are not only key combinations such as those considered by Couto-Vale, (in preparation) but also frequency effects of letters and letter bigrams which should be accounted for when assessing cognitive effort during text production. Their influence on inter key spans is at the heart of this article an will be discussed further below.

2 Modeling dynamic thresholds

For the method of pause identification, we draw upon Couto-Vale's (in preparation) category based classification system. In order to identify a cognitively effortful pause, we assign a character to the subcategories of participant, case, bigram frequency and character frequency first and then combined these to a supercategory. A character like 'a', typed by participant A1 in the German word *haben* ('have') would thus be categorized by means of the ID of the participant, the type of key pressed and frequency information which then combined to a super category with the label: *'participantA1|lower-case|high bigram frequency| high character frequency'*. A script classifies characters as upper case (shift-key + character-key), lower case , space or deletions. Deletions and space characters are grouped separately from the character keys. Especially deletions behave very differently since they almost always occur with an inter-key span of below 10 ms. The reason for this is that the backspace-key was kept pressed until the mistyped word was deleted. The keystroke logger Translog-II (Carl, 2012), which was used to record the keystrokes for the data at hand, recognizes this event as multiple key presses with a minimal inter-key span. These minimal spans would have skewed the pause analysis due to a large number of very low values. For alphabetical characters, each character press is also characterized as belonging to a high frequency group of German letters or a low frequency group. Characters receiving the same classification are grouped into the same category along with their inter-key spans in milliseconds. If an inter-key span exceeded 10,000 milliseconds, it was excluded beforehand since this pause was deemed unlikely to be linked to the complexity of the input, but rather lexical decision problems or dictionary look-ups. Inter-key spans in a super category consisting of participant, case and frequency classifications were tagged as a cognitively effortful pause when they were higher than the third quartile + three times the interquartile range of that category. Such values can be considered extreme outliers (Norman and Streiner, 2007) and it is thus likely that most of the inter-key spans below this threshold are related to normal typing activity. For the frequency classifications, character frequencies below the median of the character frequency lists are classified as 'low frequency' and above as 'high frequency'. The same is done for bigram frequencies. However, keeping the frequency types apart led to super-categories with the frequency classification of low|low, low|high, high|high or high|low. Due to the very fine grained distinctions some categories did not receive many data points. This would have made it difficult to make a reliable judgement with respect to outliers. Therefore a way to combine bigram and character frequencies was devised to reduce the number of subcategories and thus attenuate this problem.

3 Assessing possible interaction effects of frequency types

The formation of a sum score could have distorted the effects frequency on processing effort of either if character frequency would be modulating the effect bigram frequency on inter-key spans (or vice versa). In order to avoid this, a possible interaction effect of both was explored. Five translations of the same text by five different translators were analyzed statistically. Their keystrokes were logged with Translog-II (Carl, 2012). The participants were allowed to use an online dictionary, though this might affect pausing behavior and may be responsible for very long pauses not related to the processing of linguistic complexity but vocabulary problems. The five participants were German students of English linguistics enrolled in their master's and the source text was an abridged popular scientific text written in English. Prior to the translation, the participants were asked to copy a short German text, in order to familiarize themselves with the keyboard. The data consists of the participants' keystrokes and the inter-key spans associated with them. Only inter-keyspans between two characters were taken into account. Pauses after mouse-events or arrow-keys were excluded. The term inter-key span will be used instead of pause in order to differentiate it from time spans signifying a possible cognitively effortful pause (cf. the KD-files of the CRITT Translation Process Database for a similar data structure (Schaeffer et al., 2016).)

The data was analyzed by means of linear mixed regression model in R using the R-package *lme4* (Bates et al., 2015). Mixed regression models allow to control against item-specific variation. We applied the R-package *lmer-test* which calculates p-values with a combination of F-tests and likelihood ratio tests through Satterthwaite approximations (Kuznetsova et al., 2015). Cognitive effort was operationalized by the inter-keyspans for preceding each character in milliseconds. This measure was log-scaled in order to approximate a normal distribution. The model was enriched with case information (upper, lower), bigram-frequency and character-frequency (interval data) sourced from the 'Wortschatz' project (UniLeipzig, 2012) (Lyon, 2012). Since no data was available for bigram frequencies for characters preceded by 'space', the mean bigram frequency was used as a proxy for these cases. Punctuation marks and spaces were excluded from the analysis as were deletions. Since no bigram information was available for word initial characters, these were assessed with the character frequency only. Both were modelled as an interaction effect. The frequencies were log-scaled and z-scored. Each source text token was given a unique ID to control against item-specific variation in the form of a random effect. Participant specific variation was factored in by modeling each participant as a random effect as well. The final model was: *Pausetime per character ~ bigram frequency * Character Frequency + Case+ (1|Unique Source Text ID)+ (1|Participant)*

The model retreived a highly significant effect for the interaction between bigram frequency and Character Frequency (b=0.08, p<0.001). Also the main effects for Character Frequency (b=-0.21, p<0.001) and bigram frequency (b=-0.17, p<0.001) were significant as was Case (b=-2.04, p<0.001). These findings for case corroborate Couto-Vale, (in preparation) suggestion to use key combination-dependent thresholds when determining pauses. However, the results obtained here may be confounded by the fact that almost every upper case letter is also located word-initially which alone leads to longer inter-key spans (Immonen and Mäkisalo, 2010) so that it is hard to differentiate between the effect of upper case letters and word-initial characters. While the estimate of both frequency types is distinctly weaker than that of case, the results show that the higher the letter and bigram frequency, the smaller the inter-key spans. This is especially pronounced for high frequency letters in high frequency bigrams, as can be determined from the significant interaction between these two variables. These differences may become the decisive factor for being either correctly within a static pause threshold or erroneously above it. In order to account for this interaction effect the combined frequency category was formed by multiplying the z-scored frequencies for a character by the z-scored frequencies of a bigram it occured in at second position. The super-categories thus consisted of inter-key spans of key presses classified by *'participant|case|combined frequency'*.[1]

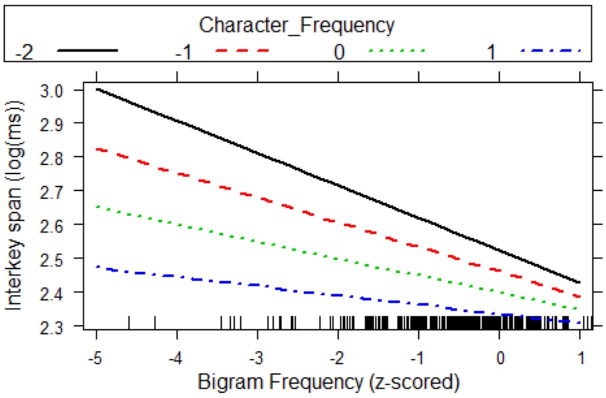

Figure 1: The interaction effect of bigram frequency and character frequency on the inter-key span. High frequency characters occuring in high frequency bigrams lead to shorter inter-key spans.

4 Comparison with a static threshold

To test if the dynamic method performs better than the static one, the cognitive effort during the translation of grammatical subjects from English into German was compared with the effort during the translation of other grammatical functions. The same materials and participants from section 2 were used. Additionally, the source texts were annotated with grammatical functions following the Cardiff Grammar (Fawcett, 2008) and aligned with the productions. For the analyses, cognitive effort was operationalized by the average time per pause while translating a grammatical function from an English source text to a German target text. This means, that the time associated with each identified pause for the translation of a grammatical element in the source text was summed up and then divided by the number of identified pauses. The minimum threshold to identify a pause for the static method was set to 1,000 ms which is an often used customary threshold (Carl and Kay, 2011), (Jakobsen, 1998). The minimum threshold for the dynamic method of pause identification was calculated in the way described above i.e. by means of category- dependent outlier identification. Inter-key spans of > 20 seconds were excluded, since they were likely to be caused by dictionary look-ups and not linguistic complexity. The participants and materials were the same as in Section 2. Eye-tracking data in form of the total reading time for the grammatical functions was used to triangulate the findings. Again linear mixed models were employed. For the pause-related models we controlled for the length of the translation of a grammatical function by counting the number of typed and deleted characters ('TT item length'). For the reading related model the length (in characters) of the grammatical functions of the source text was controlled for ('ST item length'). participant and

[1] In order to classify a key as high or low frequency, the median frequency of the combined frequency category was used as the decisive criterion as before

item-specific variation were modeled as random effects: *Operationalization of 'Cognitive Effort' ~ Grammatical Functions + Item Length (count data)+ (1|Participant)+(1|Unique Source Text ID)*

The dynamic measure of the average time spend in pause found significantly higher average pause time for the translation of conjunctions (b=1611.69, $p< 0.05$) and the translations of main verbs (b=1213.79, $p< 0.05$). compared to that of grammatical subjects.[2] The static measure, however, did not find significant effects for the translation of coordinating conjunctions (b=1286.64, p= 0.15) or main verbs (b=1213.79, p= 0.06). The eye-tracking data corroborates the results of the dynamic but not the static method of pause identification since the results for the measure of Total Reading Time for the variables Coordinators (b=-0.42, $p<0.05$) and Main Verb (b=-0.52, $p<0.0001$) are significant here, too[3]. Still, the static measurement found marginally significant results for the translations of main verbs and it is possible that with more datapoints, the static measure would have found similar results to that of the dynamic measurement. Another interesting observation is that the estimates for Coordinators and Main Verbs in the reading time measure are negative and not positive as in the pause measures. Usually, longer reading times are associated with higher processing effort in monolingual reading. However this relationship of reading time to cognitive effort does not necessarily hold true for translation since it is a very complex task. It is more likely that the translation of these grammatical functions requires excessive target text monitoring and local decision making that is not bound to the source text any more - once the necessary information is acquired. This case highlights the need to look at both eye-tracking data and keystroke logging data to draw conclusions about cognitive effort in translation.

5 Conclusion

This paper shows that additionally to participant-specific and key-combination dependent thresholds, it is worthwhile to also include frequency information for letters and bigrams in the target language to identify cognitive effort in translation, since they have a significant interaction effect on inter-key spans.

While the static threshold did not show the significant results of the eye-tracking data and the dynamic pause measure, it is very possible, that with more data points the static threshold would have found similar results. Using static thresholds with larger amounts of data might thus still be a useful approach to identify cognitive effort in translation if time and resources are scarce. For smaller data sets a dynamic pause measure seems to be a more appropriate solution to identify cognitive effort and linguistic complexity along with it.

Acknowledgments

This study was carried out as part of the TRICKLET project, supported by the German Research Foundation (DFG) under grant project number NE 1822/2-1)

References

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Michael Carl and Martin Kay. 2011. Gazing and typing activities during translation: A comparative study of translation units of professional and student translators. *Meta*, 56(4):952–975.

[2] Further variable levels and control variables (non significant): **Static Avg. Pause Time:** Adjunct; b=624.77; p=0.37 Auxiliary; b=-96.62; p=0.94 Complement; b=287.89; p=0.64; TT Item Length; b=9.66; p=0.2; **Dynamic Avg. Pause Time:** Adjunct; b=560.51; p=0.37 Auxiliary; b=218.14; p=0.85 Complement; b=69.41; p=0.9; TT Item Length; b=8.2; p=0.21; **Total Reading Time** Adjunct; b=-0.15; p=0.28 Auxiliary; b=-0.34; p=0.14 Complement; b=0.02; p=0.86

[3] The control variable ST Item Length was also significant with (b=0.03; p= <0.001

Michael Carl. 2012. Translog - II: A program for recording user activity Data for empirical reading and writing Research. In *Proceedings of the Eight International Conference on Language Resources and Evaluation, European Language Resources Association (ELRA).*

Barbara Dragsted. 2005. Segmentation in translation: Differences across levels of expertise and difficulty. *Target*, 17(1):49–70.

Robin Fawcett. 2008. *Invitation to systemic functional linguistics through the Cardiff grammar: an extension and simplification of Halliday's systemic functional grammar.* Equinox textbooks and surveys in linguistics. Equinox.

Sini Immonen and Jukka Mäkisalo. 2010. Pauses reflecting the processing of syntactic units in monolingual Text Production and translation. *Hermes Journal of Language and Communication Studies*, 44:45 – 61.

Arnt Lykke Jakobsen. 1998. Logging time delay in translation, LSP texts and the translation process. pages 73 – 101.

Alexandra Kuznetsova, Per Bruun Brockhoff, and Rune Haubo Bojesen Christensen. 2015. Tests in Linear Mixed Effects Models. Available at: https://CRAN.R-project.org/package=lmerTest.

James Lyon. 2012. German letter frequencies. Available at: http://practicalcryptography.com/cryptanalysis/letter-frequencies-various-languages/german-letter-frequencies.

Geoffery Norman and David Streiner. 2007. *Biostatistics: The Bare Essentials.* B.C. Decker.

Gabriela Saldanha and Sharon O'Brien. 2013. *Research Methodologies in Translation Studies.* St. Jerome.

Moritz Schaeffer, Barbara Dragsted, Kristian Tangsgaard Hvelplund, Laura Winther Balling, and Michael Carl. 2016. Word translation entropy: Evidence of early target language activation during reading for translation. In Michael Carl, Srinivas Bangalore, and Moritz Schaeffer, editors, *New Directions in Empirical Translation Process Research: Exploring the CRITT TPR-DB*, pages 183–210. Springer International Publishing, Cham.

Gregory M. Shreve, Isabel Lacruz, and Erik Angelone. 2010. Cognitive effort, syntactic disruption, and visual interference in a sight translation task. In Gregory M. Shreve and Erik Angelone, editors, *Translation and Cognition*, pages 63 – 84. Benjamins.

UniLeipzig. 2012. Wortschatz. Available at: http://corpora.uni-leipzig.de.