*Michael Carl & Moritz Jonas Schaeffer**

Why Translation Is Difficult: A Corpus-Based Study of Non-Literality in Post-Editing and From-Scratch Translation

Abstract

The paper develops a definition of translation literality that is based on the syntactic and semantic similarity of the source and the target texts. We provide theoretical and empirical evidence that absolute literal translations are easy to produce. Based on a multilingual corpus of alternative translations we investigate the effects of cross-lingual syntactic and semantic distance on translation production times and find that non-literality makes from-scratch translation and post-editing difficult. We show that statistical machine translation systems encounter even more difficulties with non-literality.

Keywords

translation, post-editing, non-literality, statistical machine translation

1. Introduction

Translation would be easy if every word in the source language had only one possible translation in the target language (and vice versa) and the word order would be identical in the source and the target texts. In that case, translation would be reduced to substituting source words for target words that could be enumerated and looked up in a dictionary. If translation was that easy, machine translation (MT) would work perfectly: there would be no need for human translators and not even post-editing of machine translation (PEMT) would be required, since simple, deterministic lexical substitution would produce perfect translations. However, we know that this is not the case. Such absolute literal translations are only exceptionally possible and depend on the similarities and possibilities of the languages involved. Polysemy, different semantic and conceptual representations in the source and the target languages, as well as different syntactic constraints are some of the reasons why translation is non-deterministic and difficult.

Sun (2015: 31) argues that "translation difficulty can be viewed as the extent to which cognitive resources are consumed by a translation task for a translator to meet objective and subjective performance criteria." He mentions a large number of causes, effects and factors for assessing translation difficulty. He suggests a measurement for translation difficulty where the "number of different renditions may not be an effective indicator of translation difficulty" (Sun 2015: 42). However, as Koponen (2016: 24) highlights, the number of variants available to and considered by the translator or post-editor are mentioned as indicators of cognitive effort by both Krings (2001: 536-537) and Dimitrova (2005: 26).

It is our aim to find a way of describing what makes translation difficult and non-deterministic. In order to do so, we suggest a definition of a hypothetical *absolute literal translation* which is syntactically and semantically identical to the source and develop a computational framework

* *Michael Carl*
Renmin University of China &
Copenhagen Business School
Denmark
mc.ibc@cbs.dk

* *Moritz Jonas Schaeffer*
Johannes Gutenberg University Mainz
Germany
mschaeffer@uni-mainz.de

to measure the non-literality of actual translations. We provide theoretical and empirical evidence that absolute literal translations are easier (faster) to produce and we show that the more the translation deviates from the literality criterion, the harder it is and the longer it takes to produce the translation in from-scratch translation and in PEMT. We show that the literality scores of the translation product predict behavioral measures of keystrokes and gaze times in translation and in post-editing. We also show that ambiguities in statistical machine translation and the effort spent in machine translation post-editing can be traced back to the same non-literality phenomena.

In this paper, we present an empirical approach to measure cross-lingual syntactic and semantic similarity using a corpus-driven approach. We base our investigation on a multilingual corpus which contains a large number of alternative translations, i.e., translations of the same source text produced by different translators. We take the variation in word order and the variation of lexical choice in the translations as indicators for syntactic and semantic literality. We focus on temporal effort under the assumption that shorter production times indicate less cognitive effort (see Krings 2001).

Our investigations are based on a subset of the TPR-DB,[1] a multilingual corpus of alternative word-aligned translations from which we induce the semantic and syntactic similarity of the source texts, as well as the traces of cognitive effort to produce these translations.

A substantial body of work in bilingualism studies investigates the mechanisms and constraints of human translation making use of priming studies. Priming studies measure the impact of a previously encountered stimulus on subsequent retrieval processes in order to investigate the underlying mental representations and processes. Therefore, Section 2 presents some concepts in bilingualism research (priming studies). We draw on a psycho-linguistically grounded translation model (Schaeffer/Carl 2013) which explains the findings in the context of shared representations and a recursive model of the translation process. The findings, as well as the model, indicate that more literal translations are easier to produce than less literal translations. In section 3 we operationalise the concept of translation literality and develop two metrics which are used to grade translations as more or less literal. In section 4, we introduce the *multiLing* corpus of translation and post-editing data, which constitutes the empirical basis of our study.

Findings in section 5 suggest that syntactic and lexical choices are closely related in translation and PEMT and that the degree of semantic similarity of the source and the target (i.e. the word translation entropy) has an impact on translation and on PEMT duration. However, section 6 provides evidence that PEMT leads to more literal translations and less lexical variation in the translation product, and it may also result in lower translation quality than from-scratch translation. In section 7, we trace back choices during PEMT to the internal representations of statistical machine translation (SMT) that generated the draft translations which were then post-edited. We find a transitive relation between the complexity in the search graph of the SMT system and the complexity of post-edited output. It shows that if lexical choices are difficult (i.e. plenty) for an SMT system then they are also likely to be difficult for translators. The results suggest that the underlying processes in from-scratch translation and PEMT might share common characteristics.

2. Priming studies and literal translation research

Priming studies are a powerful tool to investigate the nature of the representations in language production or language comprehension. Priming effects describe the impact of a previously encountered item or structure, mostly on the time efficiency of subsequent retrieval processes (Pickering/Ferreira 2008). Participants are usually not aware of the experimental manipulations in priming studies that they take part in, but the built-in repetitions (or non-repetitions) which lead to different response times unveil the hidden mechanisms of the underlying mental architecture.

1 The TPR-DB can be downloaded free of charge from https://sites.google.com/site/centretranslationinnovation/tpr-db

A number of priming studies (Tokowicz/Kroll 2007, Laxén/Lavaur 2010, Boada et al. 2013, Prior et al. 2013, Eddington/Tokowicz 2013) showed that translation recognition, as well as translation production is slowed down if a word has more than one translation alternative. In a translation recognition task, Eddington/Tokowicz (2013) presented unambiguous translations, *synonym translation-ambiguous*, and *meaning translation-ambiguous* source language words. A synonym translation-ambiguous word in English is for example "shy" which can be translated into German as *schüchtern* or *scheu*. In contrast, a meaning translation-ambiguous word is a homograph with several meanings: "odd" can refer to an odd number or to something strange. Depending on which meaning is used, the translations into German are different (*ungerade* or *merkwürdig*) (Eddington/Tokowicz 2013: 442). 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.

Prior et al. (2011) compared single-word de-contextualized 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) assumed 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 found that translations in and out of context overlap to some degree, but translation out of context "by no means matches [translation in context] perfectly" (Prior et al 2011: 108). Form similarity is a stronger predictor in de-contextualized translation choice, whereas word frequency and semantic salience are stronger predictors for context-embedded translation choice.

Dragsted (2012) compared eye movement measures (total reading time (TRT) and number of fixations) and pauses for words which were translated by eight participants using the same target word with words for which the eight participants used different words. Dragsted found that the TRT (total reading time) 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. Dragsted also found that the pauses prior to words with many alternatives were longer as compared to words with one or two alternatives.

Schaeffer/Carl (2013) and Schaeffer et al. (2016) explained these observations by means of a recursive model of the translation process. They distinguished between horizontal priming processes, which activate shared ST-TT representations – so-called combinatorial nodes (Pickering/Branigan 1999, Hartsuiker et al. 2004) – 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 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. Automated processes trigger literal translation correspondences, while later monitoring processes may introduce non-literal relations.

Chesterman (2011:23) argued that the *literal translation hypothesis* "has been implied or explicitly studied by many scholars, and does not seem to have a single source." Catford (1965) and Ivir (1981) introduced the *formal correspondence model*. Tirkkonen-Condit (2005: 408) reformulated 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." Toury (1995: 275) discovered the *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". All these studies imply that word-for-word, one-to-one literal translation cor-

respondences are easier to produce than translations that deviate from the structure of the source text.

3. Measuring translation literality

In order to find a way of describing what makes translation difficult and non-deterministic, we devised a corpus-driven measure for semantic and syntactic similarity to assess the literality of translations.

In section 3.1, we discuss a measure to quantify the degree of semantic similarity of a source text word and its translation(s) based on the following criteria:

1. Each ST word has only one possible translated form in a given context

Section 3.2. introduces a measure to quantify the syntactic similarity of a source text sentence and its translation(s) based on the following criteria:

2. Word order is identical in the ST and TT
3. ST and TT items correspond one-to-one

Section 3.3 (and later also section 4) show that criterion 1 correlates to a high degree with criteria 2 and 3.

3.1. Semantic similarity

Conventionally, the semantic similarity of two words or concepts is derived from the distance between the two words or concepts in a given ontology (e.g. a representation of entities; see Slimani 2013). To do this, a large number of words and concepts have to be encoded in the ontology and the relations between the encoded concepts have to be made explicit. Alternatively, corpus-based methods can be used to compute co-occurrence patterns in large texts and to generate high-dimensional vector representations of words. Distributional lexical semantics models, such as Latent Semantic Analysis (LSA; Landauer/Dumais 1997) use the distance between two vectors to quantify the semantic similarity between the two words.

Carl et al. (2016) introduced a word translation entropy metric (*HTra*) to capture the semantic similarity between a source word and its translations based on the number and distribution of different translations that are available for a given word in a given context. The entropy, ***H,*** represents the average amount of information provided by each new item. It is computed based on the sum of the probability of the items and their information. The information of a probability ***p*** is defined as $\boldsymbol{I(p) = - log_2(p).}$ The entropy ***H*** is the expectation of that information as defined in equation (1):

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

We adopt this notion to assess the entropy of word translation choices for a given ST word ***s*** into its ***n*** possible TT words $\boldsymbol{t_{i...n}}$ as shown in equation (2):

$$HTra = -\sum_{i=1}^{n} p(s \rightarrow t_i) * log_2\big(p(s \rightarrow t_i)\big) \quad (2)$$

Entropy ***H(s)*** (e.g. word translation choices for a given word in the ST) in equation (2) is the sum of all observed word translation alternatives multiplied by their information content. The word translation probabilities $\boldsymbol{p(s \rightarrow t_i)}$ of an ST word ***s*** and the possible translations $\boldsymbol{t_{i...n}}$ are computed as the number of alignments $\boldsymbol{s \rightarrow t_i}$ counted in the TTs divided by the total number of observed TT to-

kens (e.g. translations), as shown in equation (3). Thus, while in language modelling, the entropy indicates how many possible continuations of a sentence exist at any time, we deploy the metric to assess how many different translations a given ST word has.

$$p(s \rightarrow t_i) = \frac{count(s \rightarrow t_i)}{translations} \quad (3)$$

For instance, consider the English sentence fragment from Figure 2 below, '*he was given four*', which is reproduced in Table 1 below. The English sentence was translated into Spanish by 22 translators. The translations were manually word-aligned, words were lower-cased, and identical word translations $\boldsymbol{t_i}$ were counted. For each translation $\boldsymbol{s \rightarrow t_i}$ we thus obtained a number of occurrences in context, as indicated in the first column below the ST word. 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.

	H(he) = 1.95		***H(was) = 2.88***		***H(given) = 3.55***		***H(four) = 0***	
s	# **he**	***I(s→ti)***	# **was**	***I(s→ti)***	# **given**	***I(s→ti)***	# **four**	***I(s→ti)***
t1	12 le	0.87	9 le	1.29	4 dieron	2.46	22 cuatro	0.00
t2	2 recibió	3.46	2 recibió	3.46	4 condenaron	2.46		
t3	2 ---	3.46	2 dieron	3.46	2 recibió	3.46		
t4	4 se	2.46	2 condenaron	3.46	2 condenó	3.46		
t5	1 se lo	4.46	1 se condenó	4.46	1 sentenciado a condenas	4.46		
t6	1 se le	4.46	1 se condemnó	4.46	1 Se le condemnó	4.46		
t7			1 lo condenó	4.46	1 los	4.46		
t8			1 han	4.46	1 impuesto	4.46		
t9			1 fue	4.46	1 han condenado	4.46		
t10			1 declarado	4.46	1 culpable	4.46		
t11			1 ---	4.46	1 conenaron a	4.46		
t12					1 condenó a	4.46		
t13					1 aplicaron	4.46		
t14					1 ---	4.46		

Table 1. English sentence fragment and its word translation alternatives into Spanish from 22 translators. The table also shows the number of occurrences of each translation (#), the information content and the word translation entropy

We divided the number of translation occurrences of *['he→le']* by the overall number of observed translations, i.e. 22 and obtained ***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. Applying equation (2), the information content of this translation amounts to $\boldsymbol{I(he \rightarrow le) = -\log_2(p(he \rightarrow le)) = 0.874}$. Table 1 above shows the distribution of the observed translations for this English sentence fragment. 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. The dis-

tribution of translation choices is also 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. The probability of translation ***p(****four→cuatro****) = 1*** represents no information (no alternative translation is observed) and thus ***H(****four****) = I(****four→cuatro****) = 0***. In contrast, the word [*given*] has many more possible translation alternatives, entropy ***H(****given****) = 3.55*** indicates much more complex translation choices.

One other expression of the number of alternative translations is perplexity. In language 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, and the wider and the more even the distribution of probabilities are, the more difficult it is to make a decision. However, usually, models are preferred that minimize the number of possible choices, given that they have the same explanatory power. Perplexity (***PP)*** is related to entropy ***H***, as an exponential function as shown in equation (4):

(4) $PP = 2^H$

3.2. Syntactic similarity

The syntactic similarity of two sentences can be described by measuring how far they are away from a word-for-word translation. Whenever alignment links cross in a pair of word-aligned source-target sentences, we observe a syntactic re-ordering between the two languages. From a given translation and its 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 word positions in the ST and the TT, indicating the word order similarity of the two sentences.

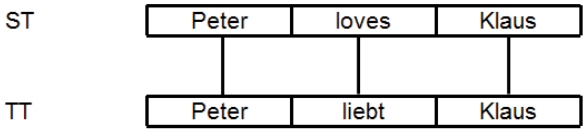


Figure 1. Absolute literal translation – each ST word is translated into one TT word and hypothetical alternative translations are all identical

Figure 1 above shows an absolute literal translation according to the literality 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 (absolute) literal translation. The *Cross* values indicate the monotonicity of the translations: in the case of an absolute literal English-German translation, as in Figure 1 above, we say that each successive word aligns with the next one in the target language, which provides the vectors *CrossS*=*CrossT*={1,1,1}. A slightly more complex translation is given in Figure 2 below:

(ST) EN: He was given four life sentences, one for each of the killings.

(TT) ES: Se le aplicaron cuatro cadenas perpetuas, una por cada asesinato.

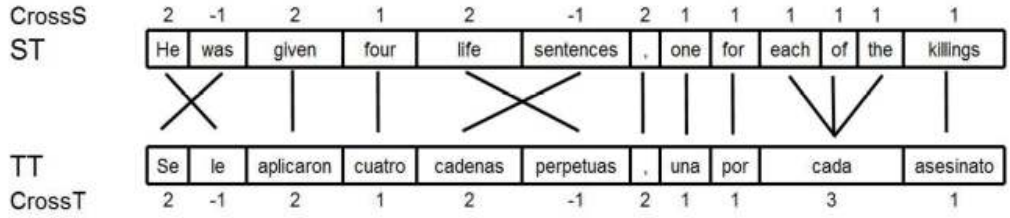


Figure 2. English-Spanish alignment with CrossS and CrossT values

Following the path of the alignment links generates two different *CrossS* and *CrossT* vectors as shown in Figure 2. The word [*He*] occurs at position 1 on the English source side while its Spanish translation [*le*] occurs one word ahead at position 2 in the translation. [*He*] thus has a *CrossS* value of 2, while Spanish [*le*] has a *CrossT* value of -1. In order to generate the translation [*aplicaron*] for the English [*given*] we need to jump from the previous alignment [*was - Se*] two words to the right, which produces a *CrossS* value of 2. In this way, *Cross* values are generated for each word position in the source and target text. This method does not penalize re-ordering of phrases if they have the same internal word order.

The word order choices that a translator has are captured in the metric *HCross,* which is calculated based on equation (5), and similar to word translation entropy in equation (2). The probability *p(Cross(s))* is computed as the number of observed *Cross(s)* values divided by the total number of observed translations for *s* (*count(s)*), where *s* is a given ST word.

$$HCross = -\sum_{i=1}^{n} p\big(Cross(s)\big) * log_2\Big(p\big(Cross(s)\big)\Big) \quad (5)$$

3.3. Correlation of lexical and syntactic choices

As a measure of semantic similarity, *HTra* measures the number of possible equivalent lexical items from which a translator can choose. *HCross* measures the number of possible syntactic renderings of the translation. Interestingly, their values correlate to a high degree (r=.79, $p < .001$). That is, semantic and syntactic variation seem to correlate highly in translation. More variation in syntactic (word-order) rendering of the translation seems to occur with more variation in lexical choices, and vice versa: low semantic cross-lingual similarity (i.e. high *HTra* values) correlate with high syntactic variation and complexity (i.e. high *HCross* values).

According to our definition of semantic similarity, a low word translation entropy entails few translation choices and a large semantic overlap between the source and the target language concepts. Similarly, low absolute *Cross* values indicate a high syntactic similarity so that literal translations are associated with low values and non-literal translations receive high scores.

4. Experimental material

The TPR-DB corpus was assembled over the past 10 years and currently contains more than 1,500 text production sessions (translation, post-editing, editing, dictation, revision, authoring and copying) in more than 10 different languages. For many of the text production sessions, keystroke and gaze data were collected and stored, and translations were semi-automatically aligned using the YAWAT tool (Germann 2008). In some cases, the STs and TTs were first automatically pre-aligned using the GIZA++ tool (Och/Ney 2003). The data were then converted into the YAWAT format, and (many of the) alignments were manually checked and corrected where necessary. Once the alignments were manually confirmed, the data were further processed into a set of summary tables, which described the text production processes, e.g. eye movements and keystrokes and the text product (TTs) by means of more than 200 different features. These features comprise, among others, the number and duration of fixations on ST and TT words, text production times,

number of insertions and deletions per word and the *Cross* and *HTra* values. A complete description of the features can be found in Carl et al. (2016) and on the CRITT website.[2]

The *multiLing* subset that is used in this paper contains multilingual translations of the same English source texts into several target languages (Danish, German, Spanish, Hindi, Japanese, Chinese) generated in different translation modes, among others from-scratch translation (translation with no additional external assistance) and PEMT. All translation sessions were recorded using Translog-II (Carl 2012) which logs all keystrokes with a time stamp. In addition, all sessions were recorded using an eye-tracker[3] and then gaze data were synchronized with the keystroke log. These data enable us to correlate properties of the translation product (e.g. translation literality) and the translation process. This allows us to assess the difficulties associated with from-scratch translation and PEMT, and to pinpoint typical patterns that expose the from-scratch translation and PEMT difficulties.

		ST:Words		ST 1:160		ST 2: 154		ST 3: 146		ST 4: 110		ST 5: 139		ST 6: 139		Total / Average		
Task	Study	TL	#part	TokT	Alt	TokT	Alt	TokT	Alt	TokT	Alt	TokT	Alt	TokT	Alt	TtokT	Ttseg	Dur
P	BML12	es	31	1733	10	2190	12	1600	10	1598	12	1229	8	1866	12	10216	431	5.22
	ENJA15	ja	39	2997	13	2359	12	2690	14	1882	12	2387	13	2132	12	14447	519	16.81
	MS12	zh	11	471	3	717	5	387	3	316	3	417	3	253	2	2561	129	3.18
	NJ12	hi	22	1371	7	2031	12	1353	8	1267	10	1802	12	1541	11	9365	409	18.20
	SG12	de	23	1227	8	1122	7	1097	7	860	8	1049	7	1115	8	6470	305	8.78
T	BML12	es	32	1955	11	1853	10	1819	8	1331	10	1820	12	1160	8	9938	411	10.34
	ENJA15	ja	39	2915	12	2464	13	2208	12	1992	13	2395	13	2160	13	14134	525	22.46
	KTHJ08	da	24	3703	24	3524	23	3440	22	0	0	0	0	0	0	10667	523	7.70
	MS12	zh	10	0		417	3	386	3	328	3	401	3	384	3	1916	89	4.12
	NJ12	hi	20	1169	7	1116	7	891	5	897	7	875	6	835	6	5783	266	14.84
	SG12	de	24	1078	6	1310	8	1296	8	888	8	1167	7	1038	8	6777	305	12.46
Total/ Average				18619	101	19103	112	17167	100	11359	86	13542	84	12484	83	92274	3912	124.11

Table 2. Properties of the TPR-DB *multiLing* corpus, which consists of a large number of alternative translations for six different English source texts into several target languages

Table 2 above shows the properties of the *multiLing* subset, which contains translations from English into six languages, Danish (da), Spanish (es), German (de), Hindi (hi), Chinese (zh) and Japanese (jp), that are used for the present purpose. All texts in this set were manually word-aligned with the YAWAT tool. A maximum of six English source texts (Text 1 to Text 6) were translated into each of the languages, in a from-scratch translation (T) and post-editing mode (P). The length in words for each of the six source texts is given in the first row. For each of the six data sets, the table indicates the number of participants and for each of the six source texts the number of alternative translations (Alt) and their total number in tokens (TokT). The total number of target words (TtokT) and target sentences (Ttsg) are also provided, together with the total production duration in hours (Dur). According to these figures, a translated sentence has on average approximately 24 words. The largest number of translations, in terms of translators, target tokens, segments and translation duration has been collected for the language pair English to Japanese; the least amount for English to Chinese.

Each ST contains between 5 and 12 sentences (segments), with an overall average of 7.3 segments per text. Each ST was translated by between 3 (MS12) and up to 24 (KTHJ08) different translators. The STs contain between 110 and 160 words, on average, 146 tokens (words) per text, which translate, on average, into 151 TT words. Not all languages behave in a similar manner. With the exception of Chinese, there is a tendency to produce more tokens in the TT than there are

2 https://sites.google.com/site/centretranslationinnovation/home

3 Various eye-trackers were used: Older experiments were recorded with Tobii 1750, T120, and more recent ones with the Tobii 300 and SMI-RED250mobile

in the ST sentences. Particularly translations into Spanish have 12% more words in the TT than in the ST. For more information on this dataset, please consult the CRITT website.[4]

5. Semantic similarity and translation duration

Translations into different languages differ with respect to their average *HTra* and *HCross* values. Average *HCross* values indicate the entropy of syntactic variation between the source and the target languages, whereas *HTra* measures the entropy of the lexical variation in the translation. As all translations were generated from the same six English source texts, we can take *HTra* and *HCross* values as indicators of syntactic and semantic similarity of the source and target languages. Taking out the numbers for Chinese,[5] the figures in Table 3 below indicate a clear distinction between the European languages (da, es, de) and the Asian languages (ja, hi).

	HTra	HCross
Da	1.51	0.99
Es	1.55	1.22
De	1.67	1.52
Hi	2.82	2.40
Ja	3.18	2.85
Average	2.15	1.79

Table 3. Lexical and syntactic choices correlate

With respect to syntactic variation, the Danish translations are closest to the English source with an average *HCross* value of 0.99 followed by Spanish (1.22) and German (1.52), while Hindi and Japanese have much higher values of 2.40 and 2.85 respectively. A similar distinction between the close languages (da, es, de) and the more remote languages (hi, ja) are also observed with respect to the *HTra* values which can be interpreted as indicators of interlingual, semantic similarity. Note that these values are computed based on the joint set of the post-edited and from-scratch translations. It was thus possible to consider larger sets, for instance for Spanish we considered more than 20 alternative translations, instead of approximately 10 in the post-editing mode and 10 in the translation mode. As Table 3 above shows, *HTra* and *HCross* strongly correlate ($r = 0.98$). According to our definition above, translations from English into Danish are thus more literal, than, for instance, translations into Japanese or Hindi.

We tested a linear mixed effects model (LMEM)[6] with the following predictors: word length (in characters), typing inefficiency (a measure of how many corrections were made, see Carl et al. 2016 for a detailed explanation), CrossS values and *HTra*. All these predictors were positive and highly significant. The random variables were ST word and Participant. Figure 3 below shows the impact of word translation entropy (*HTra*) on the (log transformed) translation production time (*Dur*) for the two tasks (Translation (T) and Postediting (P)). These translations have a much higher branching factor for translation ambiguous words than the two alternative translations used in the study by Eddington/Tokowicz (2013, see above). They are also more representative of professional translation, because the source texts were complete texts rather than single words. The graph in Figure 3 shows a strong effect for *HTra* on *Dur* for PEMT and a slightly weaker one for Translation. The main effect of HTra was highly significant (β=3.745e-01, t=34.26, p< 0.001). The interaction between Tasks P and T was also highly significant (β= -1.480e-01, t= -13.70, p<

4 sites.google.com/site/centretranslationinnovation

5 The very few alternative translations for Chinese make these numbers statistically very uncertain

6 For all the analyses in this study, R (R Development Core Team 2014), the lme4 (Bates 2014) and languageR (Baayen 2013) packages were used to perform (generalized) linear mixed-effects models ((G)LMEMs). To test for significance, the R package lmerTest (Kuznetsova 2014) was used.

0.001). What these results show is that the effect observed in the rather controlled experiments as reported by Eddington/Tokowicz (2013, see section 2 above) – where participants were slower and less accurate when responding to words that had more than one translation compared to unambiguous words – holds in a more natural environment such as the one used in our dataset and it also holds across a large number of languages and two different tasks.

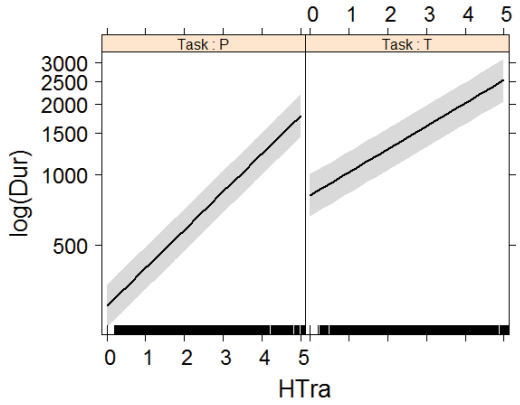


Figure 3. Effect of word translation entropy (HTra) on word production duration

6. Word Translation Perplexity

During the post-editing process, a post-editor usually only sees one single best translation produced by an MT system and then amends this output. Post-editors are heavily primed during this process so that they more easily accept sub-optimal translations which human translators, working from scratch, would otherwise not produce. The two graphs in panels A and B in Figure 4 below plot perplexity values (see section 3.1.) of English → German (A) and English → Spanish (B) translations for different part-of-speech (PoS) tags. The word translation perplexity values shown in panel A in Figure 4 below 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 (see Table 2, SG12 study). The graph plots perplexity values per PoS[7] 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) and VBN (participle) produce more variants in the target language. Note, however, that in all cases the perplexity of the post-edited texts is smaller than in the versions that were translated from-scratch.

7 http://www.monlp.com/2011/11/08/part-of-speech-tags/

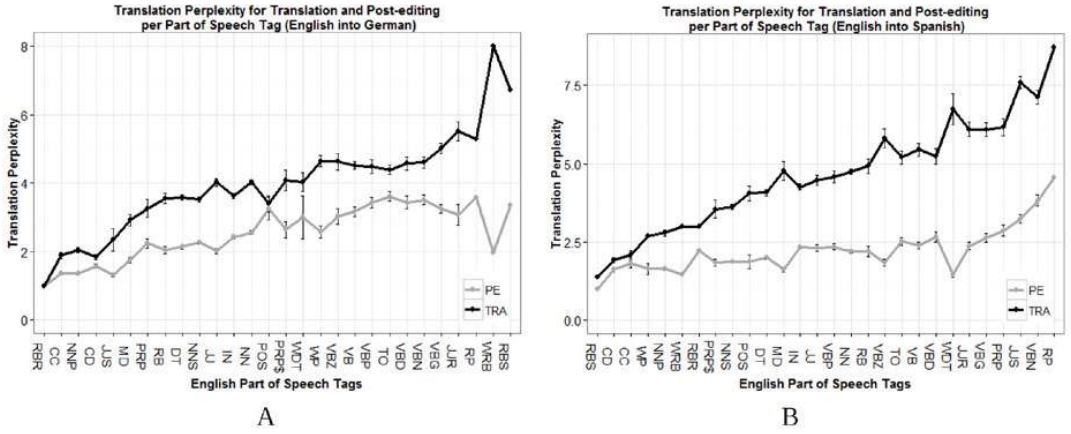


Figure 4. Perplexity of English → German (A) and English → Spanish (B) translations per word class and for from-scratch translation (TRA) and for post-editing (PE)

A similar picture is provided in the graph in panel B in Figure 4, which is based on the same English source texts translated and post-edited into Spanish by eight different translators. The number of words is thus very similar to the German translations in Figure 4 above, but perplexity values are slightly lower (see also Table 2 above), and always lower for post-editing than for translation from scratch. Note also that some PoS tags in the Spanish translation 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.

These data show that PEMT produces more literal translations 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 from-scratch translated versions of the same texts, indicating a compositional translation of expressions during post-editing that would be transferred more idiomatically during from-scratch translation. 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"* were produced. The expression "*Als Geste*", for example, has a much higher likelihood (0.000017%) of being found in a large corpus such as Google Books than "*In einer Geste*" (0.000004%)[8] – while not strictly speaking incorrect, it is much rarer than the other more idiomatic expressions produced in the from-scratch translations.

7. Word translation entropy, SMT and PEMT

State-of-the-art SMT systems can 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. Carl/Schaeffer (2014) showed that post-editing duration is closely correlated with the perplexity of the MT search graph: the more possible translation alternatives an MT

8 Figures were extracted from google ngram viewer

search graph encodes, the more post-editing time increases. In analogy with the perplexity of human word translations, we also computed the perplexity of possible word translation choices in an MT system. 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, here we are only interested in the entropy and perplexity of word translations that are encoded in a search graph.

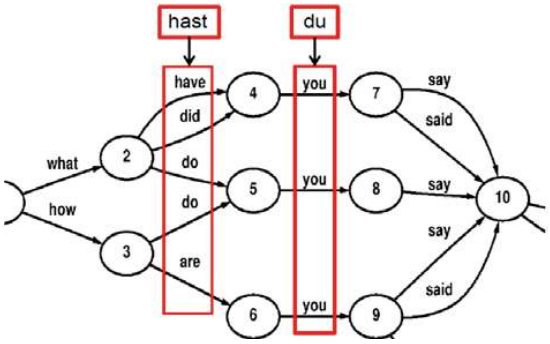


Figure 5. Machine Translation search graph (adapted from Och et al. 2003)

Figure 5 above 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 5 indicate the possible translations for [*hast*] and [*du*]:

hast → *{ have, did, do, are }*

du → *{ you }*

Similar to the previous example (see Table 1 above) in which English [*four*] has only one observed translation [*cuatro*], the search graph in Figure 5 also shows only one possible translation for [*du* → *you*], and the entropy is therefore ***H(du) = 0***.

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)*** will be higher than ***H(du)*** for which the search graph encodes [*you*] as the only possible translation.

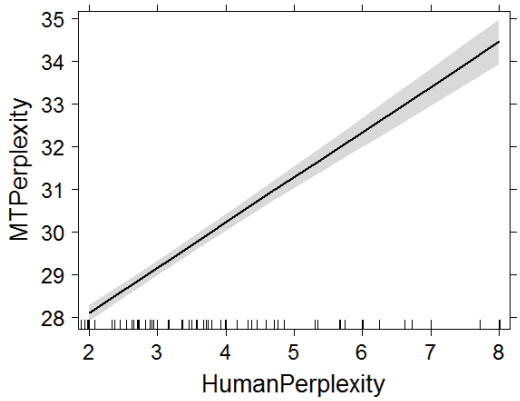
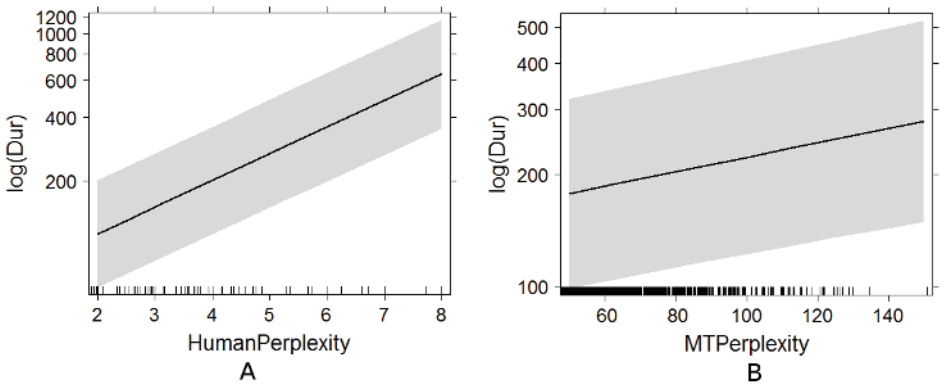


Figure 6. Simple linear regression for perplexity values based on the translations post-edited by humans (HumanPerplexity) and the perplexity values from the MT search graph (MTPerplexity)

Figure 6 above shows data acquired in the 3rd CASMACAT field trial (Carl et al. 2014). It shows the effect of perplexity values based on the translations post-edited by humans (HumanPerplexity) on the perplexity values from the MT search graph (MTPerplexity). The perplexity as generated by the search graph is considerably higher – as compared to the perplexity in the translations post-edited by humans. However, the effect of HumanPerplexity on MTPerplexity was highly significant (β=1.06, SE=0.05, t=20.16, $p < 0.001$). This effect can be understood on the basis that SMT 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.

Figures 7A and 7B below show the effect of perplexity of post-edited translations (HumanPerplexity) and MT search graph perplexity (MTPerplexity) on (log transformed) post-editing duration (Dur). The effects shown in these figures stem from a linear mixed effects model (see Section 5) with word length (in characters), typing inefficiency (Carl et al. 2016), CrossS values, MTPerplexity and HumanPerplexity as predictors. The random variables were ST word and Participant. All predictors were positive and highly significant. The effect of HumanPerplexity on (log transformed) post-editing duration (*Dur*) was relatively strong and highly significant (β=2.908e-01, SE=1.103e-02, t=26.37, $p < 0.001$). The effect of MTPerplexity on (log transformed) post-editing duration was weaker and noisier, but also highly significant (β=4.503e-03, SE=8.978e-04, t=5.02, $p < 0.001$). In summary, we can say that there is a relation between translation ambiguities (word translation entropy and perplexity), the perplexity of the search graph when an SMT system produces a translation and the final post-edited translations.



Figures 7A and 7B. The effect of translation perplexity on post-editing duration (Panel A) on the basis of translations post-edited by humans (HumanPerplexity), and the effect of translation perplexity on post-editing duration (Panel B) on the basis of the MT search graph (MTPerplexity)

8. Conclusion

In this paper we developed a translation literality metric that is based on the similarity of cross-lingual syntactic and semantic relations in the translation product and applied the metric to a multilingual corpus of alternative translations. We reported several experiments which used linear regression analyses to predict properties within and across the translation product and the translation process. We found strong correlations of cross-lingual semantic and syntactic similarities and that non-literal translations were more difficult and time consuming (e.g. higher temporal effort) to produce than literal ones. We found more lexical variation in from-scratch translations than in post-editing. Interestingly, the same properties that make human translation difficult (e.g. high number of lexical and syntactic choices) also make machine translation and consequently post-editing difficult.

References

Baayen, R. Harald 2013: *languageR: Data sets and Functions with Analyzing Linguistic Data: A Practical Introduction to Statistics* [online]. http://cran.r-project.org/package=languageR.

Bates, Douglas M./Maechler, Martin/Bolker, Ben/Walker, Steven 2014: *{lme4}: Linear mixed-effects models using Eigen and S4* [online]. http://cran.r-project.org/package=lme4.

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

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

Carl, Michael/García Martínez, Mercedes/Mesa-Lao, Bartolomé 2014: CFT13: A Resource for Research into the Post-editing Process. In Calzolari, Nicoletta/Choukri, Khalid/Declerck, Thierry/Loftsson, Hrafn/Maegaard, Bente/Mariani, Joseph/Moreno, Asuncion/Odijk, Jan/Piperidis, Stelios (eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Paris: ELRA, 1757-1764.

Carl, Michael/Schaeffer, Moritz J. 2014. Word Transition Entropy as an Indicator for Expected Machine Translation Quality. In Miller, Keith J./Specia, Lucia/Harris, Kim/Bailey, Stacey (eds.), *Proceedings of the Workshop on Automatic and Manual Metrics for Operational Translation Evaluation. MTE 2014.* Paris: ELRA 2014, 45-50.

Carl, Michael/Schaeffer, Moritz J./Bangalore, Srinivas 2016: The CRITT Translation Process Research Database. In Carl, Michael/Bangalore, Srinivas/Schaeffer, Moritz J. (eds.), *New Directions in Empirical Translation Process Research: Exploring the CRITT TPR-DB*. Cham, Heidelberg, New York, Dordrecht, London: Springer, 13-54.

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 Alvstad, Cecilia/Hild, Adelina/Tiselius, Elisabet (eds.), *Methods and Strategies of Process Research: Integrative approaches in Translation Studies*. Amsterdam/Philadelphia: John Benjamins Publishing Company, 23-35.

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

Dimitrova, Birgitta Englund. 2005: *Expertise and Explicitation in the Translation Process*. Benjamins Translation Library, 64.

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

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

Germann, Ulrich 2008: Yawat: Yet Another Word Alignment Tool. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Demo Session. HLT-Demonstrations '08. Stroudsburg*. PA, USA: Association for Computational Linguistics, 20-23.

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

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

Koponen, Maarit 2016. *Machine Translation Post-editing and Effort. Empirical Studies on the Post-editing Process*. Unpublished PhD Thesis.

Krings, Hans P. 2001: *Repairing Texts: Empirical Investigations of Machine Translation Post-Editing Processes* (Geoffrey S. Koby, ed.). The Kent State University Press, Kent, Ohio & London.

Kuznetsova, Alexandra/Christensen/Rune, Haubo Bojesen/Brockhoff, Per Bruun 2014: lmertest: Tests for Random and Fixed Effects for Linear Mixed Effect Models (lmer Objects of lme4 Package). R package version 2.0-6 [online]. http://www.cran.rproject.org/package=lmerTest/.

Landauer, Thomas K./Dumais, Susan T 1997: A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. In *Psychological Review* 104(2), 211-240.

Laxén, Jannika/Lavaur, Jean-Marc 2010: The Role of Semantics in Translation Recognition: Effects of Number of Translations, Dominance of Translations and Semantic Relatedness of Multiple Translations. In *Bilingualism: Language and Cognition* 13(2), 157-183.

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

Och, Franz Josef/Zens, Richard/Ney, Hermann 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.

Pickering, Martin J./Ferreira, Victor S. 2008: Structural Priming: A Critical Review. In *Psychological Bulletin* 134(3), 427-59.

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

Prior, Anat/Kroll, Judith F./Macwhinney, Brian 2013: Translation Ambiguity but not Word Class Predicts Translation Performance. In *Bilingualism: Language and Cognition* 16(2), 458-474.

Prior, Anat/Wintner, Shuly/MacWhinney, Brian/Lavie, Alon 2011: Translation Ambiguity in and out of Context. In *Applied Psycholinguistics* 32(1), 93-111.

R Development Core Team, 2014. *R: A Language and Environment for Statistical Computing*. Vienna, Austria.

Schaeffer, Moritz J./Carl, Michael 2013: Shared Representations and the Translation Process: A Recursive Model. In *Translation and Interpreting Studies* 8, 169-190.

Schaeffer, Moritz/Dragsted, Barbara/Hvelplund, Kristian T./Winther Balling, Laura/Carl, Michael. 2016: Word Translation Entropy: Evidence of Early Target Language Activation During Reading for Translation. In Carl, Michael/Bangalore, Srinivas/Schaeffer, Moritz J. (eds.), *New Directions in Empirical Translation Process Research: Exploring the CRITT TPR-DB*. Cham, Heidelberg, New York, Dordrecht, London: Springer, 183-210.

Slimani, Thabet 2013: Description and Evaluation of Semantic Similarity Measures Approaches. In *International Journal of Computer Applications* 80(10), 25-33.

Sun, Sanjun (2015). Measuring transation difficulty: theoretical and methodological considerations. In *Across languages and cultures* 16(1), 29-54.

Tirkkonen-Condit, Sonja 2005: The Monitor Model Revisited: Evidence from Process Research. In *Meta* 50(2), 405-414.

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

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