Chapter 6

Investigating patterns of saccadic eye movement when using Microsoft's Skype Translator between Catalan and German

Felix Hoberg
Leipzig University

This paper investigates the patterns of saccadic eye movement when using Microsoft's Skype Translator between Catalan and German. As being part of an overall evaluation of the Skype Translator on a dialogue-oriented level, a case study on 21 German-speaking participants was conducted. Despite not having any proficiency in Catalan, these participants had to text-chat with Catalan native speakers via Skype, while the Skype Translator was activated. The sessions were observed by an eye tracking system. The collected data thus represents a naturalistic starting point to evaluate how users structure computer-mediated communication situations when real-time machine translation is involved while having to rely on that output.

1 Introduction

Automatic language processing, auto-speech recognition and machine translation (MT) are considered valuable innovations by the language industry. However, progress in this field is still viewed skeptically, which in turn calls for continuous evaluation of the aforementioned systems (Ramlow 2009; Bowker & Ciro 2019). There are indeed different metrics and standards which allow for a categorical evaluation of the machine-translated output either manually or automatically (see §2.3).

Felix Hoberg. 2021. Investigating patterns of saccadic eye movement when using Microsoft's Skype Translator between Catalan and German. In Mario Bisiada (ed.), *Empirical studies in translation and discourse*, 141–163. Berlin: Language Science Press. DOI: 10.5281/zenodo.4450087

Especially when it comes to dialogic interactions between humans and MT, research has so far tackled either the interactive or the technological aspect, but seldom both of them at once. Microsoft's Skype Translator will thus serve as a central element in this case study, as it offers real-time machine translation in 10 languages in voice and video chats and 60 languages in text chats.

The general aim of this project is to highlight how MT evaluation can be applied on a dialogue-oriented level to services like the Skype Translator where all messages are displayed in a two-column-design with outgoing messages right-aligned and all MT output and incoming messages left-aligned. This study hence does not intend to offer an evaluative application of MT quality metrics on the Skype Translator's output, but to outline the users' perception and behaviour when it comes to using the machine-translated output in a real-time conversation. Thus, the present article combines research in the fields of communication research (e. g. Beißwenger 2007) and machine translation (e. g. Fišer & Beißwenger 2017).

To examine the users' behaviour, an exploratory eye-tracking-based case study was carried out. In that study, Skype Translator-mediated text chats between German and Catalan native speakers were captured in order to investigate the eye movement patterns on characteristical areas of interest of the Skype Translator, namely the entry mask and each single text chat message box (see Fig. 6.1, p. 150).

This paper's guiding research question thus is how the participants are perceiving the incoming and outgoing text messages. Based on the assumption that the MT output into German will need more attention than the other messages and that Catalan messages will be nonetheless taken into account (as the – possibly error-prone – new information is presented in both languages), it has to be investigated how participants handle this bilingual input. Special attention will be drawn upon saccadic eye movements.

For that reason, §2 introduces the theoretical background in terms of research on dialogue and computer-mediated conversation in the context of computer-mediated communication, along with previous findings on eye movements in reading tasks. §3 gives insights on the overall project conception, before explaining in detail to which extent the collected data is used for this analysis. Then, §4 presents the results of the saccadic eye tracking data. §5) situates the results against the theoretical background, before the conclusion in §6 sums up the analysis, going back to the overall project.

2 Background

2.1 Research on dialogue and conversation

Since the early 1990s, various concepts in communication research have been modelled and restructured to fit modern computer-mediated communication (Fišer & Beißwenger 2017: 7). Apart from taking a look at global concepts such as text, sender, recipient or conversation, the interest in research has now passed on to questions which reflect the transitional processes web-based communication has undergone over the last two decades: How do we interact online? How does online interaction change our ways of communicating? Can we still speak of sender and recipient after all? How do we cope with this extensive amount of data and the rising machine learning technologies? (cf. Beißwenger 2007).

These questions also implicitly refer to the phenomena of turn-taking and speaker switch or the rising use of the term *hypertext* to describe digital textual behaviour (Storrer 2001), central elements which have already been extensively studied regarding analogue, face-to-face and monolingual web-based communication, but so far have not been adopted to bilingual, machine-translated, web-based conversations such as presented in this paper. This gap might be attributed to the fact that online communication follows different rules than offline communication.

There are two obvious differences between oral, face-to-face communication and chat communication. The latter appears in written or typed form and lacks almost all non- and paraverbal elements like gesture, intonation or eye contact etc. which usually help to structure the communication act (Beißwenger 2007: 172). In contrast, an online chat message passes through more sections between sender and addressee. From the sender's mind, the message goes from typing on the keyboard to the computers' short-time memory and from there to the server the software in use is connected to. From that server it goes to the addressee's software and is subsequently processed by the computer to be displayed on screen before the addressee can spend cognitive resources on it (Beißwenger 2017: 146). Also, the additional time to send, machine-translate and receive the original message has to be taken into account for the Skype Translator. In case of high latency, this time gap can have a severe impact on communication, because while the receiving person is still answering one incoming message, the other may already have sent another text. This can result in an asynchronous communication.

Thus, the use of computer-mediated communication technology, and in this case more precisely the Skype Translator, leads to a change in the communication process of sending and receiving messages. A text chat message has to be com-

pletely written before it can be sent[1] and it has to be received and completely read before it can be reacted to. At the same time, as opposed to oral communication, the communication partners are not necessarily in the same location, nor near to each other at all (Beißwenger 2017: 146). Storrer (2001: 3) points out another important feature: even though online chatting appears mostly in written form, it follows the rules of oral production. The relationship of officially standardized language and its informal, but also widely accepted online communication use, which follows its own rules, has been object of many research projects ever since, for example for Dutch, see Verheijen (2017). This relationship might helpfully be investigated by an eye tracking study.

Consequently, the indicators explained in §2.4.2 can be taken as initial points of reference on how the participants process the information on screen when text-chatting with people, whose language they do not speak.

2.2 The Skype translator

As has already been stated in the introduction, Skype features a real-time translation engine called Skype Translator for text chats between 60 different languages and for voice and video chats between eleven languages[2]. Both the written and the video or voice real-time translation engine are based on machine learning and Microsoft's proprietary neural machine translation system, meaning that the output is supposed to enhance in terms of quality every time the feature (and any other product of Microsoft) is used. Additionally, some of the supported languages come with language detection, text-to-speech, speech-to-text, transliteration, a dictionary and the possibility of customizing the output according to individual terminology.[3]

2.3 Machine Translation Evaluation

There are several manual or automatic methods to evaluate the translation quality in general. With the expanding use of machine translation, evaluation methods are being adopted to the new environments (see e.g. *multidimensional quality metrics*[4], *LISA QA* or *SAE J2450*[5]). Automatic MT evaluation metrics are being

[1] Real-time text chat, where the text is transmitted immediately so that every user can observe the production process, will not be considered here.

[2] https://www.skype.com/en/features/skype-translator/, last accessed on 4 November 2020.

[3] https://www.microsoft.com/en-us/translator/business/languages/, last accessed on 4 November 2020.

[4] http://www.qt21.eu/quality-metrics/, last accessed on 4 November 2020.

[5] https://blog.taus.net/the-8-most-used-standards-and-metrics-for-translation-quality-evaluation, last accessed on 4 November 2020.

modelled and investigated for post-editing (Vardaro et al. 2019: 2) and for raw MT output (Doherty & O'Brien 2014).

Most metrics and standards are designed to provide results that are comparable in quality to human translations, but are based on rather subjective ground since even the most automatised metrics often compare MT output to human reference translations. Another closely-related problem is the vast amount of different aspects to account for when evaluating MT systems (name entities, lexical issues, syntactic issues etc.) (Han 2018: 2f.). In contrast, „eye tracking could remove much of the subjectivity involved in human evaluation of machine translation quality as the processes it measures are largely unconscious. " (cf. Doherty et al. 2010: 12) Furthermore, „[e]ye tracking has been used successfully as a technique for measuring cognitive load in reading, psycholinguistics, writing, language acquisition etc. for some time now" (cf. Doherty et al. 2010: 12). From another point of view, „[i]nclusion of users in evaluation of MT systems can provide benefits in both directions: such as positive influences on system development and its usability" (Doherty & O'Brien 2014: 4) to thereby improve the system's performance, output and efficiency.

2.4 Eye-tracking and machine translation evaluation

2.4.1 Machine translation evaluation

Making sense of the process that leads to a final translated product has been object of translation studies for decades. There are multiple tools and methods to acquire information on the current cognitive processes of (mostly student) translators when asked to translate something: think-aloud protocols, corpus studies, product evaluations, comprehensibility tests, stimulated recall interviews.

„Records of eye movements, however, can do this very unobtrusively" (Schaeffer et al. 2017: 23), since it has been pointed out that „[c]ertain characteristics of readers' eye movements have been shown to be sensitive to the underlying cognitive processes involved in lexically identifying words" (Schaeffer et al. 2017: 23). Additionally, as has already been stated in §2.3, MT evaluation always has to keep an eye on usability and employability of the respective system and MT output. In consequence, using eye-tracking methods in translation process research leads to a better understanding of the effectiveness, efficiency and satisfaction of the task that is completed by a specific user (cf. Doherty & O'Brien 2014: 6).

Therefore, instead of being closely guided by the quality metrics for MT evaluation, which all aim to possibly reach error-free (almost human) quality, the investigation of Skype Translator-mediated conversations focuses on the usefulness and usability of the MT output in general and the way of users making

sense of what they are reading. Doherty & O'Brien (cf. 2014: 4) state that „there are relatively few studies on the usability of raw machine translated output". Little research has been done so far on real-time chat communication – and even less on bi- or multilingual machine-translated communication. A study using eye tracking methods explores the perception of software like this.

2.4.2 Eye-tracking, saccadic eye movements and the Skype Translator

This article focuses on Skype's text chat function, that is, on written communication. Similar issues concerning voice and video chat will not be discussed here, since Catalan is not supported in those modes. That being stated, the focus moves to written text and its perception by its readers (or users), which is being investigated in eye tracking studies. Apart from fixations, saccadic eye movements can be taken as an early measure of cognitive load and mental processing. As has already been investigated, saccades vary among different kinds of reading tasks (Rayner 1998: 373). Jacobson & Dodwell (1979) for example studied left-to-right and vice versa directed saccades on (pseudo-)words, showing „that the probabilities of word components (letters, bigrams, etc.) can affect the speed with which words must be synthesized from their components before recognition occurs" (Jacobson & Dodwell 1979: 313). Schaeffer et al. (2017: 24) hypothesise that proofreading a text requires more cognitive load than reading for comprehension. They found out that saccades made during proofreading were shorter than during reading for comprehension. With respect to the Skype Translator, name entities, numbers or words of similar characters in all the involved languages may represent a similar challenge.

More precisely, studies on the matter also require fine-grained equipment to capture those high-velocity movements. In this context, Leube et al. (2017) point out the varying quality of capturing saccades with mobile eye tracking systems with a sampling rate of 60 or 120 Hz and a stationary system with 1000 Hz. This is important, since saccade duration mostly tends to range between 10 to 100ms (cf. Duchowski 2017: 40). Saccades represent movements of multiple characteristics that include blinks, regressions, corrections and glissades. All of theses have to be kept in mind and will be investigated in upcoming studies.

The present article focuses exclusively on saccade amplitude and duration as both are well described in scientific literature and thus widely used. They are defined as follows: „The saccadic amplitude (...) is the distance travelled by a saccade from its onset to the offset. The unit is typically given in visual degrees (°) or pixels (...) " (Holmqvist 2011: 312).

> During reading, for instance, saccadic amplitude is known to adapt to combined phyiscal, physiological, and cognitive factors. Reading saccades are limited in length by the visual spanwidth which is around 7-8 letters (2°) in the average reading situation. (Holmqvist 2011: 312)

Shorter saccades in terms of amplitude are made if a text is complex and thus difficult to read, which in turn can be taken as indicative for increased cognitive load. (Schaeffer et al. 2017: 24) Similarly, reduced saccade amplitude occurs when a participant inspects something carefully.

> Saccadic duration ('transition time'; not the same as transitions between AOIs) is defined as the time the saccade takes to move between two fixations or instances of smooth pursuit. (Holmqvist 2011: 321)

A longer saccadic duration can be taken as indicative for processing more difficult tasks (Holmqvist 2011: 312). „Thus, as text gets more difficult, fixations get longer, saccades get shorter, and more regressions are made" (Rayner 2009: 1460).

This article is therefore based on the assumption that, given a bilingual, machine-translated reading and text-chatting task, the saccade amplitude and duration varies depending on the different languages (Catalan vs. German) and text types (MT vs. original). It is then interesting to take a look at how the difficulty of reading MT output and foreign language differs in real-time text chat communication. The last claim on investigating saccades is the general question of how useful this indicator is in general when looking at reading behaviour in text chat communication.

3 Research design

3.1 Participants and task

For this study, 25 students with no proficiency in Catalan were recruited. The legal consent on the anonymous processing of their data was obtained explicitly before the study started and all participants were debriefed after having taken part. They also were rewarded with 10 euro each. Of those 25 participants, four had to be excluded due to insufficient data quality. Of the remaining cohort, 20 were students of the Leipzig University and one was a student of the Leipzig University of Applied Sciences (HTWK). As the call for participation was sent to almost all departments of these two universities, the participants vary in terms of programs they are enrolled in.

Three Catalan native speakers – two female and one male, ages 26, 24 and 26, respectively – were recruited as text chat counterparts for this study. All three came from different cities in the Catalan Countries: Valencia, Girona and Barcelona. All three were proficient in German since they took part in an exchange program during their studies and/or lived in Germany for a while.

Considering the amount of time each session took, the restricted access to the eye tracking system and the individual availability of the individual participants, it was impossible to have the German participants text-chatting with the same single Catalan native speaker. Recruiting only one Catalan native speaker would definitely have contributed positively to the comparability of the study, though, but this option was rejected facing the problem of recruiting students who were supposed to meet the above mentioned multiple conditions.

The task the German participants had to fulfill was split into three steps. First, they were asked to answer a questionnaire on their communication behaviour and their foreign language proficiencies. Then, they had to text chat with a Catalan native speaker via Skype, with the Skype Translator activated. This part was captured by an eye tracking system. In order to get comparable data, the participants were given an introductory instruction: To have a central theme the participants could chat about, they were told to imagine they were about to spend a year abroad in Catalonia trying to get some information in advance on where to live and how to start there. Therefore, they were contacting the Catalan native speaker. This task allowed the participants to text-chat freely in a naturalistic manner according to their individual communication behaviour. On the other hand, constraining the task was intended to produce comparable linguistic data, which can be analyzed in possibly upcoming corpus studies. Lastly, to get an impression of the participant's individual experience during the Skype session, they were asked to fill out another questionnaire afterwards concerning the output quality of the Skype Translator.

The introductory questionnaire provides additional data regarding the composition of the cohort. The students participants mean age was 23.7 (SD = 4.0, range = 20–32 years). When it comes to (foreign) language proficiency with the Common European Framework of Reference for Languages (CEFRL) as criterion, all of them indicated German as their first language with respect to use in ordinary and work life. 17 participants had English as a foreign language. As for Romance languages, French and Spanish were reported nine times each, and Italian and Portuguese once each. Possible influences of Romance language proficiencies on the participants' behaviour have to be taken into consideration in a full-range analysis, but will not be discussed in this article.

Looking at the user behaviour regarding Skype, only 17 participants reported using the software, and 13 of them less than once per month. With regards to the duration per session, four participants used Skype no longer than 15 minutes, five no longer than 30 minutes, four up to one hour and four beyond one hour.

The next part of the questionnaire was devoted to the use of alternative software, which includes all or some of Skype's functions, such as voice chat, followed by a detailed inquiry on alternatives for the individual Skype functions voice chat, video chat and text chat. Of 17 participants using alternatives, 16 used WhatsApp for voice chats, 15 for video chats and 16 for text chats. Some participants stated that they were using other alternatives such as Telegram or Discord, too. Only three of them declared Skype as their preferred and most used software for video chats. As for voice or text chat, Skype was mentioned zero times as preferred and most used software. Instead, WhatsApp was indicated to be used most times. Last, the questionnaire took into account the participants' experience of living abroad. 13 of them reported some experience living abroad for a mean of 30.53 months (SD = 36.36, range = 1–108 months).

In summary, this questionnaire draws a picture of the participants' high familiarity with communication software and their proficiency in at least one, but often even two or more languages apart from their native language. The latter observation is also supported by the high range of experience in living abroad. Taking a closer look at Skype, the software is not the primary mean of communication but other wide-spread, mobile applications such as WhatsApp. This leaves room for two opposed suggestions: Either the participants rely on their foreign language skills and foreign culture experiences and thus do not need a machine-translated communication feature like the Skype Translator or as the questionnaire insinuates, the participants are hardly aware of this feature and thus have not made use of it. Both suggestions can be used to strengthen the claim for investigating the users' behaviour when communicating via machine-translated output.

3.2 Data collection

The *EyeLink Portable Duo* eye tracking system was used to conduct the study. The sessions were recorded in the *head-free-to-move* setup with a sampling rate of 1000Hz and binocular tracing. The overall setup included an eye tracking camera on a tripod, which was placed directly between the screen and the keyboard around 60–70cm from the participants' head, a display computer with Skype and the screen captioning software packages installed, and a host computer to handle

the eye tracking system. The software in use also allowed capturing messages (buttons pressed etc.).

The core element of this study was the latest version of Skype up to that date (8.x), which already presented the Skype Translator as a built-in feature. The only requirement was to start a new conversation and add the Skype Translator service by clicking on the respective button in the user's profile one wanted to chat with. The service displayed messages in a two column structure: original messages of the user appear right-aligned, the MT output of the user, and the counterpart's incoming messages and the respective MT output appear left-aligned (see Fig. 6.1). During all sessions, Skype was displayed on maximum on screen to ensure equal quality for every participant and recording session. Nevertheless, Skype does not allow to use bigger font size in order to identify single words as AOI as is often recommended for reading studies (cf. O'Brien 2009: 261). That is why the data preparation (see §3.3) is restricted to the chat message level. The proprietary EyeLink Data Viewer-Software (SR Research Ltd. 2019) was used to process the raw eye tracking data. R version 3.4.3, (R Development Core Team 2019) and RStudio were consecutively used to analyse the processed data.

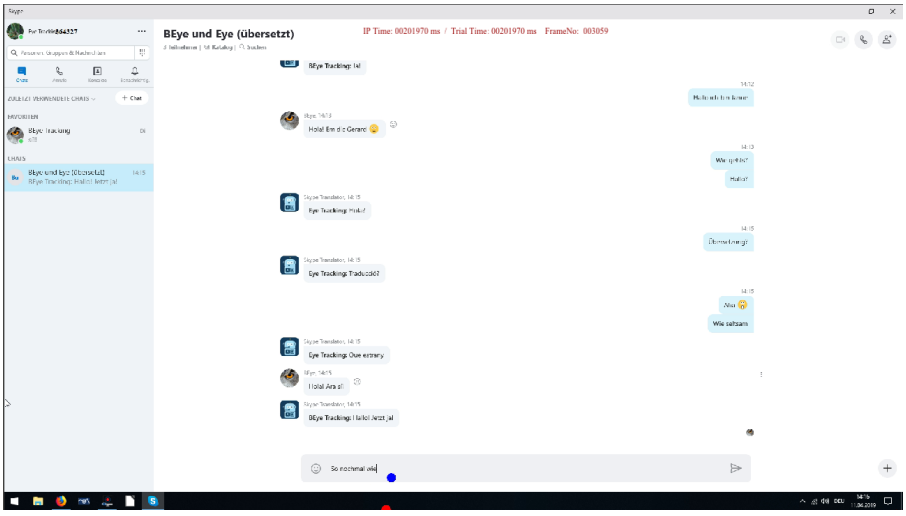

Figure 6.1: Example of text boxes in Skype. Left-aligned (grey): incoming messages and all MT output. Right-aligned (light blue): original messages of the participant.

3.3 Data preparation

There are two kinds of analyzable data that come from this study. On the one hand, there is the bilingual, authentic linguistic material produced by the participants, the Catalan native speakers and the machine translation of Skype which can be subdivided into four categories: the German and the Catalan original and the machine translated output, respectively. This kind will be spared for further research and publications.

On the other side, there are the screen captions of the eye tracking sessions. These had to be annotated with dynamic areas of interest as the single text panels in Skype move when a new message is displayed on screen. To allow for a detailed analysis of the four linguistic categories mentioned above, the text boxes of each session is marked by its own consecutively numbered area of interest (see Fig. 6.1). Following the language codes proposed by ISO-639-2[6], the following abbreviations were used to label those areas of interest: **GerO** – *German original*, **GerMT** – *Machine Translation into German*, **CatO** – *Catalan Original* and **CatMT** – *Machine Translation into Catalan*. The (static) entry mask was labelled **Entry**. Moreover, these five categories allowed for a detailed analysis of the eye tracking data as it was thus possible to create subsets sorted by participants, by label, by participant and label or other indicators.

The aforementioned 21 eye tracking sessions resulted in video material of a total duration of 375 minutes, or 18 minutes on average per trial. Taking the interest area count as measure, the mean count of German text messages is 21 (SD = 9.60, range = 6–48), of machine translated messages into Catalan 20 (SD = 9.79, range = 6-48), of Catalan text messages 27 (SD = 10.85, range = 11-49) and of machine translated messages into German 26 (SD = 10.66, range = 11-49). A diverging number of original and MT messages can be explained by the Skype Translator's MT output that was for no obvious reason automatically merged into one text box even if two original messages were written.

As one can see in Fig. 6.2 and 6.3, most attention is paid to the lower third of the screen, right above the left area of the entry mask which is where new incoming messages and the MT output are displayed. The screenshots of one participant depicted here stand for every other test person as the saccadic eye movement patterns (Fig. 6.2) and fixation heat maps (Fig. 6.3) look similar. Moreover, there are some remarkably large saccades that even reach above the recognizable screen size (cf. Leube et al. 2017: 6). One explanation might be that the participants were

[6]https://www.bib-bvb.de/web/kkb-online/rda-sprachencode-nach-iso-639, last accessed on 4 November 2020.

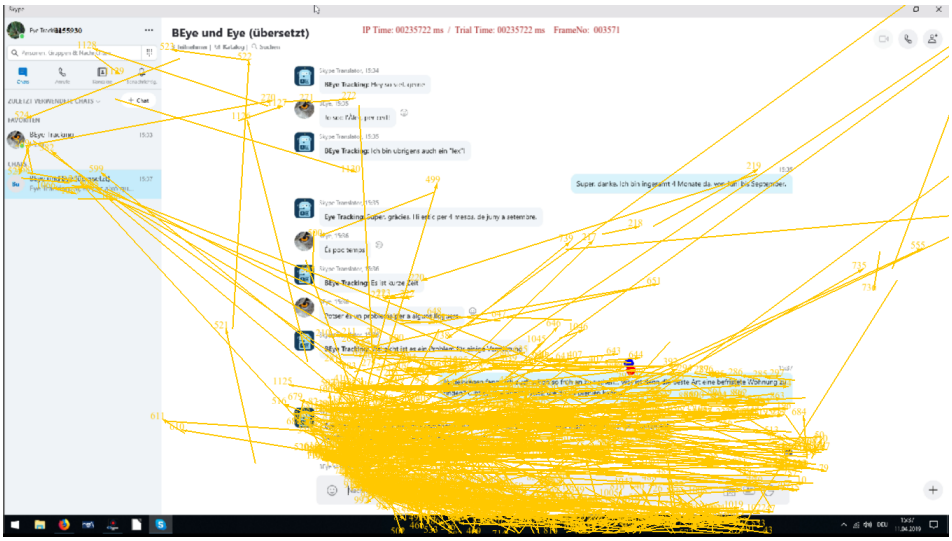

Figure 6.2: Saccadic patterns

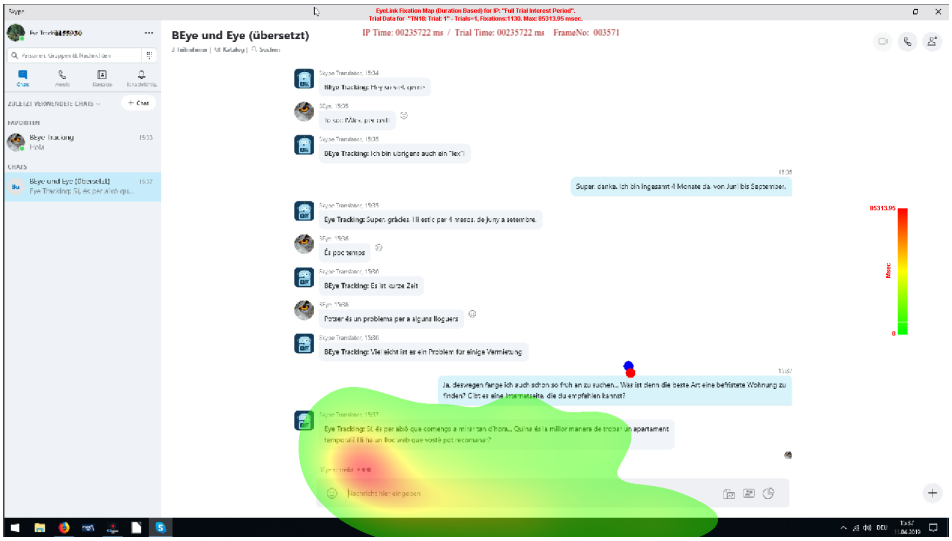

Figure 6.3: Fixation heatmap

distracted or thinking and therefore did not keep their eyes in the range covered by the eye tracking camera.

4 Results

As Holmqvist (2011: 321) points out, saccade amplitude and duration are closely related. Referencing Carpenter (1988), both parameters are correlated linearly, which can be investigated using the correlation test with Spearman's Rho, as in this present case study both amplitude and duration are not normally distributed. The correlation coefficient of Spearman's Rho is comprised between –1 and 1: –1 indicates a strong negative correlation, 0 means that there is no association between the two variables, 1 indicates a strong positive correlation[7]. The overall data set and the subsets by AOI tag turn out to be positively correlated: Global ($S = 2.9267e+11$, $p < 0.01$, $rho = 0.65$), GerO: ($S = 368854171$, $p < 0.01$, $rho = 0.76$), CatMT ($S = 3028797579$, $p < 0.01$, $rho = 0.64$), CatO ($S = 784453871$, $p < 0.01$, $rho = 0.7$), GerMT ($S = 1.0672e+10$, $p < 0.01$, $rho = 0.69$) and Entry ($S = 2144558389$, $p < 0.01$, $rho = 0.46$).

4.1 Saccade amplitude

Only saccades that start and end in one of the respective AOIs were taken into consideration. Furthermore, amplitude outliers greater than 2.5 times the standard deviation from the mean were excluded (SD = 1.43, range = 0.2–8.28). The remaining data set consisted of 1977 saccades with the label of GerO, 3627 of CatMT, 2453 of CatO, 5852 of GerMT and 3235 of Entry (see Table 6.1). That makes 17144 saccades in total.

Normal distribution of the saccade amplitude data was investigated using the Anderson-Darling-Test that can handle larger data sets than the commonly used Shapiro-Wilk-Test. As the AD-Test indicated a non-normal distribution of the overall data set ($A = 1253.1$, $p < 0.01$) and the subsets by AOI Tag (GerO ($A = 162.84$, $p < 0.01$), CatMT ($A = 305.34$, $p < 0.01$), CatO ($A = 168.12$, $p < 0.01$), GerMT ($A = 393.26$, $p < 0.01$), Entry ($A = 226.42$, $p < 0.01$)) and a logarithmic transformation did not change the data set's distribution towards normality, Kruskal-Wallis-Tests were performed to investigate the differences in the saccade amplitudes between participants and between AOI tags.

[7] http://www.sthda.com/english/wiki/correlation-test-between-two-variables-in-r, last accessed: 4 November 2020.

Table 6.1: Mean and SD of the saccade amplitude per AOI Tag

AOI tag	mean	SD
GerO	1.94	1.51
CatMT	1.82	1.47
CatO	1.90	1.42
GerMT	1.79	1.30
Entry	1.86	1.56
Global	1.84	1.43

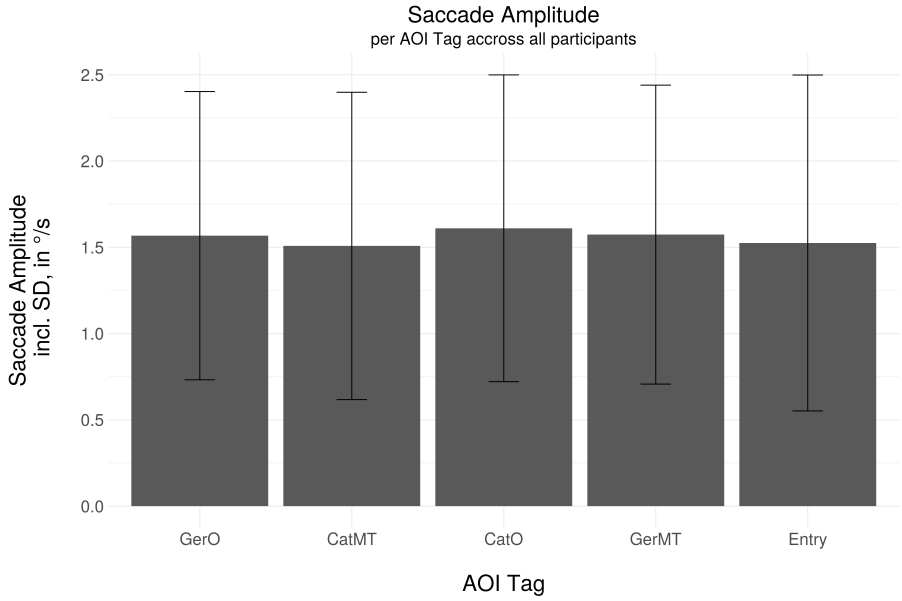

Figure 6.4: Mean saccade amplitude per AOI Tag in °/s

A Kruskal-Wallis-Test proves that there are significant differences in the saccade amplitudes between some of the participants ($\chi^2(20) = 286.86$, $p < 0.01$). A consequently performed post-hoc-test (Dunn-Benjamini-Hochberg) showed that 110 of 210 possible pairs (52.38 %) differ significantly. This is no surprise, as the amplitude varies from participant to participant, an observation that has already been stressed by Holmqvist (2011: 312).

A second Kruskal-Wallis-Test shows that there are significant differences of the saccade amplitude between the AOI tags ($\chi^2(4) = 56.56$, $p < 0.01$). A consequently performed post-hoc-test (Dunn-Benjamini-Hochberg; see Table 6.2, Asterisks indicate the significance level: * $\alpha < 0.05$) reveals that 8 of 10 possible pairs (80 %) differ significantly.

Table 6.2: Results of the Dunn-Test: Pairwise comparison of AOI tags for saccade amplitude

AOI tag pair	z-score	p-value adjusted
CatMT – CatO	-4.946410	0.0000*
CatMT – Entry	0.869816	0.2136
CatO – Entry	5.615444	0.0000*
CatMT – GerMT	-3.244902	0.0010*
CatO – GerMT	2.524951	0.0083*
Entry – GerMT	-4.090050	0.0000*
CatMT – GerO	-4.789648	0.0000*
CatO – GerO	-0.151941	0.4396
Entry – GerO	-5.427294	0.0000*
GerMT – GerO	-2.511196	0.0075*

4.2 Saccade duration

Only saccades that start and end in one of the respective AOI were taken into consideration. Furthermore, outliers greater than 250ms were excluded (SD = 37.40, range = 14–249). The remaining data set consisted of 2099 saccades with the label of GerO, 3700 of CatMT, 2491 of CatO, 5899 of GerMT and 2883 of Entry (see Table 6.3). That makes 17072 saccades in total.

Normal distribution of the saccade duration data was investigated using the Anderson-Darling-Test that can handle larger data sets than the commonly used Shapiro-Wilk-Test. As the AD-Test indicated a non-normal distribution of the

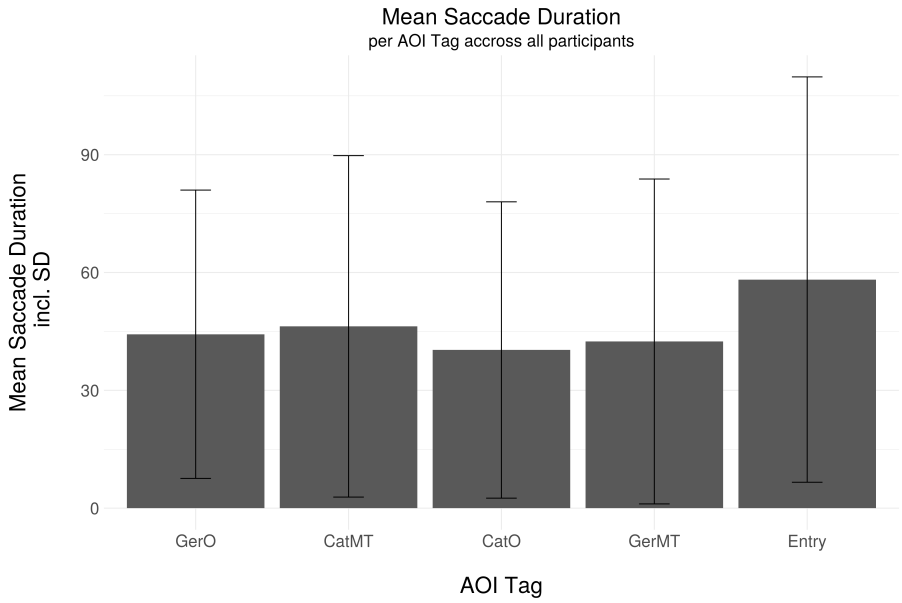

Figure 6.5: Mean saccade duration across all participants in ms

Table 6.3: Mean and SD Saccade duration per AOI Tag

AOI tag	mean	SD
GerO	31.19	31.64
CatMT	34.25	39.59
CatO	33.44	37.71
GerMT	30.60	33.97
Entry	38.12	43.76
Global	33.15	37.41

overall data set (A = 3442.2, p < 0.01) and the subsets by AOI Tag (GerO (A = 374.99, p < 0.01), - CatMT: (A = 749.75, p < 0.01) - CatO (A = 510.65, p < 0.01), - GerMT (A = 1246.5, p < 0.01), Entry: (A = 537.5, p < 0.01)) and a logarithmic transformation did not change the data set's distribution towards normality, Kruskal-Wallis-Tests were performed consequently to investigate the differences in the saccade duration between participants and between AOI tags.

A Kruskal-Wallis-Test reveals that there are significant differences in the saccade duration between some of the participants ($\chi^2(20) = 184.64$, $p < 0.01$). A consequently performed post-hoc-test (Dunn-Benjamini-Hochberg) showed that 84 of 210 possible pairs (40.0 %) differ significantly. As in the case of saccade amplitude, these differences are a natural by-participant phenomenon.

A second Kruskal-Wallis-Test shows that there are significant differences in the saccade duration between the AOI tags ($\chi^2(4) = 49.43$, $p < 0.01$). A consequently performed post-hoc-test (Dunn-Benjamini-Hochberg, see Table 6.4, Asterisks indicate the significance level: * $\alpha < 0.05$) indicates that 6 of 10 possible pairs (60 %) differ significantly.

Table 6.4: Results of the Dunn-Test: Pairwise comparison of AOI tags for saccade duration

AOI tag pair	z-score	p-value adjusted
CatMT – CatO	-1.543763	(0.0767)
CatMT – Entry	-4.700784	(0.0000)*
CatO – Entry	-2.806306	(0.0050)*
CatMT – GerMT	0.928215	(0.1963)
CatO – GerMT	2.489077	(0.0107)*
Entry – GerMT	5.995560	(0.0000)*
CatMT – GerO	-3.588229	(0.0004)*
CatO – GerO	-1.958909	(0.0358)
Entry – GerO	0.652672	(0.2570)
GerMT – GerO	-4.623838	(0.0000)*

5 Discussion

A look at both the saccade counts for amplitude and duration shows that there are nearly three times as many saccades on the MT output into German as on the German original. In other words, the fewest saccades were made on outgoing

messages of the participants compared to incoming texts regardless of language or MT. Taking the count as an indicator for reading depth, the MT into German is read by far more deeply than the German original messages.

As one can see in Tables 6.1 and 6.3 and Figures 6.4 and 6.5, mean saccade amplitude and duration by AOI tag is comparable to previous studies on reading tasks (cf. Rayner 1998: 373, Gangl et al. 2018, Nikolova et al. 2018), but the amplitude on the MT output into German is shorter than on the German original and the smallest number in general. The shorter the amplitude, the more closely the participants have read the respective AOI and vice versa. Given that the mean amplitude on both German and Catalan original messages and the entry mask is above average, it can be assumed that these AOI are read less attentively. Then again, the mean saccade duration on both German message types is below average, meaning that shorter saccades are made within these two regions compared to the Catalan messages that are above average. As for shorter saccade amplitude, a smaller duration value represents an increased reading depth and vice versa. This explanation adds up for both types of Catalan messages. Since the participants are not proficient in this language, it seems plausible that they read the messages only superficially. But when it comes to the lower duration value on German original messages, it is still questionable why these should be closely read.

Taking a closer look at the mean and SD of saccade amplitude and duration, the values reveal high dispersion in the data set. More precisely, both mean and SD values are close to each other. This can be additionally attributed to some reasons less desirable than the above mentioned. First of all, research based on naturalistic studies has to deal with by-participant variance. That is why statistical analyses in this field of research almost always have to deal with the variance that lies within the data set. Second, even the most accurate experimental set-up might miss one crucial variable which therefore deviates the results and has an impact on the interpretation. Third, a high error-rate can also be considered a reason for high dispersion in the data set. Saccades are the fastest movements the human body is capable of. Observing saccades requires therefore precise and accurate equipment. But even then, saccadic movements might go beyond the technical limits of this equipment, making them almost impossible to capture. Lastly, false-positive and false-negative results can also deviate the interpretation. The eye-tracker might detect saccades where none have been or vice versa.

Another strong reason might be the fact that the MT output is just error-prone and therefore requires deeper processing. These observations are supported visually by Fig. 6.2, which depicts saccadic eye movements during one session. Most saccades fall into the bottom third, left-aligned area of all machine-translated

and Catalan messages. This area covers the entry mask and the latest displayed messages. Only a few saccades are made above the lower third of the screen (on older messages). In addition to that, one single clear gaze path along the session info on the left upper corner of Skype can be identified (see Fig. 6.2). This can be taken as a hint that the participants seldom jump back to older messages but stick more or less to the most recent output of the text chat utterances on screen.

The pairwise calculated tests show that the German original messages cause significantly increased saccade amplitudes compared to every other AOI tag except the Catalan original (see Table 6.2). In contrast, significantly increased amplitudes in comparison to every other AOI tag except the German original can be observed on MT messages into German. The fact that the tests showed no significant results for the pairwise comparison of German original and Catalan original may lead to the conclusion that the participants are somehow noticing the incoming message in a language they are not proficient in. Given that German and Catalan share the same character system, participants may be switching between original and machine-translated utterance in search for words they can recognize. These can be names, numbers, words that share the same root in both languages or even words that can be deduced from another (Romance) language the participants are proficient in. As both pairs, German original vs. MT into Catalan and Catalan original vs. MT into German, turn out to differ significantly, this might be taken as a first indication for this hypothesis but will have to be explored in upcoming studies.

These observations may also be seen as indicators for the participants reading the German MT output more carefully due to typical MT errors in terms of syntax, semantics or orthography, which then results in shorter saccade amplitudes. The longer saccade amplitude on the German and Catalan original messages leads consecutively to the opposite assumption: the participants' reading behavior is less deep since they are already familiar to the German original. When it comes to the reason why longer saccades are made when reading the Catalan original, the participants might spend less care on reading a language they are not proficient in. One definitely would have to link the saccadic observations to their respective fixations to check for complete plausibility of this hypothesis.

When it comes to saccade duration, German original messages turn out to be significantly different compared to MT into German and into Catalan (see Table 6.4). Coming back to the observations of the relation between cognitive load and saccadic eye movements described by Holmqvist (2011: 313f.), reading (one's own) German original messages requires less cognitive capacity than processing incoming MT output into German, which in fact is new information to the participants and therefore definitely takes longer to process. In contrast, there is

no significant difference between German and Catalan original messages. Given that the participants have written the German original messages themselves and are already familiar to the information, the non-existence of any significant result might indicate that reading the Catalan original is as cognitively (non)demanding as is reading the German original. In a further step, this might be taken as a hint that the Catalan original is only read superficially. The same goes for the entry mask, where participants write, revise and send their messages. On the contrary, MT into German only differs significantly from the Catalan original and the Entry mask.

6 Conclusion

The present study tackled the question if saccadic eye movement patterns differ according to the type of text chat messages in reading tasks. Indeed, significant variance can be identified not only between participants – as is typical for saccade amplitude and duration – but also between incoming and outgoing or machine-translated and original messages. But it has to be stressed that the results represent nothing more than a first exploratory overview. It will be necessary to take a closer look at the interplay of saccade amplitude or duration and AOI size, blink patterns and the overall scan path. Further on, it has to be assumed that the limits of investigating saccadic eye movements in the context of technologies like the Skype Translator exist on the word level, as it is only marginally feasible to annotate single words with dynamic AOI on text chat message level. The text box and font size is too small, so that the impact of eye tracking indicators on (for example) orthographic information can only be deduced globally. Certainly, the evaluation of other commonly operationalised indicators in reading studies (fixation duration, dwell time, regressions, fixation count etc.) then has to be considered and linked to the findings on saccadic eye movements, too. It becomes clear that the exclusive investigation of saccadic eye movements in CMC studies therefore seems not to be enough for extracting valid results. Furthermore, the impression on the subjective quality of the communication as stated in the questionnaires at the beginning and end of each session has to be situated along the analysis.

The data set presented here is only half of the data that were collected during the overall project. The other half consists of eye-tracking data of monolingual text chats of seven German native participants via Skype. Both sets have to be linked to draw conclusions on the differences between monolingual and machine-translated text chat communication.

Nevertheless, it seems a quite promissing endeavour to observe how participants react to real time machine translation in text chats when they are not proficient in one of the involved languages. It helps to assess the requirements to better understand this type of communication technologies. In a more global context, it eventually contributes to an understanding of how all this shapes the way of communicating on the internet.

In the end, all similar endeavours have to be aware of the fast developing technology they are based on. The total count of languages featured by the Skype Translator has steadily increased since this project started. What is more, the layout of the Skype Translator changed as well, now only displaying messages in the operating system's language, leaving aside the two column comparable design this present study investigates. This fast changing environment can be taken as another argument to continuously investigate the human-machine-interaction in everyday life. This kind of technology has already penetrated every single aspect of our lives which is why it would be highly negligent to not evaluate the human behaviour when dealing with it.

Abbreviations

HTWK	Leipzig University of Applied Sciences
GerO	German Original
GerMT	Machine Translation into German
CatO	Catalan Original
CatMT	Machine Translation into Catalan

Acknowledgements

I am grateful to our institute's student assistant Tim Feldmüller who took care of the preparation of most collected research data.

References

Beißwenger, Michael. 2017. *Empirische Erforschung internetbasierter Kommunikation*. Berlin: De Gruyter.

Beißwenger, Michael. 2007. *Sprachhandlungskoordination in der Chat-Kommunikation* (Linguistik, Impulse & Tendenzen 26). Berlin: W. de Gruyter.

Bowker, Lynne & Jairo Buitrago Ciro. 2019. *Machine translation and global research: Towards improved machine translation literacy in the scholarly community*. OCLC: on1075580986. Bingley, UK: Emerald Publishing.

Carpenter, Roger H. S. 1988. *Movements of the eyes*. 2nd rev. & enlarged ed. London, England: Pion Limited.

Doherty, Stephen & Sharon O'Brien. 2014. Assessing the usability of raw machine translated output: A user-centered study using eye tracking. *International Journal of Human-Computer Interaction* 30(1). 40–51. DOI: 10.1080/10447318.2013.802199.

Doherty, Stephen, Sharon O'Brien & Michael Carl. 2010. Eye tracking as an MT evaluation technique. *Machine Translation* 24(1). 1–13. DOI: 10.1007/s10590-010-9070-9. http://link.springer.com/10.1007/s10590-010-9070-9 (22 July, 2020).

Duchowski, Andrew T. 2017. *Eye tracking methodology*. Cham: Springer International Publishing. DOI: 10.1007/978-3-319-57883-5. http://link.springer.com/10.1007/978-3-319-57883-5 (17 October, 2019).

Fišer, Darja & Michael Beißwenger (eds.). 2017. *Investigating computer-mediated communication: Corpus-based approaches to language in the digital world*. 1st edn. (Book series translation studies and applied Linguistics). Ljubljana: Ljubljana University Press. https://e-knjige.ff.uni-lj.si/ (27 October, 2017).

Gangl, Melanie, Kristina Moll, Manon W. Jones, Chiara Banfi, Gerd Schulte-Körne & Karin Landerl. 2018. Lexical reading in dysfluent readers of German. *Scientific Studies of Reading* 22(1). 24–40. DOI: 10.1080/10888438.2017.1339709.

Han, Lifeng. 2018. *Machine Translation Evaluation Resources and Methods: A Survey*. arXiv: 1605.04515. http://arxiv.org/abs/1605.04515 (2 September, 2020).

Holmqvist, Kenneth (ed.). 2011. *Eye tracking: A comprehensive guide to methods and measures*. OCLC: ocn741340045. Oxford ; New York: Oxford University Press.

Jacobson, J. Zachary & P. C. Dodwell. 1979. Saccadic eye movements during reading. *Brain and Language* 8(3). 303–314. DOI: 10.1016/0093-934X(79)90058-0.

Leube, Alexander, Katharina Rifai & Siegfried Wahl. 2017. Sampling rate influences saccade detection in mobile eye tracking of a reading task. *Journal of Eye Movement Research* 10(3). 1–11.

Nikolova, Mirela, Stephanie Jainta, Hazel I. Blythe & Simon P. Liversedge. 2018. Binocular advantages for parafoveal processing in reading. *Vision Research* 145. 56–63. DOI: 10.1016/j.visres.2018.02.005. https://linkinghub.elsevier.com/retrieve/pii/S0042698918300233 (30 April, 2020).

O'Brien, Sharon. 2009. Eye tracking in translation process research: Methodological challenges and solutions. *Methodology, technology and innovation in translation process research* 38. 251–266.

R Development Core Team. 2019. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org.

Ramlow, Markus. 2009. *Die maschinelle Simulierbarkeit des Humanübersetzens: Evaluation von Mensch-Maschine-Interaktion und der Translatqualität der Technik* (TransÜD 27). OCLC: 553597343. Berlin: Frank & Timme.

Rayner, Keith. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin* 124(3). 372–422. DOI: 10.1037/0033-2909.124.3.372.

Rayner, Keith. 2009. Eye movements and attention in reading, scene perception, and visual search. *Quarterly Journal of Experimental Psychology* 62(8). The 35th Sir Frederick Bartlett Lecture: 1457–1506. DOI: 10.1080/17470210902816461. (19 August, 2020).

Schaeffer, Moritz, Kevin B. Paterson, Victoria A. McGowan, Sarah J. White & Kirsten Malmkjær. 2017. Reading for translation. In Arnt Lykke Jakobsen & Bartolomé Mesa-Lao (eds.), *Translation in transition: Between cognition, computing and technology*, vol. 133 (Benjamins Translation Library), 17–53. Amsterdam: John Benjamins. DOI: 10.1075/btl.133. http://www.jbe-platform.com/content/books/9789027265371 (14 February, 2019).

SR Research Ltd. 2019. *EyeLink Data Viewer*. Mississauga, Ontario, Canada.

Storrer, Angelika. 2001. Sprachliche Besonderheiten getippter Gespräche: Sprecherwechsel und sprachliches Zeigen in der Chat-Kommunikation. In Michael Beißwenger (ed.), *Chat-Kommunikation. Spache, Interaktion, Sozialität & Identität in synchroner computervermittelter Kommunikation. Perspektiven auf ein interdisziplinäres Forschungsfeld.* 3–24. Stuttgart: ibidem.

Vardaro, Jennifer, Moritz Schaeffer & Silvia Hansen-Schirra. 2019. Translation quality and error recognition in professional neural machine translation post-editing. *Informatics* 6(3). 41. DOI: 10.3390/informatics6030041. https://www.mdpi.com/2227-9709/6/3/41 (20 March, 2020).

Verheijen, Lieke. 2017. WhatsApp with social media slang? Youth language use in Dutch written computer-mediated communication. In Darja Fišer & Michael Beißwenger (eds.), *Investigating computer-mediated communication: Corpus-based approaches to language in the digital world*, 1st edn., 72–101. Ljubljana: Ljubljana University Press. https://e-knjige.ff.uni-lj.si/ (27 October, 2017).