D1.3: Final report on user interface studies, cognitive and user modelling

Michael Carl, Mercedes García Martínez,
Robin Hill, Frank Keller, Bartolomé
Mesa-Lao, Moritz Schaeffer

Distribution: Public

CASMACAT
Cognitive Analysis and Statistical Methods
for Advanced Computer Aided Translation

ICT Project 287576 Deliverable D1.3

This project has received funding from the European Union's
Seventh Framework Programme for research,
technological development and demonstration
under grant agreement no 287576.

Project ref no.	ICT-287576
Project acronym	CASMACAT
Project full title	Cognitive Analysis and Statistical Methods for Advanced Computer Aided Translation
Instrument	STREP
Thematic Priority	ICT-2011.4.2 Language Technologies
Start date / duration	01 November 2011 / 36 Months

Distribution	Public
Contractual date of delivery	October 31, 2014
Actual date of delivery	November 5, 2014
Date of last update	November 5, 2014
Deliverable number	D1.3
Deliverable title	Final report on user interface studies, cognitive and user modelling
Type	Report
Status & version	Draft
Number of pages	76
Contributing WP(s)	WP1
WP / Task responsible	CBS, UEDIN
Other contributors	
Internal reviewer	Daniel Ortiz
Author(s)	Michael Carl, Mercedes García Martínez, Robin Hill, Frank Keller, Bartolomé Mesa-Lao, Moritz Schaeffer
EC project officer	Aleksandra Wesolowska
Keywords	

The partners in CASMACAT are:

University of Edinburgh (UEDIN)
Copenhagen Business School (CBS)
Universitat Politècnica de València (UPVLC)
Celer Soluciones (CS)

For copies of reports, updates on project activities and other CASMACAT related information, contact:

The CASMACAT Project Co-ordinator
Philipp Koehn, University of Edinburgh
10 Crichton Street, Edinburgh, EH8 9AB, United Kingdom
pkoehn@inf.ed.ac.uk
Phone +44 (131) 650-8287 - Fax +44 (131) 650-6626

Copies of reports and other material can also be accessed via the project's homepage: http://www.casmacat.eu/

Executive Summary

D1.3 marks the final CASMACAT report on user interface studies, cognitive and user modelling covering the completion of tasks T1.5 (Cognitive Modelling) and T1.6 (User Modelling) as part of Work Package 1.

Within tasks T1.1 to T1.4, a series of experiments have established a solid understanding of human behaviour in computer-aided translation, focusing on the use of visualization options, different translation modalities, individual differences in translation production, translator types and translation/post-editing styles. Additionally, the bulk of this experimental data has been released as a publicly available database under a creative common license and further details on this can be found in D1.4. In parallel to these more holistic studies, a second set of experiments aimed to examine some of these factors in a constrained laboratory setting. These focused on the underlying psycholinguistic processing and cognitive modelling of translators' activity to capture reading difficulty, verification and perplexity during translation and post-editing.

This deliverable combines these earlier empirical findings with experiments conducted in Year 3 of the project and grounds translation within a broader theoretical framework associated with human sentence processing and communication. As well as broadening our general understanding of bilingual cognitive processing, there were two major objectives behind the experimental investigations in Year 3. The first was to evaluate the utility of providing translators with Source-Target word alignment information through spatially-direct visual cues. The second was to determine what, if any, differences arise from expertise by comparing the results between a group of bilinguals and a group of professionally trained translators on the same translation-related tasks.

Contents

1 Introduction

This deliverable reports on cognitive and user modelling. It provides an overview over the various user interface studies with CASMACAT (2011–2014) and attempts to integrate the findings in a general relevance theoretical framework.

Relevance theory (RT) is a cognitive theory of communication which postulates an interpretive use of language, aiming at explaining how hearers arrive at the interpretations they construct, with minimum cognitive effort and maximum relevance. RT provided evidence that verbal communication always has an inferential element, and can be analysed in terms of a speaker's informative and communicative intentions.

Gutt (1990) states that translation can be explained in relevance theoretical terms. In many cases, translation is just a semiotic transposition, not fundamentally different from paraphrase and therefore does not require a proper translation theory. According to Kliffer & Stroinska (2013), RT has shown promise in explaining numerous pragmatic phenomena, such as translation, and may prove to be the most reliable tool for handling the interpretive richness of real-life data.

This deliverable discusses CASMACAT experiments in an RT framework. In section 2.1, we first present a noisy channel model of translation and post-editing which is then extended with relevance theoretical concepts. Section 2.2 provides more details on the RT framework and its relation to translation, post-editing and the investigations conducted in the context of CASMACAT.

Two different approaches to data collections were undertaken: *Lab experiments* aiming at isolating a single experimental question at a time by putting translators in a tightly controlled environment, in which only the factors under consideration were varied; and *Field studies* using the CASMACAT workbench in a real world translation environment (Celer Soluciones SL premises). This enables a wider range of questions to be addressed in a more realistic way, at the expense of being able to fully control all possible confounding factors.

After the introductory Section 2 on RT, Section 3 describes the impact of post-editing on variation during translation production. It shows that post-editors are heavily primed by the MT output, which results in a loss of variance in the final translation product. Section 4 investigates to what extent alignment visualizations can be helpful for post-editing purposes both in a lab setting and in the CASMACAT workbench. We present a series of lab experiments that establish when alignments are useful for error detection in translation output. These results led to a range of alignment visualization options being integrated into the CASMACAT workbench, which were then tested using a series of field studies. These are also reported in section 4. Section 5 reports on the results of our lab experiments with respect to user modelling; in particular, we discuss differences in the use of alignment visualization by naive bilinguals vs. professional translators. This leads us to a large-scale user-modelling study in which we show that user profiles can be predicted from translation process data using machine learning approaches. We cluster and identify translators based on their activity patterns and show that those different clusters correspond to different learning approaches. Section 6 compares post editing and interactive translation prediction settings of the CASMACAT workbench in an RT context. This section introduces a distinction between conceptual and procedural encodings and investigates to what extent interactive translation prediction can facilitate conceptual encodings. Finally, in section 7 we provide an overview of the basic configurations of the CASMACAT workbench that were tested in experimental and field studies. We conclude that different configurations are better suited to different translator profiles.

2 Theoretical framework

2.1 Noisy channel model

Recent accounts of human sentence processing have given a central role to error-correction mechanisms (e.g., Levy, 2008), formalizing sentence processing using a noisy channel model, in which an intended sentence s_i is encoded as the perceived sentence s_p, a process that is subject to noise. The goal of the comprehender is then to decode s_i by combining $P(s_i)$, the prior probability of the intended sentence, and $P(s_p|s_i)$, the likelihood that s_i will be distorted as s_p by noise (Gibson et al., 2013). This model is directly applicable to translation, if we assume that instead of the intended sentence s_i, we have the source text,

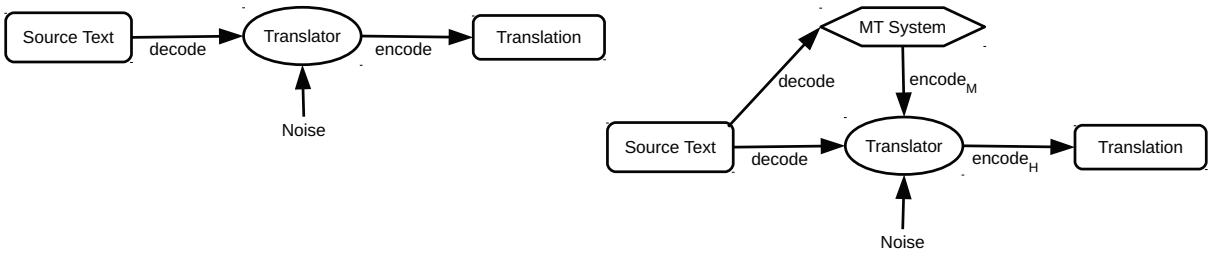


Figure 1: The noisy channel model of translation: translation from scratch (left panel); post-editing of MT output (right panel)

and instead of the perceived sentence s_p, we have the translation. The decoding and encoding process is then mediated by the translator, who is modelled as a source of noise (Figure 1, left panel).

In the context of post-editing, we have a machine translation system in the loop, which is an additional source of noise, as it uses separate encoding and decoding processes to arrive at the translation (Figure 1, right panel).

Previous work tested the noisy channel model indirectly using reading and judgment tasks (Levy, 2008; Gibson et al., 2013). A previous CASMACAT study (Hill & Keller, 2014) investigated its predictions directly by requiring participants to detect and localize errors. In an eye-tracking experiment, two groups of participants (native speakers of English, and 20 highly proficient non-native speakers) were eye-tracked while reading 120 sentences, each of which they saw in one of two versions (error condition or non-error condition). The experimental materials were drawn from the output an MT system, but only the target text was presented.

The results showed significant differences in error detection rates across error types, which confirms the prediction of the noisy channel model that error types differ in their distortion probability $P(s_p|s_i)$. Native and non-native speakers did not differ significantly in error detection rates or in total sentence reading times. However, an analysis of the reading times on the target word (which either contains an error or not) revealed differences between native and non-native speakers. First fixation times on the target word were sensitive to the presence of an error in native speakers, but not in non-native speakers. This indicates that non-native speakers have a less reliable language model $P(s_i)$, which slows down error detection.

The noisy channel model allows us to model translator type (e.g., native vs. non-native speakers, or naive bilinguals vs. experts translators), as well as making it possible to derive detailed predictions as to the effect of different error types (we will return to this in Sections 4.1 and 5.1). However it falls short to explain when and why translations become non-literal, when which external resources are used in the translation process, and how usage of translation assistance is learned.

Relevance Theory (RT) has the potential to fill this gap, by adding additional relevance theoretical constraints to the translation process, which are outlined in the next section.

2.2 Relevance theory

Originally developed as a theory of communication, relevance theory (RT) Wilson (1994) has been used to explain translation practice. RT argues that the hearer/reader/audience will search for meaning in any given communication situation. Human cognition tends to be geared to the maximisation of relevance, so that when the meaning fits their expectation of relevance, the hearer/reader/audience will stop processing (Sperber & Wilson, 1995, 260).

RT is based on two relevance principles: a *Cognitive Principle*, i.e. human cognition is geared to the maximisation of relevance, and a *Communicative Principle*, i.e. utterances create expectations of optimal relevance. The central claim of RT is that the expectations of relevance raised by an utterance are precise and predictable enough to guide the hearer towards the speaker's meaning.

Gutt (1990, 1991) applies RT to translation. He assumes that translation is an instance of human communication which involves three parts: i) the source text author, ii) the translator and iii) the target text reader. The goal of translation is to achieve adequate contextual effects for the target text reader without unnecessary processing effort, and where the translation resembles the original in relevant as-

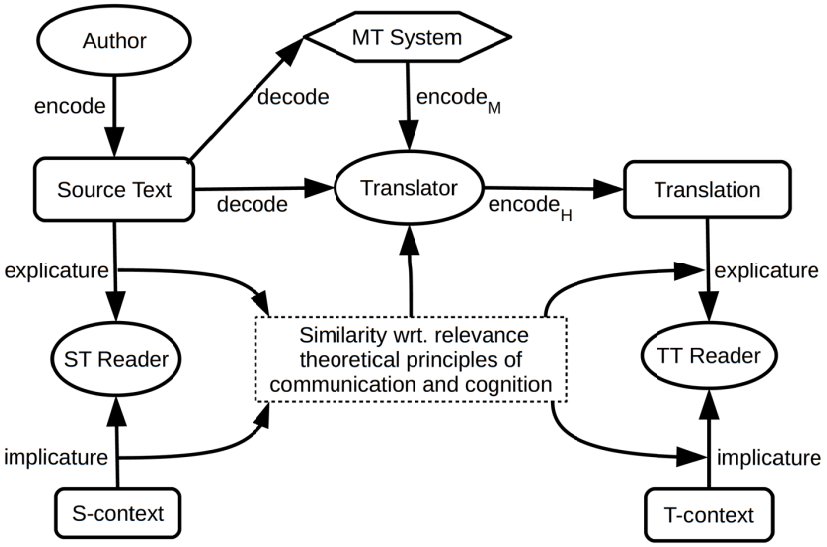


Figure 2: An integrated model of post-editing

pects. We maintain that RT complements the noisy channel model by adding constraints of causal interrelation between stimulus, context and interpretation, established by the principle of relevance, as depicted in Figure 2.

Gutt (1990) makes a distinction between *direct translation* and *indirect translation*. The former occurs if the source context (S-context) and the target context (T-context) are the same. Indirect translation, in contrast, occurs when only a subset of the S-context are represented or recoverable in the T-context. Gutt also points out that there is no clear boundary between these two modes of translation. The interpretations of translations will always be different from the original interpretations to the extent in which the target language context and the source language context make complete interpretive resemblance impossible. Nevertheless, translators (and post-editors) will seek to achieve optimum relevance in situations where the translation cannot preserve all analytic and contextual implications. Gutt points out that unnatural and even ungrammatical target language structures are acceptable, if the complications for the target audience are outweighed by an increase of relevance with respect to the intended interpretation.

Figure 2 shows a diagram an integrated model of post-editing: an author encodes a text within a cultural context of the source languge (S-context). A source text reader (ST Reader) understands this text by decoding the literal meaning (explicature) and by inferential processes over the intended meaning (implicature). The job of a translator is to encode the source text into a target language translation, constrained by relevance theoretical principles and the similarity between the explicatures and implicatures of the source text and those anticipated in the target language context. Crucially, both the author and the TT reader draw inferences based on context external to the text, but these contexts might not necessarily be the same.

The integrated model of post-editing in Figure 2 extends beyond a generative noisy-channel model by de-composing the 'noise' in Figure 1 into sets of additional relevance theoretical features. Successive sections in this deliverable describe several parameters which may likely play an important role in this post-editing model. The values of the parameters of such statistical models are usually obtained by means of a maximum-likelihood estimation method. However, the full implementation of a formalized post-editing model reaches far beyond the work within CASMACAT. In this deliverable we also present a number of parameters that are likely to play an important role in such an integrated model of post-editing.

In line with RT assumptions, we will show in section 3 that post-editing reduces translation variance

and tend to be more often *direct translation* than *indirect translation*. MT systems do not have models of optimal communicative relevance, and they do not compute intentions and implications of the source text or the target audience. That is, the communicative model of MT systems lack the ability of inference and implication all together, and hence reduce to a noisy channel model, as they are.

As Sperber & Wilson (1995, 3) point out "communication can be achieved by coding and decoding messages, and it can be achieved by providing evidence for an intended inference. The code model and the inferential model are each adequate to a different mode of communication". However, the principle of relevance can in no case be overridden, which may make it difficult for an MT system to produce valuable help for translators, if the source sentence does not fully encode the propositions independently of the context.

As discussed in D6.3, and in sections 5 and 6 below, post-editing of MT output and, in particular, post-editing with interactive translation prediction (ITP) adds to the S-context and the T-context while also constantly changing the MT system output that needs to be considered by the post-editor. We will show that different post-editors tackle this situation differently.

3 Translation priming

Priming is the response to a stimulus which is biased by a preceding stimulus. Priming is an unconscious effect which relies on the implicit memory of the previous stimulus. A number of studies have shown that the bilingual mind, and thus also human translation, is based on implicit memory effects. Hartsuiker et al. (2004) assume that entries in the bilingual mental lexicon consist of "combinatorial nodes" which connect lemmas, concepts, word forms and syntactic information, irrespective of language. These combinatorial nodes are responsible for and explain priming effects in translation. In line with numerous studies, Schaeffer & Carl (2013) argue that priming in translation extends to lexical, syntactic and semantic levels of description. It has been shown in many instances that priming works also between modalities, but best when the two stimuli are in the same modality. While a source text primes the translation into another language, we expect priming effects to be even bigger for post-editing. In post-editing the MT output effective acts as a priming stimulus which enables the post-editor to produce translations more quickly.

In order to test and quantify these hypotheses and assess their impact on the translation product, we utilise a corpus of alternative translations to determine the extent different translators and post-editors translate words and structures in the same way. As a translator reacts to the stimulus of the source sentence (or the MT output in the case of post-editing) when producing a translation, we can assess priming effects by measuring the amount of "similarity" between the source and the target texts. Carl and Schaeffer (forthcoming) introduce a quantifiable measure of *translation literality*, which measures the similarity and which is defined by the following three criteria:

1. Word order is identical in the source text (ST) and target text (TT)
2. ST and TT items correspond one-to-one
3. Each ST word has only one possible translated form in a given context

Assume a text is translated by, for instance, 20 different translators: if all 20 translators produce the same translation for a word in the context of the same text, then we can say that this word is translated literally. However, the more often the word is translated differently by the different translators, and the more the word order in the translation changes compared to the source text, the less literal is the translation.

In the following sections we will show how MT output primes the post-editor through lexical priming (section 3.1).

3.1 Lexical translation priming

Interference phenomena from the source language, such as grammatical or lexical structures of the source language that carry over into the target language have frequently been observed, in translation as

well as in post-editing of machine translations.

For example, Čulo et al. (2014) report that in the context of the English-to-German translation "In a gesture sure to rattle the Chinese Government" → "In einer Geste, die die chinesische Regierung wachrüttelt" the German expression "In einer Geste" is understandable, but literal and unidiomatic. It is a one-to-one literal translation, produced by the MT system, which was often not amended during post-editing. However, during from-scratch translations more idiomatic expressions, such as "Als Geste", "Es ist eine Geste", "Mit der Absicht", "Als Zeichen des Widerstandes", "Mit einer Aktion" would be produced. That is, post-editors are heavily primed by the MT output and accept unidiomatic expressions, while they would produce more translation variations when translating from-scratch. The priming effect in post-editors results thus often in sub-optimal translations which translators working from scratch would otherwise not produce.

An analysis of translations from English source texts into German and Spanish confirm this hypothesis. The word translation perplexity[1] values in Figure 3 are based on eight from-scratch translated versions and eight post-edited versions of a set of English source texts into German (left and Spanish (right). These texts are taken from the SG12 and the BML12 study (see D1.4, table 1) amounting to approximately 800 source text words. Table 3 shows that post-editing machine translation results in a more literal translation than from-scratch translation.

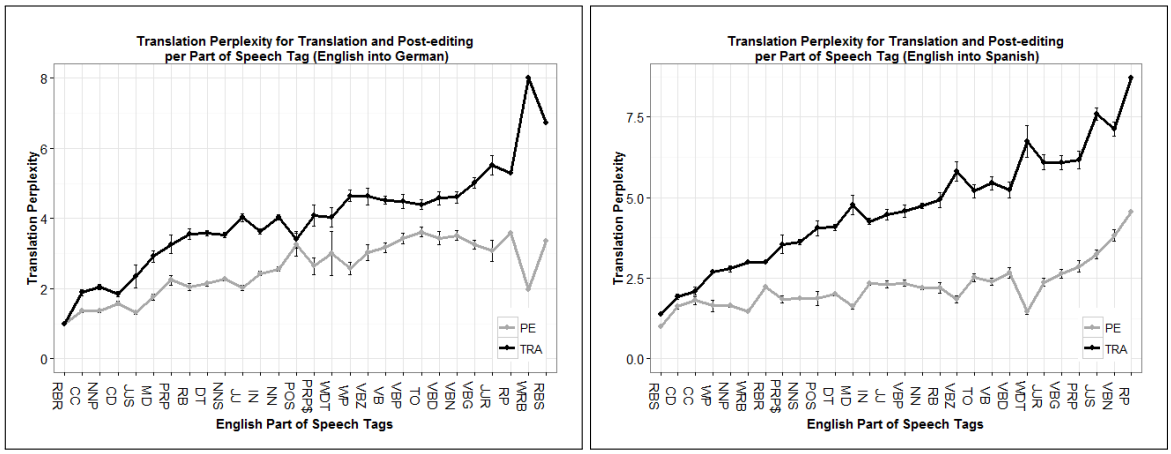


Figure 3: Priming in translation (TRA) and in post-editing (PE). Translation perplexity indicates the variation of produced translation, which is always higher in from-scratch translation than in post-editing.

The graphs plot perplexity values of English-to-German and English-to-Spanish translations for different part-of speech values, and compare variance in translation realization in from-scratch translation and in post-editing. Some PoS tags, JJS (superlative adjective), NNP (Proper names), CC (conjunctions) produce very few translation alternatives, which are during post-editing almost always accepted. Other PoS tags, such as RP (particle), VBN (participle), produce more variants in the target language.

Some PoS tags in the Spanish translations (right) show more variation in the target text, as compared to the German ones on the left. For instance superlative adjectives (JJS) are almost never touched during post-editing of the German translations while there are many translation alternatives in Spanish. Note that the difference between the post-edited and from-scratch translated texts is bigger for Spanish than for German, indicating that Spanish post-editors more easily accept the Spanish MT output than the German post-editors do. The perplexity of the post-edited texts are in all cases smaller than in the from-scratch translated versions.

It seems that, as Gutt points out, "unnatural and even ungrammatical target language structures are acceptable", particularly for post-edited texts. Maybe in this case not "if the complications for the target audience are outweighed by an increase of relevance with respect to the intended interpretation", but rather if a higher productivity rate can achieved and the costs of translations are reduced.

[1] for a full discussion of translation perplexity, see Appendix A.2.

Study	Session	TL	Task	Texts	Part	Fdur	Kdur	Tlen
TDA14	48	en	C	1-6	8	3,60	3,4924	6792
KTHJ08	69	da	T	1-3	24	6,45	5,4536	10571
SG12	45	de	P	1-6	23	5,6 0	1,9265	6352
SG12	47	de	T	1-6	24	9,39	4,6145	6632
NJ12	39	hi	T	1-6	20	13,04	7,4243	5505
BML12	64	es	P	1-6	32	2,31	0,8774	9012
BML12	63	es	T	1-6	32	8,20	5,7491	8936
Total	375	5	3	6	107	48,59	29,5378	53800

Table 1: Annotated data for the study on syntactic entropy

3.2 Syntactic translation priming

As we have seen in the previous section, priming reduces the degree of variation between different translators, as different translators are more likely to choose the same lexical item. In this section, we will investigate whether this results also holds for sentence structure. The degree of variation in structure can be measured by means of entropy: a more equal distribution of different translation realizations leads to higher entropy values while an uneven distribution or even a distinct solution produced by each translator leads to low entropy values. From this it follows that entropy is correlated with priming effects.

Priming is associated with a facilitation effect in translation. This effect has been shown to exist for lexical choice and for word order. For instance, Jensen et al. (2009) report shorter total reading time during translation if the word order of the source sentence can be maintained in the target sentence. We thus conjecture that facilitation is correlated with translation entropy. This conclusion is in line with Campbell's Choice Network Analysis Campbell (2000): the more choices and the more complex choices a translator has to consider, the more effortful is the translation of a particular item. In addition, easier translation choices are more often lead to identical realizations, while more difficult choices tend to lead to different translation realizations.

In a CASMACAT study we investigated whether priming effects can also be observed on the level of syntactic structure. For this end we manually annotated shallow syntactic parses of six English source texts and their translations into Danish (da), German (de), Spanish (es) and Hindi (hi). Table 1 gives an overview over the size of the annotated texts [2]. All together these include 107 translations with 53,800 words. Each phrase was manually annotated with triplets indicating the type of clause (main, sub), voice (active, passive), and valency (transitive, intransitive) of that clause. Source text segments and their translated target text segments were annotated using the same schema. The idea was to investigate the similarity and differences in the source and the target language structure. As in the lexical priming study (see section 3.1) we compute the syntactic entropy for a source segments based on the observed variation in the translation. A low syntactic entropy would mean that all translators produced the same target syntax, while a high syntactic entropy implies that translators produced different syntactic target language realizations.

We found a positive correlation between syntactic entropy and total reading time on source text words: on average, translators spent more time reading sentences when they had higher syntactic entropy and less time when they had lower syntactic entropy. This correlation was significant for Danish ($p < 0.01$), for Spanish ($p < 0.01$) and for Hindi ($p < 0.05$). In addition, we found that if translators maintain the same structure as in the source (adherence), most translators use the same target structure. When translators produced a target sentence with a different structure from the one of the source segment (deviation), different translators tended to use different target structures. In other words, for target sentences which had the same structure as the source sentence, syntactic entropy was significantly lower as compared to the target sentences which had a syntactic structure which was different from the source structure. This effect was significant for all languages ($p < 0.01$). Low syntactic variation of translations is likely a result of syntactic priming, influencing the translators to reproduce the same syntactic structure

[2] See deliverable D6.3 section 2.4 for a definition of the various durations, Fdur, Kdur, Pdur

that they have read before. This work has led to a number of findings which are fully described in Appendix A.2.

3.3 MT search graph entropy

In analogy with the perplexity of human translations, we also conducted a study (Carl and Schaeffer, forthcoming) to investigate the impact of the perplexity in the MT search graph on post-editing behaviour and translation choices. State-of-the-art MT systems encode possible translations in a search graph, which consists of nodes that represent target language words (and phrases) and transitions between successive nodes. Transitions are labeled with weights, which represent the cumulative costs for the translation. The task of a decoder is to find an optimum path through the search graph, which, hopefully, corresponds to the best translation in the search graph.

Weights in search graphs are trained on translations and texts that are produced by human translators. It can, thus, be expected that the entropy in the search graph transitions also reflect the entropy of the texts and translations with which the system was trained. Thus, a low entropy of search graph transitions would indicate the relative agreement between translators as how to translate the word or sequence in question, while high entropy would signal more difficult and controversial translations.

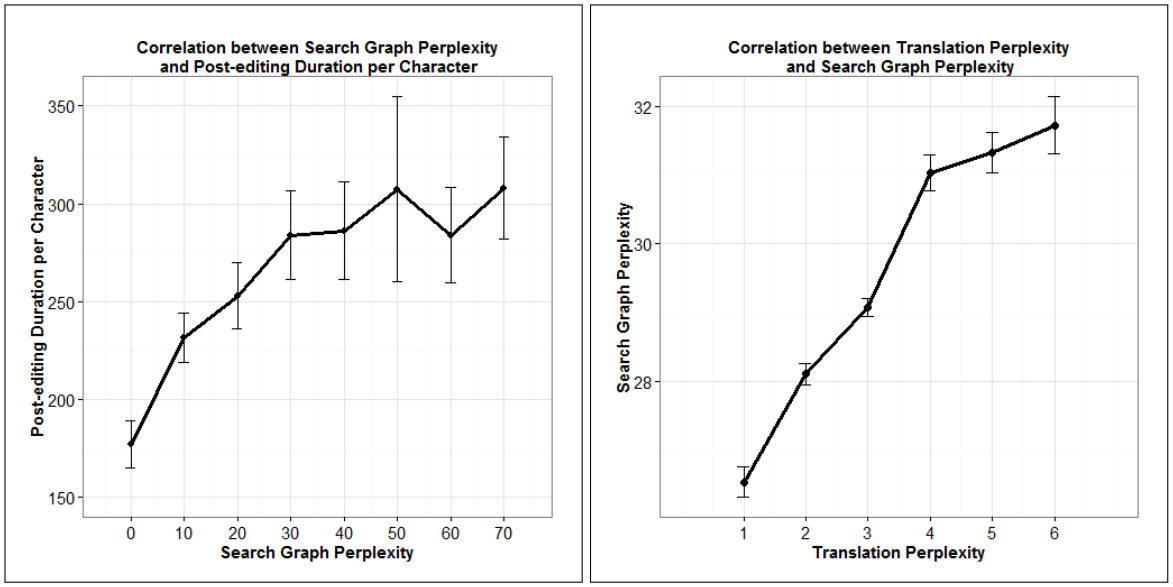


Figure 4: Correlation between word perplexity of the MT search graph and post-editing duration per character (left). Correlation between Word translation perplexity in the post-edited output and word perplexity of the MT search graph (right).

Our findings are summarized in Figure 4. Figure 4 (left) shows that post-editing duration is, indeed, strongly correlated with the perplexity values in the MT search graph: as perplexity values (i.e. number of similarly possible translation alternatives) in the MT search increase, so does the post-editing time of the produced MT suggestions. Note that post-editing duration can be seen as an indicator for the MT quality (e.g.Aziz et al. (2014)), so that we conjecture that search graph perplexity correlates with translation quality.

Figure 4 (right) shows the correlation between MT search graph perplexity and perplexity of post-edited translations. That is, there is a transitive relation of translation ambiguities in the material with which MT systems are trained via the perplexity of the search graph when the MT system produced a translation, to the finally post-edited translations.

3.4 Conclusions

Language translation is an enormously challenging task that demands unique skills of a translator. Besides bridging the linguistic aspects of the two languages, such as syntactic divergences and lexical

choices, translators have to balance the author intentions against the reader expectations, while interpreting the socio-cultural aspects of the original and the translation. In this section we have argued that this task is grounded in priming: the implicit memory of the source text segment primes the translator to produce a translation which is structurally and lexically similar to the target text. A post-editor is primed by two stimuli: the source text and the MT output. In line with the relevance theoretical assumption, post-editors accept MT suggestions even when the produced target text becomes unidiomatic or even ungrammatical, if the communicative purpose is still communicated. While PE reduces variance in the produced target texts, we also show that MT output cannot help in situations where translation entropy surpasses a certain threshold, i.e. if the translation becomes (too) context dependent, or if the source text structure cannot be easily mapped into the target language. However in CASMACAT, and in particular in the next section, we investigate whether some of the 'hidden' resources of MT systems can be instrumentalized to notify the translator about the relations in the MT output and their possible reliability.

4 Alignment visualization

4.1 Experimental results on error detection and visualization

As a continuation of the research agenda and laboratory studies described in deliverables D1.1 and D1.2, three further experiments were conducted. In the interest of continuity, these will be referred to here as Experiments 3, 4 and 5. Whereas the earlier studies had monolinguals and bilinguals examining only the translated output, the experiments reported here involve presenting the original German source text in addition to the translation. All participants were fluent German-English speakers; for the final experiment they were also professional translators.

The main discussion of expertise, along with the experimental findings of similarities and differences between fluent bilinguals and professional translators, is covered in Section 5 (User Modeling).

4.1.1 The verification task

The underlying assumption of the post-editing process is that translations can contain noise, mistakes and more subtle deviations from native-sounding language. The objective of the post-editor is therefore to reduce the level of noise, correct any errors and minimise the impact of anything being "lost in translation". However, this obviously relies on the post-editor being able to identify these problems in the first place, and remarkably little is known about the cognitive processes involved in this. The experiments described here required participants to determine whether there was a mistake in a translated sentence and to identify the problem word(s). Eye-movements were monitored while participants read to provide insight into the time-course and stages of dynamic processing that are involved in making post-editing decisions.

Five categories of errors were successfully classified in the earlier "translated output only" experiments and were subsequently used in the bilingual studies. These covered lexical, syntactic and semantic errors. Therefore while some classes of errors can be objectively determined, other classes can be considered more subjective and may be influenced by individual differences / preferences (translator style or profile). Examples are listed below, along with the original German version.

TE) Transposition (easy). This was considered the easiest error to spot and was essentially to provide a baseline for comparing the other error types against.

(1) a. We will not ovte [vote] in favour of new measures.
b. [Wir werden nicht zugunsten neuer Maßnahmen stimmen.]

TD) Transposition (difficult). While also a letter transposition, these examples involved two internal letters being switched (harder to spot than incorrect characters at the beginning or end of words) that produced an incorrect but legitimate word (and would therefore pass a spell-check or cursory examination).

(2) a. The most common cause of martial [marital] breakdown is a failure to communicate honestly.
b. [Die häufigste Ursache eines ehelichen Zerwürfnisses ist ein Mangel an ehrlicher Kommunikation.]

WO) Word order. Again, this was a transposition error, but at the word level rather than letter level. To minimise deviation in distance or alignment order, it was always the case that two adjacent words were switched rather than a random order reassignment (word salad).

(3) a. The brochure described the mountain as a park huge [huge park] in Barcelona city.
b. [Die Broschüre beschrieb den Berg als einen riesigen Park in Barcelonas Innenstadt.]

MT) Mistranslation of tense or agreement. These sentences contain a within-sentence violation in verb tense or a mismatch in gender or number agreement.

(4) a. Many of our friend [friends] are surfers and I have a great friend who lives in Tamarindo.
b. [Viele unserer Freund sind Surfer und ich habe einen großartigen Freund, der in Tamarindo lebt.]

(5) a. The cuts were [would] ultimately hit the combat troops.
b. [Die Kürzungen wurden letztendlich die Kampftruppen treffen.]

ML) Mistranslated lexical item. In these sentences, the critical words were related or semantically connected to the correct word, but contextually odd or inappropriate.

(6) a. Judge Torkjel Nesheim canceled [interrupted] Breivik during his monologue.
b. [Richter Torkjel Nesheim unterbrach Breivik während diesem Monolog.]

4.1.2 Procedure

Participants are presented with two sentences simultaneously in parallel on a computer screen. The first sentence is the original German source sentence, while the second is the English translated version. The participants are instructed to ensure they fully read the translation and to then decide if it contains an error by clicking the left mouse button for "yes" or the right mouse button for "no". Following a "yes" decision, the sentence is redisplayed on the screen and the task is then to click on the first word of any error.

4.1.3 Stimuli

Similar to the translation-only experiments, original source materials and translations were drawn from entries to the German-to-English WMT12 shared task competition (part of the EuroMatrixPlus project).[3]

24 pairs of source and translated sentences were produced for each of the five error conditions. There were two versions of the translations: one correct version and one containing an error. Two lists of stimuli were constructed: the first list contained the first half of the sentences for each condition containing errors while the second list contained the second half. Each set therefore contained 12 examples of each error type along with another 12 examples that did not contain a mistake, totalling 60 versions with errors and 60 without. As before, to ensure strict experimental control, each sentence contained a maximum of one error, word frequency was controlled and the position of the error systematically varied across the set of stimuli. Both lists also included the same 60 filler sentence pairs (fluent, error-free source and translation sentences), extracted from a similar corpus to that used for the error conditions. Following four practice items, the presentation order of the 180 sentence pairs was randomised for each participant. Half of the participants were tested using List One, while the other half saw List Two.

For Experiments 4 and 5, there was also an alternative version of the stimuli that contained a visual cue of word alignment between Source and Target sentences. For every sentence pair (irrespective of

[3] http://matrix.statmt.org/matrix/output/1692?run_id=2517]

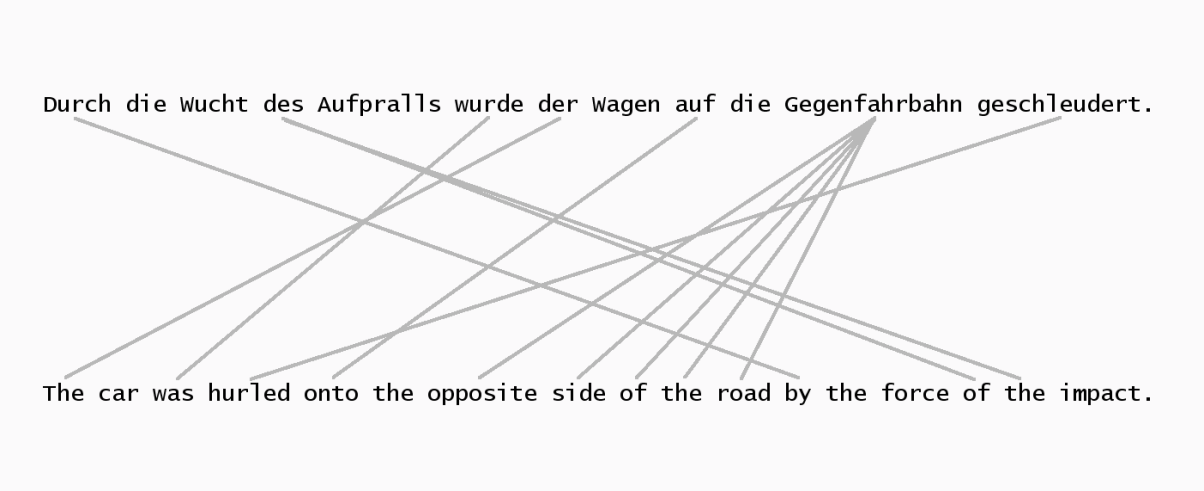


Figure 5: Example of visual alignment between Source and Target sentences

whether there was an error or not) where the cross-values between each German word and its translation was greater than one (see Carl & Schaeffer (submitted), attached Appendix) the alignment was explicitly indicated. In order to determine whether alignment information would facilitate checking the translations, a very direct static alignment link was established using clear lines between word centres. However, the lines remained in the whitespace between the sentences so that they would not impinge on the normal reading of the sentences. This non-interactive visual alignment was specifically chosen to establish the potential of this information by avoiding any usability complications associated with the interactive implementation of visual alignment in the CASMACAT workbench. An example is shown in Figure 5.

The sentences were displayed on a 22-inch widescreen monitor with the Source Text near the top of the screen and the Target Text on the centre line. Eye movements were recorded using an SR Research EyeLink 2K (binocular recording: 1KHz sample rate per eye).

Consistent with Experiments 1 and 2, a pre-experiment questionnaire covered demographic details and linguistic background, as well as familiarity with machine translation (e.g. Google Translate) and personal usage.

4.1.4 Experimental details

Participants had access to the source and the target text in all experiments. In Experiment 4 (but not in Experiment 3) they also saw lines aligning the source and the target words. Experiments 3 and 4 used naïve translators (non-experts). Experiment 5 used professional translators, who saw lines in half of the sentences, and no lines in the other half. Each experiment involved 20 participants: 40 German-English bilinguals in total and 20 German-English translators. The bilinguals were recruiting through online German-English societies and the Edinburgh University careers service, with the majority being postgraduate students. All translators were active professionally (minimum of two years of professional experience) with around half recruited through a local translation agency. Specifically, the experiments broke down as follows:

- Experiment 3: Bilinguals shown Source and Target sentences but no visual alignment.
- Experiment 4: Bilinguals shown Source and Target sentences with visual alignment (lines).
- Experiment 5: Professional translators shown visual alignment in half the sentences.

4.1.5 Analyses

Linear Mixed Effects Models were constructed and tested using the lme4 package for the R statistics environment.[4] LME models are essentially sophisticated linear regression models that simultaneously

[4] http://cran.r-project.org/web/packages/lme4/index.html

	Experiment 3 (Obs: 3690)		Experiment 4 (Obs: 3788)		Experiment 5 (Obs: 3948)	
	β	t	β	t	β	t
(intercept)	5384.60	13.677	5156.18	15.148	5838.75	16.71
With Error	-335.78	-3.309 ***	-186.37	-1.983 *	-395.73	-4.319 ***
Target Sentence	2934.09	29.821 ***	2375.03	26.092 ***	932.81	10.401 ***
Interaction	1212.79	6.163 ***	1495.22	8.213 ***	1723.48	9.608 ***

Table 2: Reduced LME table of fixed effects for Total Reading Time (* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$)

incorporate fixed effects (e.g. Errors, Visual Alignment) with multiple random effects (e.g. both participants and experimental sentences). The model description is therefore based on a predicted value obtained by a linear combination of the fixed effects plus a constant (the intercept). The intercept is essentially the grand mean of the data and all other effects are predicted on the basis of this starting value. The values associated with each fixed effect and their interactions (there is no underlying assumption of independence between variables) are estimated in the hypothesised model. These parameters (denoted as β values) are then statistically tested, producing a standard t-value for each one, in order to determine which variables contribute significantly to the proposed model.

Following the convention previously adopted, global analyses refer to sentence-wide effects while localised analyses involve effects at the sub-sentence (typically word) level. The data used was contingent on participants making a correct response on a trial, either "no" if there was no error or clicking on the correct word if there was. Two new reading measures are introduced for this dual language reading task. First Run Time investigates the initial viewing duration of each sentence (i.e. cumulative contiguous fixations on the same sentence). Shorter times indicate that less time is spent on the first reading of a sentence, typically because the reader only makes a partial examination before moving on to the other sentence or because they are selectively skimming the text. This can be meaningfully interpreted in conjunction with the second measure, the number of switches made between the Source and Target text. A higher number here indicates that the reader moved between the two sentences many times rather than simply reading each once in turn. Combined, these measures can provide detailed information about the strategies involved in human post-editing. As well as temporal measures and patterns of eye movements, the level of cognitive load or effort can also be revealed by examining pupil dilation. Increases in dilation can be another indication of processing difficulty or anomaly detection.

A summary of the key findings and comparisons between the sentences and error types are listed below. Expertise comparisons between bilinguals and professional translators are reserved for Section 5.

Global analyses The total time spent reading the sentences was actually decreased by a small but reliable amount when there was an error on the screen (see Table 2) and more time was spent on the translated sentence rather than the source. There was a significant interaction between these variables with disproportionately more time spent on the translation when there was an error. These times were consistent across the three experiments ($\beta = -369.473$, $t = -1.64$, n.s.). The presence of a Mistranslated Lexical (ML) error increased reading times for all experiments. There was also an effect of a Word Order (WO) error, but only for bilinguals shown visual alignment ($\beta = 2291.05$, $t = 2.508$, $p < 0.05$). There was no main effect of Visual Alignment on the total reading time data, although there was an interaction with expertise (reported in Section 5).

Similar to Total Reading Time, First Run Time (see Table 3) showed a significant main effect of the presence of an error and a difference between the time spent on the Target rather than Source sentence, along with an interaction (the consequence of reading a Target sentence containing an error was more than just the sum of the two main effects) for all three experiments. However, the pattern was slightly different in that the presence of an error increased the initial reading time of any sentence. Again though,

	Experiment 3 (Obs: 3690)			Experiment 4 (Obs: 3788)			Experiment 5 (Obs: 3948)		
	β	t		β	t		β	t	
(intercept)	1323.80	21.502		2960.57	20.482		1085.314	21.4	
With Error	140.38	5.641	***	224.13	3.953	***	133.667	5.242	***
Target Sentence	91.81	3.774	***	686.7	12.438	***	-193.897	-7.689	***
Interaction	223.85	4.600	***	644.2	5.838	***	135.757	2.692	**

Table 3: Reduced LME table of fixed effects for First Run Time

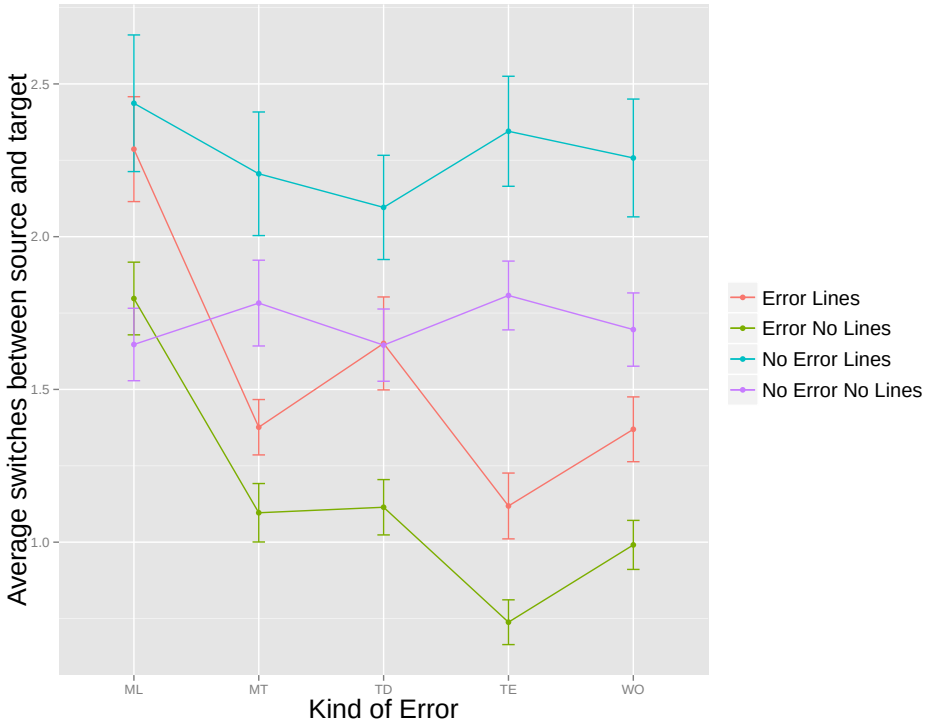


Figure 6: Mean oscillations between Source and Target sentences per trial dependent on the presence or absence of an error and/or alignment lines

the effect was particularly strong when reading a Target sentence with an error. Also in support of the Total Time findings, there was an effect of a Word Order (WO) error, but only for bilinguals shown visual alignment ($\beta = 1205.92$, $t = 2.178$, $p < 0.05$). This was the only evidence of a Visual Alignment influence on the first pass reading of any sentence.

The presence of an error resulted in an overall reduction in the number of times participants moved between Source and Target sentences for all three experiments ($\beta = -0.99476$, $t = -22.198$, $p < 0.001$). The only exception was in the ML condition where an error did not appear to alter the switching reading pattern ($\beta = 3.22215$, $t = 7.057$, $p < 0.001$), presumably as the contextual verification needs more exhaustive rereading and checking with the original source. Another consistent effect across all the participants was a reduction in switching if the Source sentence was read first ($\beta = -0.95098$, $t = -15.561$, $p < 0.001$) and this became stronger when an error was detected (interaction: $\beta = -0.51592$, $t = -5.418$, $p < 0.001$). While there is a suggestion that showing the alignment lines increases the switching between sentences (see Figure 6), this is statistically unreliable ($\beta = 0.39537$, $t = 1.353$, n.s.).

The pupillometry results were remarkably similar to the reading time findings. For all the experiments, a correctly identified error increased mean pupil dilation, and pupil size was larger when reading the Target rather than Source sentences, and there was a significant interaction between these two effects

	Experiment 3 (Obs: 3690)			Experiment 4 (Obs: 3788)			Experiment 5 (Obs: 3948)		
	β	t		β	t		β	t	
(intercept)	1690.06	8.308		504.497	12.771		1841.61	7.989	
With Error	82.187	5.935	***	12.909	8.406	***	83.722	7.946	***
Target Sentence	198.748	14.67	***	16.266	10.845	***	159.30	15.284	***
Interaction	122.66	4.527	***	9.079	3.029	**	77.62	3.724	**

Table 4: Reduced LME table of fixed effects for Mean Pupil Dilation

(see Table 4). Perhaps surprisingly, given the increase in saliency and perceptual complexity, the only effect of Visual Alignment on pupil dilation was a minor additive interactive effect for the bilinguals (see Section 5).

4.2 Field studies on CASMACAT visualization options

The lab experiments reported in the previous section showed that alignment visualization in the form of lines connecting source and target words are useful in that they reduce the time spent reading the target text, but only for naive bilinguals who are not professional translators. Building on these results, we conducted a small-scale experiment evaluating source/target text (ST-TT) alignments (ALG14 study in the TPR-DB). This study tests the alignment visualization options as they are actually implemented in the CASMACAT workbench, rather than relying on manually provided lines as the lab studies did. Furthermore, this study uses a real post-editing task, rather than a more constrained error-detection task as in the lab experiments.

The main aim of this field study was to assess if the use of visual ST-TT alignments implemented in CASMACAT workbench leads to productivity gain (shorter turnaround) during post-editing. A secondary aim of this experiment was to provide additional evidence for differences in performance between two different groups of users, i.e. professional translators vs. naive bilinguals (in this case, for English-Spanish).

4.2.1 Participant profiles

This alignment experiments involved eight participants divided in two groups: i) four professional translators (P01-P04) and ii) four bilinguals (English-Spanish) without any translator training (P05-P08). Professional translators were 36.5 years old on average (range 33-42) with more than 5 years of experience. Bilinguals were 44 years old on average (range 39-51). None of the participants had previous experience post-editing MT outputs. More specific data on the participants' age, experience, education, etc., can be found in the TPR-DB (metadata folder)[5].

4.2.2 Text type

Two texts were used for this experiment, both from the domain of general news (extracted from the News Commentary Corpus 2012). Each text had approximately 500 source text words distributed in 20 segments each. Each English source segment was pre-translated into Spanish by a statistical MT system and then loaded into the CASMACAT workbench for the participants to post-edit.

4.2.3 Experimental design

In order to assess and compare the effects of enabling visual ST-TT alignments each participant post-edited two texts in CASMACAT, each under one of the following conditions:

[5]Available online from: `http://bridge.cbs.dk/platform/?q=CRITT_TPR-db`

- *Condition 1*: Traditional post-editing without showing ST-TT alignments during the post-editing process (P).
- *Condition 2*: Traditional post-editing showing ST-TT alignments during the post-editing process (PA). In this condition, participants could hover with the mouse over either a source or target word and both the equivalent word(s) in the other language and the current word were highlighted.

Participant's performance was tracked using both key-logging and eye-tracking from CASMACAT. Texts and conditions were counterbalanced among participants.

4.2.4 Results

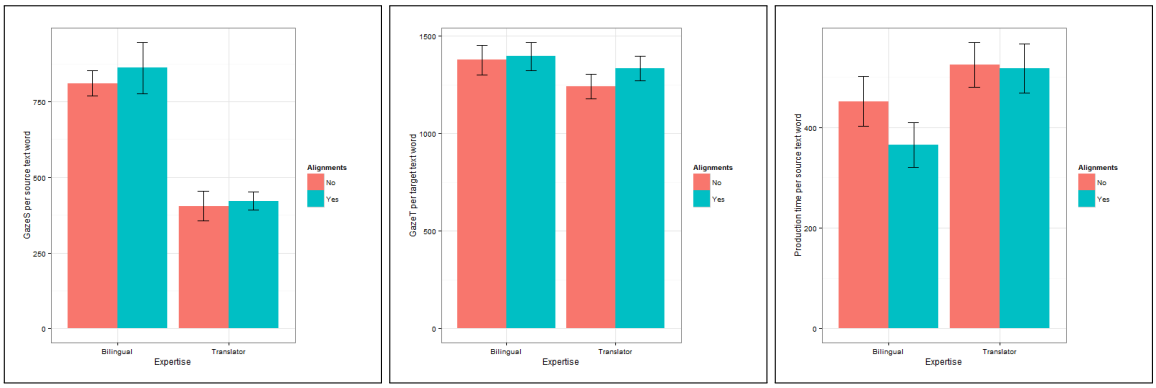


Figure 7: ALG14 Study - GazeS-TokS for professional and non-professional translators. ALG14 Study - KDur-TokS for professional and non-professional translators

Linear Mixed Model analyses for the ALG14 dataset showed that, on average, both professional translators and naïve bilinguals spent as much time (per word) reading the source text when there were alignment links as compared to when there were no alignment links and the same was true of the target text (all $t < 2$). For the production time, i.e. periods of coherent typing excluding pauses of more than 5 seconds (*Kdur*), results showed that there was no significant difference between the two alignment conditions for translators ($t = 0.103$). However, there was a significant effect for the bilinguals ($t = -2.155$): coherent typing in the condition in which visual alignment of equivalent words in the source and target text was possible was 46ms faster (per word).

The models used had the Alignment condition as a fixed effect and Participants, Text (the different texts used in the different conditions) and lexical translation entropy as random variables. Lexical translation entropy was added as a random variable given that it has a large effect on behavioural measures during translation and post-editing. Adding lexical translation entropy as a random variable significantly improved model fit ($p < .001$). In addition, we tested the overall difference in total reading time per source text word per expertise group and found that professionals spend on average 210ms less on the source text (per word). This effect was significant ($t = 2.29$). However, both professionals and bilinguals spent the same amount of time on the target texts ($t = 1.27$). Overall, bilinguals were 49ms per word faster ($t = 2.57$) than professionals in terms of coherent typing (*Kdur*).

4.2.5 Conclusion

The fact that bilinguals' production time was reduced by nearly 50ms per word in the alignment condition could be due to the fact the continuous exposure to the translation task leads to more readily available connections between items in the two languages. Priming has been identified as a learning mechanism with long lasting effects Pickering & Ferreira (2008). It is therefore possible that, upon reading items in the MT output, professionals more efficiently accessed equivalent items and therefore did not need visual aids for identifying them. Naïve bilinguals, on the other hand, might be proficient in their two languages, but given that they are less often exposed to the task of translation, connections between items in the two languages are less readily available and visual aids therefore make the process

more efficient. It is also interesting that total reading time per source word is much lower overall for professionals as compared to bilinguals. This suggests that professionals are better able to work out what needs or does not need changing mainly on the basis of the MT output, while bilinguals need to cross-reference with the source text. However, both professionals and bilinguals spend about the same amount of time on target words, suggesting that cognitive effort in terms of target text processing is the same for professionals and bilinguals.

5 User modeling

5.1 Experimental results comparing translator types

Deliverable D1.2, section 3, "Translator types and post-editing styles" proposes a taxonomy of post-editing styles which - amongst others - include style 1: first read the target text and then refer to the source text before making changes in the target text, and style 2: first read the source text and then proceed to read the target text looking for changes needed in the MT output. The data gathered in the lab experiments suggest that these differences in post-editing styles are related to expertise of post-editors.

The series of lab experiments reported in Section 4.1 manipulated not only whether participants were able to see alignment visualizations, but it also compared naive (untrained) bilinguals to professional translators in the task of MT error detection. We can therefore use this data to determine whether these two populations differ in terms of task strategy and how they use alignment visualizations.

While there was no effect of the presence of visual alignment on overall trial duration, the presence of lines did reduce the time that bilinguals spent looking at the translation ($\beta = -559.07$, $t = -4.17$, $p < 0.001$); this was not the case for the professional translators ($\beta = 111.398$, $t = 0.621$, n.s.), resulting in a significant interaction with expertise ($\beta = -670.468$, $t = -2.97$, $p < 0.01$). There was also a penalty associated with reading the target sentence first for bilinguals ($\beta = 1180.94$, $t = 4.86$, $p < 0.01$) which was not found for the translators ($\beta = 129.113$, $t = 0.584$, n.s.), giving rise to an Expertise interaction ($\beta = 1058.752$, $t = 3.25$, $p < 0.01$).

In the First Run Times, the translators actually spent less time overall on their first reading of the Target sentence compared to the Source sentence (see Table 3 in Section 4.1.5); the opposite was true for the bilinguals. This became more extreme when the Target was read before the Source: translators show a reduction ($\beta = -989.968$, $t = -14.049$, $p < 0.001$) while bilinguals showing an increase ($\beta = 679.636$, $t = 15.755$, $p < 0.001$), with a resultant strong interaction ($\beta = 1669.102$, $t = 21.37$, $p < 0.001$).

In terms of movements between sentences, the non-professionals had a lower rate of switching ($\beta = -0.55657$, $t = -2.463$, $p < 0.05$), although the difference was reduced when there was an error in the Target sentence ($\beta = 0.4598$, $t = 5.171$, $p < 0.001$) or the Source sentence was read first ($\beta = 0.30756$, $t = -2.523$, $p < 0.05$).

There is some evidence that the cognitive load or task demand was higher for the non-professionals as pupil dilation increased significantly more for them when they read the Target sentence ($\beta = 13.2736$, $t = 5.318$, $p < 0.001$). This consequence of inexperience was more pronounced when the Target sentence was read first ($\beta = 117.39$, $t = 16.487$, $p < 0.001$) and even more so when there were visible alignment lines ($\beta = 148.51$, $t = 10.917$, $p < 0.001$). In fact the professionals did not show a main effect on pupil dilation between reading the two sentences in each pair ($\beta = 0.1656$, $t = 0.114$, n.s.) whereas the bilingual participants did ($\beta = 3.44791$, $t = 2.003$, $p < 0.05$). Dilation measures for the professional group were also completely unaffected by the presence or absence of Visual Alignment. Combined, these findings could be considered as further evidence that expert translators did not find the overall task as cognitively demanding as the naïve translators did.

5.2 Predicting translator types using machine learning

In this section we describe a set of experiments that aim at 1) clustering post-editors based on their shared behavioural characteristics and 2) identifying post-editors based on their activity profiles.

We conceptualize the process of post-editing as a sequence of labeled activities, and show that these sequences of activities enable us to identify individual users' profiles. From the emergent patterns, we are able to cluster post-editors into subgroups based on the commonalities of their individual process

sequences. The identification of different post-editing styles provide insights for (a) the development and adaptation of translation tools that cater and adapt to the work style of a user, (b) the classification of individual translators based on non-process factors (translator experience, translator personality, time constraints, etc.) and (c) the most salient skills required of post-editors, which can later be applied to translator training.

Machine learning models are used to investigate the activity data tracked during the LS14 post-editing sessions to infer clustering and classification models. We use the Waikato Environment for Knowledge Analysis (WEKA 3.6 Hall et al. (2009)), an open-source toolkit for data mining and machine learning. Using several machine learning algorithms provided by the toolkit, we train various classification models. For the generative models, we use the SRI Language Modeling Toolkit (SRILMStolcke (2002)).

5.2.1 Data

The data for the current investigation on user modeling was extracted from CASMACAT longitudinal study LS14 and field trial data CFT14, as described in D6.3. This dataset is the first of its kind that implements a longitudinal approach to assess how post-editors adapt to interactive machine translation. We include three types of segmentation information derived from process unit file conventions extracted from the LS14 study: Activity units (CU), production units (PU), and translation segments (SG). While CUs and PUs are based on process data, SG units are derived from the translation product. Features used in this study for the different units are listed in Table 10 (see Appendix A.1).

Activity Units (CU) Features from the activity units serve as a baseline for translation processes. CUs describe the sequences within a translation session, in terms of typing, reading or pause activity. We employ a dichotomous model: Activity is categorized as either translation activity (Type 4) or no activity (Type 8) to follow the conventions of Carl and Schaeffer (2013). To achieve finer-grained distinctions in the activity profile, we refine the activity labels with duration information of each event resulting in five additional classes centered around the median duration (in milliseconds). Furthermore, part-of-speech (PoS) sequences extracted from the target text files are aligned with the CU data. There are in 68 unique PoS tags identified for Spanish in LS14, derived from TrEd/Treex.

Production Units (PU) PUs represent coherent sequences of typing activity (or typing bursts) and include information about the duration of the unit, duration of the preceding pause, number of edits, insertions and deletions, tokens involved in the source text and target text and average cross values. Cross values are defined as the "relative local distortion of the reference text with respect to the output text, and indicate how many words need to be consumed in the reference to produce the next token(s) in the output" Carl & Schaeffer (2013). A PU boundary is defined by a time lapse of more than one second between successive keystrokes.

Translation Segments (SG) Translation segments provide sequence information of aligned source and target text segments detailing the segment production duration, character length, insertions and deletions and gaze data, when available. Average word entropy, cross values, perplexity, and source text literality were also calculated and appended to this file type, given the level of segmentation that our analysis required.

5.2.2 Similarity of post-editing behaviour

Using the SRILM toolkit, we build n-gram models on CU-sequences and target text PoS sequences of each post-editor. We use perplexity values as scores in a k-mean clustering to find similarity between post-editors, and then validate these clusters using the metadata collected about participants in an introductory questionnaire (see deliverable D6.3).

Clustering Based on Activity Unit Sequences The original CU files included in the TPR-DB contain eight types of activities[6]. However, this classification of activity labels depends on the gaze information, which unfortunately is not available across all points in our dataset in LS14. As such, we map the original eight categories into two:

- **Type 4** (Translation activity, T4): activity units as defined by a sequence of coherent typing, which may also include gaze information; and,
- **Type 8** (No Activity, T8): boundaries between two activity units defined as a pause of 1000 ms or more without any keyboard activity.

Under this modified categorization, because there are now only two types of activity, translation activity (Type 4) is always followed by a pause (Type 8). This creates a model in which only two transitions are possible (*T4-T8-T4* or *T8-T4-T8*). Therefore, we further subdivide Type 4 and Type 8 into five categories based on the duration of these events: Five buckets centered on the median duration, further partitioning the activity and pause units into five subgroups. Table 5 illustrates the generated sequences considering the duration of the translation and pause units.

P01:	T4,1	T8,3	T4,1	T8,2	T4,5	...
P02:	T8,2	T4,3	T8,4	T4,1	T8,1	...
...	...					...
P05:	T4,1	T8,3	T4,2	T8,5	T4,2	...

Table 5: Sample user participant activity sequences bucketed by duration

We create a standard trigram language model on the activity sequences of each post-editor. The language model of one post-editor is then used to calculate the perplexity scores of the activity sequences for all the other post-editors. Perplexity, *PP*, is often used for measuring the fit of a language model to a corpus of sequences. It can be interpreted as the average number of tokens that can be produced by a model at each point in the sequence. For a test set with tokens $W = w_1, w_2, ..., w_n$, the perplexity of a trigram model on the test set is

$$PP_W = \prod P(w_i \mid w_{i-1}, w_{i-2})^{-\frac{1}{n}} \tag{1}$$

where it can be noted that perplexity is normalized by the number of tokens in the test sequence.

Table 6 shows the perplexity scores of each post-editors language model on the other post-editor's activity sequences. It illustrates that the diagonal contains the smallest perplexity value since the dataset is the same as the one used to create the model.[7]

	PE1_LM	**PE2_LM**	**PE3_LM**	**PE4_LM**	**PE5_LM**
PE1	*4.1*	4.3	4.6	4.8	4.4
PE2	4.3	*4.1*	4.4	4.6	4.4
PE3	4.6	4.4	*4.1*	4.3	4.9
PE4	4.5	4.3	4.0	*3.8*	4.9
PE5	4.0	4.1	4.5	4.8	*3.8*

Table 6: Perplexity scores for the Activity sequence LM model for all post-editors

We use the perplexity values as distance costs in a k-means clustering algorithm to produce two ($k = 2$) clusters, resulting in the following clusters: $cluster_1 : \{PE1, PE2, PE5\}$ and $cluster_2 : \{PE3, PE4\}$. When looking for possible explanations in the metadata, we found that $cluster_1$ includes the most experienced post-editors. Based on the findings provided by this clustering, it seems to be the case that

[6]Type:1 source text reading, Type:2 target text reading, Type 4: typing activity. All combination of these lead to 8 activities

[7]However, an exception as seen in Table 6, PE3's activity model has a higher perplexity score on PE3's sequence compared to that on PE4's activity sequence.

experienced post-editors produce similar kinds of activity sequences in contrast with the activity sequences of inexperienced post-editors.

Clustering Based on Target Text Part-of-Speech Sequences We extract the PoS sequences for each segment in the target text and created a n-gram language model for each post-editor (PE). Then we use this model to calculate the perplexity values of the language model for all other post-editors to measure the appropriateness of the model. Using the perplexity scores as distance metrics, we grouped the post-editors into two clusters by applying standard k-means clustering: $cluster_3 : \{PE1, PE3, PE5\}$ and $cluster_4 : \{PE2, PE4\}$. To account for this clustering, we compare the results with the participant metadata. We find that post-editor PE2 and post-editor PE4 share a not-so-positive approach for post-editing task when compared to translation tasks (they would prefer to translate from scratch), whereas the other three participants did not indicate such apprehension towards the task. Considering our data, this seems to indicate that post-editors with similar negative attitudes towards post-editing tasks tend to have similar activity patterns.

5.2.3 Identifying post-editors

In this study we sought to determine whether machine learning models are able to identify post-editors based on their activity profiles. We carried out tests on the three types of data mentioned in section 5.2.1: activity units **CU**, productions units **PU** and translation segments **SG**. The data was segmented to analyze the effect of the workbench, post-editing PE vs. interactive ITP, in the analyses. We experimented using a range of machine learning algorithms with 10-fold cross validation for classification, but found that a multilayer perceptron and classification-via-regression performed best for the task of identifying the post-editors. The baseline accuracy is 20% given that there are the same number of samples for the five participants.

	Algorithm	**PE**	**ITP**	**Combined**
CU Profile	Multilayer Perceptron	40.58 %	35.51 %	41.54 %
	Classification via Regression	42.37 %	36.82 %	42.72 %
PU Profile	Multilayer Perceptron	44.67 %	39.82 %	37.06 %
	Classification via Regression	45.83 %	47.69 %	46.48 %
SG Profile	Multilayer Perceptron	42.88 %	46.93 %	44.45 %
	Classification via Regression	44.64 %	47.51 %	45.71 %

Table 7: Results for the 5-way classification task to discriminate post-editors based on activity, production and translation segments profiles created at the segment level.

CU Profile Table 7 shows results obtained from frequencies of unigrams and bigrams of activities as features for discriminating post-editors. It illustrates that the model is able to identify post-editors better when they use the 'traditional' PE, with 42.37% accuracy, compared to 36.82% in the ITP environment. However, when the data is combined using the translation mode as an additional feature, accuracy of the model remains almost the same at 42.72%.

PU Profile A features matrix using the PU features was created as described in Table 10 (Appendix A.1 to identify post-editors. In the matrix, all text-dependent features have been normalized using *LenS* (length in character of the source sentence) to ensure that the system is not biased by differences in the length of the text. Considering there are multiple PUs for each segment, and that the number of PUs vary per post-editor, we make a sparse vector to group together the different PUs of each segment. As shown in Table 7, we achieve 46.48% accuracy while using the entire data set with translation mode as a feature (Combined). When dividing the data set depending on the translation mode (OE vs. ITP), we achieve an accuracy of 47.69% and 45.83% with the system discriminating post-editors in the traditional and interactive enabled mode, respectively.

SG Profile Translation segments (SG) have some features overlapping with PU profiles as detailed in Table 10. Nevertheless, in this file, all the information is cumulative for a segment and dependent on the text, while in the PU files, the information is created based on the post-editors' typing bursts. When testing the combined dataset including the data from the two translation modes, the SG model has an accuracy of 45.71%. When running the tests independently for the two modes, the PE dataset achieves 44.64% of accuracy, while the ITP dataset reaches 47.51% of accuracy.

5.2.4 Learning ITP

In the previous sections we have shown that post-editor profiles can be detected and single post-editors can be identified with a certain degree of accuracy, using features derived from translation process data. However, when working with a new translation assistance system over a longer period of time post-editors also show behavioural changes. As pointed out in D6.3 section 2, one of the reasons for the LS14 study was to investigate learning behaviour in a longitudinal study. In D6.3 section 2.4.2, we showed learning effects of the five post-editors when getting exposed to the ITP mode for some time. In this section we will assess what post-editors have learned in the six weeks they were using the CASMACAT ITP mode. We compare the behaviour of the post-editors involved in the LS14 study and in the subsequent CFT14 field trial.

Seven post-editors contributed to the CFT14 field trial, from which four already previously participated in the LS14 study. That is, there were three new post-editors, and four post-editors had already six weeks experiences with the CASMACAT PE and ITP modes. This fact makes it interesting to investigate how the behaviour of the four post-editors who attended both studies is different from the new post-editors who got involved with PE and ITP.

Study	TType	*H*	CrossS	CrossT	Slen	Tlen
LS14	News	0.612	1.60	1.29	25.0	27.85
CFT14	EMEA	0.445	1.44	1.23	21.0	22.93

Table 8: Comparing properties of EMEA corpus translations and the news translations

However, there were a few differences in the LS14 and the CFT14 studies. 1) For LS14, the goal was to compare the CASMACAT ITP and PE translation modes, while CFT14 aimed at comparing PE and ITP with online learning, i.e. Post-editing, Interactive, Online-learning (PIO). 2) In order to appreciate online learning capacities, texts were much longer in CFT14 than in LS14, but in turn there were only two texts for each translator, one to be post-edited in the PE (P) mode and the other in the PIO mode. 3) Whereas the LS14 data is based on an English-to-Spanish news text, the CFT14 study used a medical text extracted from the EMEA corpus [8]. As shown in Table 8, in contrast to the news text, the medical text is very rich in terminology, segments on the source side (*Slen*) as well as on the target side (*Tlen*) are on average shorter than in the news text. Translations of the medical text tend to be more literal than for news text. Table 8 shows that the EMEA translations have less lexical variability then the news translations, as measured by the translation entropy H. Translations are also syntactically closer to the source text: lower *CrossS* and *CrossT* values indicate greater syntactic similarity between the source and the target language.

Four translators $G_1 : \{P01, P02, P03, P04\}$ from the LS14 experiment also participated in the CFT14 experiment. Three additional translators $G_2 : \{P05, P06, P07\}$ joined the CFT14 study who did not use the CASMACAT workbench before. Despite the different nature of the text, it can be expected that the experience with the ITP post-editing mode that G_1 translators obtained during the six weeks of the LS14 experiment would also carry over to the CFT14 study, while the fresh translators in the group G_2 would not have this experience, and thus show different behaviour.

Indeed, we detected a difference in the ratio of coherent keystroke activities (*Kdur*) and the filtered total production duration (*Fdur*) between these two groups. Figure 8 shows that most of the post-editors in the G_1 group ({P01,P02,P04}) have a lower proportion of coherent keystroke activities (*Kdur/Fdur*)

[8] http://opus.lingfil.uu.se/EMEA.php

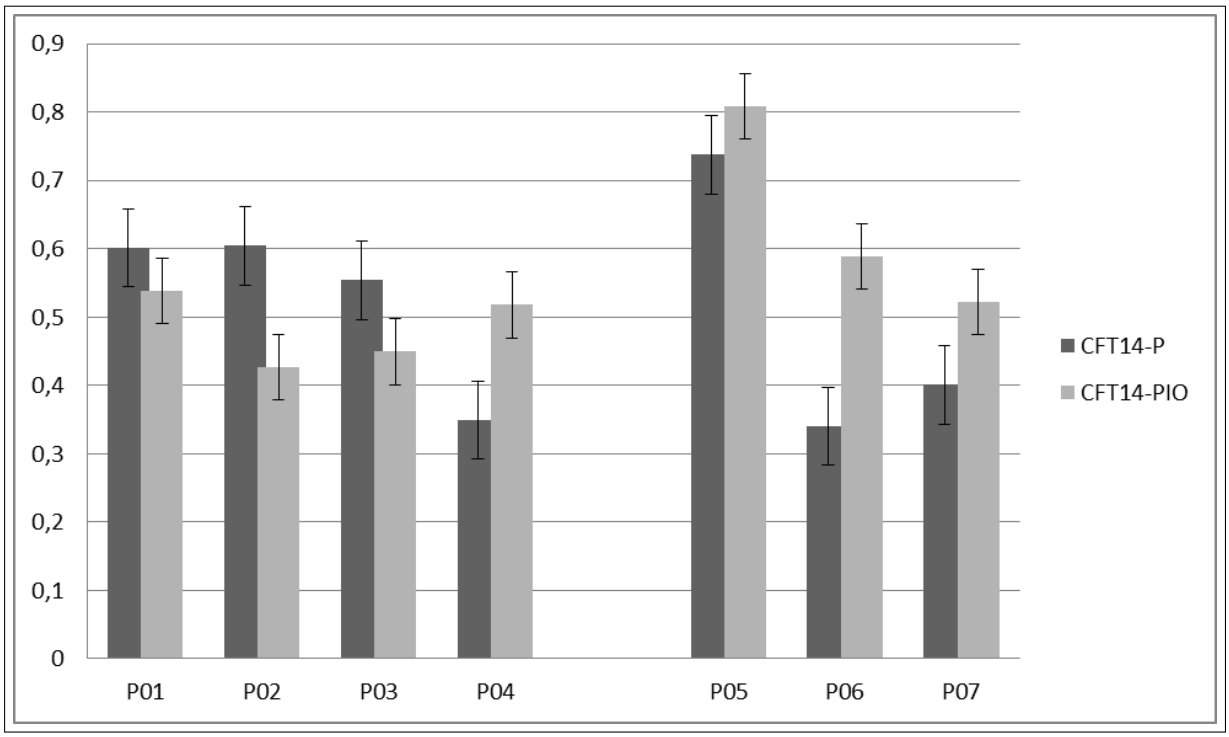


Figure 8: Ratio of typing time (*Kdur*) and production time (*Fdur*) for G_1 and G_2 translators.

in the PIO mode than in the P mode. While *Kdur* indicates the duration of proper typing activities, *Fdur* includes also thinking time, but excludes longer pauses. That is, in the interactive ITP mode these post-editors seem to have learned to accept interactive suggestions which reduces the amount of their coherent typing time, while for translators in the G_2 group that does not seem to be the case. That is, G_1 translators seemed to adopt the interactive suggestions more often than the new translators by less frequently overwriting the ITP proposals.

5.2.5 Discussion

In this section, we tested the hypothesis whether the events that make up the translation process provide enough information for the individualization of post-editor profiles. By using machine learning models, we are able to not only identify the post-editors' profiles, but also cluster and discriminate between post-editors. We found that more experienced post-editors or post-editors with a more positive attitude towards post-editing $\{P01, P02, P03\}$ are likely to adapt better to ITP than inexperienced post-editors who do not have a positive approach towards post-editing.

In contrast to the visualization of alignment relations, ITP seems to suit more experienced translators with a more positive attitude towards MT. We assume that these post-editors are also more open-minded to acknowledge and accept ITP suggestions, and thus learn how to make better use of the system in their translation work. A qualitative assessment of some post-editing sessions show that post-editors in their first encounter with the ITP mode often overwrite translation suggestions with an identical text. Presumably, once a (partial) solution for a translation problem has been found, translators (and post-editors) are engaged in a typing process which is difficult to interrupt, even if the suggested MT continuation is identical to what the post-editor had in mind. This behaviour seems to change as they become more experienced and familiar with the translation predictions.

However, since only a few post-editors participated in the LS14 and CFT14 studies, the current analysis should be considered as an initial exploration of such methods on translation process data. Considering our initial results, it would be beneficial to explore how additional features and experiments with more participants affect the models.

6 Conceptual and procedural encoding

According to RT, linguistic forms encode semantic representations which are recovered by a hearer through unconscious, automatic decoding processes (see section 3 on priming). Generative grammar provides an account for this encoding–decoding mechanism. RT goes beyond this, it is a pragmatic theory of communication seeking to explain the inferential processes that follow the encoding–decoding mechanism. Successful communication, it is claimed, depends on inferential processes. For Blakemore (2002) the distinction between semantics and pragmatics lies in the distinction between the process of encoding-decoding messages and the process of making inferences from evidence: linguistic semantics provides logical forms which are taken as input by pragmatic inferences constrained by the principle of relevance.

Linguistic forms encode constituents of conceptual representations that enter into inferential cognitive computations but they also encodes procedural information which restricts the computations in which these conceptual representations are involved. This distinction is known as *conceptual* and *procedural* encodings. Procedural encodings, thus, constrain the computations and guide the comprehension process so that the hearer ends up with a conceptual representation Blakemore (2002). In line with this, Sperber & Wilson (1993, 16) claim that "conceptual representations can be brought to consciousness; procedures cannot. We have direct access neither to grammatical computations nor to the inferential computations used in comprehension." Based on these hypotheses, Alves et al. (2014) define these terms as follows:

- **Conceptual encodings (CEN)** consist of open lexical categories such as nouns, adjectives and verbs. They convey conceptual meaning and are propositionally extendable. CEN can be enriched and contribute to the inferential processing of an utterance.
- **Procedural encodings (PEN)** consist of non-open morphological categories such as negation, tenses, determiners, word order, etc. They cannot be extended in propositional terms but contribute decisively to the cognitive processing of an utterance by imposing inferential constraints on it.
- **Hybrid encodings (HEN)** can be linguistically encoded by CEN or PEN and convey both conceptual and procedural instructions.

According to Alves & Gonçalves (2003) translators learn to consciously manipulate conceptually and procedurally encoded information, so that they can identify the inferential constraints inherent in a given statement. Only then will they will be able to meet the demands of the target audience and its context. They find that conceptual encodings show a "relatively stronger interpretive resemblance between source and target texts" (p 20), and are thus easier to encode in a translation task. Similar findings are also reported in Sekino (2012) who conduct a translation experiment from Japanese into Portuguese. Their results corroborate Alves & Gonçalves's findings, showing that processing effort is greater when dealing with procedural encodings in both manual translations and in post-editing tasks in terms of keystrokes, fixation counts and fixation duration.

6.1 The CEMPT13 study

Alves et al. (2014) investigate the allocation of cognitive effort in the CASMACAT post-editing (PE) and interactive translation prediction (ITP) mode, with the question whether and to what extent ITP can facilitate conceptual encoding during post-editing.

The translation direction was from English (L2) into Brazilian Portuguese (L1). 21 Brazilian subjects (L1) with some training on post-editing translated two short medical texts of 277 and 384 words from the EMEA corpus in the PE and ITP condition with a randomized task order. Due to data loss, 5 sessions had to be discarded, so that the data of only 16 Brazilians remains in the data set.[9] Eye-tracking data was recorded with Tobii studio to assess the impact of the CASMACAT PE and ITP modes with respect to

[9]The data is part of the TPR-DB as CEMPT13 study

how much cognitive effort is needed in each of these environments. AOIs were generated with manual annotation of HEN, CEN and PEN areas.

The English-to-Portuguese EMEA corpus was chosen as a test material, which, according to our analysis in section 5.2.4 and Table8, has literal translations, is terminologically dense, and thus potentially rich in conceptual encodings.

6.2 Findings

With a particular focus to investigate whether ITP would provide a facilitation effect for conceptually encoded texts, the findings are as follows:

- ITP requires more cognitive effort in comparison to PE in terms of Fixation Count, Fixation Duration and Fixation Mean.
- ITP improves efficiency only for CEN since the system tends to automatically correct segments which, most often, are linguistic manifestations of CEN
- Interactive post-editing may have a facilitating effect for CEN.

These findings corroborate our conclusions from section 5.2.4 which shows that interested and open-minded post-editors learn to use the ITP mode in an effective way. In the CEMPT13 study it was shown that it is likely more CEN encoded sequences in which MT output has a potential to support post-editors.

7 CASMACAT configurations

This section presents a number of different configurations of the CASMACAT workbench based on the user profiles we have defined during in the different field trials and lab tests of the project. Each configurations represents an usage scenario defined by: 1) an specific task (translation/post-editing/revision) and 2) a number of features available in the workbench deemed helpful for the task in hand.

Table 9 summarises six different configurations we tested and implemented in the framework of the CASMACAT project. All configurations show the source text on the left side of the screen (divided in segments) and the MT output (if available depending on the task) on the right side. For all different configurations, the key shortcuts in the workbench work identically. They differ with respect to whether the target window is empty for *from-scratch translation*, whether the target window is pre-filled with MT output, or whether the target window contains translated/post-edited text for final *revision*.

Section 1 in deliverable D5.4, addendum, lists three settings for conventional post-editing with CASMACAT: MT output can be generated (1) via external MT output (EXT) or (2) via the CASMACAT or the MATECAT server (MT). The interactive mode (ITP) only works in the latter case, where in addition a number of visualization options, such as alignment visualization links (AV), confidence measures (CM), the biconcordancer (BI) and translation options (TO), as well as online learning (OL) and active learning (AL) may be activated. E-pen interaction in the CASMACAT workbench has only been implemented for revision purposes.

Usage Scenario	EXT	MT	ITP	OL	AV	CM	BI	E-pen	TO
From-scratch translation	–	–	–	–	–	–	–	–	–
Text revision	+	–	–	–	–	–	–	+	–
External PE	+	–	–	–	–	–	–	–	–
CASMACAT *PE*	–	+	–	+	+	+	+	–	–
CASMACAT *ITP*	–	+	+	+	+	+	+	–	–
CASMACAT *Global Voices*	–	+	–	–	–	–	+	–	+

Table 9: Distinctive feature matrix for six different configurations of the CASMACAT workbench

These six basic configurations are described below taking into account the linguistic task involved and the user profile best suited for each configuration. For each configuration, a demo link is provided for testing purposes:

- *From-scratch translation*: No machine translation is involved in this usage scenario. An XLIFF file is pre-filled with source text segments before uploading the file to CASMACAT workbench. Only the source text segments are shown on the left-hand side of the screen; the target segment in the workbench is empty and the translator is supposed to insert her translations from scratch. This configuration was tested in the first CASMACAT field trial and it was described in D1.1. It was also used in a number of lab and user-group studies during the first and the second year of the project. This usage scenario is suited in cases where MT quality is low as well as for users facing resistance to post-editing workflows (i.e. users with very ingrained translation patterns unwilling to adapt to traditional PE or ITP mode). In this configuration, the CASMACAT workbench serves only as an online editing environment for translation purposes without any further aids.

  Demo link: From scratch translation configuration

- *Text revision*: In this usage scenario, an XLIFF file is pre-filled with source text segments and translated from scratch/post-edited text as target segments. The file is subsequently uploaded to CASMACAT and the source and target segments are plotted in the workbench. This usage scenario is suited for reviewing/proofreading purposes and it was tested in the second (June 2013) and third (June 2014) field trials, CFT13 and CFT14 studies respectively. This reviewing scenario includes additional multimodal interaction using an e-pen. The use of an e-pen can be used by reviewers willing to swap between keyboard typing and the use an e-pen featuring basic editing gestures for reviewing purposes (i.e. text deletion, insertion and text swaps).

  Demo link: Text revision configuration

- *External PE*: This mode is similar to *text revision*, with the difference that the target segments in the XLIFF file are prefilled with MT outputs coming from MT systems other than CASMACAT or MATECAT. As in the revision scenario, the XLIFF file is uploaded to CASMACAT, where the source and target segments are shown. This usage scenario does not support any of the advanced visualization options featured by the CASMACAT workbench and it is only suited for text types and language pairs for which the CASMACAT server does not produce sufficiently high quality translations in its current implementation. This configuration has been tested by some CASMACAT user group members running experiments with the workbench but feeding MT from Google Translate.

  Demo link: PE external MT configuration

- CASMACAT *PE*: In this "conventional PE mode" the workbench connects directly to the CASMACAT MT server and pre-fills the target text editing window with MT output at runtime. An XLIFF file needs to be prepared containing only the source text. This setting was tested in the first and the second field trials (CFT12 and CFT13 studies, respectively), as well as in various lab and user group studies (see deliverable 1.4, table 1 for a list of all these studies in the TPR-DB). Conventional PE has been shown to be more effective than from-scratch translation (cf. deliverable 1.1) and it is suited for professional translators with post-editing experience (touch typists).

  Demo link: CASMACAT PE configuration

  Depending on the degree of post-editing expertise of the user, additional help in the form of visual alignment information (see section 4.2 and confidence measures at the word level (see D6.2) have proved to be useful for non-experience post-editors. This user profile also benefits from the help provided by the biconcordancer tool implemented in CASMACAT (see D6.3).

  Demo link: Advanced CASMACAT PE configuration

- CASMACAT *ITP*: The interactive translation prediction mode has been one of the major research areas of the CASMACAT project. This post-editing scenario is similar to CASMACAT *PE*, but

in addition new text predictions are generated as the post-editor edits the original target language string. Two ITP versions have been implemented, but only the "inline" version was tested in the framework of the second and third field trials (CFT12 and CFT13 studies), as well as the longitudinal study (LS14 study). Although results were not entirely conclusive, ITP is most suited for professional translators with little experience in post-editing and poor typing skills (non-touch typists) when post-editing conceptually dense encoded texts, as they can benefit from the predictions provided by the system (see deliverables D6.2 and D6.3). This user profile also benefits from the visual aids provided by source-target visual alignments, confidence measure and online learning techniques (CFT13 study).

Demo link: Advanced CASMACAT ITP configuration

- CASMACAT *Global Voices*: A similar configuration has been chosen as the default configuration for volunteers translation in the Global Voices Community Translation Platform[10] (CASMACAT user group - see deliverable D6.4). This configuration features interactive MT in the form of a floating window as well as a translation options table. From this configuration, volunteer translators still have the opportunity to translate from scratch (since the MT output is only show in the floating window or in the table) but still they have the chance to engage in post-editing tasks interacting with the MT output provided either through the floating prediction window (pressing TAB) or by clicking on the different phrases shown in the translation options table. Additional help is also provided adding a biconcordancer.

  Demo link: Advanced CASMACAT Global Voices configuration

Apart from this six basic online configurations, the CASMACAT workbench can also run in a home edition. The home edition is described in detain in deliverable 5.4 addendum. It runs in a virtual machine on a local computer and implements all the features of the CASMACAT workbench. This CASMACAT *Home Edition*[11] is suited for translators with high privacy requirements, where texts to be translated cannot be sent via the Internet due to confidentiality reasons.

8 Conclusion

In this deliverable we develop a theoretical view on post-editing with CASMACAT that combined Relevance Theory and the noisy channel model. Relevance Theory is an inferential approach to pragmatics according to which a communicator provides evidence of her intention to convey a certain meaning, which is inferred by the audience on the basis of the evidence provided. Since translation is a special form of communication, RT can been applied to translation, where the translator mediates between the intentions and evidences of the source and the target text, and thus communicates an author's source text to a target language audience. The RT framework can be combined with a noisy channel that capture the encoding/decoding processes of the translator or, in a PE setting, of the post-editor and the MT system.

While traditional post-editing is based on a static text, within CASMACAT a number of advanced features were implemented that dynamically modify the automatically generated texts and/or the workbench, with the aim for better supporting the post-editor with additional information that is otherwise hidden inside the MT system. These features were tested in a series lab-based experiments and field studies in which eye-tracking and keyboard data was recorded from translators and post-editors. The collected behavioural data was used for advanced cognitive analyses and user modeling to assess these novel types of translator assistance, as they were integrated into the workbench.

In section 3 we address automated decoding processes during translation production, and show that MT output primes post-editors, and that post-edited translations are more likely literal and direct translations than from-scratch translations. However, as Gutt (1992) points our, direct translations are sometimes preferable, as they reduce interference with the translator's interpretation to a minimum.

In previous studies (D1.2) we identified a range of translation styles, some of them indicating that post-editors switch back and forth between the source and the target text, thereby matching the source

[10] `http://www.casmacat.eu/community/?action=globalVoices`

[11] CASMACAT Home Edition: `http://www.casmacat.eu/index.php?n=Installation.HomePage`

text meaning on the MT output. In section 4 we investigate whether visual support of this process can be provided in the form of a visualization of alignment links between the source and the target text. It turned out that naïve bilinguals most from this feature while it seem to have no effect for professional translators.

In section 5 we investigated whether machine learning techniques can be deployed to find similarities and differences between post-editors, as well as to identify them, based on sequences of their activity data. These approaches show only limited success, based on a small set of translators process data. However, the more surprising finding is that open-minded and positively predisposed post-editors seem to learn to better how to accept and work with translation predictions. This finding seems to suggest that a special translator training might be required, particularly for ITP, to allow users to fully appreciate the provided translation support.

References

Alves, F., & Gonçalves, J. (2003). A Relevance Theory approach to the investigation of inferential patterns in translation. In F. Alves (ed.), *Triangulating Translation: Perspectives in process oriented research*, (pp. 3–24). Amsterdam: John Benjamins.

Alves, F., Koglin, A., Mesa-Lao, B., Sekino, K., & Szpak, K. (2014). Investigating Cognitive Effort in Post-Editing: A Relevance- Theoretic Approach. In *Can these eyes lie? International Conference on Eyetracking and Applied Linguistics 26-27 September*, Warsaw.

Aziz, W., Koponen, M., & Specia, L. (2014). Sub-sentence Level Analysis of Machine Translation Post-editing Effort. In S. O'Brien, L. W. Balling, M. Carl, & M. Simard (eds.), *Post-editing of Machine Translation: Processes and Applications*, (pp. 1–31). Cambridge Scholars Publishing.

Blakemore, D. (2002). *Relevance and Linguistic Meaning: The semantics and pragmatics of discourse markers*. Cambridge: Cambridge University Press.

Campbell, S. (2000). Choice Network Analaysis in Translation Research. In M. Olohan (ed.), *Intercultural Faultlines. Research Models in Translation Studies I. Textual and Cognitive Aspects*, (pp. 29–42). Manchester: St Jerome.

Carl, M., & Schaeffer, M. (2013). The CRITT Translation Process Research Database v1.4.

Gibson, E., Bergen, L., & Piantadosi, S. T. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*, *110*, 8051–8056.

Gutt, E. (1990). A theoretical account of translation—without a translation theory. *Target*, *2*, 135–164.

Gutt, E. (1991). *Translation and Relevance. Cognition and Context*. Oxford and Cambridge: Basil Blackwell.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, *11*.

Hartsuiker, R. J., Pickering, M. J., & Veltkamp, E. (2004). Is Syntax Separate or Shared between Languages? Cross-Linguistic Syntactic Priming in Spanish-English Bilinguals. *Psychological Science*, *15*, 409–14.

Hill, R., & Keller, F. (2014). Error detection in native and non-native speakers provides evidence for a noisy channel model of sentence processing. Poster at the 27th Annual CUNY Conference on Human Sentence Processing, Columbus, OH.

Jensen, K. T., Sjø rup, A. C., & Balling, L. W. (2009). Effects of L1 Syntax on L2 Translation. In F. Alves, S. Göpferich, & M. I. M. (eds.), *Methodology, technology and innovation in translation process research: A tribute to Arnt Lykke Jakobsen*, (pp. 319–336). Copenhagen: Samfundslitteratur.

Kliffer, M., & Stroinska, M. (2013). Relevance Theory and Translation. *Linguistica Atlantica*, *25*, 165–172.

Levy, R. (2008). A noisy-channel model of rational human sentence comprehension under uncertain input. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, (pp. 234–243), Honolulu.

Pickering, M. J., & Ferreira, V. S. (2008). Structural Priming: A Critical Review. *Psychological Bulletin*, *134*, 427–59.

Schaeffer, M., & Carl, M. (2013). Shared Representations and the Translation Process: A Recursive Model. *Translation and Interpreting Studies*, *8*, 169 – 190.

Sekino, K. (2012). Investigating conceptual and procedural encodings in manual translation and in post-editing processes from Japanese to Portuguese.

Sperber, D., & Wilson, D. (1993). Linguistic form and relevance. *Lingua*, *90*, 1 – 25.

Sperber, D., & Wilson, D. (1995). *Postface to the second edition of Relevance: Communication and Cognition*. Oxford: Blackwell.

Stolcke, A. (2002). Srilm – an extensible language modeling toolkit. In *in Proceedings of the 7th International Conference On Spoken Language Processing (ICSLP 2002)*, (pp. 901–904).

Čulo, O., Gutermuth, S., Hansen-Schirra, S., & Nitzke, J. (2014). The Influence of Post-Editing on Translation Strategies. In S. O'Brien, L. W. Balling, M. Carl, M. Simard, & L. Specia (eds.), *Post-editing of Machine Translation: Processes and Applications*, (pp. 200–219). Newcastle: Cambridge Scholars Publishing.

Wilson, D. (1994). Relevance and Understanding. In G. Brown, K. Malmkjaer, A. Pollitt, & J. Williams (eds.), *Language and understanding*, (pp. 35–58). Oxford: Oxford University Press.

A Appendix

A.1 Feature for CUs PUs and SGs

Unit	Feature	Description
all	Participant	participant identifier
	Dur	duration of the unit
CU	Type	type of activity unit
	dur_cu	duration of activity units
	TokS	number of source tokens in the segment
	TokT	number of target tokens in the segment
	PoS	part of speech tag
CU, SG	LenS	character length of source segment
	LenT	character length of target segment
SG	Nedit	number of edits of the segment
	LenMT	character length of the machine translation segment
	Kdur	duration of coherent keyboard activity excluding keystroke pauses greater than or equal to five (5) seconds
	Fdur	duration of segment production time excluding keystroke pauses greater than or equal to 200 seconds
	Mins	Number of manually generated insertions
	Mdel	Number of manually generated deletions
	Ains	Number of automatically generated insertions
	Adel	Number of automatically generated deletions
	STent	average word translation entropy of the segment
	PP	perplexity score of the segment based on STent
	STlit	source text literality
	FixS	number of fixations on the source text unit
	FixT	number of fixations on the target text unit
	GazeS	total gaze time on source text unit
	GazeT	total gaze time on target text unit
	STcr2	cross value of source text token
	TTcr2	cross value of target text token
SG, PU	CrossS	cross value of source token
	CrossT	cross value of target token
	STseg	source segment identifier
	TTseg	target segment identifier
PU	Time	timestamp of the event
	ParalS	percentage of parallel source text reading activity during unit production
	ParalT	percentage of parallel target text reading activity during unit production
	Linear	degree of linear editing
	Pause	duration of production pause before typing onset

Table 10: List of feature for CUs PUs and SGs as used for translator profiling

A.2 Submission: Processes of Literal Translation and Post-editing

Submitted to an edited volume under the title *Translation in Transition: Between Cognition, Computing and Technology*" to appear in the Translation Library Series of John Benjamins Publishing Company (mid-2015).

Processes of Literal Translation and Post-editing

Michael Carl and Moritz Schaeffer

Abstract
The notion of literal translation has been defined in many different ways in a large number of studies in translation- and in bilingualism research that also address related issues. This paper introduces a novel metric to quantify literality of translations and applies it to from-scratch translation and machine translation post-editing. We develop a recursive model of translation production, which is based on the theory of shared combinatorial nodes and which explains our observations during translation production. We analyse a large corpus of translation data and show that our literality metric correlates well with production- and reading times during translation and post-editing. The introduced literality metric, together with the recursive model of translation production are powerful means to assess processes during translation and post-editing.

Introduction

A large body of studies exist in translation- and in bilingualism studies which investigate literality in translation from different points of view and with different methods and goals. Psycholinguistic research in bilingualism is usually conducted under very controlled experimental conditions to draw exact conclusions about cognitive representations from behavioural observations. Translation process studies, in contrast, prefer an ecologically more realistic experimental setup, in which experiments are closer to natural working conditions. In this chapter we attempt to fill this gap, by explaining naturalistically acquired translation process data, some of it in professional working conditions, with psycholinguistically grounded theories of translation processes, by quantifying literality and correlating it with behavioural measures.

Chesterman (2011, 23) argues that the well-known literal translation hypothesis "...has been implied or explicitly studied by many scholars, and does not seem to have a single source." Catford's (1965) and Ivir (1981) introduce the notion of *formal correspondence model,* Toury's (1995, 275) discovers the *law of interference* and Tirkkonen-Condit (2005, 408) develops *the monitor model.* All these concepts imply that one-to-one literal translation correspondences are easier to produce than translations that deviate from the source text (henceforth ST), as the latter would require more effort, and hence will take longer for a translator to produce.

Ivir (1981, 58) describes the translation process as follows:

> The translator begins his search for translation equivalence from formal correspondence, and it is only when the identical-meaning formal correspondent is either not available or not able to ensure equivalence that he resorts to formal correspondents with not-quite- identical meanings or to structural and semantic shifts which destroy formal correspondence altogether. But even in the latter case he makes use of formal correspondence.

Related to this notion of formal correspondence, is Toury's (1995, 275) "law of interference" which postulates that "...in translation, phenomena pertaining to the make-up of the source text tend to be

transferred to the target text...” Tirkkonen-Condit (2005, 408) reformulates the formal correspondence model into a monitor model: “It looks as if literal translation is a default rendering procedure, which goes on until it is interrupted by a monitor that alerts about a problem in the outcome.”

However, it is controversial what it actually means for a translation to be literal (e.g. Chesterman 2011; Malmkjær 2011) and thus difficult to assess exactly how much effort it takes to produce non-literal translations, and at what point during the translation process the extra effort occurs. To overcome this shortcoming, we define literality in translation by the following criteria:

1. Word order is identical in the ST and TT
2. ST and TT items correspond one-to-one
3. Each ST word has only one possible translated form in a given context

The first two criteria follow immediately from the word-for-word requirement. An ideal literal translation consists of the same number of tokens where each TT token corresponds to exactly one ST token, and tokens in both texts have the same order. We call this the ideal literal translation hypothesis: the ideal situation is a default which serves as a reference in the translation process. A change in word order or a situation in which one ST word is aligned to more than one TT word or vice versa violate literality criteria 1 and 2 and make the translation less literal.

In order to test the third criterion a corpus of alternative translations of the same ST is necessary: a corpus of alternative translations makes it possible to check to what extent different translators translate words in the same way. Assume a text is translated by, e.g. 20 different translators: if all 20 translators produce the same translation for a word in the context of the same text, then we can say that this word is translated literally. However, the more the word is translated differently by the different translators, the less likely it is that it will be translated literally.

In this chapter we seek to integrate findings from bilingualism research and translation studies to investigate the origin and effects of literal translation. On the basis of a large body of experimental translation data we show that and how less literal translations are more effortful to produce. We draw on a psycho-linguistically grounded translation models which explain these findings, and we investigate literality in machine translation post-editing.

In section 1, we start by describing the metrics by which we compute the literality of translations. The three literality criteria are quantified by a) the alignments links between ST and target TT tokens, where one-to-many and crossing alignment links result in less literal translations and b) the entropy and perplexity of translation alternatives. These metrics grade translations as more or less literal.

In section 2 we introduce a corpus of translation data which is the empirical grounding of our study. We extract a subset of more than 45000 words translated into 5 different languages produced by 145 different translators and investigate gazing behaviour on the source and the target text words.

Section 3 introduces a psycho-linguistically grounded, recursive model of translation production which explains automated and monitor processes based on a theory of shared combinatorial nodes, activated through bilingual priming processes.

Section 4 applies the recursive model of translation production to a six step translation process model (Jakobsen, 2011). We analyse the corpus of translation data with respect to alignment links and empirically underpin some of Jakobsen's translation steps. We show that literal translation processes play a crucial role in the analysis of these cycles.

Section 5 discusses a number of priming studies in bilingualism and investigates the corpus of translation data with respect to translation ambiguities and translation perplexity. The recursive model of translation production explains our observations of increased gaze activities for non-literal translations.

Section 6 presents a number of findings from machine translation post-editing and applies the model to translation post-editing. Section 7 points finally to open questions in user modelling. In all examples our quantification of literality correlates well with model predictions and observations in the data, and thus provides a powerful means to a assess translation processes.

Measuring translation literality

Figure 1 shows a literal translation according to these criteria. ST and TT words are translated and aligned one by one. If we further assume that in a corpus of translations all translators produce the same translation, i.e. all translators translate *'Peter → Peter'*, *'loves → liebt'* and *'Klaus → Klaus'* in the same order, then this would be an (ideal) literal translation.

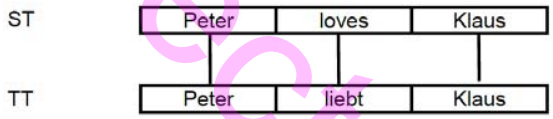


Figure 1: Ideal literal translation. Each ST word is translated into one TT word and hypothetical alternative translations are all identical.

In the next section, we describe examples which deviate from the literal translation and how we quantify non-literality. First, we describe a transducer to measure the similarity of word order in the ST and TT strings based on the length and number of crossing alignments. We then introduce the notion of translation entropy and translation perplexity, derived from the number and distribution of translation choices in a corpus of alternative translations to account for literality criterion 3.

Crossing Alignments

From a given translation and their alignment relations we compute CrossS and CrossT values on the ST and the TT sides respectively, by following the alignment links and counting the number of words between two successive alignments. We thus obtain a vector of relative distortions for ST and TT words, indicating the syntactic similarity of the two sentences.

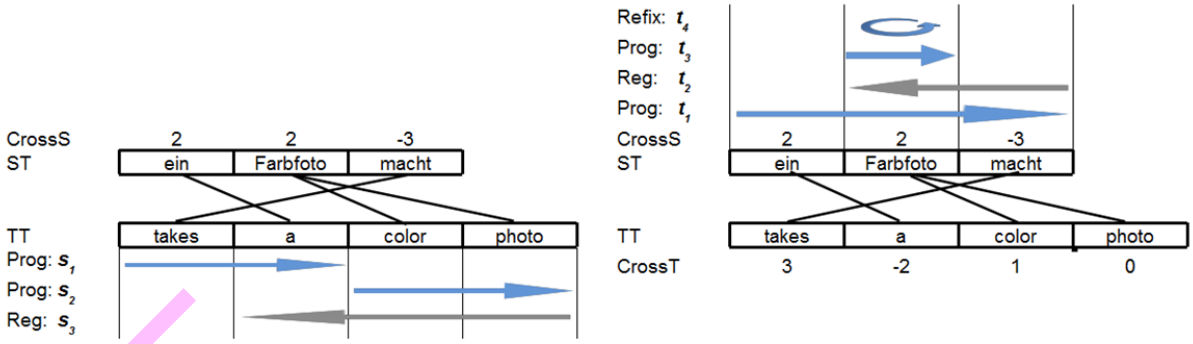


Figure 2: CrossS values (on the left) and CrossS and CRossT values (on the right). Arrows represent minimal progressions and regressions in the TT (for CrossS values) and in the ST (for CrossT values), while the word order is maintained.

Assume the German sentence [*Ich denke, dass Peter ein Farbfoto macht*] is translated into English as [*I think that Peter takes a color photo*]. An alignment of the last few words of this translation is shown in Figure 2.
In order to obtain a CrossS value for the first ST word [*ein*] we follow the alignment link to its translation [a], which is the second word of the English chunk, and thus, the CrossS value of [*ein*] equals 2. The next ST word [Farbfoto] translates into the compound noun [color photo], which requires a further progression of two words. The third ST word [macht] is aligned to the first TT word [takes] and thus a regression of 3 words to the left is required. The procedure thus produces the CrossS vector {2,2,-3} which represents the relative syntactic reordering of the German chunk into its English translation given above. The CrossS values index translation relations from the ST point of view, providing a metric for the syntactic difference of the TT compared to the ST. In other words, CrossS values describe the distance between the actual alignment and an ideal alignment in which every word has a CrossS value of 1. CrossS describes this distance by maintaining the ST word order and by reading TT words in the ST word order.

Figure 2 indicates the CrossS vector above the ST sentence, while the required movements in the TT in terms of progressions and regressions are shown below the TT sentence.
Crossing alignments can also be investigated from the target language point of view, with the aim to measure how different the ST is from the TT, and how long minimal progressions and regressions are necessary in the ST when mapping ST words into the order of the TT. The panel on the right in Figure 2 shows the same translation fragment as before, with CrossT values and progressions, regressions and re-fixations on the ST. The input device needs to scan three words in the ST, to find the translation [*macht*] of the first target word [*takes*]. Accordingly the CrossT value is 3. A regression of two words follows to map the second TT word [*a*]. The third and fourth TT words constitute the compound noun [*color photo*], which is the translation of one ST token. The assumption here is that [*Farbfoto*] produced 2 items in the target language, and therefore the input device does not need to be shifted when producing the second part of the compound [*photo*]. This is represented with a circle as (possible) refixation in Figure 2.

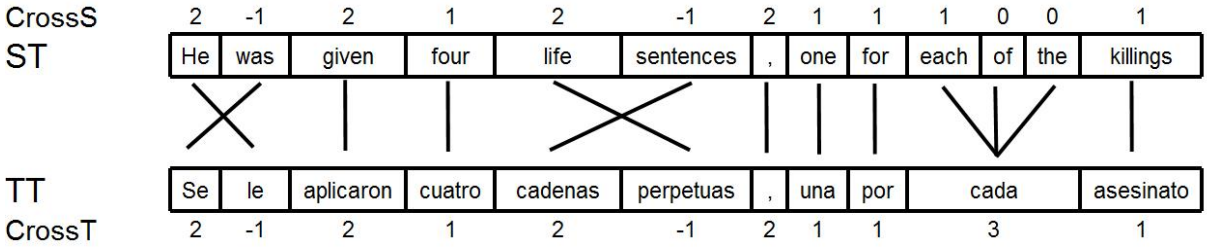


Figure 3: English-Spanish alignment with CrossS and CrossT values

An example of an English → Spanish translation is a given in Figure 3 together with the CrossS and CrossT values. Note that an alignment link is a property of a word in its translation context. A translator operates with both a representation of the ST and the TT. It is therefore likely that the translator aligns both the ST with the target text and also vice versa. For instance, in Figure 3 the word [*he*] has a CrossS value of 2, and its translation equivalent Spanish [*le*] has a CrossT value of -1. We may thus refer to [*le*] as a TT word with a CrossT value of -1 or as a translation of [*he*] with a CrossS value 2. Similarly, [*he*] can be looked at as a ST word with a CrossS value 2, or as a translation of [*le*] with a CrossT value of -1. As a translator switches back and forth between the two texts, s/he needs to develop strategies to cope with the two views. In section YYY we show that, depending on which view a translator takes, different behavioural patterns emerge. This switching back and forth between source and target, i.e. translating the source into the target and vice versa is what Ivir (1981) describes as a "check on meaning" or "back-translation". For a literal translation, the Cross values will always be 1, as shown in Figure 1.

Translation perplexity

Perplexity has been used in speech recognition (Jelinek et al, 1977) and machine translation to measure the appropriateness of language models and translation models (Popel, 2012; Sennrich, 2012). Jelinek et al. (1977) state that "...vocabulary size and static and dynamic branching factors are all inadequate" to measure the complexity of grammars. In language modelling, it is generally assumed that models with lower perplexity are also better computational models. The perplexity of a model is a measure that indicates how many different, equally probable words can be produced, and thus how many choices are possible at a certain point in time. The higher the perplexity, the more similarly likely choices exist and hence the more difficult is a decision to make. However, usually, models are preferred that minimize the number of possible choices, given they have the same explanatory power. Perplexity ***P*** is related to entropy ***H***, as an exponential function as shown in equation (1):

(1) $$P = 2^{H}$$

where entropy ***H*** represents the average amount of non-redundant information provided by each new item. The information of a probability ***p*** is defined as $\boldsymbol{I(p) = -\log_2(p)}$, and entropy ***H*** is the expectation of that information as defined in equation (2):

(2) $$H = \sum_{i=1}^{n} p_i I(p_i) = -\sum_{i=1}^{n} p_i \log_2(p_i)$$

We adopt this notion to assess the entropy of translation alternatives for a given ST word ***s*** into TT words $t_{i...n}$ as shown in equation (3):

$$H(s) = -\sum_{i=1}^{n} p(s \rightarrow t_i) * log_2(p(s \rightarrow t_i)) \quad (3)$$

Entropy ***H(s)*** is the sum over all observed word translation alternatives multiplied with their information content. The word translation probabilities $p(s \rightarrow t_i)$ of a ST word ***s*** and its possible translation $t_{i...n}$ are computed as the ratio of the number of alignments $s \rightarrow t_i$ counted in TTs over the total number of observed TT tokens, as in equation (4). Thus, while in language modelling, the entropy indicates how many possible continuations for a sentence exist at any time, we deploy the metric to assess how many different translations a given ST word has.

$$p(s \rightarrow t_i) = \frac{count(s \rightarrow t_i)}{\#translations} \quad (4)$$

For instance, consider the English sentence fragment '*he was given four*' in Table 1. The English sentence was translated into Spanish by 22 different translators. The translations were manually word-aligned, words are lower-cased, and identical word translations t_i are counted. For each translation $s \rightarrow t_i$ we thus obtain a number of occurrences in context, as indicated in the first column below the ST word in Table 1. Thus, [*he*] is translated 12 times as [*le*], 2 times as [*recibió*] 2 times it is not aligned to any target word (i.e. the symbol "---" indicates unaligned, not translated) etc. Based on this information we can approximate word translation probabilities using equation (4).

We divide the number of translation occurrences of ['*he→le*'] by the overall number of observed translations, i.e. 22 and obtain ***p('he→le') = 12/22 = 0.54***. Thus, based on our data, there is a chance of slightly more than 50% that [*he*] is translated into and aligned with [*le*] in the context of this sentence. The information content of this translation amounts to $I('he \rightarrow le') = -log_2(p('he \rightarrow le')) = 0{,}874$.

Table 1 shows the distribution of the observed translations for this English sentence fragment. Note that the number of different translations varies for each ST word. There are six different translations for [*he*], [*was*] has 12 different translations, [*given*] was aligned to 14 possible translations while [*four*] has only one translation. Also the distribution of translation choices is different for each word. The average amount of non-redundant information, as computed by entropy ***H*** in equation (3), takes into account not only the number of observed translations (i.e. the branching factor), but also the probability distribution of the different translation choices, and turns it into a number ≥0. Thus, as the probability of translation ***p('four→cuatro') = 1*** represents no additional information (no further translation being observed) and thus ***H('four') = I('four→cuatro') = 0*** and perplexity ***P('four') = 1***. In contrast, the word [*given*] has many more possible translation alternatives, entropy ***H('given') = 3,55*** and perplexity ***P('given') = 11,71*** indicates much more complex translation choices.

Source word *e*		**he**		**was**		**given**		**four**
perplexity ***P(e)***		*P('he') = 3,83*		*P('was') =8,35*		*P('given') = 11,71*		*P('four') = 1*
Distribution of translations	12	le	9	le	4	dieron	22	cuatro
	2	recibió	2	recibió	4	condenaron		
	2	---	2	dieron	2	recibió		
	4	se	2	condenaron	2	condenó		
	1	se_lo	1	se_condenó	1	sentenciado_a_condenas		
	1	se_le	1	se_condemnó	1	se_le_condemnó		
			1	lo_condenó	1	los		
			1	han	1	impuesto		
			1	fue	1	han_condenado		
			1	declarado	1	culpable		
			1	---	1	conenaron_a		
					1	condenó_a		
					1	aplicaron		
					1	---		

Table 1: English sentence fragment and its word translation alternatives into Spanish from 22 different translations. The table also shows the number of occurrences of each translation and the word translation entropy

Machine translation search graph perplexity

In analogy with the perplexity of human translations, we also compute the perplexity of possible translation choices in a machine translation system. State-of-the-art statistical machine translation systems encode possible translations in a so-called search graph. A search graph consists of nodes, which represent target language words, and transitions between successive nodes. Transitions are labeled with weights, which represent the costs. The task of a decoder is to find an optimum path through the search graph, which, hopefully, corresponds to the best translation in the search graph. While there is a plethora of ways to create and decode search graphs, we are here only interested in the entropy and perplexity of word translations that are encoded in a search graph.

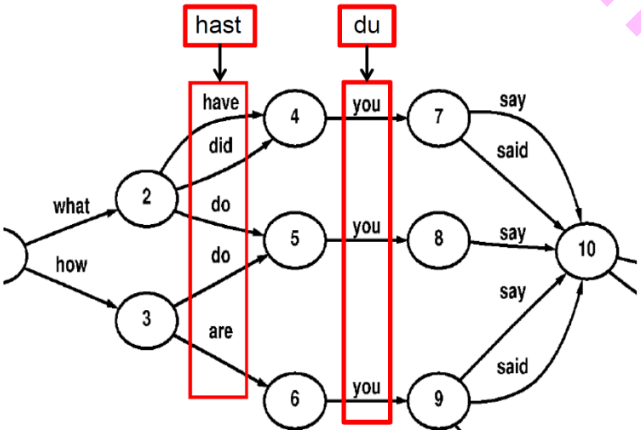


Figure 4: Machine Translation search graph (adapted from Och et al 2003).

Figure 4 shows part of a search graph which encodes several possible English translations of the German sentence: [*was hast du gesagt?*]. Two framed boxes in Figure 4 indicate the possible translations for [*hast*] and [*du*]:

hast → { have, did, do, are }
du → { you }

Similar to the previous example in which English [*four*] has only one observed translation [*cuatro*], the search graph in Figure 4 also shows only one possible translation for [*du → you*], and the entropy is therefore ***H('du') = 0*** and the perplexity ***P('du') = 1***.
Conversely, the entropy of [*hast*] in the search graph depends on transition probabilities from the preceding English words [*what*], and [*how*], to the translations {have, did, do} and {do, are} respectively. The entropy will be higher if the transition probabilities are similarly likely, and it will be lower the more uneven they are. In any case, the entropy of ***H('hast')*** and thus also the perplexity ***P('hast')*** will be higher than ***H('du')*** and ***P('du')*** for which the search graph encodes [*you*] as the only possible translation.

Experimental material

As a basis for our investigation we use a subset of the TPR-DB (Carl, 2012). The TPR-DB currently contains more than 1300 text production sessions (translation, post-editing, editing and copying) in more than 10 different languages. For each text production session, keystroke and gaze data was collected, stored, and translations were semi-automatically aligned using the YAWAT tool (German, 2008). STs and TTs were first automatically pre-aligned using the GIZA++ tool (Och and Ney, 2003). The data was then converted into the YATWAT format, and alignments were manually checked and amended where necessary. Once the alignments are manually checked and confirmed, the data is further processed into a set of tables which describe the process and the product by means of more than 80 different features.
Table 2 shows the properties of the six datasets used for the present purpose. For each of the six datasets, the table indicates the translation direction, the number of translations and translators, the number of sentences (Seg), ST and TT tokens (STT and TTT), as well as the ratio of ST and TT tokens per segment (STT/Seg and TTT/Seg) and ST tokens per target tokens (STT/TTT), the average CrossS value (CRS) and the average perplexity.

Each ST contains between 5 and 12 sentences, with an overall average of 7.3 sentences per text. Each ST was translated from between from between 3 (MS12) and up to 12 different translators (some texts in study BML12) . The STs contain between 111 and 176 words, on average, 146 tokens (words) per text (STT/Trans), which translate, on average, into 151 TT words.
However, not all languages behave in a similar manner. As can be seen from the table, with the exception of Chinese, there is a tendency to produce more tokens in the TT than there are in the ST sentences. Particularly translations into Spanish have 12% more words in the TT than in the ST.

TPR-DB study	language	translations	translators	Seg	Seg / Trans	ST Tokens	STT / Trans	TT Tokens	STT / Seg	TTT / Seg	STT / TTT	CRS	PP
KTHJ08	EN → DA	69	24	523	7.58	10571	153.20	10667	20.21	20.40	0.99	1.46	3.57
LWB09	DA → EN	40	18	329	8.23	5652	141.30	6206	24.68	27.10	0.91	1.62	2.36
ACS08	EN → DA	30	17	257	8.57	5085	169.50	5075	19.79	19.75	1.00	1.79	3.10
BML12	EN → ES	63	32	410	6.51	8936	141.84	10102	21.80	24.64	0.88	2.01	3.54
MS12	EN → ZH	15	9	92	6.13	2061	137.40	1916	22.40	20.80	1.08	2.05	2.17
SG12	EN → DE	47	24	304	6.47	6632	141.11	6777	21.82	22.29	0.98	2.45	4.09
NJ12	EN → HI	39	21	238	6.10	5505	141.15	5784	23.13	24.30	0.95	3.70	8.75
Total	sum	303	145	2153		44442		46527					

Table 2 Different language pairs and texts show different crossing values and translation perplexity.

Note also the difference with respect to cross values and perplexity. Cross values indicate the syntactic difference between the two texts and as all translations (apart from one) are from English into various different target languages we can see their relatedness from closest to more different DA, ES, ZH, DE, HI. On average, translations from English into Danish have the lowest positive and negative CrossS values. English into Chinese and Danish into English have the lowest perplexity, while German and Hindi have the highest perplexity scores and English into Danish and Spanish are situated in the middle. Perplexity is another indicator of conceptual relatedness, where a similar graduation can be seen. But since the values are based on a different number of translations per ST, a comparison across languages needs be treated with caution.

Priming in Translation

Priming describes the effect of a previously encounter item or structure on subsequent language use. The tendency to repeat aspects of previously encountered items or structures can be used to identify the nature of the representations at play during language production or comprehension. Participants in priming studies are not normally aware of the manipulations which lead to repetition. While the bulk of priming studies has investigated monolingual priming (cf Pickering and Ferreira 2008; McNamara 2005), there have been a number of studies which investigate priming between different languages (e.g. Duñabeitia, Perea, et al. 2010; Bernolet et al. 2013). Structural priming typically describes the effect of the structure of a sentence in one language on the structure in a sentence in a different language and typically, these two sentences are completely unrelated in terms of their semantic content. Nonetheless, what cross-linguistic priming studies show is that semantic and structural representations are shared across languages. During translation, the two sentences can be said to be highly related and the translation-equivalent boost observed when only one word in the two sentences is a translation equivalent (e.g. Schoonbaert et al 2007) is likely to become very strong during translation. Schaeffer and Carl (2013) present a translation production study which clearly shows semantic and structural cross-linguistic priming during translation from English into French and into German. Schaeffer and Carl (2013, 178) suggest that "...translators encode the text during … translation in a way that favors the activation of shared representations..." In another translation production experiment, Jensen et al. (2010) observe that translators gazed significantly longer at segments for which the ST word order had to be reversed in the TT. They suggest that shorter gaze times on the ST for syntactically similar

constructions in the ST and TT are due to a priming effect, while the lack of priming effects in segments which require re-ordering in the TT leads to longer gazing times.

In order to explain the results of priming studies, Pickering and Branigan (1998) propose a model of *combinatorial nodes* in which lemmas are associated with syntactic information, such as part-of-speech, number or gender, etc.. Combinatorial nodes are purely syntactic in nature (rather than mediated by meaning) and are activated whenever the speaker uses a particular construction. Priming implies thus an activation of combinatorial and lemma nodes and their links. Successive activation of the same nodes (through similar verbal input) strengthen their activation and links, while different input reduces the priming effect.

In an attempt to account for bi- and multilingualism, Hartsuiker et al (2004) extend Pickering and Branigan's (1998) monolingual model of combinatorial nodes into a shared syntax account, in which combinatorial nodes are "...connected to all words with the relevant properties, irrespective of language..." (Hartsuiker et al. 2004, 481) Given that not all languages always share the same linguistic properties, combinatorial nodes which are shared between the two languages are specific to that language combination. The model can thus explain how and why items and constructions perceived in one language can occur - and preferably are - reproduced in another language - a phenomenon that we previously referred to as literal translation.

The model of shared combinatorial nodes is also supported by recent neuroimaging findings. Buchweitz and Prat (2013) report that there is a large overlap in brain activation between two languages in proficient bilinguals: "neuroimaging and behavioral research alike show that there is a shared semantic representation in bilinguals, that is, shared concepts and shared cortical tissue" (Buchweitz and Prat, 2013, 430)

Schaeffer and Carl (2013) further develop and adapt this model for translation production, incorporating an additional monitoring process (Ivir 1981; Toury 1995; Tirkkonen-Condit 2005). They distinguish between horizontal priming processes which activate shared combinatorial nodes, and vertical problem-solving processes which act as a monitor during target text production. Both processes complement each other and are active at the same time during translation. Shared combinatorial nodes are activated early during source text reading and serve as a basis for regeneration in the target language. Shared combinatorial nodes allow the target text production to go on almost automatically, until it is interrupted by the monitor, if the produced text violates target text norms or contextual considerations of the vertical processes:

> ...the monitor needs to compare whether the source is the same as the target, but it is equally important to make sure that the target is the same as the source. Vertical processes access the output from the automatic default procedure recursively in both the source and the target language and monitor consistency as the context during translation production increases. (Schaeffer and Carl 2013, 186)

A translation process model

Schaeffer and Carl's model (2013) predicts a process of recursive ST and TT priming, interrupted by monitor activities. Monolingual monitoring is more likely and priming less likely in positions where ST

and TT differ, and where a literal translation is not possible or acceptable. Jakobsen (2011, 48) conceptualizes translation production from a slightly different - although compatible - angle, as a cycle of cognitive and motor processes that involve the following six steps:

1. Reading of a source-text chunk (and constructing a translation of it)
2. Reading the current target-text anchor word(s)
3. Typing the translation of the source-text chunk
4. Monitoring the typing of the chunk of target text and the screen outcome
5. Moving the gaze toward the next target item(s) in the source text
6. Reading source-text anchor word(s)

Jakobsen (2011, 48) stresses that not all steps are necessary and some may be repeated. The word "Monitoring" in point 4 of the list above needs to be distinguished from *the monitor* in the Schaeffer and Carl's model (2013): while *the monitor* in Schaeffer and Carl's model is a cognitive process that controls the outcome of priming processes, *monitoring* in Jakobsen's translation cycle refers to gazing at the emerging target text. However, the monitor (as proposed by Schaeffer and Carl) will react based on the assessment of the monitoring activity, possibly interrupting typing activity and resulting in a jump to one of the other states enumerated in Jakobsen's list, as discussed above. In the analysis of our data we find evidence for some of the steps which we describe as follows.

A B

Figure 5: horizontal: word alignment distance; vertical: gazing time in ms per Character. GazeS: gaze on source word, GazeT: gaze on target word, CrossS source → target alignment distance, CrossT: target → source alignment distance.

Our analysis is based on the gazing duration of translators in relation to the syntactic similarity (alignment distance) of the source and the target language as represented in Figure 5, GazeS refers to the total reading time (henceforth TRT) on a particular ST word or character. GazeT refers to TRT on the word(s) or character(s) which are aligned to a particular ST word. Both GazeS and GazeT were normalised by the number of characters in both ST and TT strings in order to control for word length and frequency effects - long words tend to be less frequent than short words and word length and

frequency have been shown to have strong effects on eye movements during reading (e.g. Inhoff and Rayner 1986).

1. A translator starts reading a source text: combinatorial nodes are activated and automatic pre-translation is initiated, pre-selecting possible translations. Chunks of the source text words which map monotonously onto the hypothesised (or partially typed) TT are processed quicker than those which require syntactic reordering. As discussed above, in instances of longer syntactic reordering, priming activities are suppressed and a more effortful translation process starts.
The graph **GazeS_Per_Character** in panel A in Figure 5 illustrates this pre-processing. It shows that TRT per character on ST words (GazeS, vertical) correlates with the ST alignment crossing distance (CrossS): GazeS durations are shortest for words which translate one-to-one into the target language (i.e. CrossS value is 1). For words with a greater (both positive and negative) syntactic reordering distance (higher CrossS values), TRT on ST words increases. This corresponds top Jakobsen's step 1.

2. The translation is typed and TT is read (GazeT). While monitoring the production of TT words, combinatorial nodes are activated, this time from the target text, and the mind maps the produced TT words back onto equivalent ST items, constantly controlling whether the emerging TT is equal to the ST chunk. Here too, priming effects are stronger if the ST and the emerging TT are syntactically similar, while longer syntactic distances between the TT and ST lead to longer GazeT.
The graph **GazeT_Per_Character** in panel B in Figure 5 reflects this. Again, TRT is shortest when ST words are translated one-by-one, and if the syntactic reordering distance to the left increases and especially to the right (positive CrossT values) TRT on the TT words also increases. This corresponds to Jakobsen's steps 2, 3 and 4.

3. In yet another processing step, which may best correlate with step 5 and 6 in Jakobsen's translation cycle, the translator verifies the produced translation and switches visual attention again back to the ST, thereby gazing at the ST segment which corresponds to the current TT words (i.e. the ST anchor word). The gaze is thus observed on the ST (GazeS), but this time following the alignment links from the TT (CrossT).
The graph **GazeS_Per_Character** in panel A in Figure 5 shows TRT when reading the ST in a TT revision mode: again less literal translations (longer syntactic distance, multi-word translations etc.) produces longer TRT on the ST words than monotonous translations.

4. From here on a recursive revision process starts in which the translator increasingly takes more ST and TT context into consideration, which is increasingly monitor-driven, and which is difficult to trace in our statistical analysis. However, one more revision loop can be observed when the translator switches in the revision mode from ST reading back to the TT. In this mode, the translator reads the TT (GazeT) following the ST → TT alignment links, checking whether the meaning of the previously produced TT sequence corresponds to ST words. The graph **GazeT-Per_Character** in the panel A in Figure 5 quantifies priming strength of this second order revision: the reduction of priming effects (i.e. longer total reading times) are only significant for regressions to TT words with CrossS values < 0.
As confirmed in other studies (Sharmin et al 2008), translators spend more time reading the TT than the ST. The strongest monitoring effect occurs during first drafting in step 2 (Figure 5 panel B). Mapping a just produced TT word onto a word ahead in the ST requires relatively more monitoring than mapping onto a past ST context. Presumably this is the case, because it requires integration of a larger amount of possibly not yet translated text in addition to possibly not yet processed ST. For all other processing steps regressive gaze movements were found to be more time consuming.

Separate simple linear regressions of the graphs in Figure 5 for negative and positive Cross values from 1 to the peak in each distribution were performed. Negative CrossS values significantly predicted GazeS (ß = -19.27, t(9) = -4.02, p < 0.01). The overall model also predicted GazeS quite well (adjusted R2= 0.60, F(1, 9) = 16.17, p < 0.01). Positive CrossS values significantly predicted GazeS (ß = 19.518, t(6) = 3.262, p < 0.05). The overall model also predicted GazeS quite well (adjusted R2= 0.58, F(1, 6) = 10.64, p < 0.05). Negative CrossS values significantly predicted GazeT (ß = -35.58, t(7) = -4.31, p < 0.01). The overall model also predicted GazeT rather well (adjusted R2= 0.69, F(1, 7) = 18.55, p < 0.01). However, positive CrossS values did not predict GazeT. Negative CrossT values significantly predicted GazeS (ß = -21.328, t(7) = -2.97, p < 0.05). The overall model also predicted GazeS relatively well (adjusted R2= 0.49, F(1, 7) = 8.81, p < 0.05). Positive CrossT values significantly predicted GazeS (ß = 27.049, t(6) = 5.25, p < 0.01). The overall model also predicted GazeS very well (adjusted R2= 0.79, F(1, 6) = 27.59, p < 0.01). Negative CrossT values significantly predicted GazeT (ß = -43.97, t(6) = -4.02, p < 0.01). The overall model also predicted GazeT quite well (adjusted R2= 0.68, F(1, 6) = 16.15, p < 0.01). Positive CrossT values predicted GazeT (ß = 89.68, t(5) = 6.75, p < 0.01). The overall model also predicted GazeT very well (adjusted R2= 0.88, F(1, 5) = 45.57, p < 0.01).

Translation alternatives

A number of studies (Tokowicz and Kroll 2007; Laxén and Lavaur 2010; Prior et al 2013; Eddington and Tokowicz 2013; Boada et al 2013) show that translation recognition, as well as translation production is slowed down if a word has more translation alternatives. In a translation recognition task, Eddington and Tokowicz (2013) presented unambiguous translations, synonym translation-ambiguous, and meaning translation-ambiguous source language words, and further separate translation-ambiguous words into dominant and subordinate translations. Bilingual participants were presented with English–German word pairs that were preceded by a related or unrelated prime and were asked to decide if the word pairs were translations. They found that translation ambiguity slows down translation recognition regardless of the source of ambiguity (synonym translation-ambiguous, or meaning translation-ambiguous). Participants were slower and less accurate to respond to words that had more than one translation compared to unambiguous words. Contrary to their initial hypothesis, Eddington and Tokowicz could not confirm that dominant translations were responded more quickly than subordinate. However, priming was significant only for the meaning translation-ambiguous words, where a related prime can facilitate translation recognition speed compared to an unrelated prime.

Prior et al (2011) compare single-word decontextualized translation choices made by bilingual speakers of English and Spanish with contextualized translation alternatives, extracted from translations that were created by professional translators. Prior et al (2011, 98) assume that translation forms which "...are most frequently appropriate in contextual real life translation should also be the strongest in translation out of context…" However, they find that the translation in and out of context overlap to some degree, but translation out of context "...by no means matches it [translation in context] perfectly…" (Prior et al 2011, 108 [our brackets]) Form similarity is a stronger predictor in decontextualized translation choice, whereas word frequency and semantic salience are stronger predictors for context-embedded translation choice.

Dragsted (2012) compared eye movement measures (TRT and number of fixations) and pauses for words which were translated by 8 participants using the same target word with words for which the

eight participants used different words. She found that the TRT and the number of fixations on words with many (5-8) alternatives target text items was significantly higher than the TRT and the number of fixations on words with only one or two different target items. She also found that the pauses prior to critical words were longer for words with many alternatives as compared to words with one or two alternatives.

Rather than investigating words out of context, we investigate translation production in context, similarly to Dragsted (2012). We investigate a translation production task and measure the TRT on ST and TT words. As shown in Table 2, translations are aligned in real contexts, which leads to many more correspondences then would be usually listed in a bilingual lexicon. Accordingly, the branching factor is much higher for translation ambiguous words than the two alternative translations used in the study by Eddington and Tokowicz (2013).

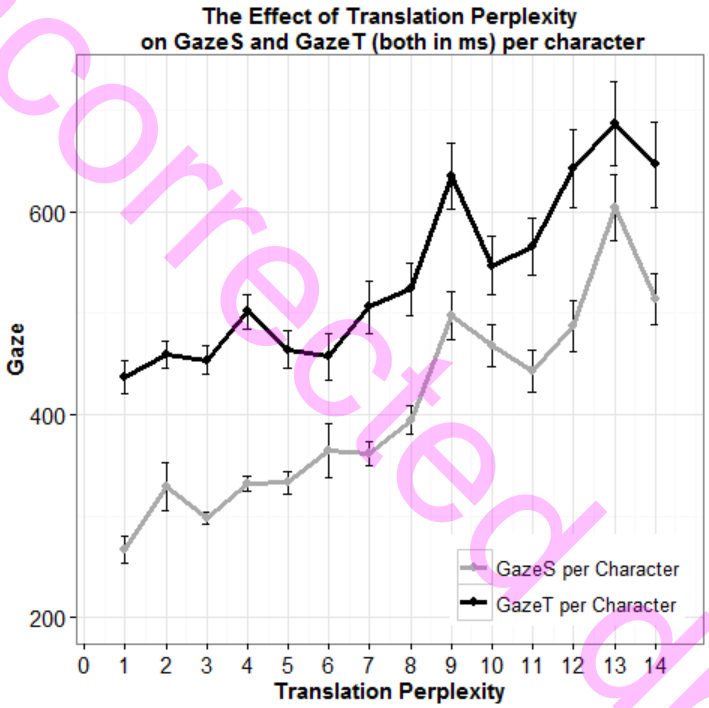


Figure 6: Correlation of from-scratch word translations perplexity and TRT (Gaze) on the 44442 source text words and their 46527 translations as in Table 2.

As argued for previously, perplexity is a better measure than simply the number of alternatives, because it takes into account the distribution of the translation alternatives. Figure 6 shows the correlation between word perplexity and TRT on ST and TT words. While TRT is on average longer on TT words (translations) than on ST words, it is interesting to observe that TRT increases with the perplexity of the translation alternatives. That is, just as for translation recognition, translation ambiguities also slow down the translation task, when reading the source text and also when gazing at the ambiguous target language words.

A simple linear regression showed that translation perplexity significantly predicted GazeS (ß = 21.56, $t(12) = 8.74$, $p < 0.001$). The overall model also predicted GazeS rather well (adjusted R2= 0.85, F(1,

12) = 76.36, p < 0.001). A second simple linear regression showed that translation perplexity also significantly predicted GazeT (ß = 18.32, t(12) = 7.41, p < 0.001). The overall model also predicted GazeT rather well (adjusted R2= 0.81, F(1, 12) = 54.89, p < 0.001).

Machine translation post-editing

Francis et al (2014) investigate the influence of language proficiency and item difficulty on translation response time. Translating into the more-proficient language (L1) is usually faster and more accurate than translating into the less proficient language (L2). This translation asymmetry, is stronger in less balanced bilinguals, and weaker in more balanced bilinguals. Less-familiar and less frequent words were more difficult to translate than more familiar and more frequent words. Francis et al also find that bilinguals who are more dominant in their L1 "...benefit more from priming in the L2 than in the L1…" (2014, 38), indicating that MT post-editing might be more helpful for less experienced translators.
A recent English → German post-editing study (Carl et al 2014) supports this assumption: 12 professional translators and 12 non-professionals (students), all of them German native speakers post-edited and translated from scratch four texts from English (L2) into German (L1). While they were almost always faster during post-editing than in from-scratch translation, 75% of the students were somewhat or highly satisfied with the post-editing task, compared to 54% of the professionals. These findings are confirmed by a large number of similar studies, showing that post-editing is, on average, faster than translating from scratch, even for experienced translators (see, e.g. Skadins 2011; Moran et at, 2014; German et al, 2014).

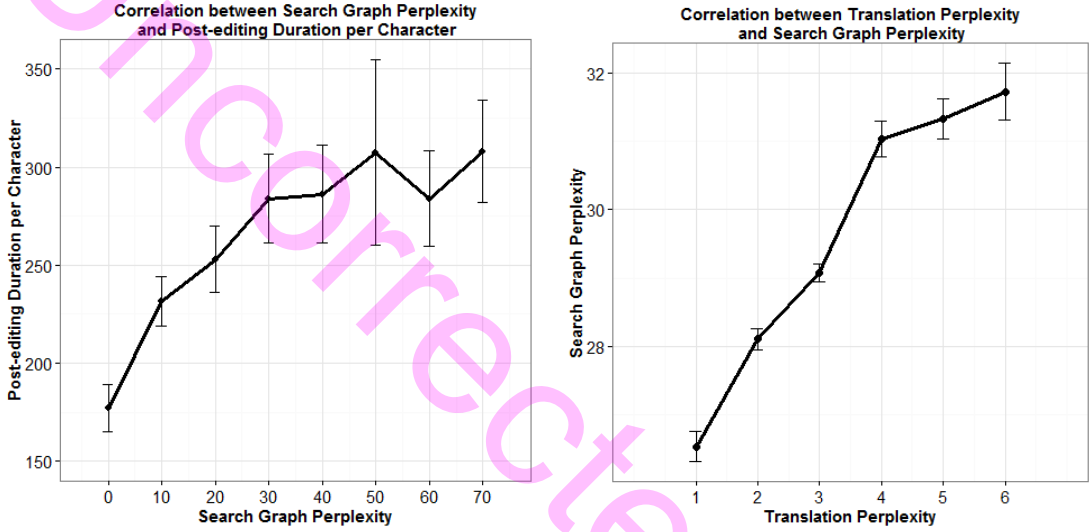
More controversial are the results with respect to final translation quality. While Depraetere et al (2014) find that post-editing does not have a negative impact on the quality of the final translation, Čulo et al (2014) observe interference phenomena from the source language in post-edited translations, i.e. grammatical or lexical structures of the source language that carry over into the target language (cf. the law of inference, referred to in the introduction). Numerous inference phenomena occur in MT output, as up-to-date machine translation systems do not produce idiomatically correct output in all cases, which remain visible in the post-edited translations if MT output is not sufficiently corrected.

For example, Čulo et al (2014) report that in the context of the English → German translation "***In a gesture*** *sure to rattle the Chinese Government ...* **→** ***In einer Geste****, die die Chinesische Regierung wachrüttelt ..."* the German expression "*In einer Geste"* is understandable, but literal and unidiomatic. It is a one-to-one translation, produced by the MT system, which was often not amended during post-editing. However, during from-scratch translations more idiomatic expressions, such as *"Als Geste"*, *"Es ist eine Geste"*, *"Mit der Absicht"*, *"Als Zeichen des Widerstandes"*, *"Mit einer Aktion"* would be produced.

Although earlier works (Underwood et al, 2001; Bernth and Gdaniec 2002) consider non-finite verbs and potentially ambiguous parts-of-speech particularly difficult to translate (in a Rule-based MT system), and thus more time-consuming to post-edit, Aziz et al (2014) find that cognitively more difficult and time consuming errors are incorrect part-of-speech, untranslated words, errors related to idiomatic expressions and word order, especially when reordering crosses phrase boundaries, while errors which require the least effort to post-edit include word form errors, synonym substitutions, and simple incorrect word substitutions with correct part-of-speech. In particular, they find that - contrary to what has been expected earlier - gerunds and other non-finite verbs require less post-editing time

than modal verbs, for example. Also Aranberri-Monasterio and O'Brien (2009) find that machine translations of English '-ing' forms into various languages are not as problematic for post-editing as they originally appeared to be.

Similar findings are reported by Čulo et al (2014), who count - in their English → German post-editing experiments of SMT output - more fixations on finite clauses as compared to non-finite ones. Note that German non-finite clauses - in contrast to finite clauses - require a verb-final position, which leads to crossing alignments similar to the situation discussed in Figure 7, and the experiments conducted by Jensen et al (2010) mentioned above. Despite a necessary syntactic reordering for English → German subordinate clauses, according to Čulo et al, the MT quality is better for the non-finite clauses than for finite clauses, and hence less interference effects or MT errors are observed.



A B

Figure 7: Correlation between word perplexity of the MT search graph and post-editing duration per character (Panel A). Correlation between Word translation perplexity in the post-edited output and word perplexity of the MT search graph (Panel B)

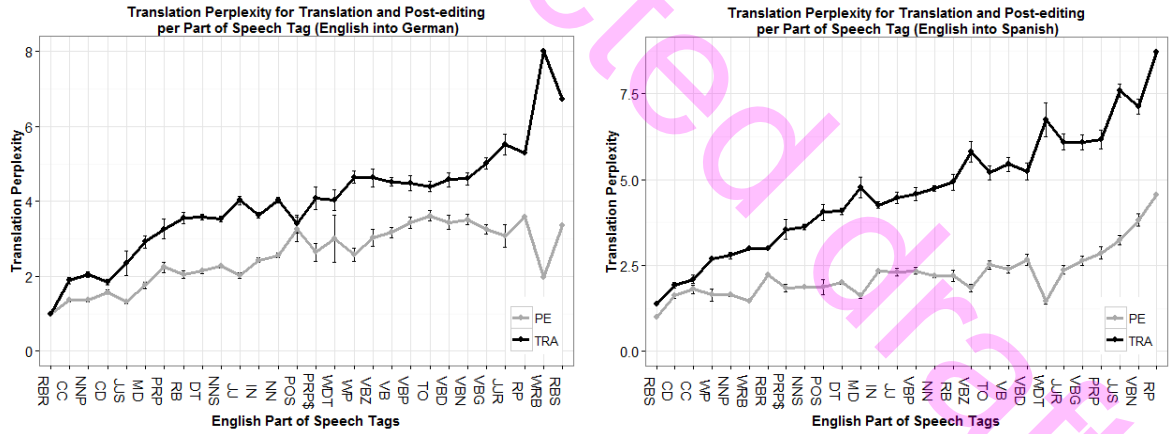
A simple linear regression showed that search graph perplexity significantly predicted post-editing duration per character (ß = 1.60, $t(6) = 4.60$, $p < 0.01$). The overall model also predicted post-editing duration per character rather well (adjusted $R2 = 0.74$, $F(1, 6) = 21.15$, $p < 0.01$). A further simple linear regression showed that translation perplexity also significantly predicted search graph perplexity (ß = 1.07, $t(4) = 7.93$, $p < 0.01$). The overall model also predicted search graph perplexity very well (adjusted $R2 = 0.93$, $F(1, 4) = 62.95$, $p < 0.01$).

Our investigation shows that post-editing duration is closely correlated with the perplexity of the MT search graph: the more similarly possible translation alternatives an MT search graph encodes the more post-editing time increases. Note that post-editing duration can be seen as an indicator for the MT quality (Aziz et al, 2014). We may conclude that search graph perplexity correlates with translation quality and hence with gaze duration.

Panel A in Figure 7 depicts this correlation on the basis of data acquired in the 3rd CasMaCat field trial (Carl, 2014). This effect can be understood on basis that statistical MT systems are trained on

translations produced by humans and, given that this training material contains more translation ambiguities, it is only natural that the search graph perplexity increases. On the other hand, the effect of higher search graph perplexity is also reflected in the variety (i.e. perplexity) of the post-edited translations. Panel B in Figure 7 shows the correlation between MT search graph perplexity and perplexity of post-edited translations. That is, there is a transitive relation of translation ambiguities in the material with which SMT systems are trained via the perplexity of the search graph when the SMT system produced a translation, to the finally post-edited translations.

However, the degree of perplexity between the training material and the MT output (the post-edited translations) decreases during post-editing, more than it would do during from-scratch translation. That is, a post-editor usually only sees only one single best translation that an MT system produces and amends this output to become a suitable translation. As previously discussed, a post-editor is primed during this process so that s/he more easily accepts sub-optimal translations which human translators, working from scratch, would otherwise not produce. Graphs in panel A and B in Figure 8 plot perplexity values of English → German (A) and English → Spanish (B) translations for different part-of speech values. The word translation perplexity values the graph in panel A in Figure 8 are based on eight from-scratch translated versions and eight post-edited versions of a set of English source texts amounting to approximately 800 source text words (SG12 study). The graph plots perplexity values per PoS (http://www.monlp.com/2011/11/08/part-of-speech-tags/) tag. Some PoS tags, JJS (superlative adjective), NNP (Proper names), CC (conjunctions) produce very few translation alternatives, and during post-editing they are almost always accepted. Other PoS tags, such as RP (particle), VBN (participle) etc and produce more variants in the target language. Note, however, that the perplexity of the post-edited texts are in all cases smaller than in the versions that were from-scratch translated.



A B

Figure 8: Perplexity of English → German (A) and English → Spanish (B) translations per word class and for from-scratch translation (TRA) and for post-editing (PE).

A similar picture is provided in the graph in panel B in Figure 8, which is based on the same English source texts and which have also been translated and post-edited into Spanish by eight different translators. The number of words is thus very similar to the German translations in Figure 8, but perplexity values are slightly lower (cf also Table 2), and always lower for post-editing than for translation from-scratch. Note also that some PoS tags imply more variation in the target text, as

compared to the German situation. For instance superlative adjectives (JJS) are almost never touched during post-editing of the German translations while there are many translation alternatives in Spanish.

Post-editing machine translation results thus in a more literal translation than from-scratch translation. The distribution of Cross values is almost identical in the combined German and Spanish post-edited and from-scratch translated texts: the difference between the percentages of CrossS values -8 to 8 for post-editing and translation from scratch were below 1%, apart from CrossS=0 and CrossS=1. The post-edited texts showed 3.6% fewer words with CrossS=0 and 6.7% more items with CrossS=1, as compared to post-edited versions of the same texts, indicating a compositional translations of expressions during post-editing that would be transferred more idiomatically during from-scratch translation. An example of such an unidiomatic but understandable translation during post-editing was shown above.

With respect to lexical variation post-edited texts do not reach the richness of from-scratch translated texts. As the amount of post-edited translations constantly increases, and semi-automatically produced text already surpasses the amount of from-scratch translation in some domains, we assume that eventually SMT systems will also be trained on post-edited translations. On the one hand, this will likely lead to better machine translation quality since the training material is less ambiguous, and on the other hand, as post-editors tend to accept also non-idiomatic translations, the resulting post-edited translations will become increasingly literal with reduced word translation perplexity. We are thus - at least for certain types of texts - in a global text-recycling loop which will lead to increasingly more literal translations as the turnover of language content on the web is cloned and only occasionally renewed in ever shorter intervals.

Post-editing styles

The post-editor is thus a facilitator in this global text recycling scenario, and must be equipped with appropriate working tools. There are a number of tools for computer assisted translation and machine translation post-editing, but there are fewer studies on user profiles and how translators actually use those tools. In order to describe different post-editing styles, Mesa-Lao (2013) suggests a post-editing cycle, similar to Jakobsen's translation cycle, which also consists of six steps:

1. Read source text segment
2. Read target text segment
3. Detect MT error in the target segment
4. Read source text segment
5. Read previous source text segment(s)
6. Fix MT error in target text segment

While there are potentially many different combinations of these processing steps, Mesa-Lao (2013) points in particular to four different post-editing cycles, which he detects in the UAD of six professional post-editors, and which he refers to as style$_1$:2 → 3 → 4 → 6, style$_2$: 1→ 2 → 3 → 6 style$_3$: 2 → 3 → 6 and style$_4$: 2 → 3 → 4→ 5 → 6. Post-editing style$_1$ seems to be the most preferred amongst his participants: the post-editor starts with reading the target text. When detecting an MT error, the post-editor cross-checks in the source segment, then corrects the error and a next post-editing cycle starts. Different post-editing styles seem to be preferred in different kinds of GUIs and by different post-editors. Mesa-Lao mentions two extreme styles:

- Style$_3$: The post-editor reads only the target text segment before fixing an MT error, and does not consult the source. This style seems cognitively the easiest post-editing pattern
- Style$_2$: The post-editor reads the target text and the source text segments and in addition refers back to previous segments. This seems cognitively more difficult

However, more post-editing styles are conceivable, for instance in cases where a post-editors accepts a target segment without any change. It might also be significant if a post-editor switches attention back and forth between the source and the target text etc. More research is required to get a better understanding of different post-editing styles and their implications. For instance: how can we explain the variance within one post-editor's behaviour and between different post-editors? Do different styles correlate with different degree of experience? To what extent are they triggered by source text properties (e.g. MT output errors), or by translator profiles? How (reliably) can we automatically detect different post-editing styles? To what extent do different post-editing styles correlate with a different post-editing brief? For instance, light post-editing may require translators to "temporarily abandon their high standards" (Wagner 1985, 2) with respect to "quality and professionalism", but - despite almost 30 years of post-editing practice - from an empirical point of view we do not know how to assess and measure this.

On the one hand, Doherty et al. (2010) show that gaze time and fixation count correlate well with quality MT output, so that worse translations trigger longer gaze time and more fixations. On the other hand, as we have shown in this chapter, machine translation output primes translation production. However, Martínez et al (2014) find that typing events contain more discriminative information to recognize the duration of translator training than fixations. Perhaps, as Aziz et al (2014) show, a sub-sentential analysis of post-editing problems and post-editing behaviour may be a step forward to finding answers to these questions.

Conclusions

In this chapter we try to bridge the gap between psycholinguistic studies, translation studies and user studies. In the context of translation studies, 20 years ago, Toury developed an interesting research framework in which he suggests:

> Since there is no inherent need for intertextual relationships [of translations and their (assumed) source texts] to always be of the same kind or intensity, the nature and extent of these relationships, as well as their correspondence to the culture's attitudes, constitute just another set of questions, to be settled through concrete research work. (1995: 144)

Twenty years later, we have now the possibility to study these relationships in a much more diverse qualitative and quantitative way than it could have possibly anticipated by Toury 20 years ago. Due to numerous studies in psycholinguistics and translation process research we have now a much more profound understanding of the bilingual's cognitive processes, and the encodings and representations that allow humans to process two different languages simultaneously. These insights have led to more advanced falsifiable theoretical frameworks, models and theories. At the same time, we have large databases of translators activity data which enable us to elicit increasingly more entangled aspects of translation processes (from scratch and post-edited), and to underpin and refine theoretical predictions.

As an empirical science, we anticipate that translation process studies will become more formalized and more mathematical: observations are interpreted as empirical evidence to validate predictions in a theoretical framework. According to wikipedia (http://en.wikipedia.org/wiki/Empirical_evidence), empirical evidence "requires rigorous communication of hypothesis (usually expressed in mathematics), experimental constraints and controls (expressed necessarily in terms of standard experimental apparatus), and a common understanding of measurement."
In addition, with the increasing usage of computer assistance, translation technologies in the everyday working practice of professional translators, and the ease of worldwide communication and information assimilation due to ready-made availability of machine translation for many language pairs, empirical translation process research needs to formulate new and innovative research programmes at the interface of human-computer interaction which are suited to answer some of Toury's questions.

According to Dillinger (2014), a key ability for post-editors (and translators) is their "ability to compare sentences (and texts) across languages, in terms of both literal meaning and the culturally determined patterns of inference and connotation that different phrasings will entail".
We show that the translation literality criterion elaborated in this chapter provides a suited metric to assess which typological differences make it difficult to judge equivalence across languages.

Acknowledgements

This work has been supported by the CasMaCat project, co-funded by the European Union under the Seventh Framework Programme, project 287576 (ICT-2011.4.2). We are grateful to all contributors to the TPR-DB for allowing us the use of their data.

References

Aranberri-Monasterio, Nora, and Sharon O'Brien. 2009. "Evaluating RBMT Output for -ing Forms: A Study of Four Target Languages." *Linguistica Antverpiensia*, New Series – Themes in Translation Studies, 8: 105–122.

Aziz, Wilker , Maarit Koponen, and Lucia Specia. 2014. "Sub-sentence Level Analysis of Machine Translation Post-editing Effort." In *Post-editing of Machine Translation: Processes and Applications*, ed. by Sharon O'Brien, Laura Winther Balling, Michael Carl, Michel Simard, and Lucia Specia, 170–200. Newcastle: Cambridge Scholars Publishing.

Bernolet, Sarah, Robert J. Hartsuiker, and Martin J. Pickering. 2013. "From Language-specific to Shared Syntactic Representations: The Influence of Second Language Proficiency on Syntactic Sharing in Bilinguals." *Cognition*, 127(3): 287–306.

Bernth, Arendse, and Claudia Gdaniec. 2002. "MTranslatability." *Machine Translation* 16: 175–218.

Boada, Roger , Rosa Sánchez-Casas, José M. Gavilán, José E . Garciá-Aleba, and Natasha Tokowicz. 2013. "Effect of Multiple Translations and Cognate Status on Translation Recognition Performance of Balanced Bilinguals." *Bilingualism: Language and Cognition* 16 (1): 183–197. Cambridge: Cambridge University Press.

Buchweitz, Augusto, and Chantel Prat. 2013. “The Bilingual Brain: Flexibility and Control in the Human Cortex.” *Physics of Life Reviews* 10: 428–443.

Campbell, Stuart. 2000. “Choice Network Analaysis in Translation Research.” In *Intercultural Faultlines. Research Models in Translation Studies I. Textual and Cognitive Aspects*, ed. by Maeve Olohan, 29–42. Manchester: St Jerome.

Carl, Michael, 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.

Carl, Michael, Silke Gutermuth, and Silvia Hansen-Schirra. 2014. “Post-editing Machine Translation - a Usability Test for Professional Translation Settings.” In *Psycholinguistic and Cognitive Inquiries in Translation and Interpretation Studies*, ed. by John W. Schwieter, and Aline Ferreira. Newcastle: Cambridge Scholars Publishing. (in print)

Catford, John C. 1965. *A linguistic Theory of Translation: An Essay in Applied Linguistics*. Oxford: Oxford University Press.

Chesterman, Andrew. 2011. “Reflections on the Literal Translation Hypothesis.” In *Methods and Strategies of Process Research: Integrative approaches in Translation Studies,* ed. by Cecilia Alvstad, Adelina Hild, and Elisabet Tiselius, 23–35. Benjamins Translation Library Volume 94. Amsterdam and Philadelphia: John Benjamins Publishing Company.

Čulo, Oliver, Silke Gutermuth, Silvia Hansen-Schirra, and Jean Nitzke. 2014. “The Influence of Post-Editing on Translation Strategies.” In *Post-editing of Machine Translation: Processes and Applications*, ed. by Sharon O'Brien, Laura Winther Balling, Michael Carl, Michel Simard and Lucia Specia, 200–219. Newcastle: Cambridge Scholars Publishing.

Depraetere, Ilse, Nathalie De Sutter, and Arda Tezcan. 2014. “Post-Edited Quality, Post-Editing Behaviour and Human Evaluation: A Case Study.” In *Post-editing of Machine Translation: Processes and Applications*, ed. by Sharon O'Brien, Laura Winther Balling, Michael Carl, Michel Simard, and Lucia Specia, 78–109. Newcastle: Cambridge Scholars Publishing.

Dillinger, Mike. 2014. “Introduction.“ In *Post-editing of Machine Translation: Processes and Applications*, ed. by Sharon O'Brien, Laura Winther Balling, Michael Carl, Michel Simard, and Lucia Specia, IX–XV. Newcastle: Cambridge Scholars Publishing

Duñabeitia, Jon A., Manuel Perea, and Manuel Carreiras, 2010. “Masked Translation Priming Effects with Highly Proficient Simultaneous Bilinguals.” *Experimental Psychology*, 57(2): 98–107.

Doherty, Stephen, Sharon O’Brien, and Michael Carl. 2010. “Eye Tracking as an MT Evaluation Technique”. In *Machine Translation* 24(1): 1–13.

Dragsted, Barbara. 2012. “Indicators of Difficulty in Translation — Correlating Product and Process Data.” *Across Languages and Cultures* 13(1): 81–98.

Eddington, Chelsea M., and Natasha Tokowicz. 2013 "Examining English–German Translation Ambiguity Using Primed Translation Recognition." *Bilingualism, Language and Cognition* 16(2): 442–457.

Francis, Wendy S., Natasha Tokowicz and Judith F. Kroll. 2014. "The consequences of language proficiency and difficulty of lexical access for translation performance and priming." *Memory and Cognition* 42(1): 27–40

Germann, Ulrich. 2008. "Yawat: Yet Another Word Alignment Tool." In *Proceedings of the ACL-08*: HLT Demo Session (Companion Volume) , pages 20–23, Columbus, June 2008

Hartsuiker, Robert J., Martin J. Pickering, and Eline Veltkamp. 2004. "Is Syntax Separate or Shared between Languages? Cross-linguistic Syntactic Priming in Spanish-English Bilinguals." *Psychological Science* 15(6): 409–14.

Inhoff, Albrecht W, and Keith Rayner. 1986. "Parafoveal Word Processing during Eye Fixations in Reading: Effects of Word Frequency." *Perception and Psychophysics* 40(6): 431–39.

Ivir, Vladimir. 1981. "Formal Correspondence Vs. Translation Equivalence Revisited." *Poetics Today* 2 (4): 51–59.

Jakobsen, Arnt Lykke. 2011. "Tracking Translators' Keystrokes and Eye Movements with Translog." In *Methods and Strategies of Process Research: Integrative Approaches in Translation Studies*. ed. by Cecilia Alvstad, Adelina Hild, Elisabeth Tiselius, 37–55. Benjamins Translation Library Volume 94. Amsterdam and Philadelphia: John Benjamins Publishing Company

Jelinek, F, R. L. Mercer, L. R. Bahl, and J. K. Baker. 1977. "Perplexity – a measure of the difficulty of speech recognition tasks." *J. Acoust. Soc. Am.* 62: S63

Jensen, Kristian T.H., Annette C. Sjørup, and Laura W. Balling. 2010. "Effects of L1 Syntax on L2 Translation." In *Methodology, Technology and Innovation in Translation Process Research: A Tribute to Arnt Lykke Jakobsen*, ed. by Fabio Alves, Susanne Göpferich, and Inger M. Mees, 319–336. Copenhagen: Samfundslitteratur.

Laxén, Jannika, and Jean-Marc Lavaur. 2010. "The Role of Semantics in Translation Recognition: Effects of Number of Translations, Dominance of Translations and Semantic Relatedness of Multiple Translations." *Bilingualism: Language and Cognition*, 13(02): 157.

Malmkjær, Kirsten. 2011. *Translation Universals*. In *The Oxford Handbook of Translation Studies*, ed. by Kirsten Malmkjær and Kevin Windle, 83–94. Oxford: Oxford University Press.

Martínez Gómez, Pascual, Akshay Minocha, Jin Huang, Michael Carl, Srinivas Bangalore, Akiko Aizawa. 2014. "Recognition of Translator Expertise using Sequences of Fixations and Keystrokes." In: *Procedings of Symposium on Eye Tracking Research and Applications*, ed. by Pernilla Qvarfordt, and Dan Witzner Hansen, 299–302. New York: Association for Computing Machinery.

McNamara, Timothy P. 2005. *Semantic priming: Perspectives from Memory and Word Recognition*. Hove, England: Psychology Press.

Mesa-Lao, Bartolomé. 2013. "Eye-tracking Post-editing Behaviour in an Interactive Translation Prediction Environment". In *Proceedings of the 17th European Conference on Eye Movements (ECEM 2013)*. Humanities Laboratory - Lund Universtiy (Sweden), August 11-16.

Och, Franz Josef , Richard Zens, and Hermann Ney. 2003. "Efficient Search for Interactive Statistical Machine Translation" In *EACL '03: Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics*, 387–393

Och, Franz Josef, Hermann Ney. 2003. "A Systematic Comparison of Various Statistical Alignment Models." *Computational Linguistics*, 29(1): 19–51.

Pickering, Martin J. and Holly P. Branigan. 1999. "Syntactic Priming in Language Production." *Trends in Cognitive Sciences,* 3(4): 138–141.

Pickering, Martin J. and Victor S. Ferreira. 2007. "Structural Priming: A Critical Review." *Psychological Bulletin*, 134(3): 427–459.

Popel, Martin and David Marecek. 2010. "Perplexity of n-Gram and Dependency Language Models." In *Text, Speech and Dialogue*, *Lecture Notes in Computer Science,* 6231: 173–180.

Prior, Anat , Shuly Wintner, BriaN MacWhinney and Alon Lavie. 2011. "Translation Ambiguity in and out of Context." *Applied Psycholinguistics* 32(1): 93–111.

Prior, Anat, Judith F. Kroll, and Brian Macwhinney. 2013. "Translation Ambiguity but not Word Class Predicts Translation Performance." *Bilingualism: Language and Cognition*, 16(02): 458–474.

Sanchis-Trilles, German, Vicent Alabau, Christian Buck, Michael Carl, Francisco Casacuberta, Mercedes Garcia Martinez, Ulrich Germann, Jesus Gonzalez-Rubio, Robin L. Hill, Philipp Koehn, Luis A. Leiva, Bartolomé Mesa-Lao, Daniel Ortiz-Martnez, Herve Saint-Amand, and Chara Tsoukala. 2014. "Interactive Translation Prediction vs. Conventional: Post-editing in Practice - A Study with the CasMaCat Workbench." *Machine Translation*. (in print)

Sennrich, Rico. 2012. "Perplexity Minimization for Translation Model Domain Adaptation in Statistical Machine Translation." In *EACL '12 Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. 539–549.

Skadins, Raivis, Maris Purins, Inguna Skadina, and Andrejs Vasiljevs. 2011. "Evaluation of SMT in Localization to Under-resourced Inflected Language." In *Proceedings of the 15th International Conference of the European Association for Machine Translation EAMT*, 35–40.

Sharmin, Selina, Oleg Spakov, Kari-Jouko Räihä, and Arnt Lykke Jakobsen. 2008. "Effects of Time Pressure and Text Complexity on Translators' Fixations." In: *Eye Tracking Research & Application Archive: Proceedings of the 2008 Symposium on Eye Tracking & Applications*, Savannah, Georgia, 26-29 March 2008, 123–126. Association for Computing Machinery.

Schoonbaert, Sofie, Robert J. Hartsuiker, and Martin J. Pickering. 2007. "The Representation of Lexical and Syntactic Information in Bilinguals: Evidence from Syntactic Priming." *Journal of Memory and Language*, 56(2): 153–171.

Tirkkonen-Condit, Sonja. 2005. “The Monitor Model Revisited: Evidence from Process Research.” *Meta: Translators’ Journal* 50(2): 405–414.

Tokowicz, Natasha, and Judith F. Kroll. 2007. “Number of Meanings and Concreteness: Consequences of Ambiguity Within and Across Languages.” *Language and Cognitive Processes*, 22(5), 727–779.

Toury, Gideon. 1995. *Descriptive Translation Studies and Beyond*. Benjamins Translation Library Volume 4. Amsterdam and Philadelphia: John Benjamins Publishing Company.

Underwood, Nancy L., and Bart Jongejan. 2001. “Translatability Checker: A Tool to Help Decide Whether to Use MT.” Proceedings of MT Summit VIII. 363-368.

Wagner, Emma. 1985. “Post-Editing Systran – A Challenge for Commission Translator.” *Terminologie et Traduction 3*.

A.3 Submission: The role of syntactic choices in translation and post-editing

Submitted to a Special Issue of *Translation Spaces* dedicated to our ABRAPT conference panel on Translation as a Cognitive Activity (September 2013).

The role of syntactic choices in translation and post-editing

Authors

Srinivas Bangalore, Bergljot Behrens, Michael Carl, Maheshwar Gankhot, Arndt Heilmann, Jean Nitzke, Moritz Schaeffer, Annegret Sturm

1. Abstract

The present study investigates the question whether the co-activation of both source and target language have an influence on the translator's behaviour. A way to measure co-activation is the comparative analysis of how the choice of producing different syntactic realizations of the target language affect reading time and production duration during translation and post-editing.

We measure syntactic choice in the translated text in terms of entropy, which quantifies the distribution of different translation realizations of a given source segment. High entropy indicates the availability of a large number of different translation options in the final translation product, which, we claim, correlates with high selection effort. We investigate the impact of syntactic entropy of the translation realizations on the cognitive effort for producing the translation. In a first step, as research by Jensen et al. (2009) suggests, source text segments which need reordering yield longer reading times than segments which do not need reordering. In a second step, based on the assumption that syntax is shared across languages (Hartsuiker and Pickering 2004), a recently activated syntactic structure is likely to influence subsequent processing, thus "priming" it. Low syntactic entropy of translation choices could, thus, be taken to be a sign of priming.

To test the hypothesis whether syntactic choices have an influence on cognitive effort, we compared four datasets comprising translation and post-editing data of the same English source texts translated into Danish, German, Spanish and Hindi under equal experimental conditions, and made available in the CRITT-TPR database (Carl 2012a). For the present investigation the data was manually annotated for syntactic structure along three relevant features: Valency of the verb, Voice and Clause type.

Our analyses reveal a positive correlation between syntactic entropy of translation realizations and total reading time per source word as well as production time in all four languages. However, no effect of syntactic entropy could be detected in the post-editing data with respect to source text reading times, suggesting that the post-editors were primed by the MT output.

Keywords: syntactic choice, eye-tracking, gaze data, behavioural measures, post-editing, MT output, entropy

2. Introduction

A translator often has more than one option when translating a particular word or syntactic structure. Sometimes, one of these options is to transpose the source text structure into the target text instead of considering all possible target options. Campbell (2000) hypothesized that translations of the same source text by different translators can be used to draw inferences about the cognitive processes during translation. He proposes a Choice Network Analysis: the more options and the more complex choices a translator has to consider, the more effortful is the translation of a particular item. These options, or the lack thereof, can be quantified and studied in terms of what Carl and Schaeffer (submitted) call translation entropy. If, out of a given number of translations of the same source text one translation unit has been translated in a different way by each translator, then the probability of each individual translation is equally small. Thus, the translation entropy is high, as the variance of target units is high. Low translation entropy occurs the lower the variety is among the target units, i.e. when only a few target options have been realized, but each by a number of translators. A translation entropy of zero would occur in cases where every translator chose the same target option.

A study by Dragsted (2012) shows that lexical translation variance correlates with production time and reading duration. The study compared these behavioral measures for words with a high number of translation alternatives for lexical items with a small number of alternatives. However, a raw count of the different target realizations of a source word does not take into account the fact that a translation option which has been chosen by more than one translator may have a greater weight as compared to an option which has only been chosen by one translator. The distribution of probabilities of different translation options, i.e. their entropy, captures the different weights much better.

Carl and Schaeffer (submitted) have shown that words with low translation entropy are less effortful to process than words with high translation entropy. The purpose of the present paper is to extend these findings to regarding lexical items to syntactical structures. The first research question to be addressed is thus whether translators, when translating a particular syntactic segment, consider more than one syntactic target expression. Taking into account several options and having to make a choice is expected to be cognitively more effortful. In a second step of our research, we investigate whether the absence of cognitive effort can also be linked to entropy. We assume that the translator will avoid engaging in unnecessarily high cognitive effort. Hence we argue that whenever a source sentence structure can be transposed into the target text, it will be transposed. In this case, the target structure is likely to be the same for all translators. Entropy should thus be low. At the same time, the simple transposition of the source text structure should be associated with less cognitive effort, as it is less demanding in terms of choice. Our second research question is thus whether the transposing of the source text structure has a facilitation effect. Both research questions will be addressed by investigating the relationship between entropy values and behavioral measures of cognitive effort (reading time and production time).

Our paper proceeds as follows: In the section 3, we lay out the background for the basic assumptions motivating our investigation, including an account of entropy in general and specifically applied to translation studies (3.1), followed by evidence of co-activation of both languages in translational settings and of priming (3.2-3.5) and concluded by statement of our specific research hypotheses (3.6). In section 4, we present the participants and the material we have made use of in the present study, including a detailed account of the annotation system we developed and applied to the translation data in order to extract entropy values. Section 5 presents these values for each language and their correlations with source text reading time, target text gaze and coherent typing activity.

3. Theoretical Background: Information entropy

3.1. Information and entropy

The term “entropy” was first used in 1865 by Rudolph Clausius to describe the valency of energy. It is an artificial word created from Greek εντροπία which could be translated as “a turning towards”, “tendency” or “potential for change”. The term became famous in Boltzmann’s Second Law of Thermodynamics which states that in a closed system, dynamically ordered states are infinitely improbable. Thus, one single dynamically ordered state is the most improbable case conceivable (Boltzmann 1886).

Shannon (1951) reformulated this idea in the context of information theory. He uses the term as a measure of the amount of information transmitted in a communication process. In this context, the term “information” is used synonymously to “variance” (Miller 1956). Many possible selections lead to a large variance, higher uncertainty and higher entropy values. The entropy concept can help to describe the behavior of a closed system: If the system gives the same response at any measure, it contains low variance, hence low information and low entropy. When the system gives a new answer at any new measure, each new item adds up to the system’s inherent variance. This high variance within the system causes a low predictability for each new answer whenever the system is measured. Thus, the system’s entropy is high.

Entropy is denoted by the symbol ***H*** and represents the average amount of non-redundant information provided by each new item. Entropy ***H*** is computed based on the probability ***p*** of the item to occur and its information. The information of a probability p is defined as $\boldsymbol{I(p) = - log_2(p)}$, and entropy ***H*** is the expectation of that information as defined in equation (1):

1. $$H = \sum_{i=1}^{n} p_i I(p_i) = -\sum_{i=1}^{n} p_i log_2(p_i)$$

TT1	TT2	TT3	TT4	TT5	TT6	H
1						0.00
0.50	0.50					1.00
0.25	0.25	0.25	0.25			2.00
0.50	0.17	0.17	0.17			1.79
0.17	0.17	0.17	0.17	0.17	0.17	2.58
0.30	0.14	0.14	0.14	0.14	0.14	2.51

Figure 1: Example probability distributions of hypothetical translations. TT1 - TT6 exemplify the effect of probability distributions on entropy (***H***).

Figure 1 describes the effect of probability distributions on entropy (***H***): if all of six translators choose the same translation realization for a given segment, the probability of this translation is at its maximum (1) and entropy is at its minimum (0), but as soon as translators opt for different target realizations, entropy increases: if one option has a probability of 0.30 and five other options have each a probability of 0.14, then entropy is relatively high (2.51). If there are four different options, but all four options have the same probability (0.25), entropy is higher than when one of the four options has a higher probability (0.50) than the other three (0.17).

As a transfer process including the reproduction of an initial source message in another context, every translation is a selection of a final target formulation out of many possible target formulations (Neubert and Shreve 1992). However, the details of this selection process and the factors influencing it are largely unknown. Whenever a source text ST is translated by *n* translators producing TT_n translations, each single translation TT_i is selected out of many possible target texts. Each selection of the actual elements of TT_i is determined by the characteristics of the target language, its morphology, syntax, pragmatics and stylistics, the translation brief and target audience etc., but also by the individual translator, her background and experience. Each final target text TT_i is thus a selection from possible text options in the target language which were available to one particular translator at one particular point in time. Comparable to Boltzmann's definition of entropy, it is highly unlikely that any two translators produce exactly the same translation of the same source text. In cases where every translator produces a different translation, one would assume that the selection process was cognitively demanding, as all possible realizations of TT elements are assumed to have been potentially available to all translators. In cases where all translations of a given source text unit are identical, this can be taken as a sign of lacking choice, as there might have only been a single correspondence in the target language. Consequently, the translation was comparatively easy as the translator did not have to make any choice. Entropy as the number of choices for a given translation unit can thus be understood as a basic measure of selection effort in the translation process. Although translation competence can be defined in terms of selection and selection effort, namely as "the ability to generate a series of more than one viable target text (TT_1 , TT_2 … TT_n) for a pertinent source text (ST)" and "the ability to select only one viable TT from this series, quickly and with justified confidence." (Pym 2003, 489), this does not imply that the selection process is the same for all competent translators. Whereas entropy allows for measuring the cognitive effort involved in the selection process, possible factors which may influence this selection process are presented in the following sections.

3.2. Co-activation and translation

One of the first questions to be addressed is the onset of the selection process. At which point in time during the translation process does the translator start with the mental production of the target text, and to which degree is does this mental production process interfere with the source text comprehension? Studies suggest that both languages of a bilingual are always active. Grosjean (1997) argued that activation of the bilingual's two languages are situated on a continuum which has a relatively monolingual state at one extreme and highly co-activated bilingual state at its other extreme. Grosjean argued that it is the context of the language use which determines where on the continuum the bilingual is currently situated: if both interlocutors speak the same two languages, it is more likely that both languages are active, while when only one interlocutor speaks two languages or two interlocutors do not speak the same two languages, it is more likely that the bilingual(s) are situated closer to the monolingual mode. Translation would situate the bilingual firmly towards the very extreme of the bilingual state. A range of studies supports this hypothesis. Macizo and Bajo (2006) presented professional translators and naïve bilinguals (Spanish/English) with single sentences containing interlingual homographs. In a masked self-paced reading paradigm, participants were instructed to either read the sentence for comprehension or for

translation. According to the condition, participants had to read the Spanish sentences and either translate them into English, or answer comprehension questions. The homographs made the sentence ambiguous when their meaning in the other language became activated: e.g. (presente) in Spanish is very similar to the English (present). While the Spanish word is not ambiguous in the sentence, when translating it into the English word (present), it could either refer to the present moment or to a gift. Macizo and Bajo found that the ambiguous homograph slowed down reaction times, but only when the reading purpose was translation. This effect was more pronounced for naïve bilinguals than for professional translators. Ruiz et al (2008) used essentially the same experimental design, but manipulated the frequency of the equivalent target word. They kept the monolingual frequency of critical words in the Spanish source sentence constant while the equivalent target words had either a high or a low frequency. Ruiz et al found that reaction times were slowed down when the equivalent English target word had a low frequency, but again, this was only the case in the translation condition. The authors interpret their findings in terms of online parallel activation of source and target items during translation; i.e. both languages are active to a high degree during translation. Schaeffer et al (submitted) used a similar experimental design: this study compared reading for comprehension with reading for translation, but instead of self-paced reading, they used an eye-tracking paradigm. Furthermore, the authors manipulated the number of target words required to translate a single source word embedded in the same sentence frame. For example [worry] and [laugh] were embedded in the same sentence frame (Many of the fishermen will [worry/ laugh]). Whereas the translation of (worry) into German requires three words (sich Sorgen machen), (laugh) can be translated by a single word (lachen). Schaeffer et al found that the first fixation duration was 23 ms longer when more than one target word was needed for the translation. Again, this effect occurred only in the reading for translation condition. This study further supports the idea that translation occurs online and that target items are activated early during source text reading. Thierry and Wu (2012) lend further support to the automatic co-activation of the two languages which they observed even though the experimental design discouraged it. In their ERP study, participants were asked to press a button in response to the presentation of circles or squares. Participants were told that sometimes words would appear on the screen, but were instructed to ignore these. 15% of these words were interlingual homophones, i.e. their Chinese translation would sound similar to either of the words [circle] or [square]. Thierry and Wu found an N200 effect for these homophones, suggesting that participants had to inhibit their spontaneous reaction of pressing the button anytime the English word activated the Chinese words for either [square] or [circle]. Thus, co-activation could be detected in an environment where it was explicitly discouraged and even irrelevant to the task. We therefore assume that both the ST and TT language are simultaneously activated during the entire translation process. That means that the translator becomes engaged in exploring and selecting potential target text elements as soon as she starts reading the source sentence. As both languages may be activated to the same degree, it is likely that they influence one another during this selection process. One form of this mutual influence is priming.

3.3. Priming and translation

There is evidence from a range of cross-linguistic priming studies which suggest that semantic and / or syntactic representations may be shared between two languages when these aspects are similar in the two languages (e.g. Duñabeitia, Perea, et al. 2010; Bernolet et al. 2013). In translation, priming has been associated with a facilitation effect. Jensen et al. report lower processing times due to a possible "automatic transfer of L1 syntax to all types of L2 processing" (Jensen et al. 2009, 333). The authors report shorter total reading times in cases where the word order of the source sentence has to be transposed in the target sentence. Priming effects have been observed to be at work on various levels.

In the case of post-editing, the translator may in turn be primed by the machine translation-output. Alternative translation options would then become inaccessible to the post-editor. Evidence for this hypothesis is presented by Carl and Schaeffer (submitted) who have shown that the translation entropy of post-edited texts is consistently lower than the translation entropy of translations from scratch of the same texts.

3.4. The study of syntax during reading for comprehension and translation

Surprisal theory has shown that sentences are understood incrementally by predicting upcoming words (Hale 2001, Levy 2008). The likelihood that participants guess a particular word correlates with eye

movement measures: words which are highly predictable are read faster than words which are less predictable, when other factors which influence eye movements are held constant. Surprisal theory incorporates not only lexical aspects, but also the syntactic category of upcoming words. Levy (2008) shows that the statistical likelihood that a word of a particular syntactic category occurs at a given point within a sentence can predict behavioural measures such as eye movements. As several different structures are of course often possible at any given point in the sentence and their probabilities are described by surprisal theory in terms of their entropy. Processing difficulty occurs when the entropy of the actually encountered word differs from the entropy of the predicted word. Levy (2008, 1129) describes this degree of surprisal as relative entropy: the more unexpected the upcoming syntactic entity is, the greater the surprisal and the greater the processing difficulty of this word. Surprisal theory has been successful in accounting for a variety of empirical observations (e.g. Levy and Keller 2013). Surprisal theory predicts processing difficulty at particular points during sentence processing and can therefore argue that at any moment more than one parse is active. If this is true of monolingual sentence reading, it is likely that the same is also true of reading for translation. The model proposed by Schaeffer and Carl (2013) suggests that these processes both occur in parallel. Nevertheless, as reported in section 3.2, current evidence suggests that translation occurs already early and online. The processing effort of entertaining several parses of the source and target language simultaneously is likely high. We would therefore predict longer reading times on the source sentence for items which have a high syntactic entropy than for source sentences which have a low syntactic entropy. While surprisal theory predicts bottlenecks at certain points during the sentence, based on predicted structures, for the current purpose we will only report on global measures, i.e. the syntactic entropy of the entire sentence, the sum total of all fixations on a given sentence and the total time required to produce the sentence.

3.5. The literal translation hypothesis

Like many other concepts in Translation Studies, the concept of literal translation is the object of various definitions (Chesterman 2011, 24). However, it is important to be able to quantify literality if the aim is to show that whatever effect that is observed is not language specific, if the aim is to produce a model of translation which is language independent. Carl and Schaeffer (submitted) propose a definition of literality which allows for quantification of the phenomenon. According to their definition, a translation is literal when the three following literality criteria are fulfilled:

1. Word order is identical in the ST and TT.
2. ST and TT items correspond one-to-one.
3. Each ST word has only one possible translated form in a given context.

Literality criterion 3 is of particular interest as it refers to translation entropy. Expanding this criterion to syntactic features, it implies that two ST sentences may have the same structure in the source language, but the literality criteria are already fulfilled as long as each of them is mapped into one single TL structure. In other words, syntactic entropy measures the extent to which different translators produce the same TT structure for one ST sentence. Syntactic entropy is an indicator for the literality of translations on a syntactic level, and syntactic literality evolves as a consequence of the three literality criteria above:

4. All translations of one source sentence are mapped into the same target structure.

This means that an ideally syntactical literal translation would be one with a syntactic entropy of 0. By means of entropy measures, this definition allows for a quantitative approach to the phenomenon of literality. In line with Ivir's (1981) notion of formal correspondence, literality has been associated with less cognitive effort than non-literal translations. Ivir (1981, 58) describes the translation process as follows:

> The translator begins his search for translation equivalence from formal correspondence, and it is only when the identical-meaning formal correspondent is either not available or not able to ensure equivalence that he resorts to formal correspondents with not-quite- identical meanings or to structural and semantic shifts which destroy formal correspondence altogether. But even in the latter case he makes use of formal correspondence.

Equally related to this notion of formal correspondence as employed by Ivir is Toury's (1995, 275) "law of interference" which postulates that "(…)in translation, phenomena pertaining to the make-up of the source text tend to be transferred to the target text." Similar to Ivir, Toury used this law of interference to posit that less cognitive effort is involved in the production of literal translations as they are a kind of "default setting" in the translating mind. In sum, we argue that the default option for a translator is to consider a literal translation which is more likely to be activated first due to a priming effect and we further argue that if the default is not acceptable or if other, less literal options are activated, this leads to more cognitive effort.

3.6. Hypotheses

- High entropy reflects high cognitive effort and is thus reflected in behavioural measures such as gaze time and production time for translation.
- Translations with low syntactic entropy are easier to produce and the ST segment easier to process. Again, this should be reflected in behavioral measures.
- If the TT sentence has the same structure than the ST sentence a facilitation effect occurs.
- Priming and entropy effects are language independent, i.e. they occur in all language combinations.

4. Methodology

To address the question whether high syntactic entropy is correlated with behavioral measures, in translation and post-editing, we compared the gaze data for translations of six English texts into four different target languages (Danish, German, Hindi, Spanish) with a monolingual copying task as a baseline.

4.1. Translation condition

4.1.1. Participants

The German data comprises translations produced by 24 translators (13 students, 11 professionals), the Danish dataset contains translations from 24 translators (12 students, 12 professionals), the Hindi data was produced by 21 translators (all professionals), and the Spanish data was collected on 32 translators (27 students, 5 professionals).

4.1.2. Material

For the translation task, the four relevant datasets (SG12 for German, KTHJ08 for Danish, NJ12 for Hindi and BML12 for Spanish) were extracted from the CRITT-TPR database (Carl 2012 a). The four datasets contain translations of the same six English source texts. BML12, NJ12 and SG12 contain translations of all six texts, whereas KTHJ08 contains translations of the first three texts only. Each set contains gaze and key-stroke data which were recorded with a Tobii T120 eyetracker, Translog (Jakobsen and Schou 2000) and Translog II (Carl 2012 b). Detailed features of each dataset are presented in figures 2 and 3. Source texts were matched for number of words and readability measures (Jensen 2009).

Figure 2 lists the number of tokens (words) and sentences for the six English source text sentence that were used in this study and the average length (in words) of their translations into the four target languages. Each source text has a length of between 110 words (Text 4) and 160 words (Text 1), and the sum of all source text words is 847. These English texts were translated by several translators (see Figure 3) into German (de), Danish (da), Hindi (hi) and Spanish (es). The abbreviation in brackets indicates the study from which the data was taken. For each of the target languages, the average length of the six target text is shown. For instance, the English source Text 1 has 160 words, but the average length of the German translations is 157,18 words, while for Hindi it has 176,42 words. Notice that translations are on average longer than the source text. The total number of target language words produced for for each language is given in the last column, and the total amount of target words in all four languages is 47008.

	Text 1	Text 2	Text 3	Text 4	Text 5	Text 6	Total

ST tokens	160	154	146	110	139	138	847
ST sentences	11	7	5	5	6	7	41
de (SG12)	57,18	160,78	158,38	108,13	148,57	137,62	12984
da (KTHJ08)	54,29	153,22	156,36	---	---	---	10571
hi (NJ12)	169,33	165,63	160,29	127,29	148,78	139,76	5505
es (BML12)	176,42	183,42	161,31	132,62	152,43	154,13	17948
Total Av.	164,30	165,76	159,08	122,68	149,92	143,83	47008

Figure 2: Number of tokens and sentences of the six source texts and the average number of tokens in their four translations.

Figure 3 contains a detailed overview of the produced target texts: it indicates the translation *Task*, text copying (C), translation from-scratch (T) or post-editing (P), the number of participants (*Part*) involved, the number of translation sessions (i.e. target texts produced), as well as the duration and the total number of target language tokens for each translation mode. Translation (and copying) duration is measured in two different metrics:

- *Fdur*: total production time for all segments, excluding pauses > 200 seconds.
- *Kdur*: total duration of coherent keyboard activity excluding keystroke pauses > 5 seconds.

The BML12 study, for instance, contains 63 from-scratch translations and 64 post-edited translations which were produced by 32 translators (participants). Note that the same translators participated in the translation and the post-editing tasks. That is, each participant had to edit, post-edit and to translate two texts in each mode, and texts were distributed in a randomized order. As shown in figure 3, the post-edited and translated texts amount together to 17948 target text words from which 9012 are produced in the 64 post-editing sessions and 8936 words are produced in the 63 translation sessions. The total duration of the post-editing sessions was 2,31 hours while for the translations were needed 8,2 hours in terms of *Fdur* time. That is, gaps of keystroke activity for more the 200 seconds (almost 2,3 minutes) are excluded, under the assumption that translation activities are interrupted in such instances. However, no such pauses were observed in these studies.

Note that post-editing (in BML12 and SG12) is much faster than from-scratch translation, and the amount of coherent typing activity is much less during post-edition than during translation. Thus, for BML12, post-editing is almost 4 times faster than from-scratch translation (*Fdur*), while roughly 70% of the time in from-scratch translation was spent on typing activities (*Kdur*), compared to only 40% of the time for post-editing.

Study	Session	TL	Task	Texts	Part.	Fdur	Kdur	Tlen
TDA14	48	en	C	1-6	8	3,60	3,4924	6792
KTHJ08	69	da	T	1-3	24	6,45	5,4536	10571
SG12	45	de	P	1-6	23	5,6	1,9265	6352
SG12	47	de	T	1-6	24	9,39	4,6145	6632
NJ12	39	hi	T	1-6	20	13,04	7,4243	5505
BML12	64	es	P	1-6	32	2,31	0,8774	9012
BML12	63	es	T	1-6	32	8,2	5,7491	8936
Total	375	5	3	6	107	48,59	29,5378	53800

Figure 3: Properties of the target texts under the C, T and P conditions into the five target languages: and of tokens and sentences of the six source texts and the average number of tokens in their four translations.

Figure 3 also contains information concerning a copying task (TDA14). The copying data serves as a baseline and allows for contrasting the results of the translation task to a similar text production task without a language switch. Six participants in the copying task were native speakers of German, one of Turkish and one of French. All of them have learned English at school and/ or university for 10 to 18 years. Five of them were students currently enrolled in a translation programme and two have a degree in translation. Note that during copying more than 96% of the text production time is spent on coherent typing (*Kdur*).

Post-editing data was only available in German and Spanish.

4.3. Overall procedure

Syntactic choices within each dataset were quantified with the help of the manual annotation system presented in section 4.4. below. Entropy values were calculated on the basis of this annotation for each TT sentence. During analysis, entropy values were correlated with source text reading time, target text gaze and coherent typing activity and compared across languages.

4.4. Baseline task: Monolingual copying

The copying task serves as a baseline and allows for contrasting the results of the translation task to a similar text production task without a language switch to find out about the translation specific effects.

4.4.1. Participants.

Nine participants took part in the study. Due to calibration problems, one participant had to be excluded. Eye-tracking and key-logging data were thus collected from eight participants. Six participants were native speakers of German, one of Turkish and one of French. All of them have learned English at school and/ or university for 10 to 18 years. Five of them were students currently enrolled in a translation program, two have a degree in translation and one is a not a language expert.

4.4.2. Material

The same six English source texts were used as in the translation test condition (cf. figure 2).

4.4.3. Procedure

Participants were asked to fill out a questionnaire before the experiment and had to answer three comprehension questions upon completion of the task. They were instructed to copy the English text and were informed that comprehension questions would follow the task. During the task, key-stroke and gaze data were recorded with Translog II (Carl 2012b) and a Tobii T120 eye-tracker respectively.

4.5. Annotation

We have made a shallow syntactic parse of the ST and TT segments, distinguishing first of all between independent and dependent clauses. Simple sentences and main clauses are tagged as independent (I). Subclauses, whether finite or non-finite (infinitival, gerundial, participial (adjuncts)), are dependent (D). They include clausal adverbials, clausal complements of prepositions, clausal arguments (subjects or objects) to verbs, clausal complements to nouns or other clausal modifiers (finite or non-finite relative clauses). Each clause is annotated for valency (transitive(T)/intransitive(I)). To make the analysis as simple as possible, we have included copular clauses in the class of transitives.

features/ labels for levels				
valency	I:intransitive	T: transitive	D:ditransitive	Imp: impersonal
voice	A: active	P: passive		
clause type	I: independent	D: dependent		

Figure 4: Labels used in the syntactic annotation

As intransitives do not passivize, these features gave rise to 14 possible combinations applicable across the languages under study: TAI, TPI, IAI, DAI, DPI, ImpAI, ImpPI, TAD, TPD, IAD, DAD, DPD, ImpAD, ImpPD, RPI, RPD. We have added one other feature, the Spanish Passiva Refleja, as it does not seem to fall under either of the categories we have. A problem has sometimes arisen with respect to proper assignment, as it is sometimes hard to decide whether for example a preposition is part of a prepositional verb (in which case we have a transitive structure) or a preposition introducing a prepositional phrase (in which case the verb would be intransitive). These cases have been annotated according to mutual agreement among the four annotators. We include in our analysis clauses that we term Impersonal (IMP). These are clauses with dummy subjects, such as for example the first clause of a cleft construction (e.g. *It was Norris* in *It was Norris who killed them*), or a clause with a clausal extraposed subject (e.g. *Es comprensible* in *Es comprensible que los paises en desarrollo se opongan a comprometer sus posibilidades(…)*). IMP clauses are not assigned a value for valency.

The clauses are also annotated for Voice (active (A), or passive (P)). The annotation leaves us with a characterization of a segment (sentence) according to its syntactic complexity (its number of clauses), with each clause being assigned a triplet of clausal features : Valency, Voice, Clause Type (e.g. TAD (transitive, active, dependent), TPI (transitive, passive, independent) etc. Each source segment structure is then correlated with the set of triplets its target translations are made up of. An example is given below for illustration:

ST	Only the attention of other hospital staff put a stop to him and the killings.	TAI
de	Nur die Aufmerksamkeit der anderen Krankenhausmiterbeider setze ihm und den Morden ein Ende. *(Only the attention the(Genitive) other hospital staff set him and the murders an end)*	TAI
da	Det var udelukkende opmærksomhed fra andre hospitalsmedarbejdere, der fik stoppet ham og mordene. *(It was only attention from other hospital staff that got stopped him and the murders)*	IMPAI,TAD
es	Solo el hecho de que el personal reparara en ello pudo parlarle los pies y detener los asesinatos. *(Only the fact that the personnel noticed him could stop his feet and end the murders)*	TAD DAI TAI

hi	यह अस्पताल के स्टाफ की जागरूकता ही है कि इसे पकड़ा जा सका। (*It is the awareness of hospital staff that he could be caught*)	TAI

Figure 5: Translations of the same source sentence into the four target languages and associated triplet labels.

In the example above, the German translation has retained the structure of the source, with a transitive, active simple (independent) structure (TAI). The Danish version is a cleft structure with an impersonal active main clause (analysed as independent) (IMPAI), and a transitive, active dependent clause (TAD). The Spanish translation has a dependent clause complementing the subject (*hecho*) of the ditransitive main clause (DAI) ,and a coordinated clause with a transitive verb. Coordination to the main clause counts as independent. If a clause is coordinated with another subclause, it counts as dependent, along with the subclause with which it is coordinated. An example from Danish is:

(1) Analytikere har advaret om at priserne vil stige yderligere, hvilket gør det vanskeligt for den enegelske bank at sænke renten, fordi den kæmper med inflation og med at holde økonomien under kontrol.
(Analysts have warned about that prices will increase further, which makes it difficult for the English bank to cut the interest rate, because it fights with inflation and with to keep the economy under control.)

The sentence is a translation from the following English source:

(2) Analysts have warned that prices will increase further still, making it hard for the Bank of England to cut interest rates as it struggles to keep inflation and the economy under control.

The English source segment consists of a transitive main clause (*Analysts have warned….)* with a nominal object intransitive that-clause, a (dependent) adverbial non-finite participial clause (*making it…*), a verbless impersonal dependent clause (*it hard….),* an infinitival, transitive clause (extraposed subject: *to cut interest rates*), a finite, subordinate, intransitive clause (*as it struggles*..) and a non-finite (infinitival) transitive clause (*to keep inflation and the economy under control*).
The Danish version is very similar in structure but for the last segment, which is a coordinated clause: "*and with to keep the economy under control*". We consider the preposition 'med/*with'* here to be part of a prepositional verb, the noun '*inflation*' is the first object of that clause, and the infinitival clause is a coordinated object clause. As it is coordinated to the object of a subclause, we count it as a dependent clause.
The example in Figure 5 shows only one translation option in each language for one source segment. Our annotation allows us to see the variation in syntactic constellations for each language, as per translator. In Spanish, for example, the source segment in Figure 5 yields several structures, alternating between the TAD-DAI-TAD (as in Figure 5), a simple active ditransitive (DAI) and an IMPAD-DAI combination, viz.:

(3a) *Solo el conocimiento de otro personal del hospital puso fin a los asesinatos*
(3b) *Solo el conocimiento de que había otro personal en el hospital puso fin a los asesinatos*

The segment, then, has three options in Spanish, according to our annotation system. In Danish, the same segment shows over 10 different options, from a simple TAI structure or a TPI structure, to embedded structures of four clauses of various kinds (IMPAI-TAD-IAD-TPD or IMPAI-TAD-TPD-IAD).

The languages differ wrt how the triplets distribute. Table 6 below gives the triplets for text 1 for Danish, German and Spanish as an example:

		INDEPENDENT CLAUSES (txt 1)							DEPENDENT CLAUSES (txt1)								
triplet	RPI	TAI	TPI	IAI	DAI	DPI	ImpAI	ImpPI	TAD	TPD	IAD	DAD	DPD	ImpAD	ImpPD	RPD	
ST	-	8	3	0	0	0	0	0	4	0	1	1	0	0	0	-	
DA (24)	-	156 6.4	47 2	14 0.6		35 1.5	11 0.5	2 0.1	92 3.8	5 0.2	48 2	24 1	1 0.04	1 0.04	3 0.13		
DE (7)	-	46 6.6	24 3.4	2 0.3	3 0.4	0	1 0,14	0	6 0.85	4 0.7	11 1,6	4 0.7	0	0	0	-	
ES (11)	21 1.9	63 5.7	21 1.9	8 0.7	2 0.2	0	2 0.2	6 0.5	35 3.2	1 0.1	34 2.8	14 1.3	0	1 0.1	0	1 0.1	
ES T5	1	58	8	19	1	0	0	0	60	2	4	2	2	0	0	0	

Figure 6: Triplet combinations for the Danish, German and Spanish translations of the first three texts

Source text 1 has 17 clauses, distributed into 11 main clauses, all of which are transitive clauses, but distinguished between 8 active and three passive clauses. The Danish, German and Spanish translations show more variety in the main clause structures: Danish and Spanish having more impersonal constructions, German and Danish have somewhat higher scores on passives. The subordinate clauses also show more variety in the translations: Danish again showing cases of impersonal constructions, while German and Spanish showing a stronger use of intransitives.

5. Analysis

The annotation gives us the options the translator is assumed to have available when she selects her preferred translation. Our question is whether the number of options measured by our annotation and assumed to be the available options for the translator, is reflected in the time it takes for the translator to make her choice and spell out her target sentence (as measured by the behavioral measures).

We note that some translators have chosen to fuse two source segments into one target segment, or split one source segment into two segments in the target language. When a translator fused two segments into one, we had to exclude these from our statistical analysis since the behavioral measures used are reported per source segment.

In order to correlate entropy values with behavioral measures, simple linear models were produced with the behavioral measure as a function of syntactic entropy. Analyses contain the total reading time per word for the source and for the target text, as well as the duration of coherent typing behavior, excluding periods of typing excluding pauses longer than five seconds. The total reading time is the sum of all fixations on a particular segment. These measures were normalized by the number of words per segment.

The first segment of each text was excluded from the analysis as it might be considered as prone to noise. First segments, titles in particular, tend to have very long total reading times. This might partly be the case because fixations are wrongly mapped onto the title of a text in particular, while they should actually be mapped onto subsequent sentences.

In a first step, we correlated syntactic entropy and behavioral measures for the control condition, the translation task as well as post-editing.

In addition to the labels described above a further label with two levels was generated: one label describes instances where syntactic structures were kept the same during translation or post-editing which we termed adherence and the other label describes instances of target text structures that were changed during translation or post-editing which we termed deviation. This was done in order to investigate the relationship between syntactic entropy and adherence / deviation. The distributions of adherence / deviation in relation to syntactic entropy were compared with the help of a Mann-Whitney-U-Test across all four texts in the translation condition. Additionally, the probability of a sentence structure to re-occur in a translation was correlated with total reading time. As mentioned above, we assume that if a translator maintains the same structure as in the source (adherence), we would expect a facilitating effect as compared to when translators produced a target sentence with a different structure (deviation). In other words, we would expect a structural priming effect for sentence which have the same structure in the source and the target sentence.

6. Results

6.1. Syntactic Entropy and Behavioral Measures in Translation

The four basic simple linear models showed a slight positive correlation between syntactic entropy and all behavioural measures (Total reading time on the source and target sentence and coherent typing) and all languages for the translation condition. No correlation was found for the control condition (copying).

The models turned out to be statistically significant for the measure of total reading time per word on the source text (termed GazeS in the database) in Danish ($F(1, 440)=28.14$, $p<0.01$) and Spanish ($F(1, 389)= 6.753$, $p<0.01$), but not in German nor Hindi. An additional analysis for Hindi and German was carried out in which outliers were excluded. The positive correlation between syntactic entropy and total reading time per word on the source text became significant for Hindi ($F(1,355)=4.355$, $p<0.05$) but did not reach significance for German. Data points exceeding a total reading time per word higher than 1.5 times the interquartile range above the upper quartile were excluded.

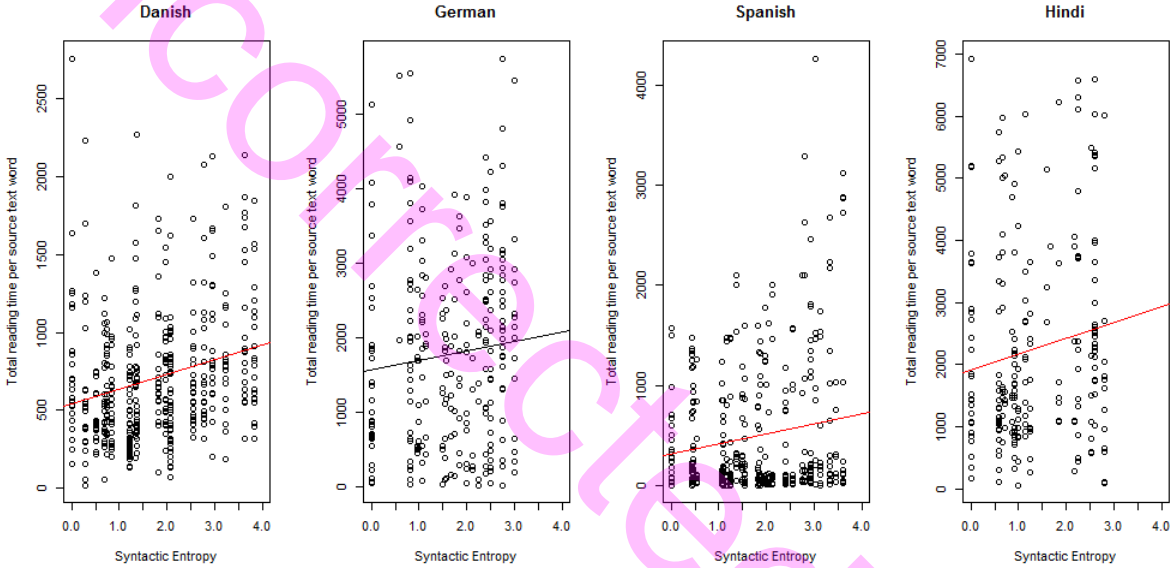


Figure 6: Total reading time per word on the source text in milliseconds and syntactic entropy of the translation condition

In contrast to the translations from scratch, the post-editing condition showed a slightly negative trend. Even after outlier elimination, this trend is not significant for either of the two languages ($p>0.35$ for German and $p>0.9$ for Spanish).

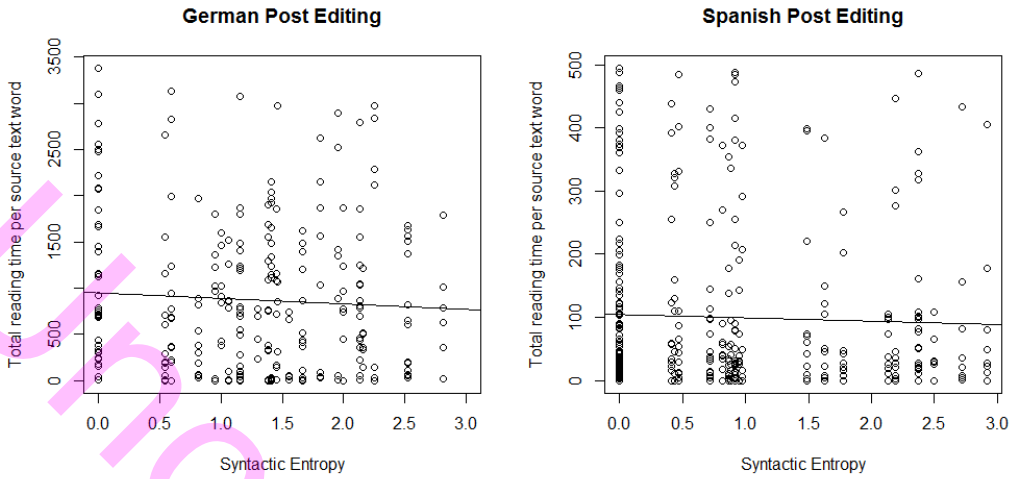


Figure 7: Total reading time per source text word in milliseconds and syntactic entropy of the post editing condition

For the total reading time per word on the target text, a significant correlation could be found in the Danish dataset ($F(1, 440)=10.26$ $p< 0.01$), but not in the German, Spanish nor Hindi data. We analysed these three data sets again and excluded outliers, using the same procedure as before. The positive correlation between Syntactic Entropy and reading time per target word became significant for Hindi ($F(1,204)=5.207$, $p<0.05$) but did not reach significance for Spanish or German.

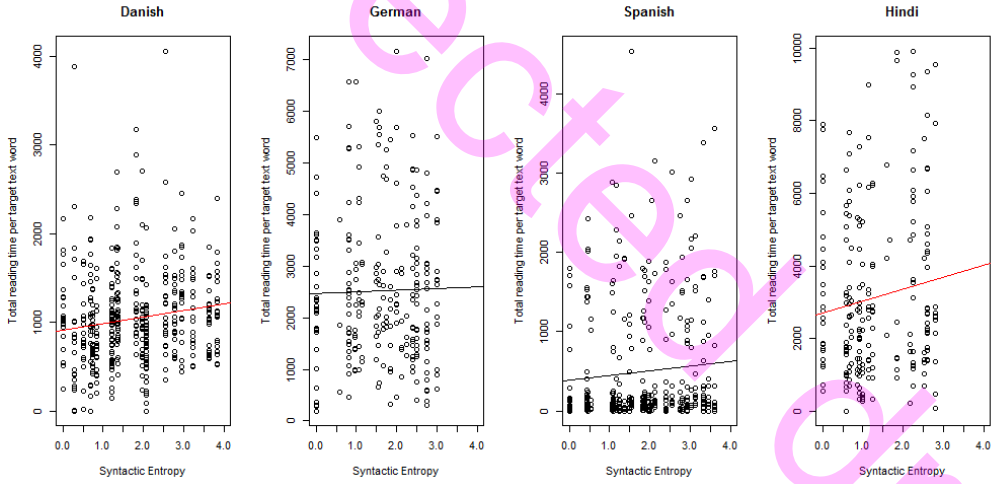


Figure 8: Total reading time per word on the target text in milliseconds and syntactic entropy of the translation condition.

In the post-editing condition, the trends are similarly steep for German and slighlty less steep for Spanish, but in both cases far from significant (p> 0.35 for German and >0.9 in Spanish).

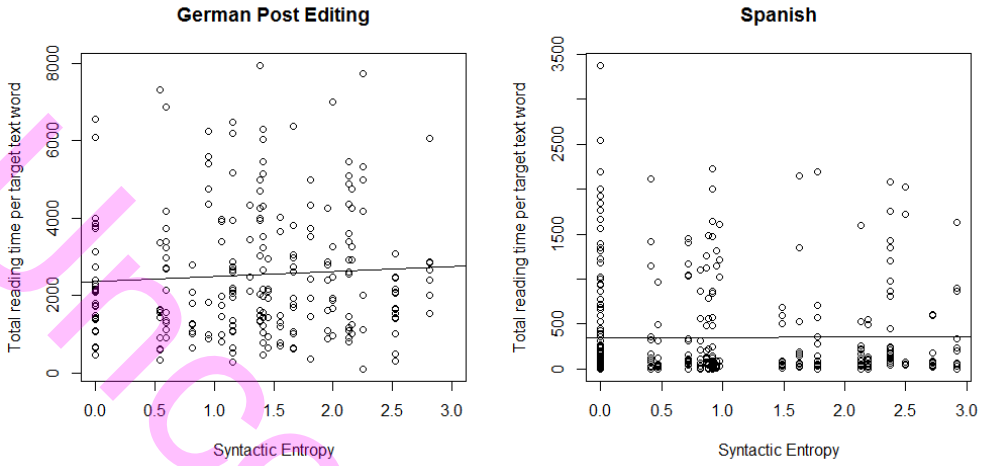


Figure 9: Total reading time per word on the target text in milliseconds and syntactic entropy of the post-editing condition.

The analysis of coherent typing activity mirrored the positive correlation between syntactic entropy and cognitive effort of the before mentioned results. Significant correlations could be observed for Danish ($F(1, 44= 5.516)$, $p< 0.05$) and Spanish ($F(1, 331)= 4.704$, $p< 0.05$), and a marginally significant correlation was found for German ($F(1, 245) = 3.763$, $p= 0.054$). Only for Hindi no significant correlation could be observed between coherent typing activity and syntactic entropy. The results will be discussed in section 7.

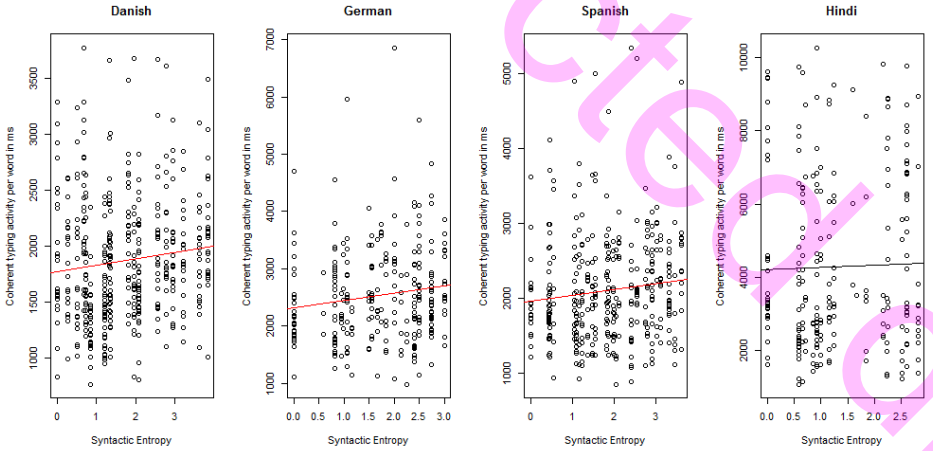


Figure 10: Coherent typing activity per word in milliseconds and syntactic entropy of the translation condition

In the post editing condition for German and Spanish positive trends occur too. The trend for Spanish post edited translations even reaches significance ($F(1,359) = 9,78$ and a p-value < 0.001).

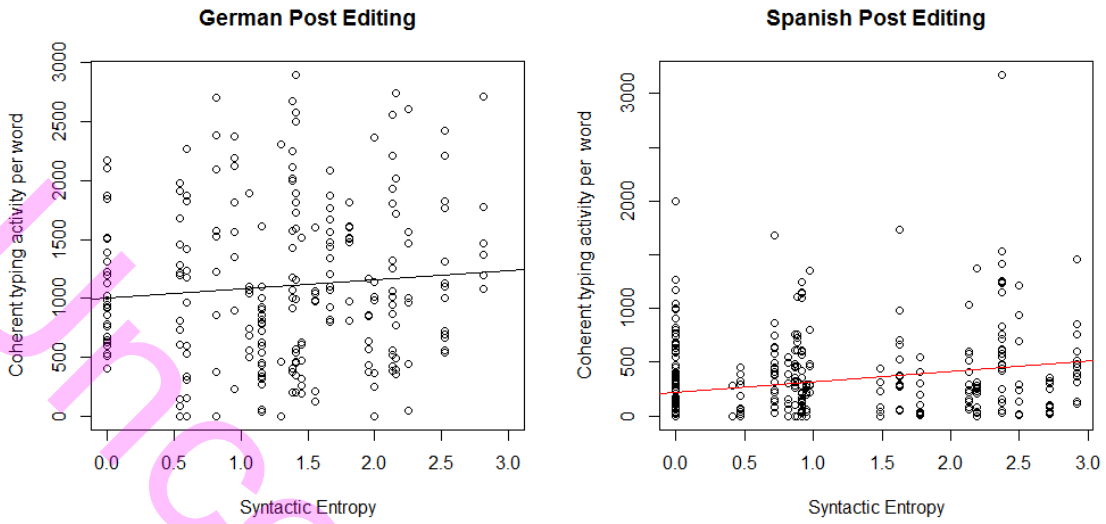


Figure 11: Coherent typing activity per word in milliseconds and syntactic entropy of the post editing condition

We also correlated the average lexical translation entropy with syntactic entropy. For this analysis all languages and tasks were collapsed, i.e. syntactic entropy values and lexical translation entropy values for all four languages and post-editing and translation were analysed jointly. Results showed a highly significant positive correlation ($F(1, 2138) = 13.23$, $p < 0.001$). This result is not surprising given that different structural realisations often require different lexical items. But it also shows that lexico-semantic aspects are difficult to isolate from purely syntactic aspects.

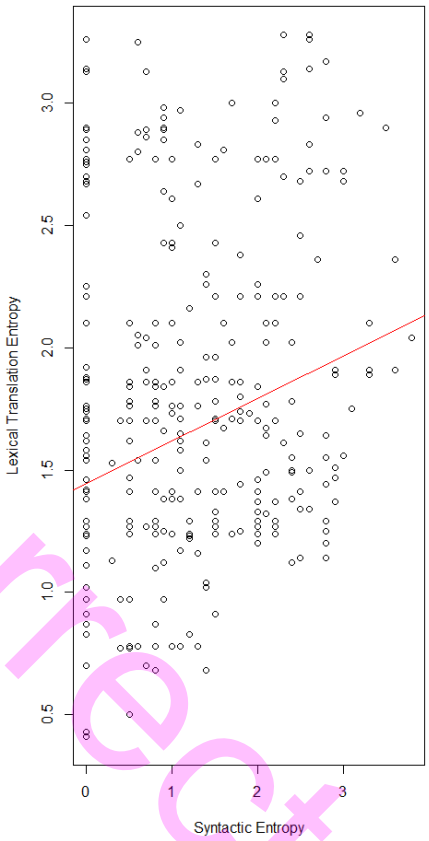


Figure 12: Correlation between syntactic entropy across all languages / tasks and lexical translation entropy.

6.2. Syntactic entropy and structural priming

Source text structures that remained the same during target text production (adherence) were associated with lower syntactic entropy across all languages, while changed structures (deviation) were associated with higher syntactic entropy. The difference between both groups was significant across all four languages with $p < 0.01$.

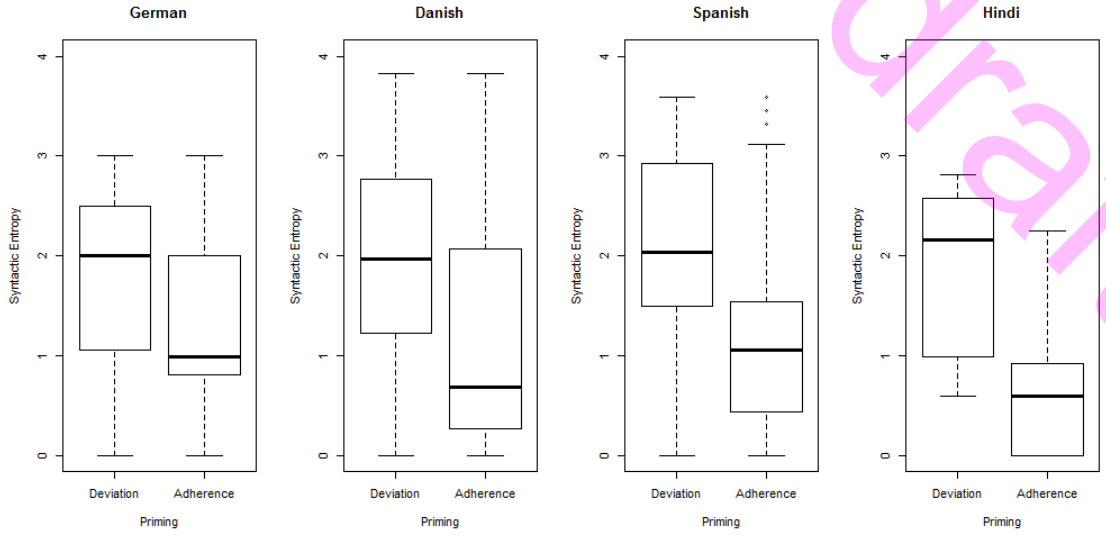


Figure 13: Syntactic entropy of the translation condition by structural deviance and adherence.

In a further step we analyzed how priming might affect behavioral measures. The probability that a source text structure is retained is computed by dividing the number of structures that were kept the same during translations by the total number of translations. A priming probability of 1 means therefore that all translators used the same structure as in the source, while a priming probability of 0 means that none of the translators used the same structure as in the source. Given that syntactic entropy was higher for those sentences which did not adhere to the source structure, we can say that, if translators deviate from the source structure, they are likely to produce many different target structures. However, when translators adhere to the structure, it is likely that all translators produce the same target structure. We correlated this probability with the total reading time on source text words. After outliers were removed in German and Hindi due to the noise created by the long translation time, Danish, Spanish and Hindi showed a negative correlation, which was marginally significant in the Danish case ($F(1, 440)= 3.299$, $p= 0.07$) and significant after outlier elimination ($F(1,416)= 5.47$, $p< 0.05$).In the German condition a positive trend could be observed for the reading time per source text word. Interestingly, translations into German also show negative trend, like the other languages up to a priming probability of 0.65, when a sharp rise of reading time occurs.

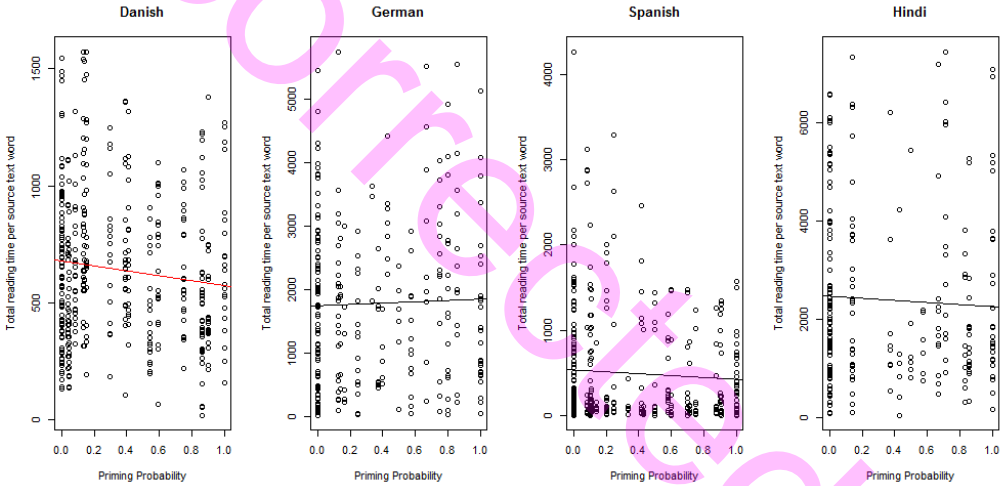


Figure 14: Total reading time per word on the source text as a function of the priming probability

7. Discussion

7.1. Syntactic entropy, behavioural measures and structural priming

The fact that we observe an effect of syntactic entropy on total reading time of the source text segments supports co-activation of the two linguistic systems during translation. In other words, translators activate target language specific aspects during source text reading. It could be argued that sentences with high syntactic entropy are more difficult to process because of source language properties. However, syntactic entropy has no effect on behavioural measures during copying which indicates that syntactic entropy captures processes which are specific to translation. We only see a correlation of syntactic entropy with Kdur during Spanish post-editing suggesting that during post-editing, co-activation plays much less of a role. The fact that we observe a less consistent effect of syntactic entropy on total reading time on the target texts suggests that target language specific process are more often resolved during source text reading than during target text reading, or at least while visual attention is on the source text. In other words, these results suggest that target text reading is a slightly more monolingual process than source text reading. Jakobsen (2011, 48) suggests that the translation process can be broken down into six micro-cycles:

1. Moving the gaze to read the next chunk of new source text (and constructing a translation of it),

2. Shifting the gaze to the target text to locate the input area and read the current target-text anchor word(s),
3. Typing the translation of the source-text chunk,
4. Monitoring the typing process and the screen outcome,
5. Shifting the gaze to the source text to locate the relevant reading area,
6. Reading the current source-text anchor word(s).

On the basis of the evidence presented here, it is likely that steps 1, 5 and 6 are processes which involve a high degree of co-activation while step 4 is likely to involve co-activation to a lesser degree, given that the effect of syntactic entropy on total reading time on the target text was less consistent. Step three is also likely to involve a high degree of co-activation: the effect of syntactic entropy on Kdur supports such a view.

Low entropy, as discussed above, is a result of low degree of variation within the translations, i.e. most or even all translators choose the same syntactic structure. A possible factor that is limiting the potentially infinite choice of the translators, are priming effects. Structural priming effects occur for example in monolingual comprehension when the syntactic structure in one language activates the same structure in the other language, so that subsequent production makes use of this structure. This effect has also been noted for bilinguals in dialogue situations, suggesting that they rely on a shared structural representation (Hartsuiker, Pickering, Veltkamp 2004, 409–414). A similar mechanism might explain the results seen in Figure 13. Low syntactic variation is likely a result of syntactic priming, influencing the translators to reproduce the same syntactic structure that they have read before. This effect is persistent across all analysed languages and results in lower syntactic deviation in the translations. When priming fails and translators cannot reproduce the same source text structure in the target language, entropy is significantly higher.

Priming has been associated with a facilitation effect in translation. Jensen et al report lower processing times due to a possible "automatic transfer of L1 syntax to all types of L2 processing" (Jensen et al 2009, 333). They report shorter total reading time if word order of the source sentence can be maintained in the target sentence. While our shallow annotation does not capture word order, this effect should also be prevalent on other levels of linguistic description, since priming effects are at work on various levels.

The results in figure 14 capture this effect for Hindi, Spanish and Danish, though this facilitation effect, when interpreted as lower total reading time per source text word, is only significant for the Danish translations. In the German condition, we see a facilitation effect only up to a priming probability of about 0.65. Dissonances between primed structures and target language norms might be the reason for this observation. Vandepitte and Hartsuiker (2011) observe this effect for translations from English to Dutch. In their study, adherence to syntactic structure of the source, in some cases results in non-conformity with target language norms. These were attributed to a non-prototypical i.e. unfamiliar use of inanimate subjects in Dutch which resulted in longer translation latencies if the source text structure was maintained in the translation (Vandepitte and Hartsuiker 2011).

We are well aware that our object languages differ in that not all triplet types in one language are shared by the other languages. One such case is the category *passiva refleja* in Spanish, which is non-existing in the other European languages we analyse here. We have included this category in the analysis of valency and voice for Spanish RP (pasiva refleja, reflexive passive).

We should keep in mind that a deeper parse would also distinguish between the different types of sub-clauses and their level of embedding. Here, we do not make a distinction between structures in which part of a segment in the source is translated as a relative clause, an adverbial clause or an argument clause. Moreover, the translator may sometimes have the choice between a finite or a non-finite structure, which may indicate a choice of explicitation or style. This distinction is also not reflected in our analysis. As the present study is a first attempt to measure entropy values against the cognitive effort spent on translation, we wanted to keep the variants as shallow as possible, and yet as deep as needed, i.e. the aim was to annotate a (hypothesized) sufficient minimum of syntactic categories to give us a first reliable answer to our research question.

The number of variants (sets of triplets) in our analysis thus only partly reflects the number of alternative structures the translator has at hand to translate a particular segment.

8. Conclusion

The results presented here extend findings regarding the effect of lexical choice (i.e. lexical translation entropy) to syntax by quantifying syntactic literality and by observing how this affects behavioural measures. The current study lends support to the literal translation hypothesis and, by extension, to the law of interference (Toury 1995). We observe an effect of syntactic entropy on individual languages and across languages suggesting that these effects are not language specific and possibly universal. Results also suggest that post-editing is different from translation, given that the initial machine translated output plays an important role during the process by limiting the choices a post-editor may consider.

9. Acknowledgments

This work was supported by EU's 7th Framework Program (FP7/2007-2013) under grant agreement 287576 (CASMACAT).

References

Bernolet, Sarah, Hartsuiker, Robert J. and Martin J. Pickering. 2013. "From Language-specific to Shared Syntactic Representations: The Influence of Second Language Proficiency on Syntactic Sharing in Bilinguals." Cognition, 127(3): 287–306.

Boltzmann, Ludwig. 1886. The Second Law of Thermodynamics. In: Populare Schriften, Essay 3, address to a formal meeting of the Imperial Academy of Science, 29 May 1886, reprinted in Ludwig Boltzmann, Theoretical physics and philosophical problem, S. G. Brush (Trans.). Boston: Reidel.

Campbell, Stuart. 2000. "Choice Network Analysis in Translation Research." In Intercultural Faultlines. Research Models in Translation Studies I. Textual and Cognitive Aspects, edited by Maeve Olohan, 29–42. Manchester: St Jerome.

Carl, Michael. 2012a. "The CRITT TPR-DB 1.0: A Database for Empirical Human Translation Process Research" In AMTA 2012 Workshop on Post-Editing Technology and Practice.

Carl, Michael. 2012b. "Translog-II: a Program for Recording User Activity Data for Empirical Reading and Writing Research" In LREC.

Clausius, Rudolf. 1865. Über die Wärmeleitung gasförmiger Körper. In: Annalen der Physik 125: 353–400.

Clausius, Rudolf. 1865. Über verschiedene für die Anwendung bequeme Formen der Hauptgleichungen der mechanischen Wärmetheorie: vorgetragen in der naturforsch. Gesellschaft den 24. April 1865. p. 46.

Clifton, J. C., Staub, A., and Rayner, K. 2007. Eye Movements in Reading Words and Sentences. In R. Van

Dragsted, Barbara. 2012 ."Indicators of difficulty in translation — Correlating product and process data" In Across Languages and Cultures 13(1).

Duñabeitia, Jon A., Manuel Perea, and Manuel Carreiras. 2010. "Masked Translation Priming Effects with Highly Proficient Simultaneous Bilinguals." Experimental Psychology, 57(2): 98–107.

Ehrlich, Susan F., and Keith Rayner. 1981. Contextual Effects on Word Perception and Eye Movements during Reading. Journal of Verbal Learning and Verbal Behavior, 20(6), 641–655.

Frazier, Lyn. 1979. On comprehending sentences: syntactic parsing strategies. Unpublished Dissertation. University of Massachusetts.

Frazier, Lyn and Janet Dean Fodor. 1978. The sausage machine: A new two-stage parsing model. Cognition, 6(4), 291–325.

Grosjean, François. 1997. The Bilingual Individual. Interpreting - International Journal of Research and Practice in Interpreting, 2, 163–187.

Hale, John. 2001. A Probabilistic Early Parser as a Psycholinguistic Model. In Proceedings of NAACL, Vol. 2, 159–166).

Jensen, Kristian Tangsgard Hvelplund, Annette C. Sjørup and Laura W. Balling. 2009. Effects of L1 Syntax on L2 Translation. In Fabio Alves, Susanne Göpferich, & Inger M. Mees. (Eds.), Methodology, technology and innovation in translation process research: A tribute to Arnt Lykke Jakobsen (pp. 319–336). Copenhagen: Samfundslitteratur.

Levy, Roger P. 2008. Expectation-based syntactic comprehension. Cognition, 106(3), 1126–77.

Levy, Roger P. and Frank Keller. 2013. Expectation and Locality Effects in German Verb-final Structures. Journal of Memory and Language, 68(2), 199–222.

Macizo, Pedro, and Maria Teresa Bajo. 2006. Reading for Repetition and Reading for Translation: Do they Involve the Same Processes? Cognition, 99(1), 1–34.

Miller, George A. 1956: The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information. In: Psychological Review Vol. 63, 81-97; reprinted in Psychological Review1994, 101, 343-352.

Neubert, Albrecht and Gregory Shreve. 1992. Translation as Text. Kent State University Press.

Ruiz, Carmen, Natalia Paredes, Pedro Macizo and Maria Teresa Bajo. 2008. Activation of Lexical and Syntactic Target Language Properties in Translation. Acta Psychologica, 128(3), 490–500.
Pym, Anthony. 2003. Redefining Translation Competence in an electronic Age. In Defence of a Minimalist Approach. Meta: Translators' Journal, 48(4), 481–497.

Schaeffer, Moritz Jonas and Michael Carl. 2013. Shared Representations and the Translation Process: A Recursive Model. Translation and Interpreting Studies, 8(2), 169 – 190.

Shannon, Claude. E. 1948. A Mathematical Theory of Communication. Bell System Technical Journa. Vol. 27, pp. 379- 423, 623-656.

Shannon, Claude. E. 1951. Prediction and Entropy of Printed English. The Bell System Technical Journal.

Vandepitte, Sonia and Robert Hartsuiker. 2011. Metonymic language use as a student translation problem: towards a controlled psycholinguistic investigation. In Cecilia Alvstad,
Adelina Hild and Elisabet Tiselius (Eds.) Methods and strategies of process research: integrative approaches in translation studies. Benjamins Translation Library Vol. 94. pp. 67-92.