Statistical Machine Translation Support Improves Human Adjective Translation

Gerhard Kremer, Matthias Hartung, Sebastian Padó, and Stefan Riezler*
Institute for Computational Linguistics, University of Heidelberg
{kremer,hartung,pado,riezler}@cl.uni-heidelberg.de

In this paper we present a study in computer-assisted translation, investigating whether non-professional translators can profit directly from automatically constructed bilingual phrase pairs. *Our support is based on state-of-the-art statistical machine translation (*SMT*), consisting of a phrase table that is generated from large parallel corpora, and a large monolingual language model. In our experiment, human translators were asked to translate adjective–noun pairs in context in the presence of suggestions created by the* SMT *model. Our results show that* SMT *support results in an acceptable slowdown in translation time while significantly improving translation quality.*

1 Introduction

Translating a sentence adequately from one language into another is a difficult task for humans. One of its most demanding subtasks is to select, for each source word, the best out of many possible alternative translations. This subtask is known, in particular in computational contexts, as *lexical choice* or *lexical selection* (Wu and Palmer, 1994).

Bilingual lexicons which are commonly used by human translators contain by no means all information that is necessary for adequate lexical choice, which is often determined to a large degree by *context*. Often, dictionaries merely list a small number of translation alternatives, or a small set of particularly prototypical contexts is provided. The provided translations are neither exhaustive, nor do they provide distinguishing information on which contexts they require.

In this study, we ask whether the shortcomings of traditional dictionaries can be evaded by directly using a data structure used in most current machine translation (MT) systems, namely *phrase tables* (cf. Koehn, 2010b). Phrase tables are merely bilingual lists of corresponding word sequences observed in parallel corpora, and thus provide a compact representation of the translation information inherent in a corpus, complemented with statistical information about the correspondences (e. g., frequencies or association measures). Together with the orthogonal information source of a monolingual language model, phrase tables build the core components of state-of-the-art statistical machine translation (SMT). While phrases serve the purpose of suggesting possible translations found in parallel data, the purpose of the language model is to fit the phrase translations into the larger context of the sentence. In our experiment, we will extract bilingual phrase pairs from the SMT output of *n*-best translations of the input sentence. In this manner, we directly deploy the information available from SMT to support human translators.

*Credit for the implementation of the experiment GUI goes to Samuel Broscheit. We are grateful to our significant others for participating in the quality ratings.

The current study focuses on one construction, namely the translation of adjectives in attributive position (preceding a noun). This task is fairly simple and can be manipulated more easily than sentence-level translation. At the same time, it is complex enough to be interesting: adjectives are known to be highly context-adaptive in that they express different meanings depending on the noun they modify (Sapir, 1944; Justeson and Katz, 1995). They also tend to take on figurative or idiomatic interpretations, again depending on the semantics of the noun in context (Miller, 1998). Lexical choice is therefore nontrivial, and context-dependent translations are seldom given systematically in dictionaries. For example, consider the adjective *heavy*. In noun contexts like *use*, *traffic*, and *investment*, its canonical translation as German *schwer* is inappropriate. It might be translated as *intensiv(e Nutzung)*, *stark(er Verkehr)*, and *groß(e Investition)*.

Another reason for the restricted experimental setup is to control for translation complexity explicitly. While previous experiments on computer-aided translation could show a significant increase in productivity and quality for machine-assisted translation (especially for less qualified translators), they can only demonstrate a weak correlation between translation times and translation quality. This is due to the varying complexity of test examples and the varying degree of expertise of human translators. In our experiments, we aim to control the variable of translation complexity better, by restricting the task to translations of adjectives in noun contexts, and by providing machine assistance for these pairs only. Furthermore, the human translators in our experiments were all native speakers of the target language, German, with a similar level of expertise in the source language, English. The goal of our experiment is to provide a basis for re-interpretation of results by using a clear and simple experimental design which allows us to analyse the contribution of each variable.

Our experimental results show that, at least for translation from English into German by native German speakers, phrase table support results in an acceptable slowdown in translation time while significantly improving translation quality. This confirms the conclusions drawn in previous studies through evidence from a rigidly controlled experiment.

2 Related Work

Interactive MT systems aim to aid human translators by embedding MT systems into the human translation process. Several types of assistance by MT systems have been presented: *translation memories* (Bowker, 2012) provide translations of phrases recurring during a project. Such phrases have to be provided by the translator the first time they appear, and they are typically restricted to a document, a project, or a domain (cf. Zanettin, 2002; Freigang, 1998).

A closer interaction with human translators is explored in the TransType system of Langlais et al. (2000). Here, the machine translation component makes *sentence completion predictions* based on the decoder's search graph. The interactive tool is able to deal with human translations that diverge from the MT system's suggestions by computing an approximate match in the search graph and using this as trigger for new predictions (Barrachina et al., 2008).

Other types of assistance integrate the phrase tables of the MT systems more directly: Koehn and Haddow (2009) and Koehn (2010a) deploy a phrase-based MT system to display word or phrase *translation options* alongside the input words, ranked according to the decoder's cost model. Finally, full-sentence translations can be supplied for

Variability class	Translation support condition			Noun context
	None	Adjective unigrams	Adjective–noun bigrams	
High	5	5	5	} × 4
Low	5	5	5	

Table 1: Partitions of the set of 30 adjective stimuli presented to each participant for the factors variability *and* support. *Factor* context: *Each adjective was shown in 1 out of 4 sentences. Each context combines the adjective with a different noun.*

post-editing by the user.

Our approach is most closely related to the display of translation options alongside input words. Similarly to Koehn and Haddow (2009), we use a web applet to display options and record reaction times. However, our experiment is deliberately restricted to translations of adjectives in noun contexts, in order to explicitly control for translation complexity, an aspect that has been missing in previous work.

3 Experimental Approach

This section presents an overview of the experimental design and describes how the set of stimulus items was assembled.

The study comprises two experiments. In the first experiment (cf. Section 4 on page 108), participants performed a translation task with different types of supporting information provided by the machine translation system (no suggestion, best unigram translation of the adjective, best bigram translation of the adjective–noun pair). In order to test the impact of presenting phrase tables on translation speed, we measured reaction times between specific time points during each of the participants' translation tasks, using time gain/loss as a measure for the usefulness of machine-aided human translation as discussed in Gow (2003).

The second experiment complements the time aspect with a measure of the translation's quality (cf. Section 5 on page 112).[1] We collected human judgements for all translations from experiment 1 on a simple three-point scale. This appears to be the only feasible strategy given our current scenario which focuses on local changes, i. e., the translation of individual words, which are unlikely to be picked up by current automatic MT evaluation measures like BLEU (Papineni et al., 2002) or TER (Snover et al., 2006).

Participants in the experiment were asked to translate an attributive adjective in sentential context (e. g., *bright* in "The boy's *bright* face, with its wide, open eyes, was contorted in agony."), given one of our set of translation support types. With German participants, we investigated translations from English into German, the participants' native language. This is the preferred type of translation direction in professional human translation, as the translator's experience of commonly used words in a particular semantic context is more extensive in the native language. In this experiment we assumed four factors to interact with translation speed and accuracy (cf. Table 1): adjective (30 different items), noun context (4 sentences per adjective, each sentence with a different adjacent noun), variability class (2 levels), and translation

[1] Note that there have been ongoing debates on how translation quality can be assessed objectively (cf. House, 1998). For example, see Reiß (1971) for a discussion on factors to consider when evaluating a translation.

support (3 conditions), all of which are described in more detail below.

Given these considerations, each experimental item is an instance of an adjective in sentence context combined with some type of translation support. As shown in Table 1, we sampled a total of 120 experimental items for 30 adjectives. To avoid familiarity effects, we ensured that each participant saw only one instance of each adjective. Consequently, we showed each participant exactly 30 experimental items. Each participant saw 3 differing sets of 10 adjectives in one of our three support conditions.

3.1 Variability Classes

Stimuli for the translation experiment have been collected by examining the most frequent adjectives from the British National Corpus (BNC), many of which are polysemous, i. e., showing high context-dependent variability in translation (cf. Section 1 on page 103).

To verify this postulated relationship between corpus frequency and degree of polysemy, 200 high-frequent adjectives from the BNC were used in a measurement of translation variability. We defined the variability as the number of times an English adjective lemma in a two-word phrase was translated into a different German lemma[2] according to the EUROPARL v6 phrase table (see Koehn, 2005). Two-word phrases should roughly account for adjectives in noun context (please note that the translated phrases were constrained to consist of exactly two words, but neither correspondence of nouns nor word order was checked). All translations that occurred only once for a given target lemma in the phrase table were considered spurious translations and thus were excluded.

The set of high-frequent adjectives from the BNC showed a highly significant correlation (Spearman's $\rho = 0.5121$) between corpus frequency and variability in translation (operationalised as the number of unique translations in the EUROPARL v6 phrase table). We divided adjectives into two classes and collected our targets from both extremes: one set that shows a particularly high variability in unique translations, and one set with a relatively low translation variability.

Hypothesis Highly variable adjectives are more difficult to translate, but translators will profit more from the presentation of phrase table information.

3.2 Adjectives and Contexts

For each of the two variability classes (according to the phrase table) we selected 15 adjectives (see Appendix A.1). For each English adjective, we randomly sampled four full sentences from the BNC (Burnard, 1995) parsed with the C&C parser (Clark and Curran, 2007) as experimental items, with the adjective in attributive position directly preceding a noun so that the modified noun was different for each sentence.

In order to further minimise variation in translation times, we imposed some constraints on the sentences. Their length was restricted both in terms of words (15–20) and characters (80–100). Also, sentences with HTML tags were excluded and sentences were manually checked for tagging errors and cases where the noun was part of a compound expression. Selecting a set of four sentence contexts for each of the full set of 30 adjectives, our set of experimental items summed up to 120 (see Table 1 on the preceding page).

[2] Bernd Bohnet's parser (Bohnet, 2010) was used to lemmatise the German words.

Clearly, our setup leads to a *domain difference* between the sentences to be translated (sampled from the BNC) and the phrase table (drawn from EUROPARL). This makes the task of the model more difficult, and we might fear that the BNC bigrams we want to translate are very rare or even unseen in EUROPARL.

We made the decision to adopt this setting nevertheless, since it corresponds to the standard situation for machine translation. There is only a very small number of domains (including newswire, parliamentary proceedings, and legal texts) in which the large parallel corpora exist that are necessary to train SMT models. In the translation of texts from virtually all other domains, the models are faced with new domains. Being able to show an improvement for this across-domain scenario is, in our opinion, significantly more relevant than for the within-domain setting.

3.3 Translation Support

Finally, we provided three kinds of translation support to participants: (a) no support, (b) the list of translations for the adjective unigram produced by the SMT system, and (c) the list of translations for the adjective–noun bigram produced by the SMT system. In addition to adjective translations proposed by the system in the unigram condition, suggested noun translations for the target sentence might further aid the human translator in finding the most appropriate adjective in that context, in particular for collocation-like phrases.

We presented three distinct candidate translations as supports. We chose three as a number which is high enough to give translators at least some insight into the polysemy of target adjectives but still not enough to overload them and to slow down the translation process too much. The candidate translations were shown in the order in which they were extracted from the *n*-best list (with $n = 3,000$) produced by the Moses[3] (Koehn et al., 2007) MT system (trained and tuned on EUROPARL v6) that decoded each target sentence. See Example 1 for an illustration (target adjective: *bright*).

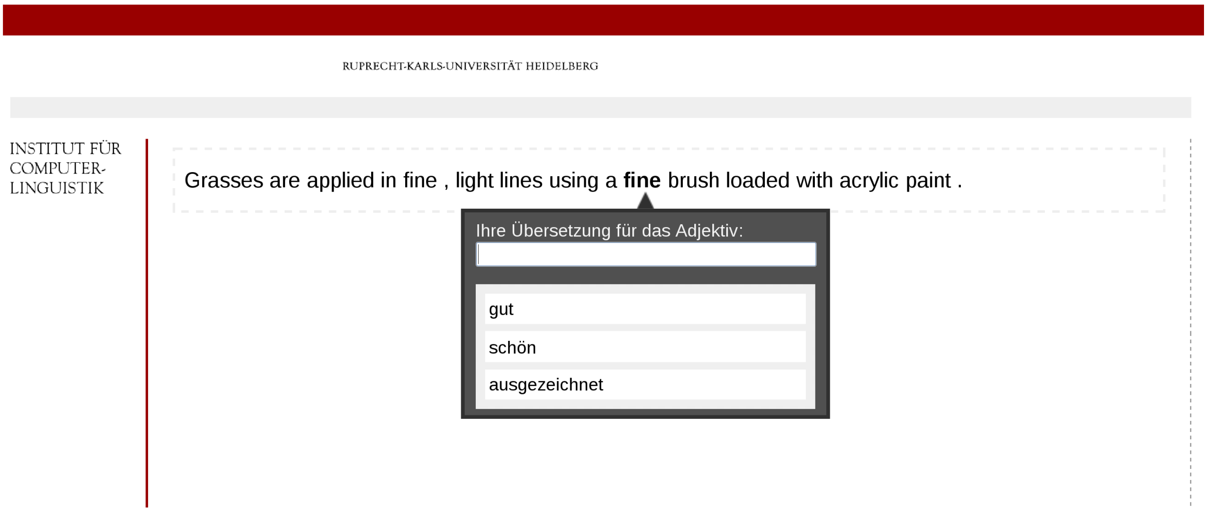
(1) The boy's ***bright*** *face, with its wide, open eyes, was contorted in agony .*

Unigram support:	Bigram support:	
verheißungsvoll	*verheißungsvolles*	*(Angesicht)*
positiv	*positives*	*(Angesicht)*
gut	*verheißungsvolles*	*(Gesicht)*

Specifically, phrase alignments were looked up for each *n*-best sentence given as output for a target sentence, and the corresponding translated adjective and noun was used for the unigram or the bigram list, respectively. In case of phrase alignments containing multiple words (instead of just one), word alignments were looked up in the phrase table and if in this manner English target words could be uniquely paired with translated German words, these pairs were chosen. Three differing unigrams and three differing bigrams were selected in order of appearance in the *n*-best list and lemmatised manually.

In case this procedure yielded less than three differing unigrams, the missing adjective unigrams were chosen from the unigram list of the adjective in the other three sentence contexts. Similarly, in case less than three bigrams were found, adjective unigrams produced by the MT system for that sentence were combined with nouns in the bigram list of that sentence (in order of appearance in the list). Candidate words for unigrams and bigrams were only selected from the *n*-best lists if they plausibly could have been tagged as adjectives or nouns, respectively.

[3] http://www.statmt.org/moses



RUPRECHT-KARLS-UNIVERSITÄT HEIDELBERG
INSTITUT FÜR COMPUTER-LINGUISTIK

Grasses are applied in fine , light lines using a **fine** brush loaded with acrylic paint .

Ihre Übersetzung für das Adjektiv:

gut

schön

ausgezeichnet

Figure 1: Screen shot of translation setup

Hypothesis Presenting unigram translations leads to faster and more appropriate translations. Bigram phrases will produce the most appropriate translations, even if translating in this condition might be slower due to the need to read through more complex translation suggestions.

4 Experiment 1: The Time Course of Machine-Supported Human Translation

4.1 Experimental Procedure

The experiment was realized as a dynamic web page, using an internet browser as our experimental platform and administering the experiment over the internet. The advantage of this method is that we have quick access to a large pool of participants. In psycholinguistics, the reliability of this type of setup for reading time studies has been demonstrated by Keller et al. (2009). Our setup is also similar to crowdsourcing, a recent trend in computational linguistics to use naive internet users for solving language tasks (Snow et al., 2008; Mohammad and Turney, 2011). Unlike almost all crowdsourcing work, however, we did not use a crowdsourcing platform like Amazon Mechanical Turk and were specifically interested in the time course of participants' reactions.

The 30 experimental items were presented in three blocks of ten items each. Each block corresponded to one support condition (none, unigram, bigram). The participant could take a break between blocks, but not between items. Both the order of the blocks and the order of the items within each block were randomised.

For each item, the experiment proceeded in four steps:

1. Sentence is shown to participant (plain text, no indication of the target adjective).
2. When the participant presses a key, the target adjective to be translated is marked in boldface. Concurrently, the translation support is shown as well as a window for entering the translation (shown in Figure 1).
3. The participant starts to type the translation.
4. The participant marks the current item as finished by pressing return. The experiment proceeds directly to step 1 of the next item.

The central question in this procedure is how to measure our variable of interest, namely the length of the period that participants require to *decide on* a translation. The total time of steps 2 to 4 is a very unreliable indicator of this variable. It involves the time for reading and the time for typing. Since participants can be expected to read and type with different speeds, the total time will presumably show a very high variance, making it difficult to detect differences among the support conditions. Instead, we decided to measure the time from the start of step 2 to the start of step 3. We assume that this period, which we will call *response time*, comprises the following cognitive tasks: (a) reading the bold-faced target; (b) reading the translation suggestions; and (c) deciding on a translation. We believe that this response time, which corresponds fairly closely to the concept of *décalage* in sight translation, is a reasonable approximation of our variable of interest. This assessment rests on two assumptions. The first one is that at the time when a participant starts typing, they have essentially decided on a translation. We acknowledge that this assumption is occasionally false (in the case of subsequent corrections). The second assumption is that it is not practicable to separate translation time from reading time for the target adjective and the translation suggestions, since presumably the translation process starts already during reading (John, 1996; Carl and Dragsted, 2012).

To avoid possible errors introduced into the time measurements by a remotely administered experiment, all time stamps during the course of an experiment are measured by the participant's machine, similar to Keller et al. (2009). It is only at the end of each experiment that these time stamps are transmitted back to the server and evaluated. In this manner, the time measurements are as accurate as the users' machines, which usually means at least a millisecond resolution. We also applied the usual methods to remove remaining outlier participants (cf. Section 4.3).

4.2 Participants

We solicited native German speakers as participants mostly through personal acquaintance; no professional translators participated. Participants were not paid for the experiment. We had a total of 103 participants. 87 of these were from Germany, 13 from Switzerland, and 1 each from Luxembourg and Austria.[4] 47 were male and 56 female. The mean age was 32, and the mean number of years of experience with English (comprising both instruction and practical use) was 16.1. Thus, the participant population consisted of proficient speakers of English. This is also supported by the participants' self-judgements of their proficiency in English on a five-point scale (1: very high, 5: very low), where the mean was 1.8.

4.3 Analysis of Response Time

We removed outliers following standard procedure. First, we completely removed 18 participants from consideration who did not complete all experimental items. From the response times for the remaining 85 participants, we removed all measurements below the 15th percentile ($t < 2.4$ s) and above the 85th percentile ($t > 12.9$ s) for each experimental item. These outliers have a strong chance of resulting from invalid trials. Participants with a very fast response time may have used their computer's copy–paste function frequently to simply copy one of the suggested translations into the response field. Participants with very slow response times may have been distracted.

[4] One participant declined to state their country.

	Low variability	High variability	Overall
No support	5.512	5.603	5.558
Unigram support	5.885	5.335	5.615
Bigram support	6.118	6.120	6.119

Table 2: Mean response times for all support conditions × translation variabilities

Recall that each of the 85 participants saw one instance of each of the target adjectives, and that our materials contain 12 experimental items for each adjective: 3 support conditions combined with 4 context sentences. Having further discarded 30 % of our measurements, we were left with an average of (85 / 12) * 0.7 ≈ 5 measurements for each experimental item. In our analysis, we use the mean of these individual measurements.

Our data set contains independent variables of two distinct classes (Jaeger, 2008). In the first class, we have two variables (variability class and the support condition, cf. Table 1 on page 105) which are *fixed effects*: we assume that these variables explain variation in the response time. The second class comprises a number of *random effects* which we expect to introduce variance but whose overall effect should be essentially random. This class includes the context sentence and the identities of adjective, participant, and context.

We therefore analysed our data with a linear mixed effects model (Hedeker, 2005). Linear mixed effects models are a generalisation of linear regression models and have the form

$$\text{(2)} \quad y = X\beta + Zb + \epsilon \qquad \text{with} \quad b \sim \mathcal{N}(0, \sigma^2\Sigma),\ \epsilon \sim \mathcal{N}(0, \sigma^2 I)$$

where X is a set of variables that are fixed effects, Z a set of variables that are random effects, and ϵ an error term. The first term in the model ($X\beta$) corresponds to a normal regression model—the coefficients β for the variables X are unconstrained. The second term, Zb accounts for the nature of random effects Z by requiring their coefficients b to be drawn from a normal distribution centred around zero. The model was implemented in the R statistical environment[5] with the package `lme4`[6].

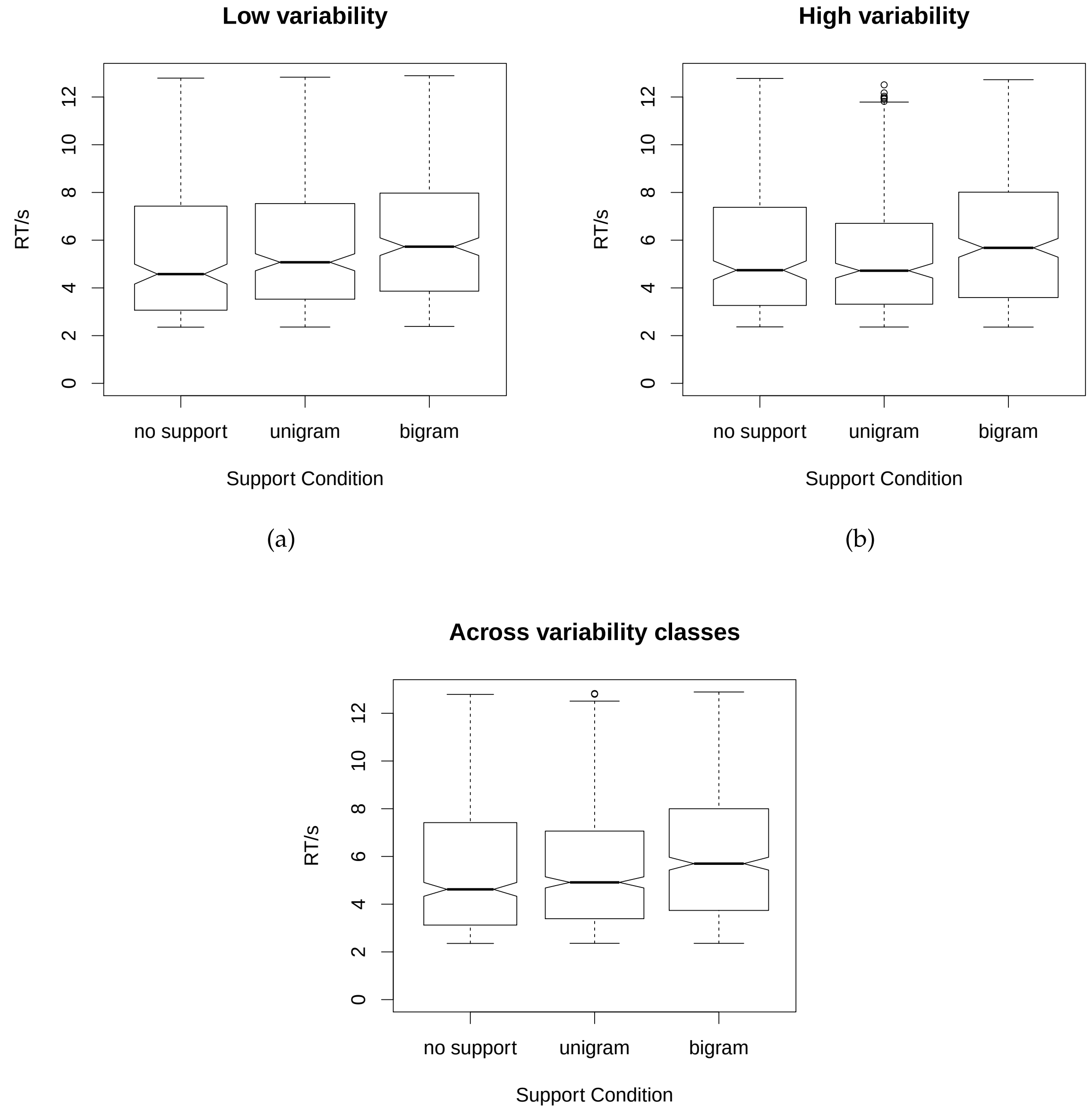
4.4 Results and Discussion

Table 2 shows mean response times for the six conditions corresponding to all combinations of the levels of the fixed effects, variability and support. All conditions result in mean response times between 5.5 and 6.1 seconds. Figure 2 on the facing page visualises robust statistics about the data in the form of *notched box-and-whiskers-plots* (McGill et al., 1978). The box indicates the median and the upper and lower quartiles, and the whiskers show the range of values. The notches (i. e., the "dents" in the boxes) offer a rough guide to significance of difference of medians: if the notches of two boxes do not overlap, this offers evidence of a statistically significant difference (95 % confidence interval) between the medians.

We make two main observations on these boxplots: (a) comparing Figure 2(a) with Figure 2(b), there does not appear to be a significant influence of variability; (b) comparing the different conditions in Figure 2(c), there appears to be a significant

[5] http://R-project.org

[6] http://lme4.r-forge.r-project.org



(a)

(b)

(c)

Figure 2: Distribution of response times in all experiment conditions for (a) low- and (b) high-variability adjectives, and (c) across variability classes

influence of the support condition. In all three boxplots, we find that bigram support leads to significantly longer response times than no support and unigram support, which in turn are not significantly different.

These observations were validated by an analysis of our mixed effects in which we determined the significance of the individual coefficients using a likelihood ratio test. Selecting the condition "high variability/no support" as the intercept, the coefficient for bigram support (0.69, SE: 0.15) is significantly different from zero ($p < 0.001$) while the coefficient for unigram support (0.11, SE: 0.15) is not. The coefficient for low variability (0.13, SE: 0.24) is also not significantly different from zero.

In sum, one of the two hypotheses we formulated in Section 3 on page 105 does not hold, while the other one holds at least partially. Contrary to our expectations, we do not find an effect of variability. That is, the adjectives with many possible translations are as difficult to translate as those with few possible translations. We believe that this effect is absent because we present all adjectives in a rich sentence context, as a consequence of which usually just a fairly small number of translations is reasonable, independent of whether the adjective, as a lemma, has a very large number of translation candidates or not.

Regarding the influence of the different levels of translation support, there is no significant difference between no support and unigram support: reading three additional words does not seem to interfere greatly with the time course of translation (although note that there is a tendency towards a difference between the low and high variability adjectives for this level). Bigram support, on the other hand, does add a statistically significant delay to the response time. However, the overall size of this effect, namely 0.5 to 0.6 seconds per translation, accounts for just 10 % of the response time, and only a very small percentage of the total translation time. Therefore, this effect should not be an obstacle to presenting translators with bigram support, should it be beneficial for the quality of the outcome.

5 Experiment 2: Translation Quality Rating

The second experiment investigates possible effects of different support conditions on translation quality. For this purpose, we elicited quality ratings from human annotators for all translations and support suggestions from the first experiment. We first describe the experimental procedure of this survey in Section 5.1, before we thoroughly analyse and discuss the obtained quality rating data in Section 5.2 on the next page.

5.1 Experimental Procedure

We elicited quality ratings for all translations collected in the first experiment after eliminating the reaction time outliers (cf. Section 4.3 on page 109). This includes the union of all translations entered by participants and all suggestions provided by the system. The full set consisted of 1,334 adjective instances to be rated, including inflected forms and incorrect spellings of the same adjective.[7] The sentences were presented to all raters in the same randomised order. For each sentence, the corresponding adjective translations and support adjectives were shown in alphabetical order alongside the sentence and the target adjective's head noun translation (which had been manually produced by one of the authors). The English target adjective was explicitly marked

[7] If the same adjective lemma occurred in various forms as a translation in the same sentence due to inflection or spelling mistakes, the raters were instructed to assign the same rating to all these forms.

(surrounded by stars: '*') in the sentence context. See Examples 3 and 4 for an illustration.

(3) As they reached the [. . .] tunnel , fresh air drifted in and Devlin took a *deep* breath .
tief Atemzug
tiefen Atemzug
tiefer Atemzug

(4) But after three weeks of this Potter claimed to have lost nothing but his *good* humour .
frohe Stimmung
gute Stimmung
positiv Stimmung

Each adjective instance was judged by eight human raters who were native speakers of German with a (computational) linguistics background. They were asked to rate the quality of each adjective translation in the given sentence context and for the predefined head noun translation. For their judgements, we instructed our raters to apply a three-point Likert scale according to the following conventions:

- 3: perfect translation in context of sentence and noun
- 2: acceptable translation, while suboptimal in some aspect
- 1: subjectively unacceptable translation

Our notion of "suboptimal translation" (level 2 on the scale) includes two aspects: core semantic mismatches (the meaning of the adjective does not fully reflect all aspects of the best translation) and collocational incongruence (the translation of the adjective does not yield a well-formed collocation in combination with the respective noun). The second translations listed for the two following examples illustrate semantic mismatch (Example 5) and collocational incongruence (Example 6):

(5) But there is a *common* belief that low-rise building will increase the urban sprawl.
verbreiteter Glaube (3.00)
allgemeiner *Glaube (2.67)*

(6) Until now, he had managed that, with a *heavy* hand and crude peasant humour.
harte Hand (3.00)
starke *Hand (2.50)*

Numbers in parentheses state the average quality of the translation as given by our human raters. For our detailed rating guidelines see the appendix (Section A.2 on page 125).

5.2 Analysis and Discussion

The basis for all analyses in this section are the experimental items without reaction time outliers (as described in Section 4.3 on page 109) and the quality ratings of these experimental items (as described in Section 5.1 on the preceding page).

Recall that in our translation experiment translators were always free to choose a translation from the support items or, alternatively, choose a translation on their own. We will use the terms *support translations* and *creative translations* to refer to these two options. *Support suggestions* denote all support items provided in a specific experimental

	No support	Unigram support	Bigram support
Support suggestions	—	groß breit hoch	großer (Unfall) großes (Unglück) große (Katastrophe)
Support translations	groß (2)	groß (4)	groß (4)
Creative translations	riesig (1) schwer (1) schwerwiegend (1)	schwer (1) weitreichend (1)	schlimm (1)

*Table 3: Example translations of different types for the sentence: "In other words, it is a measure of the scale and likelihood of a *large* accident." Numbers in parentheses: the number of participants who produced an item.*

condition, irrespective of whether or not one of these candidates was selected by the participants as a translation. Table 3 illustrates these three terms by example for a sentence taken from the experiment data.

More specifically, for the experiment conditions "unigram support" and "bigram support", *support translations* are defined as those items that both appeared as *support suggestions* (in the respective support condition) and were also selected as translations by participants. Items that were produced by participants, but did not appear in the *support suggestions*, are considered as *creative translations*.

The "no support" condition is a special case, as in this condition all translations were freely produced by the participants, i. e., without the possibility of relying on any support. To maintain the distinction between creative and support translations, we computed the union of all adjectives contained in the unigram support and bigram support and compared the freely produced translations against this set. Thus, the translations found in this union were considered as support translations, all the other translations as creative. Given these differences in calculation, an exact comparison of the ratio of creative translations will be possible for the unigram and bigram condition only. Nevertheless, we consider the proportion of creative translations in the "no support" condition as defined above to be meaningful in that it provides an impression of the range of the spectrum of human translations that is not covered by SMT support material.

5.2.1 Inter-Rater Correlation

We started by analysing the agreement among the raters. We computed an inter-rater correlation coefficient using leave-one-out re-sampling (Weiss and Kulikowski, 1991). For this analysis, we first (manually) mapped all inflected word forms and incorrect spellings to the same adjective lemma. This should reduce the influence of morphological variation on the magnitude of the correlation coefficient. Second, as proposed by Mitchell and Lapata (2010), we correlated the judgements of each rater with those of all the other raters to obtain an averaged individual correlation coefficient (ICC) for each rater in terms of Spearman's ρ. This resulted in an overall correlation coefficient of $\rho = 0.43$ for the eight raters. As we found substantial deviation of two raters from all others[8], we decided to discard their judgements. Averaging the ICCs of

[8] Their ICCs are the only ones below 0.4, while the coefficient of their pairwise correlation is extremely low ($\rho = 0.24$; cf. the full IRC matrix in Section A.3 on page 126).

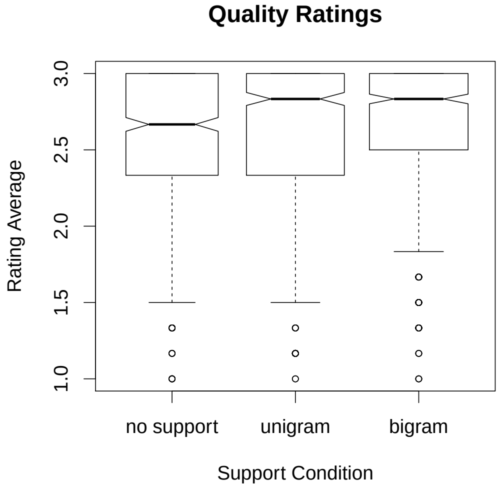


Figure 3: Distribution of averaged translation quality ratings

	Translation quality mean
No support	2.53
Unigram support	2.60
Bigram support	2.65

Table 4: Overall translation quality rating means for all experiment conditions

the remaining six raters resulted in an overall inter-rater correlation of $\rho = 0.47$. This outcome indicates that translation quality rating is a difficult task, but that our raters still produced reasonably consistent ratings. We then computed the average quality rating for each adjective instance by including the judgement scores of the best six raters. We use these averages as the basis for analysing the overall translation quality between experiment conditions in the next section and for all subsequent analyses.

5.2.2 Overall Translation Quality

We next consider the overall translation quality for the different support conditions. Figure 3 visualises the translation quality data as a boxplot (cf. Section 4.4 on page 110). The medians of the quality ratings for no support and unigram support differ substantially, with non-overlapping notches, indicating a statistically significant difference in average quality ratings between these two conditions. Comparing the conditions "unigram support" and "bigram support", their medians are almost identical. However, the variance is smaller in the bigram condition (smaller box), and there are noticeably fewer outliers at the lower end (shorter whisker). Thus, although there is no significant difference in terms of average translation quality, there is a tendency of bigram support to produce fewer medium and low quality translations. The corresponding means are shown in Table 4.

These findings are corroborated by our mixed effects model analysis: analogously to the analysis of response times (see Section 4.3 on page 109), we assumed that the factors "variability class" and "experiment condition" are fixed effects. We used the same factors as in the response time analysis as random effects and added rater identity. But, as in

	No. participants with ≥ 1 creative translation	
No support	62	(72.9 %)
Unigram support	50	(58.8 %)
Bigram support	46	(54.1 %)

Table 5: Number (and rate) of creative participants in each experiment condition

the present analysis the "quality rating" (1–3) was used as the dependent variable in the model, we applied a model tailored to categorial response variables, namely the cumulative link mixed model (Christensen, 2011), provided by the R package `ordinal`[9]. Selecting unigram support as the base level, the model yields significant differences both when compared to no support ($p < 0.001$) and bigram support ($p < 0.01$).

These results suggest that the quality of our participants' translations, while being already rather high in the absence of any support, benefits from more detailed support material. Unigram and bigram support tend to have a slightly different influence, however: unigram support primarily seems to trigger better translations as compared to no support, while there is still a number of bad translations that cannot be ruled out in this condition. Admittedly, bigram support does not yield a further quality improvement, but contributes to a reduction of poor translations.

5.2.3 Ratio of Creative to Support Translations

An essential fact for interpreting the results of Section 5.2.2 on the preceding page is that participants were always free to forgo the support suggestions and enter their own translations. Thus, the analysis is still inconclusive, since it does not take into account how many support suggestions were actually accepted or overridden by the participants, and what exactly contributed to the augmentation in translation quality for unigram and bigram support. In fact, the quality gains observed under unigram and bigram support might be artefacts due to exhaustive use of creative translations (although creativity might have been triggered by presenting support suggestions). In that case, the direct contribution of the support suggestions to the participants' translation performance would be questionable.

For this reason, we investigate the ratio of creative translations from different perspectives, starting from the level of participants. Afterwards, we broaden the scope to include the levels of sentences and individual translations.

Analysis by Participants We first investigated the proportion of participants who produced at least one creative translation. Table 5 shows that in the absence of any support, more than 70 % of the participants occasionally produced a translation that is not contained in the unigram and bigram support suggestions. In the unigram condition, the proportion of creative participants amounts to 58.8 %, decreasing with more extensive support material to 54.1 % in the bigram condition.

To obtain a more detailed picture, we also considered the individual creativity rate per participant: did participants systematically accept (or reject) the support suggestions, or did they make use of them in an intelligent manner? To address this issue, the creativity rate was measured as the number of creative translations of the respective participant in relation to all their individual translations under unigram and bigram

[9] http://www.cran.r-project.org/web/packages/ordinal

	Rate of creative translations per participant							
	0 %	1–10 %	11–20 %	21–30 %	31–40 %	41–50 %	51–60 %	> 60 %
No. Participants	23	13	25	15	8	0	1	0

Table 6: Creativity rate per participant in the "support" conditions (unigram and bigram)

	Sentences with ≥ 1 creative translation	
No support	71.7 %	(86)
Unigram support	36.7 %	(44)
Bigram support	39.2 %	(47)

Table 7: Proportions of sentences with creative translations in each experiment condition

support. Table 6 shows that 23 (about 27 % of the whole group of) participants never produced a creative translation, but always used a translation that is included in the set of support suggestions. The other participants exhibit creativity rates that are distributed within a region of moderate creativity (with one outlier, a participant who came up with creative translations in more than half of the items she translated).

Combined with the data presented in Table 5 on the preceding page, this indicates that in both "support" conditions (unigram and bigram), only little more than half of the participants ever decided to override the support material, without individually overusing this opportunity. On the other hand, we do not observe any participants who systematically reject the support material provided.

Analysis by Sentences On the sentence level, we are primarily interested in whether some sentences show a stronger tendency to evoke creative translations than others. Therefore, along the lines of our analysis on the level of participants, we first investigated the proportion of sentences with at least one creative translation, before taking a closer look on the creativity rate per sentence.

In the "no support" condition, our group of participants produced translations that are neither contained in the unigram nor in the bigram support in more than 70 % of the sentences (cf. Table 7). In the "unigram support" condition, 36.7 % of the sentences provoked a creative translation. Interestingly, however, this proportion is slightly higher in the "bigram support" condition.

We believe that this effect is not just random variation: we encountered 15 sentences in the data which triggered at least one creative translation in bigram support, but none in unigram support. Analysing these sentences, we discovered two major reasons for their higher disposition towards creative translations in bigram support. First, some of the support suggestions contained in the unigram set are not included in the bigram set—Example 7 illustrates this phenomenon, where *angemessen* would be categorised as a creative translation based on bigram support (on the right), but not based on unigram support (on the left).

	Rate of creative translations per sentence				
	0 %	1–25 %	26–50 %	51–75 %	76–100 %
No. sentences	61	28	23	6	2

Table 8: Creativity rate per sentence in the "support" conditions (unigram and bigram)

(7) *The show was the best it had ever been , and its *proper* length , for once .*

Unigram support:	Bigram support:	
richtig	*richtige*	*(Zeit)*
ordnungsgemäß	*richtige*	*(Dauer)*
angemessen	*ordnungsgemäße*	*(Länge)*

Second, on the one hand, in the context of ambiguous or abstract nouns that are hard to translate when given just unigram support, some participants apparently tended towards accepting one of the unigram suggestions without reasoning too much about its collocational fit with the best translation of the context noun. On the other hand, in some cases the bigram support suggestions include a good translation of the noun in combination with an incongruous adjective suggestion. Consider Example 8, where all participants translated *great* as *groß* in the "unigram support" condition, while during bigram support, we also encountered the creative translation *hoch* (high), which is a better collocational match for *Genauigkeit* (accuracy) and *Präzision* (precision) in German than *groß*.

(8) *Someone who hits the ball with *great* accuracy on the volley and with [. . .] .*

Unigram support:	Bigram support:	
groß	*große*	*(Genauigkeit)*
großartig	*große*	*(Sorgfalt)*
riesig	*große*	*(Präzision)*

The creativity rate per sentence measures the fraction of creative translations in all translations that were collected for the respective sentence in both the conditions "unigram support" and "bigram support". Table 8 summarises the results. For about half the sentences, no creative translation was produced at all, i. e., the participants were satisfied with the support material being provided. 75 % of the sentences exhibit a creativity rate of 25 % or below. For only eight sentences, the majority of translations (> 50 %) was found to be creative. Apparently, the availability of support limits the need for creative translations, regarding both the number of sentences that exhibit creative translations and the creativity rate within these sentences.

Analysis by Translations Finally, we investigated the creativity rate on the basis of individual translations. The results of this analysis are shown in Table 9 on the facing page.[10] Comparing the creativity rate across the three experimental conditions, we can observe a pattern that is in line with our preceding analyses: for unigram and bigram support, only 12.3 % and 13.4 % of the translations, respectively, were found to be creative. Considering freely produced translations, we encounter a relatively

[10] Note that the absolute number of translations as stated in the first column of the table differs across the experimental conditions due to the elimination of response time outliers (cf. Section 4.3 on page 109).

	No. translations	Creative translations
No support	546	41.6 %
Unigram support	614	13.4 %
Bigram support	624	12.3 %

Table 9: Overall creativity ratio for experiment data without response time outliers

	Creative translations	Support translations	Support suggestions
Unigram support	2.40	2.64	2.46
Bigram support	2.42	2.68	2.52

Table 10: Average quality ratings for complete data set

high creativity rate (41.6 %). The latter percentage is also interesting from a different perspective, as it provides an estimate of the coverage of the support material: almost 60 % of the translations produced by our participants in the "no support" condition are covered either by the unigram or the bigram suggestions.

Given that the support material in the translation experiment for each target adjective comprised only the three most likely translations as extracted from the SMT *n*-best list (cf. Section 3.3 on page 107), the question arises whether support coverage would improve if more suggestions from the MT system were included in the translation support. To tackle this question, we also extracted the five-best and ten-best unigram translations for the test adjectives from the Moses output.[11] As expected, the creativity rate drops from 13.4 % for the top 3 support to 10.4 % for the top 5 support (64 creative translations) and finally to 7.7 % for the top 10 support (47 creative translations).

Summary Our creativity analysis based on participants, sentences and individual translations yields a coherent pattern: (a) translators use support translations for both unigram and bigram support in a total of almost 90 % of the cases; (b) translators use creative translations only for a subset of sentences (less than 40 %) when translation support is given; (c) about 60 % of the participants exhibit moderate individual creativity rates of between 11 % and 40 %. These findings suggest that creative translations, despite their sparsity, are used deliberately in particular cases. This leads to the question whether creative translations have an effect on translation quality, i. e., whether the quality of individual creative translations is higher compared to the corresponding support suggestions.

5.2.4 Translation Quality of Creative Translations and Support Suggestions

Our latest analysis compares the overall average quality of creative translations, support translations and support suggestions in both "support" conditions. The results are shown in Table 10. Our first observation is that bigrams outperform unigrams in all the three categories, which is in line with the results of our overall quality analysis in Section 5.2.2 on page 115.

Next, we compare the results for the different columns. The third column, "support

[11] This required consulting a 50,000-best list to obtain enough distinct translations for most cases. Still, for 8 items (≈ 6.7 %) we found less than five translations, and for 81 items (67.5 %) less than ten translations.

	No. instances (creative trans.)	Creative translations	Support translations	Support suggestions
Unigram support	82	2.40	2.17	1.83
Bigram support	77	2.42	2.15	1.95

Table 11: Average quality for experimental items that triggered creative translations

suggestions", can be considered as a baseline of randomly picking one of the support suggestions. Such a strategy would achieve an average quality of 2.46 (with unigram support) or 2.52 (with bigram support). These numbers indicate that the support material provided to our participants was of good average quality. In fact, the quality of the support suggestions is only slightly below the average of our human participants translating without support (2.53, cf. Table 4 on page 115).

The "support translations" column shows that our human translators did a good job picking out the best translations from all support suggestions, increasing the quality by 0.18 (unigram condition) and 0.16 points (bigram condition). In contrast, and somewhat surprisingly, the average quality of all creative translations taken together falls slightly below the baseline in both the unigram (2.40) and the bigram (2.42) condition. Thus, it appears that creative translations cannot be assumed a priori to be of high quality.

A possible explanation for this finding is that creative translations were produced in particular for difficult adjectives to be translated. If this were true, we would expect that the support translations for these sentences should perform even worse. To test this prediction, we repeated our analysis for the *creativity-triggering experimental items* (i. e., the subset of experimental items for which at least one participant produced a creative translation). The results in Table 11 show that this is indeed the case: the quality of all support suggestions for these sentences is below 2, and even picking the best candidates (column "support translations") yields an average quality of below 2.2. The creative translations, with an average quality of around 2.4,[12] outperform the support suggestions and translations significantly ($p < 0.001$ for both contrasts—as determined by an approximate randomisation test, cf. Noreen, 1989).[13] This means that, overall, translators not only use good supports when appropriate, but they are also able to recognise bad supports and replace them with better suited creative translations. For illustration, consider the following two examples where creative translations outperform the support translations (i. e., support suggestions that were actually selected by at least one participant):

(9) *What does a *large* attendance at Easter communion imply?*

Support translations:		Creative translations:	
groß	*(2.00)*	*zahlreich*	*(2.17)*
hoch	*(1.83)*		
breit	*(1.83)*		

[12] Note that in our experimental setting three support suggestions were provided for each experimental item. To compare the average qualities of creative translations and support suggestions, we triplicated the rating score for each creative translation.

[13] The significance analysis was performed on a slightly smaller number of experimental items (69 for unigram support, 71 for bigram support), as for some of the items, none of the participants selected a support suggestion. Average quality of the creative translations in these cases: 2.38.

(10) *He delivered a *great* kick backwards at Terry's shins, the edge of his boots like iron.*

Support translations:		Creative translations:	
groß	*(1.5)*	*kräftig*	*(2.67)*
		heftig	*(2.50)*
		großartig	*(2.33)*
		gut	*(2.33)*
		gut gelungen	*(2.33)*
		fest	*(2.33)*
		schwer	*(2.17)*

These examples show all support translations (left column) and creative translations (right column) for the respective sentence in all conditions (and their average qualities).

5.2.5 Summary

Across all analyses, we clearly see a positive effect of SMT support on human translation performance. Our initial hypothesis is largely confirmed, as we found a significant gain in translation quality for unigram support compared to the "no support" condition. Beyond that, bigram support does not yield a further increase in translation quality, but still tends to help excluding poor translations.[14] We found that the generally high quality of the SMT suggestions is the primary source of this effect, as our participants relied on the provided support suggestions in almost 90 % of the cases.

However, high quality support material is not sufficient on its own to explain the improvement in translation quality in the two "support" conditions. We found that the human translators need to review the support suggestions to (a) pick the most appropriate of the suggestions and (b) if there are no appropriate ones, suggest a creative translation. Even though the latter case occurred only for a relatively small subset of the data, in these cases the participants' creative translations turned out to be significantly superior to the support suggestions. At the same time, (b) appears to be a difficult task, given that a fraction of about a third of our participants never produced any creative translations at all. It seems, therefore, that the decision when to accept and when to override the support suggestions is the most challenging task for many participants in computer-aided translation. In contrast, (a) appears quite feasible, as the quality of our participants' selections is well beyond a "random selection" baseline and consistently so across participants.

6 General Discussion and Conclusion

In this study, our goal was to investigate the usefulness of adjective–noun translations generated by MT systems and presented to non-professional human translators as unigram or bigram suggestions during the translation of individual adjectives in sentence context. This choice makes for an interesting translation task, due to the meaning variation of adjectives in context, while allowing us to control translation variability fairly strongly.

The first variable we measured was translation time. In presenting three suggestions in both the unigram and bigram conditions, we found a statistically significant increase in response times for the bigram support condition but not the unigram support condition. Even for the bigram condition, however, the mean response time increased

[14] Translating text segments of more than one word as natural "translation units" is exactly what is proposed in translation studies (see, e. g., Toury, 1995), and which our study corroborates.

only by around 0.6 seconds (i. e., by ≈ 10 %) compared to no support. Contrary to our intuitions, the level of translation variability as defined by phrase table counts had no statistically significant influence on response times. However, in interaction with the support condition "unigram" we partly observed an effect we had predicted: highly variable adjectives were translated faster than low-variability adjectives in the unigram condition.

The second variable of interest in the translation process was translation quality. We elicited judgements on a three-point scale from human annotators. Although the inter-rater correlation in the judgement experiment was only mediocre, the average quality ratings in the two support conditions were statistically significantly higher than without support. Furthermore, in the bigram condition, participants produced the least amount of low-quality translations. Further analysis established that the SMT-produced support suggestions were generally of high quality, and were accepted well by human translators, who were consistently able to pick the best translations from among the candidates.

In summary, we found a strong case in favour of supporting non-professional translators with SMT support, provided that the quality of the support material is high enough that just choosing between support suggestions is a reasonable strategy. In terms of the choice between unigram and bigram support, there is a substantial improvement in quality already for unigram support without a significant accompanying translation delay. For bigram support, the time to read through the suggestions becomes a significant (although still small) factor, but pays off with a further reduction in poor translations.

Recall that we obtained these results by presenting three support candidates for each adjective to be translated. This is of course not the only possible choice. We found that longer *n*-best lists will cover a larger fraction of translations (90 % for 5 suggestions), but we would expect that more suggestions will slow down the translation process considerably, clutter the translation interface, and make translators even more reluctant to dismiss poor suggestions.

Machine-supported human translation is an open field with ample potential for creative strategies to combine the complementary strengths of man and machine. In future work, we would like to explore ways to generalise our experimental setup to larger phrases without giving up the control over translation complexity that we have utilised in this experiment.

7 References

Barrachina, Sergio, Oliver Bender, Francisco Casacuberta, Jorge Civera, Elsa Cubel, Shahram Kadivi, Antonio Lagarda, Hermann Ney, Jesus Thomas, Enrique Vidal, and Juan-Miguel Vilar. 2008. "Statistical Approaches to Computer-Assisted Translation." *Computational Linguistics* 35(1): 3–28.

Bohnet, Bernd. 2010. "Top Accuracy and Fast Dependency Parsing is not a Contradiction." In *Proceedings of COLING*, 89–97. Beijing, China.

Bowker, Lynne. 2012. *Computer-Aided Translation Technology – A Practical Introduction*. University of Ottawa Press.

Burnard, Lou. 1995. *User's Guide for the {British National Corpus}*. British National Corpus Consortium, Oxford University Computing Services.

Carl, Michael and Barbara Dragsted. 2012. "Inside the Monitor Model: Processes of

Default and Challenged Translation Production." *Translation: Computation, Corpora, Cognition* 2(1).

Christensen, Rune Haubo Bojesen. 2011. "Analysis of Ordinal Data With Cumulative Link Models – Estimation with the *ordinal* package." Retrieved from http://www.cran.r-project.org/web/packages/ordinal, R package version 2011.09-14.

Clark, Stephen and James R. Curran. 2007. "Wide-Coverage Efficient Statistical Parsing with CCG and Log-Linear Models." *Computational Linguistics* 33(4).

Freigang, Karl-Heinz. 1998. "Machine-Aided Translation." In *Routledge Encyclopedia of Translation Studies*, edited by Mona Baker, 134–139. New York: Routledge.

Gow, Francie. 2003. *Metrics for Evaluating Translation Memory Software*. Master's thesis, University of Ottawa.

Hedeker, Donald. 2005. "Generalized Linear Mixed Models." In *Encyclopedia of Statistics in Behavioral Science*. Wiley, New York.

House, Juliane. 1998. "Quality of Translation." In *Routledge Encyclopedia of Translation Studies*, edited by Mona Baker, 197–200. New York: Routledge.

Jaeger, T. Florian. 2008. "Categorical Data Analysis: Away from ANOVAs and toward Logit Mixed Models." *Journal of Memory and Language* 59(4): 434–446.

John, Bonnie E. 1996. "TYPIST: A Theory of Performance in Skilled Typing." *Human–Computer Interaction* 11: 321–355.

Justeson, John S. and Slava M. Katz. 1995. "Principled Disambiguation. Discriminating Adjective Senses With Modified Nouns." *Computational Linguistics* 21: 1–27.

Keller, Frank, Subahshini Gunasekharan, Neil Mayo, and Martin Corley. 2009. "Timing Accuracy of Web Experiments: A Case Study Using the WebExp Software Package." *Behavior Research Methods* 41(1): 1–12.

Koehn, Philipp. 2005. "Europarl: A Parallel Corpus for Statistical Machine Translation." In *Proceedings of the Tenth Machine Translation Summit*, 79–86. Phuket, Thailand. http://mt-archive.info/MTS-2005-Koehn.pdf.

Koehn, Philipp. 2010a. "Enabling Monolingual Translators: Post-Editing vs. Options." In *Proceedings of the Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL*. Los Angeles, CA.

Koehn, Philipp. 2010b. *Statistical Machine Translation*. Cambridge University Press.

Koehn, Philipp and Barry Haddow. 2009. "Interactive Assistance to Human Translators Using Statistical Machine Translation Methods." In *Proceedings of Machine Translation Summit XII*. Ottawa, Ontario, Canada.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Birch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. "Moses: Open Source Toolkit for Statistical Machine Translation." In *Proceedings of the ACL 2007 Demo and Poster Sessions*. Prague, Czech Republic.

Langlais, Philippe, George Foster, and Guy Lapalme. 2000. "TransType: A Computer-Aided Translation Typing System." In *Proceedings of ANLP-NAACL Workshop on Embedded Machine Translation Systems*. Seattle, WA.

McGill, Robert, John W. Tukey, and Wayne A. Larsen. 1978. "Variations of Box Plots." *The American Statistician* 32(1): 12–16.

Miller, Katherine J. 1998. "Modifiers in WordNet." In *WordNet. An Electronic Lexical Database*, edited by Christiane Fellbaum, 47–67. MIT Press.

Mitchell, Jeff and Mirella Lapata. 2010. "Composition in Distributional Models of Semantics." *Cognitive Science* 34: 1388–1429.

Mohammad, Saif and Peter Turney. 2011. "Crowdsourcing a Word–Emotion Association Lexicon." *Computational Intelligence* To appear.

Noreen, Eric W. 1989. *Computer Intensive Methods for Testing Hypotheses. An Introduction*. New York: Wiley.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. "BLEU: A Method for Automatic Evaluation of Machine Translation." In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, 311–318. Philadelphia, PA, USA.

Reiß, Katharina. 1971. *Möglichkeiten und Grenzen der Übersetzungskritik: Kategorien und Kriterien für eine sachgerechte Beurteilung von Übersetzungen*, vol. 12 of *Hueber Hochschulreihe*. München, Germany: Max Hueber Verlag.

Sapir, Edward. 1944. "Grading. A Study in Semantics." *Philosophy of Sciences* 11: 83–116.

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. "A Study of Translation Edit Rate with Targeted Human Annotation." In *Proceedings of AMTA*, 223–231. Cambridge, MA.

Snow, Rion, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. 2008. "Cheap and Fast – But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks." In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, 254–263. Honolulu, HI, USA. http://www.aclweb.org/anthology/D08-1027.

Toury, Gideon. 1995. *Descriptive Translation Studies and Beyond*, vol. 4 of *Benjamin Translation Library*. Amsterdam, The Netherlands: John Benjamins.

Weiss, S. M. and C. A. Kulikowski. 1991. *Computer Systems that Learn. Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning and Expert Systems*. San Mateo, CA: Morgan Kaufman.

Wu, Zhibiao and Martha Palmer. 1994. "Verb Semantics and Lexical Selection." In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, 133–138.

Zanettin, Federico. 2002. "Corpora in Translation Practice." In *Proceedings of the LREC Workshop Language Resources for Translation Work*, 10–14. Las Palmas de Gran Canaria, Spain.

A Appendix

A.1 Adjective Stimuli Set

Low variability	High variability
lovely	final
bright	essential
formal	hard
dark	large
complex	common
fresh	proper
ordinary	real
rich	main
deep	present
recent	strong
heavy	serious
immediate	major
domestic	clear
separate	great
likely	good

Table 12: The set of 30 adjectives used as stimuli in the translation support experiment

A.2 Guidelines for Quality Rating

- If more than one inflected form of the same adjective lemma occurs as a translation in the same sentence: assign the same rating.

- In case of spelling mistakes: rate the adjective as if it was spelled correctly.

- If more than one word has been produced as a translation: consider only the (first) adjective.

- If the only translation produced is not an adjective, but a noun: rate the appropriateness of the noun as a translation for the adjective in the given context (e. g.: *major* → *Haupt-*).

- Rate the appropriateness of each adjective only in combination with the translation given for its head noun.

- Try to use the full scale (1–3) to rate the quality of all adjective translations per sentence. However, in case of sentences with only a few different adjective translations: if all of them are bad, it is not necessary to exhaust the full scale.

- Try to work swiftly.

A.3 Quality Rating: Inter-Rater Correlation Matrix

	R1	R2	R3	R4	R5	R6	R7	R8	ICC
R1	1.00	0.40	0.36	0.24	0.38	0.38	0.36	0.41	0.36
R2	0.40	1.00	0.49	0.42	0.50	0.47	0.41	0.48	0.45
R3	0.36	0.49	1.00	0.38	0.44	0.47	0.51	0.45	0.44
R4	0.24	0.42	0.38	1.00	0.41	0.42	0.37	0.43	0.38
R5	0.38	0.50	0.44	0.41	1.00	0.50	0.39	0.49	0.44
R6	0.38	0.47	0.47	0.42	0.50	1.00	0.49	0.49	0.46
R7	0.36	0.41	0.51	0.37	0.39	0.49	1.00	0.47	0.43
R8	0.41	0.48	0.45	0.43	0.49	0.49	0.47	1.00	0.46

Table 13: Inter-rater correlation matrix for our full set of raters in the judgement experiment