Universität des Saarlandes

Philosophische Fakultät 4.7 Allgemeine Linguistik

Computerlinguistik

Master's Thesis

Post-Editing of Statistical Machine Translation

A crosslinguistic analysis of the
temporal, technical and cognitive effort

Lisa Beinborn

September 2010

Supervisor: Dr. Pirita Pyykkönen

Correctors: Prof. Hans Uszkoreit, Dr. Pirita Pyykkönen

Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel verwendet habe.

Saarbrücken, September 2010

Lisa Beinborn

Thank you...

... Pirita for the incredible support. You provided so much valuable feedback and always took the time to discuss all my questions. I learned a lot from you.

... Hans Uszkoreit for raising my interest in machine translation and bringing me in contact with Sharon O'Brien and Horst Liebscher.

... Mirko Plitt and Francois Masselot from Autodesk for providing the data, for answering all the questions and for the helpful comments.

... Sharon O'Brien from Dublin City University for introducing me to the field of post-editing.

... Horst Liebscher from Euroscript for giving insights into the translation industry.

... Casey for patiently checking punctuation and giving stylistic advice.

... my flatmates and friends for all the motivating coffee and lunch breaks.

... Stefan for being you.

Abstract

Human inspection and correction of machine translations is still indispensable to ensure accurate and stylistically acceptable output. This process of retrospective modification of machine translation output is called Post-Editing. In this thesis, the post-editing process is analyzed under temporal, technical and cognitive aspects. Crosslinguistic data from a prior productivity test constitutes the basis for the analysis. English source segments had been machine-translated by a statistical system into Italian, Spanish, French and German, and then post-edited by professional translators. Segments with higher temporal, technical and cognitive effort are identified in order to detect crosslinguistic negative translatability indicators. The results show that the translatability of the source segment is not the only factor that has influence on the post-editing effort. A comparison of post-editing and translation effort highlights the difference between the two processes. Finally, possible improvements for the post-editing working conditions are proposed.

Contents

List of Figures

List of Tables

1. Introduction

In the last ten years statistical machine translation has made a lot of progress. Faster processing mechanisms enable systems to explore big amounts of increasingly available parallel language data in the search for the best translation. Though the previous forty years of machine translation research had failed to convince the broader public, newer statistical research systems now also attract business clients from the localization industry. Unfortunately, machine translation systems still cannot guarantee a high quality output. Despite the great improvements in translation quality, subsequent human inspection and correction is indispensable to ensure accurate and stylistically acceptable translations. This process of retrospective modification of machine translation output is called *Post-Editing*. During post-editing, the translators correct errors, insert missing words, and improve stylistic elements to enhance the accuracy and the readability of the machine translated text. Instead of autonomously creating a translation from one language, the source, into a foreign language, the target, post-editors work with a given basis and assure a reliable quality of the outcome by the appropriate modification of this basis. The linguistic operations that have to be performed depend on the machine translation quality and the requested final quality. If the content is created mainly for informational purposes like internal notes, stylistic flaws of the machine translation output might be neglected. More sophisticated usage scenarios, such as marketing brochures, on the other hand, may require a complete restructuring of the translated sentence. The provided machine translation basis already fulfills parts of the work of a translator, thus the overall task is reduced to the revision and potential modification of the machine output.

Vasconcello appraised the advantages of post-editing already in 1986.

> "Post-Editing gets to be more fun and more relaxing than translating from scratch. For the same number of words, post-editors are less fatigued at the end of the day." ([68], p.145)

However, her positive assessment did not find its way from research to translators' work routine. The assumption that post-editing is more productive and time-efficient than standard human translation has been confirmed by various independent evaluations (e.g. [25], [53], [20]), but translators do not personally experience this shift. They widely agree that machine translation does not work [19] and reject it: "Post-editing has become by now one of the most disliked tasks by translators" ([26], p.1). Some years ago a similar rejection towards the use of translation memories could be observed. Translation memories store all the translations for a post-editor. These memories are able to recognize previously translated segments and suggest translations based on these matches. The time efficiency and the quality consistency that is guaranteed by this reuse of former translations have greatly increased the acceptance of translation memories and have become a common tool today. One reason to react overly critical towards new technology might be the fear of getting replaced by a machine. Yet, there are also other factors which interfere, such as the need of having to learn new methods. Employing a new technology implies a change in the working process that requires translators to adapt their habits and strategies. In the long run, this change can only succeed if it is supported by the majority of the translators.

Previous research of the post-editing process is mainly based on bilingual experimental data (e.g. [67], [6]). Taking more target languages into account allows for a crosslinguistic analysis that abstracts from properties specific to a certain language pair. We can thus distinguish between language-specific details and more general principles. English gerund forms, for example, might be more difficult to translate into German where an equivalent of this grammatical form does not exist. Therefore, a slight restructuring of the sentence is necessary. Translating a gerund form into Spanish, on the other hand, can be done very intuitively because the Spanish *gerundio* is structurally comparable to the English gerund ("está caminando" ≈ "is walking"). Two languages that share similar structures probably cause less challenges for the translation of complex sentences. Additionally the linguistic and also cultural aspects of the two languages might play a role especially for the translation of idioms or metaphors [23]. Some idioms can be translated literally from one language to another (e.g. to have eagle eyes = Adleraugen haben = avere occhi d'aquila), others have a very different meaning in another language (e.g. to be blue ≠ blau sein (= to be drunk)) and some are impossible to translate because a correspondent concept does not exist in the target language (e.g. the Chinese concept of "feng shui" [23]).

In this thesis the post-editing process is analyzed by examining data from four different target languages, namely Spanish, French, Italian and German. The purpose of the analysis is two-fold; the crosslinguistic comparison of the data under temporal, technical, and cognitive aspects will lead to a better understanding of the post-editing process. In a second step the findings of the analysis can be used to propose changes in the post-editing working environment for an improved work flow.

For the crosslinguistic analysis I particularly want to address two research topics:

(1) Which properties of the source text increase the post-editing effort crosslinguistically?

(2) Are translation and post-editing effort negatively influenced by the same source segments?

In order to get a coherent picture of the challenges translators face when post-editing output from statistical machine translation engines, different measures are combined. The temporal complexity of a segment indicates how much time and effort a segment causes for the post-editor. The technical annotation classifies the editing operations into different types and shows how often each type occurs in the different target languages. Relating the edit categories to the temporal measure reveals which operations require more effort than others. This allows us to draw conclusions on the problematic source segment properties. Additionally, I evaluate the translator's subjective feedback to understand their experience with the new technology. An analysis of the pause time also gives evidence for the cognitive effort of the modifications. Formatting operations might be cumbersome and time-consuming, but are probably not very challenging. Combining the pause time measures with the editing operations provides further insights into this relation. The established methods can be used to compare the translation and the post-editing effort. In order to better understand the difference of the two processes, it is useful to examine whether the segments causing increased effort correlate for the two procedures. Terminology issues can be problematic for both activities, while other phenomena are related specifically to post-editing or translation. Understanding the differences and similarities between post-editing and translation makes it possible to adapt the findings from translation research and project them on the post-editing process.

In order to provide the background for the understanding of the thesis, the theoretical underpinnings are given in section 2. Machine-assisted translation is perceived as a serial process consisting of preprocessing, machine translation and human correction. The details for these three processes are explained in the corresponding subsections. In addition, post-editing is compared to human translation and in particular to the revision of human translations. Finally, the possibility of lowered quality expectations for the output is discussed.

The data under study was collected in the context of a productivity test for the company Autodesk. It was earlier presented with different analyses by Plitt and Masselot in [53]. A description of the data and the collection procedure can be found in section 3. Several temporal and technical calculations had already been performed on the data for the evaluation of the productivity test. These measures are summarized in section 3.4.

The analysis is divided into three different aspects. The temporal analysis in section 4.1 evaluates the temporal complexity of a segment. For the technical analysis (section 4.2), I establish an annotation scheme that captures the modifications a post-editor performs on the raw machine translation data. The data will be analyzed according to this scheme to detect crosslinguistic differences and similarities. Combining the temporal complexity measure with the annotation data facilitate the detection of more complex and more time-consuming edits. In the cognitive analysis in section 4.3., the exploration of cognitive cues in the data gives insight into the underlying process. The data was not collected with the purpose of a cognitive analysis, so some important factors are missing. However, a pause time analysis and a summary of the translators' subjective feedback were possible and

gave interesting insights into the cognitive process of post-editing. For all three analysis aspects, the identification of crosslinguistic negative translatability indicators and the comparison of post-editing and translation are evaluated.
The results of the analysis are summarized and discussed in section 5. Additionally, the conclusions drawn from the analysis of the linguistic process lead to proposals for possible improvements in the post-editing work flow.

2. Background

Recent work in human translation research perceives the translation activity as a process (e.g. [4], [66]). The task is split into three stages [33]: orientation in the source text together with the cognitive segmentation of the text into suitable units, draft translation of the segments into the target text, and finally, revision and correction of the draft. Usually these stages occur consecutively, but occurrences of regressive steps or loops are common. For example, when creating the draft translation of a segment, the translator might reread the source text for a better understanding. When comparing human translation to machine translation, it can be noted, that even though the employment of machine translation technology radically changes the translation process from a text creation to a text correction task, the perception of a serial process remains. The process can be segmented into three different subtasks (Figure 1). First, the text is preprocessed into a suitable textual format for the machine translation engine, and then it is automatically translated and is finally post-edited by a human translator.
In the next sections I will elaborate more detailed on the properties of each subtask. Additionally I will explain how post-editing differs from the revision of human translation. In 2.5 it is described how lowered output quality expectations can facilitate the overall process.

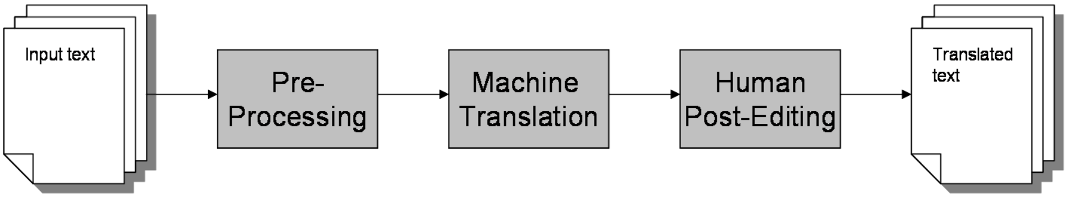


Figure 1: The translation process including post-editing

2.1. Preprocessing

Preprocessing refers to the modification of the source text before applying the machine translation system. Standard preprocessing only operates on the surface of the text. For example, data type or format changes might be required to match the input constraints of the machine translation engine. Figures or images usually have to be removed from the source text and processed independently. The input files should typically contain only one segment per line and each segment needs to be tokenized. Tokenization is the process of splitting a stream of text into individual items, in practice this usually means to separate punctuation by whitespaces from the preceding or following word. These requirements might vary depending on the system, but they are important to be fulfilled. Otherwise the input data cannot be processed properly and the post-editor has to deal with faulty output.

In addition, more advanced preprocessing steps specific to machine translation could also be applied. Due to exhaustive training processes, statistical machine translation systems provide a very detailed vocabulary and good context-sensitive lexical selection. They work particularly good on standardized input from narrow domains such as technical reports [30]. Erroneous output is, however, still unavoidable on all linguistic levels, especially for texts from unlimited domains such as newspaper articles. The disambiguation of polysemous words or structures remains a big challenge. Additionally, system inherent failures such as incorrect punctuation or the accidental omission of words degrade the output. Some of these errors, such as missing named entity recognition, formatting mistakes or problems with subordination are predictable as de Camp [14] showed in a classification of common errors of machine translation technology. Her data observations confirmed a commonly accepted presumption: Machine translation quality highly depends on the difficulty of the input. The *translatability* [67] of a source sentence determines its suitability for machine translation. More simplistic sentences have a higher probability to be machine translated correctly, they have a better translatability. Underwood and Jongjean [67] described linguistic phenomena that have a negative effect on the machine translation performance when they occur in the source text. In their English-Danish machine translation system, sentence-initial adverbs, multiple coordination, and sentences containing potentially ambiguous subphrases like prepositional phrases caused serious translation problems. Bernth and Gdaniec[6] identified similar source text properties that negatively affected the English-German system *Logos*. These phenomena determine the translatability of a segment and are therefore called *Negative Translatability Indicators* (NTIs). O'Brien additionally [47] distinguished between grammatical indicators like the use of the passive voice and stylistic indicators like parenthetical statements in the middle of a sentence. In example (1) from Underwood and Jongjean ([67], p.3) four grammatical indicators are highlighted.

1. **However**, **in practice**, this is not the way **that** such containers are filled.

Underwood and Jongjean did not describe exactly how these indicators degrade the machine translation. They mention word order problems caused by subclauses, the preposition "in" and the subordinating conjunction "that" as examples for this phenomenon. The word "practice" is a noun-verb homograph, how it is translated depends on the context. Sentence-initial adverbs like "however" are generally difficult to translate adequately, because they can have various different and often vague meanings.

To compute a translatability score for a sentence, most researchers rely on the number and the type of negative translatability indicators appearing in the sentence (e.g. [67], [6], [47]). Different types of indicators receive different weights, and for each occurrence of an indicator the weight is accumulated to determine the translatability. How to actually set the different weights and whether to apply confidence values for simple sentences [6], or penalty values for complex sentences [67], is still under debate. These proposed methods for a translatability measure have mainly been based on experiments with only one specific language pair, to what extent the findings can be generalized to other languages has remained unresolved. The occurrence of negative translatability indicators has been shown to influence the quality of the machine translation output. This easily leads to the assumption of a correlation between the type and amount of negative translatability indicators in the source sentence and the post-editing effort for the machine translation output. O'Brien [49] tried, but was not successful to confirm this correlation by comparing pause durations for sentences with and without negative translatability indicators. Her results showed that difficult source segments are always preceded by a pause. However, the correlation between the duration of a pause and the editing of the corresponding element appeared to be only minor. Sentences with few or no of the determined negative translatability indicators, also showed frequent occurrences of pauses. There might be two possible reasons for this, one being that the analysis of pause time alone is not a sufficient indicator for cognitive processing and the other lying in the definition of negative translatability indicators. The occurrence of pauses does not seem to be coupled with the post-editing difficulty. A combination of pauses time measures with technical and temporal measures could help to get a more complete picture of the post-editing process and to identify which edit types are correlated with longer pause times. Another explanation could be that the determined negative translatability indicators did not affect the post-editing effort as much as expected. They might be perceived as degrading the machine translation quality, but the influence on the correction of the output could be only minor. A solution to avoid this dependency between source text complexity, machine translation quality and post-editing effort could be to directly analyze the complexity of the post-editing effort and determine the influencing parameters backwards. This means to analyze the segments that have caused difficulties in the post-editing process and detect their common properties.

The translatability discussion revealed that some sentences are more likely to be machine translated correctly then others. This gave rise to the idea of constraining the machine translation input to avoid difficult elements. Huijsen [28], Mitamura [45] and more recently Aikawa et al [1] proposed the policy of *Controlled Language* (CL), a set of authoring guidelines that aims at simplifying the input for machine translation to avoid negative translatability indicators. The guidelines partly address correctness issues like correct spelling, accurate punctuation and the adequate use of capitalization. These aspects are standard quality indicators that should generally be respected for all kinds of texts. In contrast the controlled language style principles try to overcome complex structures and ambiguities by giving suggestions like the following [1]:

2. Category: Formal Style, "Don't use slang or colloquial expressions"
   a) Our next bit of magic was to increase the number of storage groups.
   b) Our next **improvement** was to increase the number of storage groups.
3. Category: Relative Clauses, "Avoid reduced relative clauses"
   a) Use only fonts optimized for display on the Web.
   b) Use only fonts **that are** optimized for display on the Web.

The example in 2.a) can be rewritten as in b) to avoid the idiomatic expression "bit of magic". This modification does not change the meaning of the sentence, but it reduces it to a less figurative expression. In 3.b) the reduced relative clause of a) is expanded to reduce the ambiguity and thus improve the translatability of the sentence. The controlled language rules do not impair the comprehensibility of the text, in fact they might even improve it [28]. However, the excessive use of such simplified structures might degrade the perception of the quality of the text and might take the reader to get used to. Reuther [58] has shown that controlled language rules aiming at translatability are more restrictive than rules defined for improving the readability of a text. Avoiding parenthetical statements, the cautious use of relative pronouns and reducing ambiguities promote a very factual style that affects the naturalness of a text.[1] Repetitions and rewritings might be disturbing, especially when the context already provides sufficient information for disambiguation. Applying controlled language rules thus can lead to a trade-off between the translatability of a text and its stylistic level: the more rules constrain the text, the less fluent it reads.

The controlled language guidelines can only be followed while actually producing the source text. In practice, localization companies and translation agencies currently have to accept their clients' content as it is and cannot perform stylistic changes. However, detecting which sentences cause augmented post-editing effort can provide means to effectively improve the working process for these sentences. If the same source text properties cause problems among multiple languages, it might be worthwhile to develop rules together with the content writers to avoid these properties already in the production process.

For this purpose a crosslinguistic analysis is necessary to evaluate the actual impact of complex source structures on the post-editing effort. The described approaches to translatability determined the negative translatability indicators based on the machine translation quality of bilingual studies. O'Brien then determined whether the post-editing effort increased for exactly these previously defined negative translatability indicators. Another possible approach is to work the other way around. Crosslinguistic negative

[1] compare the comments about "general English" by MacDonald in [44], p. 15

translatability indicators can be detected by actually examining sentences that cause high post-editing effort. The translatability is thus not determined by the machine translation quality alone, but by the effort it causes to transform the machine translation output into a correct translation. This approach also avoids the set of problems related to the estimation of machine translation quality that I will address in the next section.

2.2. Machine Translation

Machine translation systems take a source text as input and compute its translation into a target language. There exist different approaches to machine translation relying on a different understanding of language. The major directions are statistical systems, rule-based systems and hybrid systems. In order to understand the machine translation output used in the current study I will give a short introduction into the underlying paradigms of these directions.

Statistical systems work with reference translations and compute the probability of each phrasal translation. The term phrase comprises meaning units from whole clauses down to single words. In the training phase a phrasetable is computed that determines how likely it is that a target phrase is the correct translation of a source phrase. During translation the phrases with the highest probabilities are combined to form the target sentence [43]. The system usually tries to match bigger units, only if no correspondent is found the source segment is broken down into smaller parts. A language model then assigns a score to evaluate whether the found combination of phrases is a probable target sentence. The language model has no knowledge of the source language and can therefore not judge the translation, it only checks the validity of the target sentence. The additional knowledge about the target language helps the system to select the most probable translation by penalizing sentences containing wrong word order and sentences with missing elements. Statistical systems usually succeed well in lexical selection when applied to texts from the same domain they were trained on. In new domains vocabulary gaps complicate a successful translation. On the syntactic side the systems have more weaknesses in forming the correct structure [64] than rule-based systems. The language model can only indicate a possible error, but not exactly determine the syntactical failure. When combining phrasal translations from different texts the agreement of distant words (e.g. subject-verb agreement) often cannot be maintained and sometimes words might even be dropped (e.g. verbs) [63] as shown in the following example taken from Theison ([62], Appendix). Sentence 4 is the German source, the next is the reference translation and sentence 6 is the machine translation output produced by the statistical open-source system Moses [37].

4. Source: Er setzt diesen Völkermord derzeit fort und versucht außerdem, Medienfreiheit und Rechtsstaatlichkeit in Russland, die ohnehin kaum existieren, ganz zu beseitigen.

5. Reference: He is currently continuing this genocide and moreover he is also trying to completely eliminate the freedom of the media and the rule of law in Russia ,which ,as it is, barely exist .

6. Statistical MT: He is currently trying to continue this genocide, and also, media freedom and the rule of law in Russia, which already exist, very little .

7. Rule-based MT: He puts this nation murder currently away and in addition attempts completely to remove media freedom and rule of law in Russia that anyhow hardly exists.

In the statistical translation the main verb "trying" has moved to another place and therefore the subclause remains without verb. The adverb "completely" is also missing. Sentence 7 is the output of the commercial rule-based machine translation system Lucy.[2]

Rule-based systems follow a more linguistic approach. The source sentence is transformed into an internal representation capturing the syntactical and semantic properties of the sentence. How exactly the internal representation is realized varies from system to system, usually an underlying grammar formalism such as lexical functional grammar (LFG) [35] or head-driven phrase structure grammar (HPSG) [54] determines the realization. The internal representation of the source has to be translated into a representation of the syntactical structure of the target language that reflects the same semantic properties by following a set of transformation rules specific to each language pair. Suitable lexical items have to be selected from a dictionary that stores translations for words and terms to properly represent the semantic content. To distinguish the different meanings of ambiguous words lexical constraints, verbal selection preferences and semantic types have to be respected. Adapting a rule-based system to a new domain thus requires adding lexical items to the dictionary and updating the lexical constraints. Unless this procedure is automatized, it is a very time-consuming process. The output of rule-based systems usually delivers a good syntactical structure with only very few grammatical mistakes. However, this only holds, if the analysis of the source sentence succeeded. For complex structures the system might fail to build the internal representation properly and thus can only return a partial translation. Another weakness is the selection of dictionary entries. If several translations for one word are available, the system has only restricted means to choose the correct entry [63]. Very limited context awareness and insufficient

[2] Lucy is owned by the company *Lucy Software and Services*. Its architecture is described in [3].

selection routines can often not achieve a successful disambiguation of lexical entries. As can be seen in example 7 the rule-based system Lucy fails to choose the correct translation for "Völkermord" (= genocide) and for "setzt fort"(= continues). The syntactical structure, in return, is much better than in the statistical output in 6 and the sentence reads more fluent.

The idea of combining the previous two approaches into *hybrid systems* originates in the almost complementary strengths and weaknesses of statistical and rule-based machine translation (e.g. [64]). Ideally the weaknesses annul each other in the combination whereas the strengths accumulate. In the worst case the disadvantages of each approach lead to an even worse quality when combined. There are many different ways to combine rule-based and statistical systems. Rule-based systems can give a skeleton translation that is enriched with statistical procedures or the statistical output is augmented with linguistic knowledge to improve the grammaticality. Multi-engine combinations [12] take hypotheses from several systems - both rule-based and statistical - and either select the best translation or combine the best parts of the different hypotheses into a new translation, that is assumed to be better than the original proposals. The selection procedure can again be either statistically driven or be performed in a rule-based setting. The details of each system are unnecessary here, only the *statistical correction of rule-based output* is introduced. This architecture is relevant for the post-editing discussion as it can be perceived as a simulation of the human post-editing process. Some of the errors in rule-based machine translation output occur predictably and could be automatically corrected. Statistical systems can be trained on reference translations to transform rule-based output into a more correct translation [61]. As these are two separate processes, the statistical correction can be seen as automatic post-editing of the rule-based output.[3] Especially the lexical selection quality, a known weakness of rule-based systems, gets increased by the employment of statistical techniques [16]. Statistical methods could thus be exploited for a faster domain adaption of rule-based systems (e.g. [31], [18]). However, the statistical correction might also introduce new errors, it could impair the syntactical structure or degrade the accuracy [17]. Statistical correction techniques can only partially improve rule-based machine translations. It still cannot be guaranteed that the final output is a correct translation.

Doyon et al. [15] experimented with automatic learning of post-editing techniques from human post-editors for an Arabic-English system. Human post-edits were classified into the categories "easy to automate", "difficult to automate" and "impossible to automate/should only be performed by a human". 63% of the post-edits fall into this last category, most of them being word insertions or deletions. Nevertheless, Doyon et al.

[3] In several articles the process is called "post-editing" [61]. To avoid confusion with human post-editing the term "statistical correction" is used in this thesis.

applied commercial automatic tools that were initially developed for second language learning in order to improve the texts. These products can correct wrong spelling and simple grammatical errors. To judge the quality of the output evaluators were asked to rate the grammaticality of a text on a scale from -3 (extremely unacceptable output) to +3 (extremely acceptable output). Additionally they should indicate at what level of the rating scale a document becomes useful to them. None of the correction tools could reach the established level of usefulness. The automatic correction of the machine translation could not at all improve the evaluators' perception of quality of the output. Texts post-edited by human editors in contrast all scored at or above this level. This suggests that statistical correction techniques could be applied to improve the lexical selection of rule-based systems in a hybrid architecture and facilitate the task for post-editors, but they cannot fully take over the post-editors' work.

Directly improving or comparing different machine translation systems is out of the scope of this thesis, but there is a lot of research focusing on it (e.g. [64], [71]). The particularities of a certain system have to be considered when evaluating the post-editing effort. I will discuss how the results of this thesis can be generalized to other systems in section 5.1.

Related approaches

Machine translation is a fully automatic process, due to its persisting weaknesses so far *computer-assisted translation* has played a bigger role in the translation industry. Computer-assisted translation refers to technological means supporting the translator's work. The most important tools are translation memories and terminology databases.[4] Newer approaches try to integrate machine translation into computer-assisted translation environments (e.g. [60]), so the boundaries are blurred.

Translation memory content is generated by human translators, but extracted automatically. Previous translations are stored and the software screens the source for segments that have already been translated before. These findings are considered *exact matches*, only partial accordance on a subset of the words is called a *fuzzy match* and comes along with a percentage value indicating the degree of the match. The reuse of earlier translations assures the coherent use of terminology because translation memory content usually consist of translations from similar text types and genres. Good matches from regularly reviewed and quality-checked translation memories are often preferred over machine translation output because the quality is more reliable. In practice many translators adhere to the principle that applying machine translation technology makes only sense if the translation memory returns matches below a threshold of 75-80% [10]. Otherwise the translation memory quality is perceived to be more dependable and easier to adapt. However, this assumption is rather a widespread custom than an empirically supported

[4] García [21] also mentions hive translation and "translation-on-tap", these are still on an experimental level.

finding. Translation memories might also contain errors [9] and for fuzzy matches new editing effort arises. Guerberof [25] has found that editing fuzzy matches with a percentage value in the range of 80-90% already leads to worse quality and smaller productivity than editing machine translation.

Another helpful aid for translators are terminology databases. They are simple electronic tools that provide all relevant technical terms, abbreviations and specialized vocabulary for a certain domain. They range from simple lexicons and glossaries up to structured thesauri. Terminology databases help the translator to find the correct terms and to avoid inadequate translations. Databases can be shared among a group of translators and therefore facilitate the coherent use of terms in a whole project. Approaches for automatic terminology extraction exist, but as they are known to also introduce wrong terminology pairs, many databases are still hand-built [57]. Computer-assisted translation environments combine translation memories and terminology databases. Source words that exist in the terminology database can be highlighted and the translation memory concordance function automatically shows all previous translations of the source word in context which facilitates the selection from multiple possible translations. The provided features in computer-assisted translation environments are manifold, also depending on the tool in use.[5]

A machine translation system might probably return better results than computer-assisted translation software especially when working on unknown domains. The phrasal approach of statistical machine translation systems allows for a better generalization of the data than the segmentation into bigger units in translation memories. Several proposals have been made to effectively combine the advantages of translation memory matches and machine translation technology. Simard and Isabelle [60] experimented with selection procedures that decide whether to use the translation memory match or the machine translation hypothesis based on the similarity between the source and the translation memory match. They also evaluated possibilities to enrich the machine translation system by adding feature functions that include knowledge from the translation memory to the phrase table and to the language model. If the translation memory contains suitable matches, these methods significantly improve the machine translation quality.

Independent of the technology in use the quality of the machine output is assumed to affect the post-editing effort (e.g. Krings). The resulting hypothesis is simple: the better the output, the less corrections have to be performed, the less effort arises for the post-editor. Evaluating this hypothesis yet is not so intuitive. How to judge the quality of a machine translation is still an open debate. Many researchers try to find means to automatically evaluate the quality of translations, for example the widely used BLEU-score

[5] A detailed description of features for computer-assisted translation can be found in [8]

[52] computes the similarity of the machine translation output to a reference translation by counting the matching subsequences. This measure works appropriately when judging translations that use similar words as the reference translations. Alternative translations using different words are assigned a particularly bad score, though they may express the same content. The METEOR score [5] tries to overcome this weakness by including stem comparisons and word similarity scores into the measure. However, for a confident assessment human judgment is still considered the best tool, though it is always subject to individual differences. These quality assessments all refer to the final product of the translation process. Which quality aspects of the intermediate machine translation outcome in fact influence the post-editing process has not yet been thoroughly investigated. Guerberof [25] has found that seemingly good machine translation results or translation memory matches mislead the editors to overlook terminology errors. Detecting surface errors in a structurally good translation might be more difficult than correcting obvious severe errors. Categorizations of machine translation errors according to their degree of influence on the post-editing effort mainly stem from the 1980s (e.g. Lavorel [39], Green [22]). Since then machine translation quality has changed significantly and problems like the "inappropriate one-to-one lexical translation" of words that could better be expressed in a multi-word expression are for most systems not a problematic issue any more.

2.3. Post-Editing

Post-editing is the process of the human correction of machine translation output. In the correction task four different levels of final quality can be sought as described by Allen [2]: no editing, rapid post-editing, minimal or partial post-editing and full post-editing. The level of post-editing depends on the purpose of the target text. Two purposes of machine translation can be distinguished, inbound and outbound translations.

If a text is only needed for brief information and shallow orientation in a topic like in internal communication, rough translations are sufficient. This practice is called inbound translation. For inbound translations *unedited* machine translation output might already be satisfactory. Especially in small domains the machine translation output can be a sufficient basis for information. Machine translations of more sophisticated topics require at least a short human revision. In this case, *rapid post-editing* that corrects only "blatant and significant" errors without accounting for style can be utilized. Allen [2] does not further determine the category of these errors, he probably refers to grave mistranslations and incomprehensible sentences. This technique usually applies for urgent texts like work papers or technical reports that are not intended for public use.

Outbound translation refers to texts that are determined for publication and therefore have to fulfill standards of higher quality. *Minimal or partial post-editing* tries to keep the structure of the machine translated output and remove the errors while performing only a minimal number of changes. The term "minimal" can be seen as a continuum that is defined by a company's expectations and guidelines. Typically all surface errors like grammar, orthography or formatting mistakes are corrected. The graver stylistic errors are then adjusted, but only if they generally degrade the quality and the comprehensibility of the target text. Stylistic modifications should not be performed just to match an editor's personal preferences. The upper bound of this post-editing continuum is also called *full post-editing*, then the final quality should be indistinguishable from a human translation.

Different measures have been developed to better understand the post-editing task and the cognitive process. The most striking property of post-editing is a significantly higher productivity rate compared to standard translation. Post-editors easily reach a higher data throughput than translators for the same type of texts. Experimental comparisons between post-editing and translating all confirmed this result, but with varying degree ranging from 13% to 74% improvement. This large variance arises from the combination of subjective differences and the use of different machine translation systems. The productivity is measured in translated *words per hour*. Guerberof [25] has shown that post-editing machine translation is slightly faster than post-editing translation memory fuzzy matches and 13-25% more efficient than translating from scratch for an English-Spanish system. In a study by Flournoy and Duran[20] the participants improved their productivity by 20-51% when employing machine translation systems. Plitt et al. [53] compared translation and post-editing productivity for four target languages (Spanish, French, Italian, German) and confirmed a significant higher productivity (74% on average) in post-editing than in translation for all of them. Their study constitutes the basis of this thesis and their findings on temporal issues will be discussed in more detail in section 3.4.

In the translation business, translation speed is directly reflected in financial expenses, so measuring the translated words per hour is the most important factor for determining the commercial benefit of post-editing. The average productivity improvement can, of course, account for a general tendency. All cited researchers agree, however, that the processing time also depends on subjective differences such as work experience or previous use of technology and might be biased by very fast or very slow translators. When studying the influence of source text phenomena on the post-editing effort, the temporal variety of the subjects has to be respected. In section 4.1 one way to accomodate individual differences is described. In general, temporal measures alone can only give limited insights into the post-editing process. An increased productivity provides no evidence about which aspects actually fastened the process. Did the machine translation component mainly reduce the typing time and leave the cognitive process of translation unchanged, or did it affect the process on a deeper level? The average productivity might also conceal possible delays for certain source properties. It is therefore important to not only evaluate the temporal measures in isolation, but combine the findings with technical and cognitive insights.

For the determination of the technical post-editing effort the actual modifications of the text are considered. The *Levenshtein distance* [40] compares the machine translation output to the final translation by counting all deletions, insertions and reorderings on a character basis, the *word error rate*(WER) computes the same on word level [51]. The position-independent error rate is a more robust measure that computes the same as the word error rate, but neglects the word order [65]. These editing measures show the concrete modifications, yet the difficulty of the operation is neglected and thus it is not possible to distinguish between different edit types. It can be assumed that, for example, adjusting a wrong case marking is probably easier than improving the translation of an idiomatic expression (compare [23]) because there exists only one solution for the correct case. The edit distance measures do not differentiate between surface operations and deeper changes of the syntactic or semantic structure. Manual annotations of post-edited data (e.g. Guerberof [25]) could shed more light on this approach, but unfortunately they usually put the focus on the machine translation errors and not explicitly on the actions undertaken by the post-editor. Categories like "Mistranslation" or "Accuracy" only highlight the wrong aspect of the source text, but do not indicate which steps are necessary for the post-editor to resolve the error. For a deeper understanding of the post-editing process, technical annotations should reflect the post-editors decisions and not the properties of the data. Relating the findings of this technical effort to the temporal and cognitive measures make it possible to identify the operations which challenge a post-editor. Examining the process under these three different aspects facilitates to distinguish between time-consuming edits, technically complex edits and cognitively challenging edits. Then, it might be possible to draw inferences about the source text properties that provoke increased post-editing effort.

In translation research, the occurrence and the duration of *pauses* have been identified as an indicator of cognitive processing (e.g. Jakobsen [32] and Hansen [27]). Jakobsen observed a "systematic syntagmatic distribution of delays" during the translation process and Hansen classified different categories of pauses. O'Brien's work on pause time analysis [49] has already been partly introduced in section 2.1. She transferred the pause analysis to the post-editing task and conducted a *choice network analysis* to compare differing decisions of post-editors. In a choice network analysis several post-editors work on the same text and afterwards the results are compared. Whenever the post-editors made different decisions in their correction, the respective region is considered to be a difficult property of the source text. O'Brien then analyzed the occurrence of pauses in exactly these regions. The results show that more complicated decisions are always preceded by a pause. However, a correspondence between the duration of the pause and the editing of more difficult elements could not be detected. In other words, the occurrence of a pause can indicate processing difficulties, but the anticipated complex regions did not cause longer pauses than regions that were considered to be easier. The choice network analysis detects phrases that allow multiple correct translations, whether the post-editing of this phrase has caused difficulties for each individual translator remains unresolved. A deeper analysis of the technical effort might help to correlate pauses with the editing complexity. A choice network analysis is only possible when post-edited texts from different editors are available. However, there also exist other cognitive measures that are applicable to individual participants. Krings [38] was one of the first researchers who empirically analyzed the post-editing process. He used *think-aloud protocols* to determine how the cognitive process of post-editing is different to translation and revision. For these protocols post-editors verbalize their decisions for the researcher and motivate each step while actually performing the edits. They speak freely and post-edit at the same time; this allows an online insight into the translator's decision process. The researcher might point the participant to specific aspects, but usually he does not interfere into the verbalization process. The protocol is recorded so that a detailed analysis combining think-aloud data and the post-editing product can be performed. Krings discovered that the verbalization slows down the post-editing process by one third. He also noticed that the additional task of having to utter explanations impeded continuous correction acts like reordering phrases and favoured singular edits on individual words, like deletion or insertion. Jakobsen and O'Brien questioned the use of think-aloud protocols for exactly these reasons. The slow-down effect shows that the verbalization interferes with the translation process. This leads to a change in the actual process, which is reflected in the different edit patterns. The use of think-aloud protocols directly influences the temporal measures, thus only one of the procedures can be applied.

Since the development of the key-logging software *Translog* [34] by 1999 it was possible to avoid these side effects. Translog monitors the translation or post-editing process by recording all performed keystrokes and mouse movements as well as the time spent on the task. These actions can then be played back to the editor, which allows retroactive protocols subsequent to the post-editing task. Retroactive protocols can simulate think-aloud protocols without interfering in the actual post-editing process. Participants first terminate the post-editing task and can then watch their performance. They can thus comment on their decisions and explain their motives without interrupting the actual process. However, this retrospective view might change the perception of the process.

Another possibility is to couple Translog with *eye-tracking measures*. With eye-tracking it has become possible to investigate the eye movement behaviour in milliseconds accuracy, while the post-editor is working on the task. Recent studies in translation research have triangulated eye-tracking data with Translog results to get further insights into the unconscious cognitive processes that cannot be verbalized (e.g. [46], [11], [59]). O'Brien has used eye-tracking measures to show that the cognitive load for correcting translation memory fuzzy matches and for editing machine translation matches appear to lie in the same region [48]. Further eye-tracking studies on post-editing could provide a deeper understanding of the process.

When studying the post-editing process, one primary interest is, of course, the quality of the final outcome. Machine translation output itself is not very reliable and the quality of the product might vary from sentence to sentence. Some phrases might receive perfect translations, while others cannot be translated at all. During post-editing the human translator always has the final responsibility for the translation, in order for quality expectations to be usually sufficiently met. Fiederer and O'Brien [19] evaluated post-editing output and human translations according to the criteria clarity, accuracy and style. The post-editing output scored higher or equal in clarity and accuracy aspects, but human translations were preferred in the style criterion. Plitt and Masselot [53] also compared quality judgments for post-edited and translated data. In their study[6] the post-editing data scored even slightly better than the standard human translations.

2.4. Postediting and Translation

Post-Editing and Human translation are two tasks that strive for the same outcome; a translation of a source text into a target text. Due to the partial automatization of the translation in the post-editing task, the realization of this goal results in very different processes. O'Brien compared translation and post-editing under practical and cognitive viewpoints and analyzed the different objectives [50].

Practically, translation and post-editing differ in the number of resources. Translators work with one source text and create a target version, while post-editors work with two texts; they correct the machine translation output to correctly convey the source text. These practical differences have an influence on the *cognitive* processes. As post-editors can work with an already given basis, Krings ([38]) concludes that post-editing is a significantly more linear process than traditional translation. In translation the three stages of orientation, drafting and revision might interleave each other if they are performed by the same person, so drafting and correction of the draft occur alternately. The post-editor only has to focus on the correction, thus the task can be fulfilled in a more linear fashion. The difficulty of having to handle two texts is compensated by the information gain of the machine translation proposal. The post-editor can focus on the already given translation and only needs to check the reference for content analogy. Guerberof [25] assumed that in post-editing language errors like wrong agreement or formatting mismatches do not require source text consultation to be corrected properly.

[6] This study refers to the Autodesk data used in the thesis. More details can be found in section 3.

For the *objectives* of post-editing and translation O'Brien names accuracy as the major difference. Translators try to project the textual and cultural properties of the source text as accurately as possible on the produced target text. Depending on the respective guidelines, this accuracy might not be necessary for post-editing tasks. According to O'Brien, the traditional translator training might even act "as a hindrance to post-editing where the aims are frequently different" ([50], p.101).
In the post-editing process the machine translation component fulfills the major translation part. Human post-editing is therefore often compared only to the human revision of a draft translation and not to the complete translation process.

> Post-editing is logically parallel to revision of human translation (Koby in [38], p.4).

Instead of a human translation draft, the post-editor is revising a machine translation draft. This technical difference of the draft has an impact on the nature of the revision task. A human translation is an almost complete product. The reviser only checks the translation for inadvertent mistranslations and accidental lapses. Koby (in [38]), Vasconcellos ([69]) and Simard ([61]) all agree that the difference between revision and post-editing lies in the repetitiveness of the errors. Machine translations contain significantly more errors than human translation drafts, and the types of errors are different. Revision of human translation might, for example, detect spelling errors (e.g. "a**dd**ress" becomes "A**d**resse" in German) or wrong translations of so-called "false friends", words that seem to be translations of each other, but denote different meanings (e.g. eventually $\neq$ "eventuell" (= potentially). Both examples are not an issue for statistical machine translation as long as the training material was correct. Reversely, frequent machine translation errors, such as wrong word order, occur only rarely in human translations.

> "In other words, the emphasis [in post-editing] is on an ongoing exercise of adjusting relatively predictable difficulties rather than on the discovery of any inadvertent lapse or error. The passages that clearly require corrections, though many of them are minor and local, are more frequent than in traditional revision." (Vasconcellos in [69], p.411)

This partial transformation from the translation task into a correction task might lower the demand for absolute bilingual proficiency of the translator. Judging whether a translation is correct is an easier task than producing the translation. It might be more important to sensitize the post-editor to the occurrence of errors specific to the respective language pair and to the machine translation system in use rather than extensive, active knowledge of the languages. Therefore, it is important to understand how the edit patterns differ across languages.

2.5. Output

In evaluation tasks of machine translation, translators judge the quality of the translation usually considerably more critical than potential end users[19]. For many tasks, Guzmán [26] considers a low quality translation as already adequate. Lowering the quality expectations of the post-editing output directly influences on the previous subprocesses. As the final output can be less accurate, post-editors can tolerate minor errors of the machine translation. Post-editing guidelines can be less restrictive and the omission of a correction of an overlooked error does not severely impair the outcome. Guzmán proposes to keep post-editing effort minimal and restrict changes to mistranslations, grammatical and orthographic errors. Stylistic or language-specific preferences can be ignored because they do not add to the comprehensibility of the text. This corresponds to Allen's "fast editing" which he restricted to inbound translation, yet Guzmán wants to extend the technique to outbound translations: "[...] it could be agreed that readers of manuals and user guides can tolerate a certain level of 'artificial' language as long as it is intelligible, accurate and grammatically correct" ([26],p.2). The following sentence triples are taken from Guzmán's examples for corrections that can be omitted. The first item is the English source, the second is the Spanish machine translation output and the third sentence shows the "unnecessary" post-edit.

8. a) Documentation version 1.0

   b) Versión **1.0** de la documentación

   c) Versión de la documentación **1.0**

9. a) You need to supply the Access Server name and user password to connect to the Access Server.

   b) **Usted necesita** proporcionar el nombre del servidor de acceso y la contraseña del usuario para conectarse al servidor de acceso.

   c) **Es necessario** proporcionar el nombre del servidor de acceso y la contraseña del usuario para conectarse al servidor de acceso.

In the first sentence triple, the word order is not correct. However, as long as the sentence remains understandable, which is clearly the case here, the correction is not considered as relevant. The second example refers to a language-specific detail. In Spanish, personal pronouns like "Usted" (="you") are often omitted because they are already indicated by the verb. A more impersonal translation like in 2.c) is considered more formal. The correction does add a slightly different perspective to the sentence, but it does not change the content. Thus, it is considered unnecessary.

Performing only the minimum number of operations will probably increase the post-editors speed and also facilitate the task compared to full post-editing. Stylistic operations often correspond to the translator's individual preferences and do not add to the comprehensibility of the text. Supporters of the controlled language paradigm (see section 2.1) argue similarly. Yet, stylistic means do have an important social function for the perception of content. When using web services or inbound translations, the user is probably aware of the possible faultiness of the text, and can grasp the correct information. For manuals and user guides, this is not necessarily the case. When confronted with a technological problem, the "artificial" language of a user manual can be a problem for the user as it might hinder the finding of a solution and will affect his perception of the product. Technical writers have often highlighted this importance to motivate and engage with the readers of a manual ([44], [42]) to improve the effectiveness and the user satisfaction. The accepted style impairment proposed by Guzman might be sufficient for private users who accept a cheap, but possibly poor translation. Business clients, on the other hand, will always expect high quality translations from an agency, independent of the technology used.

In the previous sections I gave the theoretical background for the post-editing process. To improve the overall process each of the three subtasks preprocessing, machine translation and post-editing can be modified individually, but as it is a serial process subsequent tasks depend on the outcome of the previous step. A better translatability of the source text which can be achieved in the preprocessing phase leads to a better quality of the machine translation output. The question of how the machine translation quality influences the post-editing effort needs to be further investigated. Generally, it is assumed that a higher quality of the output reduces the post-editing effort because only minor modifications have to be performed. However, improved machine translation output also complicates the detection of the errors. A direct correlation between the translatability of the source sentence and the post-editing effort has been proposed [49], but has not yet been sufficiently confirmed by experimental data.

For the crosslinguistic analysis of this thesis, data from a productivity test have been provided. The data set and the collection procedure are described in the following section.

3. Data

Plitt and Masselot from the American company Autodesk conducted an internal productivity test comparing human translation from scratch with post-editing machine translation [53]. Their data will be the basis for the following crosslinguistic analysis. In the next subsections, the experimental procedure and the preliminary results are summarized.

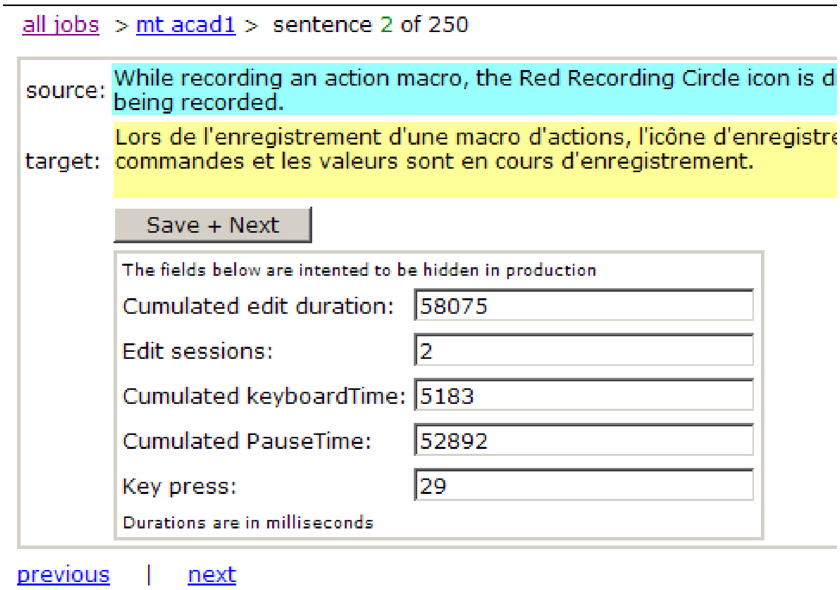
3.1. Test set

The test set mainly consisted of randomly chosen software tutorials and documentation of newly developed Autodesk products written in English that were to be translated into French, Spanish, German and Italian. The test was designated to explicitly measure the efficiency of the deployment of machine translation technology on previously unseen, "new" data. Therefore Plitt and Masselot intentionally selected only data that yielded less than 75% translation memory matches. The effect of translation memory matches on the productivity was excluded to avoid the influence of an additional factor. The data had been machine translated with the open-source statistical machine translation system Moses [37]. Moses is a statistical decoder that is trained on reference translations to build up a phrase table. For the test, Moses had been trained on parallel Autodesk data from previous years. The training resulted in a Moses translation model that was used to automatically translate the test data. Moses expects the input to be tokenized and only contain one segment per line. In translation research the term "segments" is used to describe translation units. It comprises full sentences and also shorter meaning units like bullet points, headings etc. [49]. Additional preprocessing procedures such as controlled language rules were not applied. The test data segments were split into different "jobs" and were then grouped according to the described product. The original order was preserved to provide at least minimal context, though some segment gaps were unavoidable.[7] All jobs were distributed for both tasks, translation and post-editing, but it was assured that no translator worked on the same text twice.

3.2. Procedure

The test consisted of two stages performed on two days. In the first phase, the translators had to manually translate the source texts in a specific workbench developed deliberately for this test. In the second phase of the test, the machine-translated data was post-edited by the same translators using the same workbench as for the translation task. This workbench (Figure 2) consisted of an interface displaying the source and the target segment in fields. For the post-editing task, the target segment field was prefilled with the

[7] These segment gaps occurred due to the exclusion of segments returning translation memory matches higher than 75%.



all jobs > mt acad1 > sentence 2 of 250

source: While recording an action macro, the Red Recording Circle icon is d being recorded.

target: Lors de l'enregistrement d'une macro d'actions, l'icône d'enregistre commandes et les valeurs sont en cours d'enregistrement.

The fields below are intented to be hidden in production

Cumulated edit duration:	58075
Edit sessions:	2
Cumulated keyboardTime:	5183
Cumulated PauseTime:	52892
Key press:	29

Durations are in milliseconds

Figure 2: Autodesk Workbench - the recording fields were hidden from the user

machine translation proposal. A terminology list for the specific products was available for reference. The participants had not received a specific training; they were only administered post-editing guidelines (see below). The guidelines demand to transform the machine translation output into a correct translation by keeping the technical effort minimal. This resembles Allen's definition of "minimal post-editing". The guidelines also point the post-editors to possible terminology errors that should not be ignored. The task for the productivity test was to correctly complete different jobs in the translation and the post-editing category. The interface recorded the number of keystrokes and the edit time. The edit time was divided into keyboard time when the translator was typing and pause time for the rest.

Post-editing guidelines for the Autodesk productivity test

The below high-level guidelines are based on the *Post Editing Guidelines For GALE Machine Translation Evaluation* . Please read them carefully before starting post-editing.

- Read the source first. Then read the MT output and decide whether it is worthy to be kept or should be discarded. Don't waste your time by trying to fix segments that are clearly of low quality. Only if the MT output looks usable for post-editing go on with editing it. If the output isn't worthy to be corrected, just discard it.
- The aim is to create a correct sentence with as little number of edits as possible. There are several ways to correct a raw MT sentence, but the objective is to edit manually as little as possible - and still make the sentence a correct translation.
- The final translation has to be a complete translation. The edited version should have the same meaning as the source; it shouldn't add or omit any information compared to the source.
- Do not try to make the final translation more understandable than the source.
- Do not change correct machine translations just because you prefer something else.
- Terminology. The MT engine picks up the most frequent translation of the term in a given context, which isn't necessarily the product-related (correct) translation. Follow the software bundles and the term database closely and not be misled by seemingly correct translations.

3.3. Participants

Twelve translators participated in the task, three for each target language. Plitt and Masselot did not request a specific participant profile like previous experience with translation technology or the like. The selection of candidates was performed by three independent vendors, each of them providing one translator for each target language.[8]

3.4. Results

In the two testing days the translators manually translated 4842 segments and post-edited 7878. This already indicates that the post-editing task could be processed faster. Plitt and Masselot performed a quantitative analysis of the productivity and the post-editing effort. I summarize their results in this section. A detailed description and supporting figures can be found in [53].

The purpose of the test was a productivity measure, so the primary interest was to find

[8] As all participants in the post-editing tasks were translators, the terms editor and translator are used interchangeably.

out to what extent post-editing quickened the translation process. This is measured in terms of *word throughput* per hour. All translators improved their throughput when post-editing machine translated output compared to translation from scratch, though in varying degree. On average the use of a machine translation system improved the translator's productivity by 74%. Slower translators had more benefit from the support of the system than faster translators. Still, all of them showed a significant improvement. Plitt and Masselot assume, that "fast translators presumably have a smaller margin of progression because they have already optimised their way of working." ([53], p. 11)

The influence of the *sentence length* on the post-editing throughput was also examined. For both tasks, translation and post-editing, the processing time grows linearly in relation to the segment length with a bit more variation for longer sentences. The productivity maximum is reached for a sentence length of around 22 words.

The duration of a segment can be divided into *keyboard time*, when the translator is typing, and *pause time* for the rest. During the pause time the translator reads the target and source text and reflects about possible translations. Typing might be interrupted to reconsult the text or to think about difficult translation units. Pause time is thus often considered as an indicator for cognitive processing. The use of the machine translation system reduced both; the keyboard time by 70% on average and the pause time by 31%. This indicates that post-editing not only reduces the temporal effort, but also cognitively facilitates the task of translation. For this data, unfortunately, only the overall pause time can be accessed; the duration of each pause is not available. Therefore, the location of the exact unit causing the processing load cannot be identified.

The translated and post-edited segments of the test suite where checked by a *quality* assurance team. All jobs passed this test reaching either good or acceptable quality. Post-edited sentences scored even higher than translated segments, they contained less translation errors. The large productivity gain in the post-editing process thus does not lead to a loss of quality. This shows that the post-editing task not only reduces the typing work, using a machine translation draft is a real facilitation of the translation process.

The *post-editing effort* was only analyzed by calculating the mean edit distance per translator. The edit distance was calculated by four different measures. The BLEU score [52], the word error rate [29], the position independent error rate [65] and the ratio of unchanged sentences and edited sentences. Though the computation of these measures is very different, the results were similar when comparing the edit distances of the different post-editors. The results did not show a clear correlation between the edit distance a translator achieves with the modifications and his productivity measured in words per hour. Performing less edits thus does not automatically lead to a higher productivity rate. This supports the idea that not all edits are equally expensive.

3.5. Data subsets

For the crosslinguistic analysis in section 3, two data subsets are used. A total of 454 segments had been machine translated and post-edited in all four languages. For all other segments in the data, the post-editing task had not been completed for at least one language and cannot be considered for the crosslinguistic analysis. 21 of these 454 segments had an overall duration of more than five minutes in one of the languages and were therefore excluded.[9] Thus, **subset A** consists of 433 segments in four languages which accumulates to a total of 1732 segments. Subset A is used for the detection of crosslinguistic negative translatability indicators.

Subset B is smaller and contains segments that have been completed in both the post-editing and the translation task for all four languages. Excluding sentences with an overall duration that exceeds five minutes results in 74 segments, all related to one particular product. These 74 segments are available in all four languages for the two categories of post-editing and translation, which leads to a total of 600 segments. Subset B is used for the comparison of post-editing and translation.

These two subsets form the data pool for the crosslinguistic analysis described in the following section. They are examined under temporal, technical and cognitive aspects to give insights into the post-editing process.

[9] Five minutes is an unrealistically long duration for a segment, the post-editors presumably forgot to lock the program while doing something else.

4. Analysis

This crosslinguistic analysis follows a classification of the post-editing processes into three different levels by Krings. He distinguished between temporal, technical and cognitive aspects.

> [Time effort is] undoubtedly the most important aspect of post-editing from an economic perspective. But the time effort is ultimately only the obvious external form of post-editing effort. The issue of the defining variables of post-editing effort arises. The cognitive effort and the technical post-editing effort have to be clearly separated from the time effort [...]."

Krings further divided the post-editing process into very small processing units. He identified 85 different subprocess types, that are connected to the source text, the machine translation and the target text. This analysis will operate on a less detailed level. The focus is on two research topics, that will address more general aspects of the post-editing process. The goal is to detect generic tendencies that are generalizable across all four target languages.
(1) Which source segment properties increase the post-editing effort? Do there exist common properties across all four target languages that degrade the translatability of a segment?
(2) Do the same source segments cause an increased effort in post-editing and translation?
These questions are analyzed according to temporal, technical and cognitive measures. Temporal effort denotes the overall processing time that is necessary for the correction of the machine translation output. Technical effort refers to the performed changes during post-editing. Cognitive effort describes the difficulties the correction of the segment poses on the editor.

4.1. Temporal analysis

As time is the key factor in the translation industry, this issue becomes an important point to be studied when evaluating translation processes. Recent investigations on the efficiency of post-editors have increased the interest in the use of machine translation. The productivity, as introduced in section 2.3, is usually measured in words per hour. For translation tasks, the number of words refers to the number of whitespace separated items in the source text. In the post-editing process, the translators mainly work on the already machine translated text, and only use the source text as a reference to check the intended meaning. The machine translation of a source sentence will be of unequal length for different target languages and for different machine translation engines. Counting the number of words in the source sentence cannot account for different machine translation quality. It would not make a difference whether the post-editor is working with a full

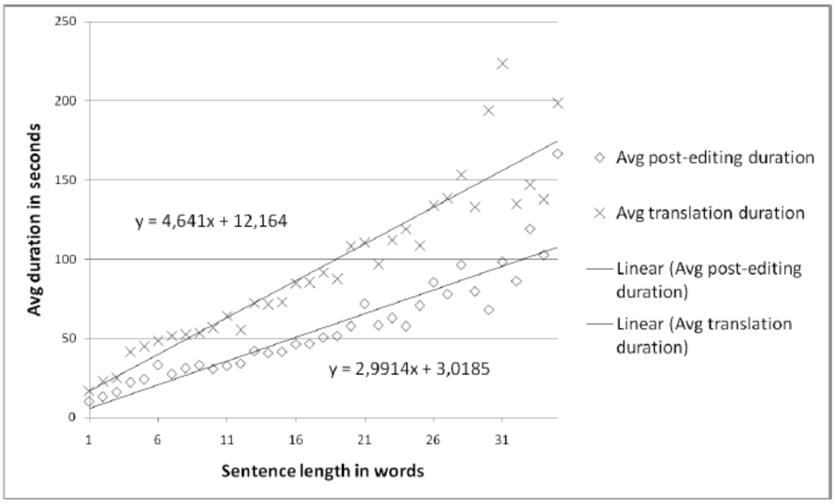


Figure 3: Duration and segment length as visualized in [53]

sentence or only with a partial translation. Therefore, it is more reasonable to take the number of words of the raw machine translation as a basis for measuring the processing time for the post-editing task. This is done in the current study.

The overall productivity measures how much data a translator can process over time, in this analysis the reverse direction is of interest. How much time is required to process the different source segments? The focus is on investigating how the segments differ in the processing difficulty they pose on the editor in order to understand which segments are more difficult than others. Does the processing time only depend on source segment properties or does the target language have an influence on the editing difficulties which a source segment causes?

The processing duration for each segment reveals information about this, but it neglects the length of the segment. The segments in the data range from length 1 to length 39. Post-editing a sentence containing 39 words naturally takes longer than editing a single word, but this does not reflect the difficulty of the operation. For the Autodesk data, the processing duration grows almost linearly to the segment length (see Figure 3). In order to abstract from the segment length, the processing duration of each segment is normalized by the number of words resulting in the reverse productivity measure in milliseconds per word.[10] It is assumed that the processing time a post-editor needs to post-edit a segment provides a measure that indicates the post-editing complexity of the segment.

Subjective differences also raise challenges that need to be considered. In the available data each segment had only been edited by one translator per language. Due to this lack

[10] Processing time of a segment henceforth refers to the normalized processing time measured in milliseconds per word.

	IT1	IT3	ES1	ES3	FR2	FR3	DE2	DE3
Mean	2343	5687	5490	2096	4973	2813	2163	4501
Standard deviation	3280	7011	9022	2670	7707	3956	1949	7172

Table 1: Mean and standard deviation for the normalized processing time of post-editors in ms/word

of extensive data from several translators in each language a direct comparison of the processing times of the different editors is not possible. Plitt and Masselot [53] already showed that the mean editing times reveal significant processing differences between the individual post-editors. This is in line with findings by O'Brien [49].

These differences are of course also reflected in the mean processing time (see Table 1). However, the individual processing times of each segment deviate highly from this mean. The standard deviation reveals that the time spent on an individual sentence varies strongly. It ranges from very long (81965 ms/word) to extremely short (183 ms/word). This confirms the assumption that certain segments are significantly more difficult to post-edit than others. In order to abstract from the actual processing times, a ranking scheme is established. The segments of each language are ranked according to the normalized processing time, with the longest processing times being ranked highest.

	IT	ES	FR	DE
Editor 1	28	25	27	26
Editor 2	22	25	23	24

Table 2: Contribution of Post-Editors to 50 highest ranked segments

Table 2 shows that the contribution of two different post-editors to the fifty highest ranked segments is almost balanced. For the fifty lowest ranked segments, the picture is similar. The subjective differences are thus resolved because of the high intra-editor variance. The ranking scheme allows to explicitly compare the post-editing complexity of different segments and to distinguish between fast and slowly edited segments.

4.1.1. Crosslinguistic NTIs - the temporal perspective

Research by O'Brien [47] and Vasconcellos [68] suggests that the post-editing effort mainly depends on the translatability of the source segment. This indicates that a source segment containing many negative translatability indicators will cause more difficulties in the editing process of the machine translated target text than a segment without these indicators. Previous research has only worked with bilingual text samples to detect negative translatability indicators. As negative translatability indicators are defined as a property of the source text only, the hypothesis should be generalizable to all target languages. A sentence is found to have a worse translatability if it contains complex struc-

tures like parenthetical statements or passive constructions. These structures remain the same independent of the target language of the translation. Thus, it should be possible to determine crosslinguistic negative translatability indicators. Instead of defining the properties which cause a worse translatability in advance, the negative translatability indicators are determined by the post-editing effort. I want to examine whether there exist source segments that are more difficult to edit in any language by comparing data from all the four available target languages.
For subset A, the proportional processing duration was calculated as described before, and then the segments of each language were ranked from 1 to 433 according to this measure. Assuming that difficult source segments are complicated in any target language, it can be expected to find a set of complicated source segments being ranked high in all languages. Table 3 shows the ranks for the five segments that were ranked highest in Italian. It can be seen that the distribution of ranks varies significantly across languages.

IT	ES	FR	DE
1	164	3	248
2	112	25	34
3	6	253	48
4	77	8	180
5	264	208	96

Table 3: Highest Italian Ranks

The intersection of the 50 highest ranked segments of all languages is a set of only five segments (segments 10 to 14). This set is relatively small, extending the focus to the 100 highest ranks only adds three more sentences to the intersection of the four target languages. This variety in the rank distribution indicates that the source segments that cause increased temporal effort differ depending on the language. Yet, the intersection set of all languages is very informative.

10. Minimum command

11. EXPORTPAGESETUP

12. License timeoutall

13. Polyline subobjects

14. License Borrowing Content Reference

These segments are all extremely short and only consist of noun compounds referring to named entities or headings. Plitt and Masselot already assumed that longer sentences are "probably more likely to be semantically self-contained than shorter sentences, thus requiring fewer context checks." [53] As these segments do not contain a verb or other content, it is difficult to understand to which objects the compound nouns are referring. Context awareness and a good familiarity with the domain are necessary to find a precise translation. Proper nouns are often technical terms, the translation of which has to be cross-checked in a glossary or terminology database.

Editing a new segment requires an almost constant orientation phase (refer to Figure 3) including navigation in the workbench which "plays also a proportionally bigger role for shorter sentences." [53] As the processing time is normalized by the length of the segment, for shorter segments this initial orientation phase is apportioned to only very few words. For German, the effect is even bigger as English multi-word expressions can often be translated into one single compound noun. Nevertheless, it can be concluded that shorter segments require a proportionally longer processing time than longer sentences when post-edited. When considering the temporal aspects alone, short segments containing only compound nouns should thus be regarded as crosslinguistic negative translatability indicators. However, the technical effort for these segments might be particularly small, as only few words have to be corrected. This aspect will further be investigated in the technical analysis.

4.1.2. Post-Editing vs Translation - the temporal perspective

In section 2.4. post-editing and translation have been described to be very different tasks. For subset B, the translators fulfilled both of these tasks for the same source segments. This provides a good basis to compare the two processes. It is particularly interesting to examine whether the same or different source segments cause increased processing times in the two tasks in order to analyze how the processes differ. Machines and humans have different weaknesses when facing complex problems. In the post-editing task the human contribution to the translation process is smaller and occurs posterior to the actual translation. This is expected to be reflected in the segments which cause difficulties. Yet, in both tasks, the goal is a proper translation of a source segment into a target segment; it might be possible that segments containing ambiguities or challenging structures cause problems independent of the particular activity. As indicated before, the processing time for translation is longer than for post-editing. Relying on the previously introduced ranking instead of the processing times makes it possible to abstract from these differences. Hence, the ranking provides a means to investigate whether the analyzed differences of the two processes are reflected in the source text properties that increase the temporal effort. For this purpose the segments of subset B have been ranked for each language and for each category in post-editing and translating. In Figure 4 the distribution of the ranks for the two tasks is visualized for each language. The distribution seems to be very unstructured initiltally. For a more objective evaulation of the ranking correlation, a

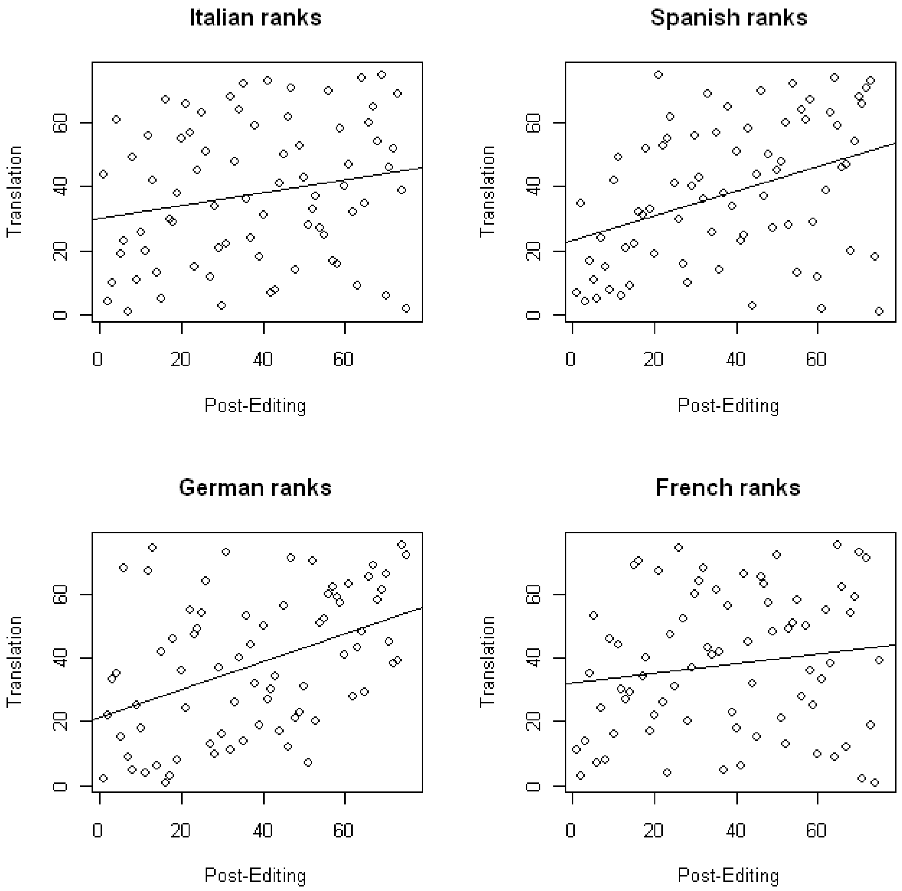


Figure 4: Correlation of ranks in the translation and the post-editing task

pearson product-moment correlation coefficient was calculated. A correlation coefficient of 1 would indicate a perfect positive linear relationship, -1 indicates a negative linear relation and 0 signals no correlation. The results show that the ranks for post-editing and translation correlate significantly, except for French (Italian: r = 0.2005405, df = 73, p = 0.0845; Spanish: r = 0.3838122, df = 73, p = 0.0006757; French: r = 0.1512091, df = 73, p = 0.1953; German: r: 0.4378947, df = 73, p = 8.532e-05). This indicates that the challenges post-editors and translators face are not completely converse. The segments that require proportionally longer processing in translation and post-editing seem to overlap except for French in which correlation was not found to be significant. Identifying the segments that are temporally challenging in both translation and post-editing helps to understand the relation between post-editing and translation.
Segments 15-18 were ranked high (<=15) in at least six of the eight categories (translation and post-editing into four target languages).

15. Polyline subobjects

16. License Borrowing Content Reference
17. Create Annotative Multileader Style
18. EXPORTPAGESETUP

Segments 15, 16 and 18 are similar to those identified as temporal crosslinguistic negative translatability indicators for post-editing. Only the presumptive headline "Create Annotative Multileader Style" is new in this set, but the pattern - short segment with noun compound - remains the same. This reveals that the terminology problems related to noun compounds are not only challenging for post-editors, but also for translators.

19. To attach a PDF underlay

Segment 19 was found to be challenging for all translators, but not for post-editors. The difficulty here lies in the correct transformation of the infinitive construction. The translator has to find a corresponding construction in the target language, while the post-editor only needs to check the machine translation proposal. For this example only minor changes were required so the post-editor saved time.
Although post-editing and translation are very different processes, the comparison of the ranks has shown that applying the correct terminology is a time-consuming task for both of them. In contrast, structural challenges can often get at least partially resolved by the current machine translation system and thus have less impact on the post-editing process than on the translation task. The translation and post-editing ranks correlated significantly in German, Italian and Spanish, but not in French. This difference is surprising and there cannot be found an obvious reason why post-editing and translation differ more in French than in the other three languages. In order to further analyze this finding, it would be necessary to consider more participants for each task. The subjective differences between the French post-editor and the French translator might have had more influence on the ranking than in the other three cases.

4.1.3. Summary of the temporal analysis

For the temporal analysis a temporal ranking according to the normalized processing time measured in milliseconds per word has been established. The distribution of these ranks differed highly across languages. A crosslinguistic intersection set of the fifty highest ranked segments comprised only five segments. This indicates that the translatability of the source segment is probably not the only factor determining the temporal complexity of the post-editing task. The variety of the temporal rank distribution across languages shows that the target language probably has a bigger influence than expected. The results have been obtained by a very small sample of post-editors. Thus, another

explanation for the variability across languages could be the individual differences between the subjects. Individual characteristics or post-editing strategies might influence the results and appear as crosslinguistic differences. This variable could be examined by an experimental approach that repeats the test with a bigger sample of participants.
Besides the crosslinguistic and individual differences, the negative translatability of short noun compound segments was found to be an important factor for all four target languages. The five segments provoking the highest temporal effort across the target languages all exhibited this property. This finding indicates that the post-editors had problems in determining the correct terminology for the noun compounds. Integrating machine translation into a computer-assisted translation environment and coupling it with terminology tools would provide support for the post-editors, and facilitate this terminology issue.
The comparison of the post-editing and the translation ranking of source segments revealed that the segments that challenge the user also overlap for the two tasks. The ranking distribution of a small set of 74 segments correlated significantly for post-editing and translation in German, Italian and Spanish, but not in French. Terminology problems are temporally challenging in both tasks whereas structural problems can be solved more efficiently by post-editors.

4.2. Technical Analysis

In the technical analysis, the focus is on the changes the editor is actually performing on the raw machine translation. It determines the types of edits that are necessary to transform the machine translation output into a correct translation. The technical effort might vary depending on the source segment properties. In the temporal analysis, short segments consisting of noun compounds have been found to cause increased temporal effort. However, the temporal aspects are not the only factor determining the post-editing effort. The noun compounds probably require only few correction edits once their meaning in the context is determined. Longer sentences or more complex structures may need considerably more edits to guarantee a correct translation. On the other hand, these edits are not necessarily very time-consuming. In the technical analysis the goal is to detect the most frequent post-editing patterns by annotating the technical changes. In a second step these edits are examined in relation to the ranking that was established in the temporal analysis. Correlating technical and temporal measures makes it possible to classify edit types in function of the time they require. Segments requiring several quick edits can thus be distinguished from those needing only few, but time-consuming edits.
Machine-translated data is often evaluated by annotating the detected errors. The LISA categorization scheme ([41], used by Guerberof in [25]) aims to standardize the error categorization of translations. The scheme comprises the error categories *Mis-*

translation, *Accuracy*, *Terminology*, *Language*, *Style*, *Country*, *Consistency* and *Format*. These distinctions classify the translation or machine translation output, but they are not very clear cut. *Accuracy*, for example, mainly refers to omissions and additions of words [25], but a missing word could also indicate a mistranslation of a multi-word expression. *Terminology* covers mistranslations of glossary terms and *Consistency* refers to coherence in terminology, both could be easily confused with each other and with the *Mistranslation* category. These categories only reflect the flaws of the translation, but they do not represent the required changes to transform the error into a correct translation. For a better understanding of the post-editing effort, the performed modifications on the raw machine translation are of primary interest. Groves and Schmidtke [24] developed an automatized process for the annotation of post-edits. As a result, the categorization is very finegrained and returns patterns like the German article "die" is replaced with "der" or even more complex structures such as (FITTED(NP (AVP(NOUN)(CHAR))(NOUN)) (NP(NOUN)(NAPPOS(NOUN)))) becomes (NOUN)(NAPPOS(NOUN)))(NP(AVP(NOUN) (CHAR)). These findings are very specific and difficult to generalize. Additionally, this automatic technique is based on structural parse trees of the machine translation engine that cannot be obtained from the Moses system used in the current dataset.

In order to facilitate the generalization of the findings and to overcome the lack of parsing structures, I established an edit annotation scheme for the Autodesk data. It will be described in the following section.

4.2.1. Annotation Scheme

The developed scheme for manual annotation focuses on technical changes like Groves and Schmidtke, but generates more general categories oriented to the LISA categorization scheme. The scheme was created for annotating the Autodesk data, but it should be generally applicable to post-edited texts.

The scheme consists of eleven edit categories: *Insertion*, *Deletion*, *Retranslation*, *Change of POS*, *Translate UNK*, *Detranslation*, *Reordering*, *Agreement*, *Recase*, *Formatting* and *Orthography*. The meaning of these categories will be described below. For an overview including examples, see Table 4. All categories can apply to either single words or full phrases. During the annotation, both the edit category and the part-of-speech tag of the changed element are stored. The part-of-speech tags can be grouped into phrasal and lexical categories. This makes it possible to distinguish between phrasal changes and edits on word level. A third dimension are tag-specific edits. Tags are usually numbers in curly brackets (e.g. "{2592}" in example 16) which serve as placeholders for HTML- and XML-tags and are left untranslated. Tags do not need to be inserted or deleted as they are directly transferred from the source text, but they frequently cause reordering problems. It is important to distinguish between these three different elements because they cause different levels of post-editing effort. Retranslating or reordering a full VP is more complex than substituting individual words because it changes the internal structure of the segment (example 20 below, target language Italian). Tag-specific edits instead (as in example 21, target language French) can be considered as surface operations.

20. Source: You have helped Viola to create annotative multileaders.
    Raw MT: È possibile creare multidirettrici annotative helped viola.
    Post-Edited: Viola è stata aiutata a creare multidirettrici annotative.

21. Source: {2592} Show Me: Use ShowMotion to Transition to a Saved View
    Raw MT: Démonstration: utiliser {2592} ShowMotion pour passer à une vue enregistrée
    Post-Edited: {2592}Démonstration : utiliser ShowMotion pour passer à une vue enregistrée

The eleven categories can be distinguished into two different groups, namely deep and surface operations. *Insertion*, *Deletion*, *Retranslation*, *Change of POS*, *Translate UNK* and *Detranslation* directly change the machine translation and are therefore considered to be deep operations. *Insertion* adds missing elements and *Deletion* removes redundant items. A *Retranslation* edit corrects the inaccurate translation of a word or phrase. *Change of POS* is a milder form of *Retranslation*, here only the word class of the element is modified. *Translate UNK* refers to words that could not be translated by the machine translation system (UNK = unknown[11]). The editor thus has to come up with a proper translation for the word. *Detranslation* captures the opposite direction; words that have been translated by the machine translation engine though they should have remained in the source version. This usually concerns named entities or technical terms that are standardized across languages. The post-editor has to detranslate the erroneously translated word back into the source form. The category *Reordering* captures changes in the serial order of the elements. Word and phrase reordering can have a strong effect on the content of the segment because a changed word order can result in a different meaning ("grand homme" = famous man, "homme grand" = tall man)[12]. Tag reordering in contrast only changes the surface appearance of the sentence and thus belongs more to the second group. *Recase*, *Agreement*, *Formatting* and *Orthography* are mapped under the concept of surface operations, they do not modify the structure or the content of the segment. In most cases they do not even require the consultation of the source text for comparison, it is simply a correction edit of the raw machine translation. *Recase* changes a word

[11] often also referred to as out of vocabulary (OOV) words (Masselot, 31.8.2010, personal communication

[12] I thank François Masselot for suggesting this example.

from uppercase to lowercase or vice versa. *Agreement* makes sure that the concordance between the elements of the segment (e.g. subject and verb or noun and adjective) is correctly adjusted. *Formatting* takes care of formal errors such as wrong punctuation, missing or incorrect whitespaces or language-specific styles (e.g. in French it is common to insert a whitespace before a colon or a semicolon, in other language this is considered wrong). *Orthography* refers to the correction of spelling errors.
The annotation scheme is applied to the final outcome of the post-editing process, thus, the assessment of the intermediate steps performed by the post-editor can only be done retrospectively. The editor might have performed additional actions (e.g. inserting a word and directly deleting it), which cannot be guessed from the final version. In the annotation, only the differences between the raw machine translation and the post-edited segment are visible and allow to draw inferences about the actions the editor had performed. To guarantee a certain constancy and uniformity among the annotations, several compromises were established.

- It is difficult to judge whether a word has been retranslated or has only undergone a change of the part of speech or has been completely deleted and another word had been inserted. For this annotation scheme, the term *Retranslation* is used when the two words have the same part of speech. If the part of speech changes, but the word stem is not modified, it is considered to be a *Change of POS*. If both the part of speech and the word stem are modified, a *Deletion* of the previous word and an *Insertion* of the new word are annotated.

- The categories *Word Reordering*, *Tag Reordering* and *Phrase Reordering* are only annotated once per segment, independent of the number of reordering steps because it already implies moving more than one element. For a reordering operation the post-editor has to keep track of the whole sequence of words anyway, independent of the exact number of words that are being changed. Formatting or casing, in contrast, is being annotated for each single operation, because those are usually individual edits that do not depend on other words.

- In order to understand how the post-editor changes a word or phrase from one form to another, the edits are annotated on a very finegrained level including linguistic knowledge. For example, if the French phrase "le dessin" in the raw machine translation becomes "du dessin" in the post-edited version, it is assumed that a preposition has been inserted ("de le dessin") and then the preposition-determiner agreement has been modified ("du dessin"). An N-gram based automatic measure like BLEU perceives this as one single change.

This annotation scheme has been applied to all 1732 segments of subset A in order to get a detailed picture of the performed edits. The large amount of annotated segments from four different languages allows to draw conclusions about the post-editing patterns. The focus is on finding out whether the necessary edits only depend on the source segment properties and the results are similar across languages or whether different tendencies can be found depending on the target language.

Insertion	insert a word/phrase	Asegúrese de que el Administrador de propiedades de capas abierto. Asegúrese de que el Administrador de propiedades de capas **esté** abierto.
Deletion	delete a word/phrase	**Los** flujos de trabajo, tubos y tuberías flujos de trabajo, tubos y tuberías
Retranslation	substitute a word/phrase	Toutefois, vous pouvez les tracer pour les revue de conceptions. Toutefois, vous pouvez les tracer pour les **révisions** de conceptions.
Change of POS	change word class	Copia di questa linea in basso 1.25 "", 10.75"" e 12 "". **Copiare** questa linea in basso 1.25 "", 10.75"" e 12 "".
Translate UNK	translate word/phrase that has been left untranslated	Usage Summary Report Ausgabe **Nutzungsübersicht - Berichtausgabe**
Detranslation	change translated word back to source word, leave it untranslated	# cols [# righe] nome del set di dati #cols [#**rows**] nome del set di dati

Recase	change to uppercase or lowercase	{10194}onglet Annoter{10195}cotes{10196}inspecter{10197}
		{10194}**O**nglet Annoter{10195}**P**anneau Cotes{10196}**I**nspecter{10197}

Reordering	reorder words, phrases, tags	Barra del panel Tabla de clavos Tabla de clavos (barra del panel)
Agreement	subject-verb-agreement noun-adjective-agreement, etc	Par défaut, une contrainte dynamique est créé. Par défaut, une contrainte dynamique est **créée**.

Formatting	change punctuation,white spaces, language-specific style, measure	{3144}Dateiname{3145}:{3146}Detail _i_start.dwg{3147}
		{3144}Dateiname{3145}**: {**3146}Detail _i_start.dwg{3147}

Orthography	correct orthography	Esta es la vista en planta de la plataforma. **Ésta** es la vista en planta de la plataforma.

Table 4: Edit Categorization Scheme

4.2.2. Results

This section presents the results of the technical annotation. The annotation had been applied to segments from four different languages, so extensive tables are required. In order to facilitate the orientation and the understanding of the results, the most frequent and the rarest edit categories will be highlighted in extra tables and discussed. Additionally, the proportion of phrasal edits and the relation between deep and surface edits have been examined.

Before having a deeper look at the technical edits, it is also worthwhile to briefly mention the unedited segments. These segments were judged as already being an adequate translation for the corresponding source segment and did not require any correction. As Table 5 shows, only about 30% of the machine translated sentences were acceptable without any human interference. Though this is actually a quite good result for a machine translation system; it shows how unstable the output still is. A human correction step is absolutely necessary to provide accurate and adequate translations. This holds equally for all four languages. Almost half of the sentences even require multiple modifications. Table 6 explicitly shows the results for each editing category.

	All 433 segments			
	IT	ES	FR	DE
No edit	121	137	122	135
Multiple edits	234	213	207	203

Table 5: Unedited segments and multiple edits

	All 433 segments			
	IT	ES	FR	DE
Insertion	185	152	131	230
Deletion	140	149	119	196
Retranslation	177	90	111	162
Change of POS	39	19	15	22
Translate UNK	15	9	1	7
Detranslation	16	10	10	9
Reordering	120	115	94	93
Recase	171	159	200	67
Agreement	101	97	97	90
Formatting	56	97	83	119
Orthography	0	2	3	1
All edits	1020	899	864	996

Table 6: Annotation Results

Interestingly, the distribution of edit types is rather similar across languages. The most frequent edit categories are *Insertion* and *Recase*. Insertion edits accumulate to a high sum in all languages, but especially for German. Casing plays a minor role for German, but in all other languages the *Recase* edits occurred very often (up to 200 edits of only 864 edits in total for French). The rarest editing operations for all languages are *Orthography* and *Translate UNK*. After summarizing more general results, I will come back to these specific cases and elaborate on them in more detail, focusing on the crosslinguistic similarities and differences.

Table 6 shows that the Italian post-editors had performed considerably more edits (1020) than the Spanish (899) and French (864) editors and even more than the German editors (996). German machine translations are known to often contain more errors due to the more complex structure of the language (e.g. [55]). The fact that the Italian editors performed 34 edits more than the German editors and more than a hundred edits more than the Spanish and French editors in only 433 segments is therefore surprising. The additional edits come mainly from the group of deeply modifying operations. This can have two reasons. Either the Italian editors did not adhere to the post-editing guideline to perform as few edits as possible as closely as the others, or the Italian machine translation output had been of lower quality. The surprisingly high number of unknown words (15 times *Translate UNK*) or wrongly translated words (16 *Detranslations*) supports the impression of bad machine translation quality, probably due to a smaller amount of good training data. Plitt (14.06.2010,personal communication) confirmed that the Italian training corpus had been smaller than the corpus for German and French, however it was still bigger than the Spanish training corpus. Though size is not the only criterion for a usable corpus, the number of seen training instances is an important factor for the machine translation quality.

Insertion and *Recase* have been the most frequent edit types across all four languages. However, there are some differences between the languages which are discussed below.

	All 433 segments			
	IT	ES	FR	DE
Insertions	185	152	131	230
V/VP Insertions	51	29	18	112

Table 7: Insertions

Insertions. The German editors added considerably more words than the other editors, almost half of the insertions concern the verb. In German, the verb or a part of it can move to the end of the segment (example 22). Statistical machine translation systems work phrase-based and therefore currently cannot capture relations between distant tokens. The post-editors have to manually correct this.

22. Nach der Anzahl der beteiligten Sprachen **können** monolinguale und bilinguale Ansätze **unterschieden werden**. ([57], p.71)

	All 433 segments			
	IT	ES	FR	DE
Recase	171	159	200	67

Table 8: Casing operations

Recase. The Spanish, French and Italian machine translation output often contained incorrectly lowercased words, especially at the beginning of a segment. German is more case-sensitive than the other three languages (e.g. "Anzahl", "Sprachen" and "Ansätze" in example 22), so the training data had been more discriminative. This explains the low number of wrongly cased words in German.

The edit types *Orthography* and *Translate UNK* depend on the machine translation quality. They both occurred very rarely and showed less crosslinguistic variance (Table 9 and 10). The training corpora all consisted of previous translations of Autodesk texts from the same domain. They are therefore assumed to be of comparable quality across languages. This high training quality hence assures a relatively stable quality of the machine translation.

	All 433 segments			
	IT	ES	FR	DE
Translate UNK	15	9	1	7

Table 9: Translate UNK

Translate UNK. The occurrence of untranslated words in the machine translation output depends on the coverage of the training data. Moses leaves words it has never seen before untranslated, assuming that they are proper names. As already indicated before, the Italian coverage seems to be worse than in Spanish, German and French.

	All 433 segments			
	IT	ES	FR	DE
Orthography	0	2	3	1

Table 10: Orthography

Orthography. Orthographic errors depend on the quality of the training corpus. This category had only been introduced for some particular spelling errors. Orthography should generally not be a problem for a machine translation system as long as it receives correctly spelled input in the training phase.

The previous tables have shown the most frequent and the rarest edit categories. In the following, the results are discussed from a more general perspective. The proportion of phrase-level edits gives further insights into the machine translation quality and the necessary post-editing effort (Table 11). In addition, grouping the edit types into deep edits and surface edits allows to draw more general conclusions about the technical effort (Table 12).

	All 433			
	IT	ES	FR	DE
All edits	1020	899	864	996
Phrase-level edits	169	94	117	174
Segments with phrase-level edits	110	68	93	116

Table 11: Phrase-level edits

Phrase-level edits. While annotating the edits, word-level, tag-level and phrase-level operations have been distinguished. Phrase-level edits change the segment on a deeper layer and therefore require more linguistic processing. When deleting, inserting or retranslating a full phrase, the whole structure of the sentence gets changed and usually the other elements also have to be adjusted. Phrase-level edits constitute less than 17% of all edits, the post-editors mainly had to operate only on word- or tag-level (83%). However, if editing on the phrase-level is required, this often affects more than one phrase. The corresponding segments are probably not very usable machine translations and have to be entirely reconstructed. Thus, the technical post-editing effort varies from only dealing with word-level errors, in most of the cases, up to the complete reconstruction of some segments.

	All 433			
	IT	ES	FR	DE
Deep edits	572	429	427	626
Surface edits	328	355	383	277
Reordering	120	52	94	93

Table 12: Deep and surface edits

Deep and surface edits. Table 12 shows the relation between deep and surface edits. Deep edits comprise *Insertion*, *Deletion*, *Retranslation*, *Change of POS*, *Translate UNK* and *Detranslation*, surface edits refer to *Agreement*, *Recase*, *Orthography* and *Formatting*. *Reordering* is listed apart because it cannot be clearly categorized as a deep or a surface error. The Italian and French editors performed more deep edits than surface edits, but this difference is less pronounced than in the other two languages. The Italian and especially the German post-editors performed almost twice as many deep edits as surface edits. This is explained with a lower number of casing errors and a higher amount of insertions in German. The Italian post-editors generally performed considerably more edits in the categories *Retranslation*, *Translate UNK* and *Detranslation*, all of them indicate a worse machine translation quality. It can be seen that both, surface edits and deep edits, play an important role in the post-editing process. The post-editors have to be sensitive towards deep language errors, as well as to less evident formatting errors.

The results of the technical annotation show that the editing patterns resemble each other across languages. Some language-specific characteristics like a smaller number of casing operations in German and an increased number of deeply modifying operations in Italian are highlighted, but the overall picture of a comparable distribution of edit types across languages remains.

4.2.3. Combination of technical and temporal measures

The annotation of edit categories helps to understand what is actually happening during the machine translation correction. However, the technical effort does not reveal which of the edits are more challenging than others. Combining the technical measures with the temporal effort helps to identify the more time-consuming editing operations. For this purpose the temporal ranking of section 4.1 is combined with the technical annotation. The occurrences of each edit in the 50 highest ranked segments (those that require a proportionally long edit time) are compared with the 50 lowest ranked segments (those that require a proportionally short edit time) of each language. The presentation of the results is similar to the previous section. First the distribution of edit types is given for both the first fifty and the last fifty segments. Comparing these two distributions with the overall distribution of all 433 segments enables the identification of faster and slower editing operations. The edit types showing the greatest alternation are then further discussed. Finally, the proportion of phrasal edits and the relation of deep and surface edits are compared for the three categories.
The first thing to note is the significant difference of unedited segments between the two categories (see Table 13). Almost 80% of the lowest ranked fifty segments are left unedited whereas this number is particularly lower for the first fifty segments. This sup-

	First 50				Last 50			
	IT	ES	FR	DE	IT	ES	FR	DE
No edit	4	8	11	14	41	43	39	41
Multiple edits	35	30	25	26	6	3	3	3

Table 13: Unedited segments and multiple edits in first and last 50 segments

ports the assumption that a better machine translation quality reduces the post-editing effort. Validating a correct machine translation is faster than modifying an erroneous one. However, unedited segments also occur in the highest ranked segments. For German and Italian this rate is surprisingly high. The editors might have considered alternatives before deciding to leave the machine translation unchanged, which increased the processing time. The high number of segments requiring multiple edits also suggests that the amount of necessary edits highly correlates with the temporal effort. More than half of the fifty highest ranked segments require more than one modification. For the last fifty segments, multiple edits only play a minor role. The detailed distribution of edit categories in the first and last fifty segments is listed in Table 14. In the first fifty segments, all edit categories are present (except for *Orthography*) with an almost comparable distribution as in the overall analysis of all 433 segments. The more drastic changes can be examined among the last fifty segments. Only few edit types occur at all, and some totally disappear from the post-editor's repertoire for this subset of fast processed sentences.

	First 50				Last 50			
	IT	ES	FR	DE	IT	ES	FR	DE
No edit	4	8	11	14	41	43	39	41
Insertion	23	28	16	36	1	0	0	0
Deletion	12	19	9	21	0	0	0	2
Retranslation	34	19	15	23	1	0	0	2
Change of POS	11	2	2	3	1	0	0	0
Translate UNK	8	3	0	1	0	0	0	0
Detranslation	2	2	0	2	6	3	2	2
Reordering	19	20	11	14	0	0	0	0
Recase	27	24	27	8	25	0	22	2
Agreement	9	11	12	12	3	1	3	0
Formatting	9	12	7	17	2	27	6	6
Orthography	0	0	0	0	0	0	0	0
All edits	154	140	99	137	39	31	33	14

Table 14: Comparison of technical edits for segments with high and low processing time

In general, these results confirm the assumption that a higher number of required edits increases the post-editing effort whereas segments needing only few modifications are processed particularly fast. The sum of all edits is, for all languages, at least three times higher in the fifty slowest processed segments than in the fifty fastest processed segments. For German, the increase is almost tenfold (137,14).

However, not all edits are equally time-consuming. The comparison allows to draw inferences about the complexity of the different post-editing categories. In the following, the edit types that show the biggest variation in the three distributions are discussed.

Edit categories that occur more frequently in the last fifty segments than in the first fifty segments required only a short processing time and can therefore be considered as fast operations. Accordingly, the edit types that are frequent in the first fifty and rare in the last fifty are time-consuming operations. In order to enable a better comparison of the distribution of edit types among all 433 segments with the first and last fifty segments, the absolute values in Table 14 are replaced with percentage values. The percentages reflect the proportion of one particular edit type among all performed edits.
Recase (Table 15), *Formatting* (Table 16) and *Detranslation* (Table 17) are by far the most frequent edits in the last fifty segments. In the first fifty segments and in the overall distribution they constitute a smaller part.

Recase				
	IT	ES	FR	DE
All 433	16.8	7.7	23.1	6.7
First 50	17.5	17.1	27.3	5.8
Last 50	64.1	0	66.7	14

Table 15: Casing operations in percentages of all edits

Recase. The amount of casing operations is on a comparable level for all 433 segments and for the first fifty segments. For the last fifty segments the proportion of *Recase* edits among all edits is more than twice as high for French, Italian and German. This suggests that changing a word from uppercase to lowercase, or the other way around, is only a minor modification that is not very time-consuming. For French and Italian the *Recase* category constitutes more than 64% of all edits in the last fifty segments. For German, this category was on a lower level, but still present for the last fifty segments. In contrast, the Spanish editors performed a total of 24 casing operations on the time-consuming segments, but for the last fifty segments, this edit type vanished completely. Thus, for the Spanish editors, the casing operation seems to be a more time-consuming operation. Another explanation could be, that casing errors co-occur with more complex operations in the Spanish segments and can therefore not be processed as fast as in the other three languages.

Formatting				
	IT	ES	FR	DE
All 433	5.5	10.8	9.6	12
First 50	5.8	8.6	7.1	12.4
Last 50	5.1	87.6	18.2	42.9

Table 16: Formatting edits in percentages of all edits

Formatting. Formatting edits also occur in the last fifty segments. Here, the picture is reverse to the *Recase* category. For Spanish editors *Formatting* is an operation that can be performed very quickly, 87.6% of the edits in the fifty fastest processed segments are *Formatting* edits. For German and French, this edit type also increases in importance, but to a smaller extent. The Italian editors generally performed relatively few formatting edits. *Formatting* as well as *Recase* operations require only minor source text consultation, if any. They are surface errors and are therefore particularly easy to detect. Exhaustive knowledge of the source and target language is not necessary for a correction of these errors.

Detranslation				
	IT	ES	FR	DE
All 433	1.6	1.1	1.2	0.9
First 50	1.3	1.4	0	1.5
Last 50	15.4	9.7	6.1	14.3

Table 17: Detranslations in percentages of all edits

Detranslation. Of the very few detranslation edits, many occur in the last fifty segments. *Detranslation* is the only deep edit that has been performed for all languages in the modification of the last fifty segments. Wrongly translating a named entity or fixed term probably results in an absurd sentence. The error is therefore easy to detect, and, as the correct term is already available in the source text, the modification can be performed considerably fast. This indicates, that detranslations might actually be considered as surface operations. However, the absolute values are very small (refer to Table 14), so the percentages for *Detranslation* might be slightly biased and should be interpreted carefully.

The edit categories, that occurred frequently in the first fifty segments and lost importance in the last fifty segments are time-consuming editing operations. They are combined in Table 18 because the tendency is similar for all three edit types.

	Insertion				**Retranslation**				**Reordering**			
	IT	ES	FR	DE	IT	ES	FR	DE	IT	ES	FR	DE
All 433	18.1	16.9	15.2	23.1	17.4	10	12.9	16.3	11.8	12.8	10.9	9.3
First 50	14.9	20	16.2	26.3	22.1	13.6	15.2	16.8	12.3	14.3	11.1	10.2
Last 50	2.7	0	0	0	2.6	0	0	14.3	0	0	0	0

Table 18: Insertion, Retranslation and Reordering edits in percentages of all edits

Insertion, Retranslation, Reordering. The three edit categories *Insertion*, *Retranslation* and *Reordering* have a major impact on the temporal processing of a segment. Together they constitute half of the edits in the fifty highest ranked segments. In the fifty lowest ranked segments, these edit types almost completely disappear. The three operations all cause a change in the structure and/or the content of the machine translated segment. For these types of changes, both the source text and the proposed machine translation have to be revised properly to find the best modification of the machine translation so that it accurately reflects the source text. Therefore, they are challenging edits that require more temporal processing.

Comparing the amount of phrasal edits and the relation between deep and surface edits for the three categories helps to generalize the findings of more and less time-consuming edits.

Phrasal edits				
	IT	ES	FR	DE
All 433	13.1	8	11.3	14.6
First 50	14.9	12.1	11.1	19
Last 50	2.6	0	0	0

Table 19: Phrasal edits in percentages of all edits

Phrasal edits. The proportion of phrasal edits is higher in the first fifty segments than in all 433 segments for all languages, except for French. In the last fifty segments, in contrast, phrasal edits do not occur (one exception for Italian). These results confirm the assumption that phrasal modifications are more complicated than word-level edits. Phrasal edits modify the segment on a deeper level, and often the remaining elements of the segment also have to be adjusted. The segment has to be revised properly after the modification to assure that the structure and content remained intact. Thus, phrase-level edits can be considered as very challenging corrections.

The percentages of deep edits, surface edits and reordering operations always add up to 100%. In order to better visualize the relation between these three groups, the orientation of Table 20 is flipped.

	All 433				First 50				Last 50			
	IT	ES	FR	DE	IT	ES	FR	DE	IT	ES	FR	DE
Deep edits	56.1	47.7	49.4	62.9	58.4	52.1	42.4	62.8	23.1	9.7	6.1	42.9
Surface edits	32.1	39.5	44.3	27.8	29.2	33.6	46.5	27	76.9	90.3	93.9	57.1
Reordering	11.8	5.8	10.8	9.3	12.3	7.9	11	14.1	0	0	0	0

Table 20: Deep and surface edits in percentages of all edits

Deep and surface edits. As the previous results already indicated, the edit repertoire for the fast processed segments consists to a very high degree of surface edits. For Italian 76.9%, and more than 90% for French and Spanish show a very clear tendency. Reorderings seem to be closer to the category of deep edits, as they do not occur at all among the last fifty segments. The percentages for German are slightly biased by the small number of overall edits (14), so the percentages are skewed. Only six deep edits and eight surface edits do not reflect the strong preference for surface edits in the other three languages.

In summary, three tendencies could be observed from the results of the combination of the temporal ranking with the technical annotation. The technical and the temporal effort correlate according to the following points:
(1) The more technical changes are required, the higher the temporal effort.
(2) The temporal effort is higher for deeply modifying edits than for surface edits.
(3) The temporal effort is higher for phrasal edits than for word-level edits.
It can be assumed that edits which demand more temporal effort are cognitively more challenging. This hypothesis will be deeper evaluated in the cognitive analysis in section 4.3.

4.2.4. Crosslinguistic NTIs - the technical perspective

A higher number of edits per segment leads to a higher technical effort for the post-editor. The technical-temporal combination confirmed the intuitive assumption that performing more edits negatively influences the processing time. In order to detect crosslinguistic negative translatability indicators from the technical perspective, the focus is placed on segments that required three or more edits. This is the case for about one third of the examined segments for each language (see Table 21 for the absolute values).

	All 433 segments			
	IT	ES	FR	DE
3 or more edits	148	165	129	150

Table 21: Segments requiring three or more edits

When building the intersection of these segment sets, a total of 32 sentences remain. These 32 segments are decided to be categorized as technically challenging because they provoked higher editing effort from the post-editors of each language. The full list of 32 segments can be found in the Appendix, but some examples are given below. In a comparison of the segments four general source text properties have been identified to cause the increased technical effort.

Long segments. The segments that crosslinguistically cause increased technical effort are very long (see example 23 and 24). They have a mean length of 19 words, the mean length of all 433 segments (12 words per segment) is considerably lower. The shortest segment (8 words) still contains more words than each of the compound noun segments causing increased temporal effort (see the temporal analysis in 4.1.1). Longer sentences contain more words and therefore naturally introduce more opportunities for potential errors.

23. Alternatively, the entire expression, sqrt (Length * Width / PI), could have been assigned to the {11941} Radius {11942} dimensional constraint, defined in a user variable, or some other combination.

24. In order to choose a language for an individual product, you first must click the Select Language for Individual Products check box, then select the language from the drop-down list.

Tags. Among the 32 segments, 15 segments contain tags. These segments frequently consist of lists of menu items (example 26), introducing several tags. Overall 42 tags occur in the 32 segments. As already described in the results of the technical annotation, tags often caused reordering effort. Additionally tags should be separated by whitespaces which was often not realized by the machine translation system and therefore had to be corrected by the post-editor.

25. A full list of available functions is documented in the {13009} AutoCAD User's Guide {13010} Help topic, {13011} Constrain a Design with Formulas and Equations {13012}.

26. {10952} Annotate tab {10953} Dimensions panel {10954} Baseline {10955}

Technical instructions for the user. Many of the 32 segments directly explain technical procedures to the user. This is reflected in the use of imperatives (example 27) in six segments, and the use of you references (example 28) in nine additional segments. In the English source, imperatives have the same inflection as infinitives and conjugated present tense verbs (except for third person singular). In the four target languages, in contrast, the imperative inflections are usually more specific. Therefore the English source verb had to be disambiguated to select the correct target verb inflection. This decision often failed because statistical systems usually do not possess grammatical or morphological knowledge.

27. Click once inside the cell, and enter {4224} Pipes {4225} as the Display Name.

28. When you do not need an underlay in the current drawing session, you can improve performance by temporarily unloading it.

Descriptions. Six segments neither contained tags nor technical instructions for the user. They provide detailed descriptions of complex technical procedures (example 29 and 30). These descriptions are very important to be precise and at the same time understandable. The post-editors have to adjust the machine translation such that it correctly reflects the source. Both example segments contain initial subclauses, a complex structure that has been considered as a negative translatability indicator for machine translation systems by almost all previous studies on translatability (e.g. [6]). Machine translation systems often fail to capture the complex content links introduced by these conjunctions and cannot correctly capture the long-distance dependencies a translation of the subclauses causes in the target languages (e.g. German verb movement).

29. If a report is filtered on user==User1 and user==User2, the resulting report contains usage of features by either User1 or User2.

30. Since a drawing file is normally compressed, the final size of a saved drawing file on disk will vary based on the size and number of objects in a drawing.

Four source text properties have been found to increase the technical effort crosslinguistically - long segments, tags, technical instructions and complex descriptions. Improving the handling of these segment would probably reduce the technical effort. The correct ordering of tags might be adjusted by automatic methods using the alignment information. This will be explained in section 5.4. For the other three properties the machine translation quality is crucial. If the machine translation system consistently fails to properly translate these segment types, it might be worthwhile to filter them in the preprocessing phase and directly translate them by hand. On the other hand this procedure disregards the advantages of the machine translation draft. Thus, a sensitization and improved training of the post-editors for the correction of these segment types could better improve the overall performance.

4.2.5. Post-Editing vs Translation - the technical perspective

The technical analysis confirmed that in the post-editing task, surface corrections like formatting, agreement and casing operations occur very frequently. During translation these aspects are of little significance. The translator almost intuitively applies correct casing and transfers the source format directly to the target text. Accidental agreement errors might occur during the drafting phase, but to a smaller extent than in machine translations. Major reordering problems (a very common error in machine translations) are very unlikely to occur in human draft translations. For the revision of a human translation, the focus is on the correct content representation and on the detection of unconsciously produced errors. Post-editors can pay less attention to these aspects as computers do not produce errors accidentally; machine translation errors are caused by systematic deficiencies. These systematic errors occur repetitively and with a good knowledge of the system, they can even be predicted. This provides the possibility that a post-editor gets accustomed to a specific system and detects the elements requiring a correction even faster.

Only 15% of the annotated edits were phrase-level edits, all the others were performed on word level. When correcting a single word, the influence of this correction on the whole segment structure is limited. Phrase-level corrections often require a complete restructuring of the segment to adjust the remaining elements. As this is not very often necessary during post-editing, the translation task is reduced to a correction task for the majority of the segments. The ability of finding a proper translation and phrasing the content adequately loses importance for post-editors.

4.2.6. Summary of the technical analysis

The technical analysis has provided important insights into the technical effort that is required from post-editors. More than 70% of the segments had to be modified, and this shows the importance of the post-editing phase to ensure an acceptable quality. Only 15% of the edits were performed on phrase level, the majority only concerned word level changes. Crosslinguistically the technical effort is comparable. The distribution of edit types is similar except for some specific cases. Casing plays a minor role for German than for the other languages and Italian post-editors generally performed more deeply modifying operations. Very frequent edit types are *Insertion*, *Deletion*, *Retranslation*, *Reordering*, *Agreement*, *Formatting* and *Recase*. The other four categories *Change of POS*, *Detranslation*, *Translate UNK* and *Orthography* occurred less often. Generally, the amount of required surface edits is smaller than the need for deeply modifying operations, but it is still surprisingly high.

The combination of temporal and technical measures revealed three major correlations. The more technical changes are required, the more increases the temporal effort. Deep edits and, in particular, phrase-level edits have the biggest negative impact on the temporal processing of segments.

The crosslinguistic analysis of negative translatability indicators detected four source properties that provoke high technical effort in the correction. Long segments, segments containing tags, technical instructions and detailed descriptions of procedures can often not be machine-translated properly and require substantial human correction.

The finding that post-editors perform considerably many surface corrections and only few phrase-level edits constitutes a major difference between post-editing and translation. Post-editors work less on the content and the deeper structure of the segment as this is already given by the machine translation. The major challenge for the post-editor lies in correcting and polishing an already existing text, whereas the translator autonomously creates the target text.

4.3. Cognitive Analysis

In the cognitive analysis of the data, cues revealing cognitive processing are examined. The technical analysis had shown that some types of errors require more temporal effort than others. The increased temporal effort might also indicate increased cognitive effort, but the cognitively challenging edits cannot easily be distinguished from those that simply take technically long. The cognitive analysis examines cognitive cues in the data in order to understand which of the required corrections actually pose cognitive challenges on the editor. The data provide two properties that can reveal cognitive cues: pause time and subjective feedback.

Pause time analysis. Pauses are usually analyzed as an indicator for cognitive processing. In the pause time analysis, the main focus will be on identifying what kind of changes are associated with longer pause times. These pause-expensive changes are assumed to reflect higher cognitive load for the post-editors.

Evaluation of the subjective feedback. After the experiment the translators where asked to give feedback about the post-editing task. This subjective view gives an insight into the translator's experience of the new working procedure. The comments are analyzed in relation to the previous findings in order to examine whether the translators perception of the encountered difficulties reflects the technical annotation.

4.3.1. Pause Analysis

The pause time is a subset of the overall processing time. The experimental data provides three temporal measures: duration, keyboard time and pause time. The duration captures the whole processing time the editor spends on the segment. The overall duration is then split into keyboard time and pause time. The keyboard time comprises the milliseconds the editor actually spends on typing the corrections. The pause time covers the remaining processing time, when the editor is not typing. The pause time thus comprehends reading times for the source and the raw machine translation, cognitive decision processes, consulting of references, but also possible lacks of attention when the post-editor is distracted. Subtracting the keyboard time allows to abstract from the actual technical effort of typing and can give hints on the cognitive effort. However, the pause time does not reflect the pure cognitive effort because it encompasses different activities that cannot be separated from each other on the basis of the available data. Furthermore, the measurements only provide the overall pause time, the number and duration of pauses cannot be assessed. This complicates the localization of the cognitively challenging items. The pause time just like the overall processing time is influenced by the segment length and by individual differences of the post-editors. Thus, the same normalizations are calculated. The pause time is divided by the segment length which results in the proportional pause time milliseconds per word.[13] According to this pause time measure the segments are ranked, the longest pause times being ranked highest. The ranking allows to identify the segments causing increased pause times and in com-

[13] Henceforth, the use of the term pause time refers to the proportional pause time unless indicated otherwise.

bination with the technical annotation the relation between edit types and pause times can be examined.
In order to detect crosslinguistic negative translatability indicators from the cognitive perspective, segments requiring phrasal edits are further examined. For the cognitive comparison of post-editing and translation, the previously explained pause ranking will be determined for post-edited and translated segments in order to identify a possible correlation.

4.3.2. Combination of technical and cognitive measures

The established ranking makes it possible to distinguish between segments causing long pause times and those requiring very short pause times. Comparing the technical annotations for the fifty highest and lowest ranked segments, enables to identify the relation between the edit types and the required pause time. Table 22 shows how the distribution of edit types differs for the two categories.

	First 50				Last 50			
	IT	ES	FR	DE	IT	ES	FR	DE
No edit	5	12	11	15	35	40	37	37
Insertion	23	25	15	29	2	0	0	4
Deletion	12	15	9	21	0	1	0	3
Retranslation	33	15	15	20	1	0	1	2
Change of POS	10	1	2	2	1	0	0	0
Translate UNK	8	3	0	1	1	0	0	0
Detranslation	2	2	0	2	7	3	2	3
Reordering	18	19	11	12	2	2	0	0
Recase	27	21	27	8	29	1	22	2
Agreement	9	12	12	11	3	1	4	1
Formatting	8	8	7	13	10	29	6	7
Orthography	0	0	0	0	0	0	0	0
All edits	150	121	98	119	56	37	35	22

Table 22: Comparison of technical edits for segments with high and low pause time

The results of the combination of the technical edits with the cognitive ranking are very similar to those obtained by the combination of the technical edits with the temporal ranking. Table 22 and Table 14 differ only in a couple of concrete numbers, the tendencies remain the same. Unedited segments form a major part of the fifty lowest ranked segments and occur significantly less frequent in the first fifty segments. Generally the number of performed edits is in almost all categories higher for the segments with a proportionally long pause time. As the results strongly resemble those presented in section 4.2.2, the discussion of the results is kept less detailed here. Only the numbers for phrasal edits and for the relation of deep and surface edits are given in Table 23 and 24 to highlight the general tendencies.

Phrasal edits				
	IT	ES	FR	DE
First 50	16	14.9	13.2	16
Last 50	1.8	5.4	2.9	4.5

Table 23: Phrasal edits in percentages of all edits - the cognitive perspective

Phrase-level edits. The number of phrasal edits (Table 23) is considerably higher for the fifty highest ranked segments. In the last fifty segments only one phrasal error occurs in each language. The differences in the percentage values result from the varying overall sum of edits per language. A correction of a full phrase often requires a restructuring of the segment. The post-editor needs to assure that the changed phrase still fits to the other elements structurally and content-wise. Thus, the phrasal edits very likely cause bigger cognitive effort than the correction of an individual word.

	First 50				Last 50			
	IT	ES	FR	DE	IT	ES	FR	DE
Deep edits	58.6	50.4	41.8	63	21.4	10.8	8.6	54.5
Surface edits	29.3	33.8	46.9	26.8	75	83.7	91.4	45.5
Reordering	12	15.7	11.2	10	3.5	5.4	0	0

Table 24: Deep and surface edits in percentages of all edits - the cognitive perspective

Deep and surface edits. Especially deeply modifying edits like insertions, deletions and retranslations occur extremely rarely among the last fifty segments. Table 24 highlights this relation between deep edits and surface edits. It can clearly be observed how the relation changes from the focus on deeply modifying edits in the first fifty segments to a strong prevalence of surface edits in the last 50 segments. Reordering edits behave similar to deep edits in this analysis. This dominance of surface edits in the last fifty segments also explains the lack of phrasal edits described previously. Surface edits are usually applied only on words not on phrases, thus surface edits and phrasal edits exclude each other.

In total, the results of the pause analysis in combination with the technical annotation reveal exactly the same three correlations as the results of the technical-temporal analysis in section 4.2.3.
(1) The more technical changes are required, the higher is the cognitive effort.
(2) Deeply modifying edits increase the cognitive effort more than surface edits.
(3) Phrasal edits increase the cognitive effort more than word-level edits.
In general, the fact that these three general tendencies hold for both the temporal and

the cognitive effort, is not surprising. Deep edits and, in particular, phrasal edits change the structure of the segment. The post-editor has to reconsider the source segment, find an appropriate translation and adjust the remaining elements. This is a challenging procedure which requires increased cognitive processing load and therefore takes longer to perform. The almost identical results even on the more specific level of the technical-temporal and the technical-cognitive combination, however, have not been expected. They allow two different conclusions.

1. Temporal effort and cognitive effort are similar.
When building the intersection of the highest fifty ranks across the four target languages, exactly the same five segments remain that were described in the temporal analysis in section 4.1.1. This indicates that the temporal and cognitive ranking do not only promote the same type of edits, but actually rank the same segments high. The subtraction of the keyboard time does not seem to have a major effect on the proportional processing time. However, the conclusion that temporal and cognitive effort are equal in post-editing is not very intuitive. It seems logical that cognitively challenging edits take long because they require more careful consideration. Yet, the reverse direction, edits that take long have been cognitively challenging, does not necessarily hold. Reordering tags in the machine translation, for example, requires many mouse movements in the text that are not captured by the keyboard time. This might take longer for technically less experienced users. The cognitive effort for this manipulation in contrast should be minimal, as the segments only have to be changed back into the source format and no alternative solution needs to be considered. Thus, temporal and cognitive effort do refer to different concepts, but could not be properly separated for this data. This supports conclusion 2.

2. Pause time is not a sufficient indicator for cognitive processing effort.
The measure of the pause time for this data does not properly capture the cognitive processing load. The time that has been denoted as pause time by Plitt and Masselot [53] is actually simply non-typing time. There is no separation of the initial reading time of the segment from the processing time for the segment correction. The reading time is probably even longer for post-editing than for translation as two segments (the source and the machine translation) have to be read. The pause time is shorter for post-editing than for translation[14] and the reading time is presumably longer, hence the cognitive processing load is supposed to be significantly smaller. However, O'Brien has observed [49] that participants used the arrow keys to move around in the text while actually thinking about a translation. These phases of cognitive processing are not recognized as such at all because they are considered as keyboard time. This points to another

[14] as described in the results of section 3.4

weakness of the setting. That is, the participants were not observed while performing the translation and post-editing task to guarantee a very realistic setting. The translators were hired by a translation vendor and simply delivered the final product, including the measures provided by the workbench. There is no evidence of what the translator was actually doing during the task. The translated segments certainly reveal that the participants were working on the task, but they might as well have included extra pauses while looking out of the window or answering the phone. Additional video data, the use of think-aloud protocols or eye-tracking measures could have provided information about the translator's activity, but they would have turned the task into a more artificial setting. Finally, it should be highlighted again, that the related pause analysis by O'Brien [49] failed to confirm the relation between the occurrences of pauses and editing difficulty. In her setting, O'Brien could rely on more finegrained pause data due to the use of Translog. The software allows to locate the exact occurrence of the pause during the task, so the initial reading time could be discarded. Additionally, the duration of each single pause was calculated, so it was possible to distinguish between longer and shorter pauses. However, a correlation between the duration of a pause and the correction of a difficult element could not be established. This supports the finding that pauses are not a sufficient indicator for cognitive processing.

The results of the technical-cognitive combination have not revealed any new insights into the post-editing process, that had not been covered by the technical-temporal combination. The overall processing time and the pause time seem to correlate very well. This indicates that cognitive effort and temporal effort are closely connected. In order to examine the cognitive effort independent of the temporal processing, more elaborate methods would be necessary.

4.3.3. Crosslinguistic NTIs - the cognitive perspective

Crosslinguistic negative translatability indicators determine source properties that cause increased post-editing effort. In the previous sections, segments causing a long processing time and segments requiring more than three editing operations have been examined to determine temporal and technical negative translatability indicators. Phrasal edits increased the overall temporal processing as well as the pause time. They change the segments on a deeper level, therefore segments requiring phrasal modifications from the post-editor in all four languages have been decided to categorize as cognitively challenging.

Table 25 shows that about one third of the segments provokes phrasal edits from the post-editor in each language. The intersection of the four languages returns a set of only seven segments that require phrasal edits in all four languages. This reveals, that the sets of cognitively challenging segments vary significantly across languages. Only the

	All 433 segments			
	IT	ES	FR	DE
Segments with phrasal edits	110	68	93	116

Table 25: Segments with phrasal edits

very small intersection set of the segments below has been found to be crosslinguistically cognitively challenging.

31. Alternatively, the entire expression, sqrt (Length * Width / PI), could have been assigned to the {11941}Radius{11942} dimensional constraint, defined in a user variable, or some other combination.

32. In the {1418}plan.dwg{1419} file, ensure that the Elevation and Floor Plan layout tab is active.

33. Click inside the table to select it and to display the just-in-time (JIT) Table toolbar.

34. The Non-Uniformly Scaled Blocks dialog box appears.

35. If the New Features dialog box appears, select Maybe later and OK to close it.

36. After you click the Configure button, the following dialog boxes and options are displayed:

37. Create a link to cost estimate data stored in a spreadsheet

Determining the source properties that are common to all these segments and negatively affected the translatability such that phrasal edits were necessary is challenging. Some properties are found to be shared by a couple of the segments and are discussed below, but a clear tendency like the identification of compound nouns as temporal crosslinguistic negative translatability indicators cannot be derived.

The above set of segment partially overlaps with the crosslinguistic set of technically challenging segments (segments 31, 32 and 34). This is explained by the previous finding that segments requiring phrasal edits very frequently have to be edited multiple times to adjust the remaining elements.

One observable property is again the length of the segment. The mean length (15) is shorter than those of the technically challenging segments (19), but still significantly longer than the mean of all segments (12). The seven segments listed above are all full sentences, except for segment 32 which might be a heading as the full stop at the end is missing. Shorter segments consisting of single bullet points or menu paths are not among the examples because they usually contain only series of words instead of real phrases.

More than half of the segments (31, 32, 35, 36) contain subclauses which complicate the structure of the segment. This might cause ambiguities that are wrongly resolved in the translation and have to be corrected on phrasal level. Another property causing this kind of ambiguities are reductions. Segment 31 and 37 contain reduced relative clauses ("defined", "stored") and in segment 33 and 35 occur ellipses. Both constructions make the sentence more compact and hinder a correct analysis. Controlled Language rules as introduced in 2.1 recommend to avoid these structures because they have often been identified as negative translatability indicators. This finding seems to hold crosslinguistically.

Only few segments cause phrasal corrections in all four languages and no clear property was found to be common to all of them. This indicates that the patterns provoking phrasal errors differ across languages. The relation between the source and the target languages seems to play an important role for cognitively challenging edits.

4.3.4. Post-editing vs Translation - the cognitive perspective

The employment of machine translation technology changes the pure translation task into a correction exercise. This change in the working conditions also influences the cognitive processes that are involved in the completion of the task. The previously described pause ranking has been applied to subset B to enable a comparison of the post-editing and the translation task from the cognitive perspective. The typing effort is much higher for translation than for post-editing because the target text has to be produced from scratch. Furthermore, the results from Plitt and Masselot ([53]) showed that the pause time also decreases when turning from translation to post-editing. If the pause time is taken as an indicator for cognitive processing, this reduction signifies that the cognitive load is higher for translation than for post-editing. The established ranking allows to abstract from these absolute pause time values. The ranking correlation reveals whether the same source segments cause increased pause times in both tasks. Irrespective of the discussion whether the increased pause times categorize as cognitively challenging, the ranking makes it possible to compare the post-editing and the translation task. The technical effort of the two tasks is different, this is assumed to be reflected in the source segments that cause longer processing times independent of the individual typing time. The correlation plots in Figure 5 show that the similarity between the pause ranking and the temporal ranking does not only hold for the post-editing ranks, but also for the translation ranks.

The Pearson correlation coefficients of the post-editing and translation pause ranks are slightly higher than for the temporal ranking, but the tendency remains the same (Italian: r = 0.246771, df = 73, p = 0.03282; Spanish: r = 0.4372404, df = 73, p = 8.768e-05; French: r = 0.2033570, df = 73, p = 0.08014; German: r = 0.4294737, df = 73, p = 0.0001205). The ranks correlate significantly for all languages, except for French. The

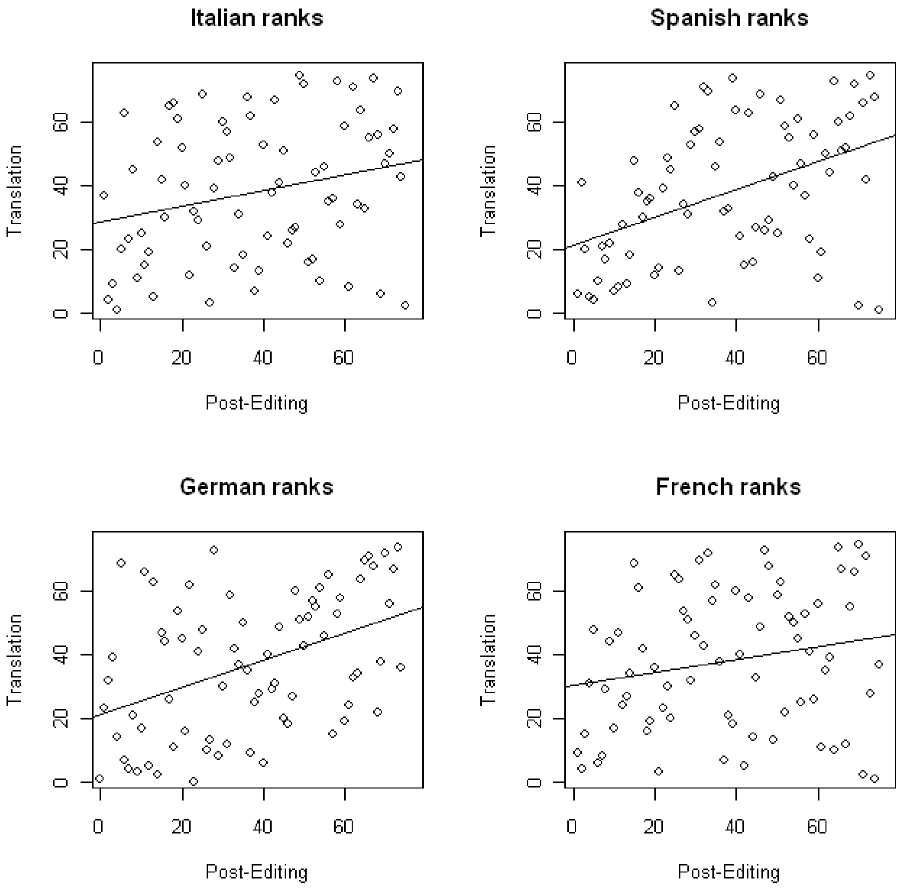


Figure 5: Correlation of pause ranks in the translation and the post-editing task

segments provoking high pause times in translation and post-editing in all four languages are exactly the same segments as those that were described to be temporally challenging in both task (section 4.1.2.).

The temporal and the cognitive ranking have been found to differ only slightly. Thus, it is not surprising that the correlation of translation and post-editing ranks, that had been analyzed for the temporal ranking is now confirmed for the pause time ranking. Post-editing and translation are different tasks, but the segments that produce increased pause times overlap.

4.3.5. Subjective Feedback

After the post-editing task the translators where asked to give feedback about this new experience. Most of them focused on technical feedback explaining the problems and challenges they encountered. These reviews had been collected in an open, non-structured way. Therefore I summarize their comments and correlate them with the findings of the technical analysis.

In general, the translators where positively surprised by the machine translation quality. Previous experiences with machine translation engines had left them skeptic about the success of the technology. Only one Spanish translator insisted on strongly preferring the more traditional tools for computer-assisted translation, because the final quality would be higher. Interestingly, the use of machine translation reached the highest productivity gain (more than 120%) for exactly this translator. Furthermore, the quality check in [53] does not confirm the negative assumptions about the post-editing quality. The quality of the post-edited segments was even assessed higher than the quality of the translated segments. The other participants rated the machine translation output in this test as "surprisingly good and encouraging" so that "only minimal corrections where required". This positive feedback supports the hypothesis that higher machine translation quality reduces the post-editing effort and increases the user's contentment. Despite this overall positive assessment many detailed remarks concerned the actual occurring errors and proposals for improvement. Though the majority of the translations was considered good, "others had to be translated from scratch". This confirms the previously stated assumption that translations requiring several phrasal edits are considered to be of inadequate quality.

Most criticism refers to surface errors which require formatting modifications. As confirmed by the technical analysis, the translators noticed the frequent wrong handling of tags. Tags are often placed in the wrong order and the formatting gets almost arbitrarily changed by the machine translation engine. The correction of these "minor" errors was experienced as "very annoying". Similar comments where reported about punctuation errors. Some of the errors were categorized as language-specific like the French convention to use a space before a colon or the use of old German comma-rules, others hold crosslinguistically like non-standard capitalization and the irregular insertion of quotation marks. Almost all translators mentioned these formatting errors that do not have a big impact on the quality of the translation, but are tedious to correct.

Another big issue was the consistency of terminology. Several technical terms had not been found in the glossary, therefore the technical accuracy was not guaranteed. Terminology handling by the machine translation system was generally judged as error-prone. The consistent use of vocabulary was not guaranteed and the hindered context look-up of the workbench made it difficult for the translators to decide on the correct translation. Several participants mentioned the lack of a concordance function usually available in translation memories that allows to compare previous translations of a term. These terminology issues are reflected in the high number of retranslations found in the technical analysis and the finding that the post-editing of short noun compounds is temporally challenging.

In addition to the terminology difficulties more linguistic problems where also observed. The translators recognized that the machine translation system did not treat headings differently and therefore often translated them as sentences including a full stop at the end. Other erratic components where idiomatic expressions and noun composites in German and French. The high number of insertions found in the technical annotation was also reflected in the comments about missing articles, words or even full phrases. The German editors mentioned the numerous necessary insertions of segment-final verbs. One Italian editor referred to the often incorrect gender agreement as an example of the minimal corrections that had to be frequently performed. Long segments, that were analyzed to be both, technically and cognitively challenging, were also named as a problematic issue by the translators.

In general, the translators' remarks are confirmed by the analysis. However, the translators focus more on surface and terminology errors while the number of performed deep errors was actually higher. This supports the assumption that the amount of tedious correction errors is higher than expected. These errors constitute the biggest difference between post-editing and translation. In human translation or revision, formatting or casing errors are of minor importance, in post-editing on the contrary, they are more relevant. The correction of surface errors usually does not require language-specific knowledge and is very easy to realize, but the identification of these errors is often not so simple. This is one reason why the translators consider them as an annoying and unnecessary burden.

The translators also proposed possible improvements for the workbench. These proposals are integrated into the ideas for practical realizations in section 5.4.

4.3.6. Summary of the cognitive analysis

For the cognitive analysis, pauses were considered as an indicator for cognitive processing. A ranking of the segments was established according to the normalized pause time in milliseconds per word. This ranking is comparable to the temporal ranking, but abstracts from the typing time.

The technical annotation was combined with the cognitive ranking in order to detect, which edit operations are more cognitively challenging than others. The results are very similar to the results found for the technical-temporal combination. Deep edits and in particular phrasal edits were identified to increase the cognitive effort just like they also increase temporal effort. These apparent similarities raised doubts about the distinctness of temporal and cognitive effort and the suitability of pause time analysis to reveal cognitive effort. Temporal and cognitive effort are considered to be closely related, but cannot be seen as equal concepts. The pause time measure as realized for this data was therefore categorized as an inadequate indicator for cognitive processing load.

For the determination of crosslinguistic negative translatability indicators, segments causing phrasal edits in all four languages were analyzed. Subclauses and reductions are assumed to negatively influence the cognitive effort crosslinguistically, but only few examples were found and the similarities were too vague to generalize a pattern. This indicates that the influence of the target language plays an important role for the cognitive effort; the source segment properties are not the only parameter affecting the translatability. Again, the individual cognitive differences of the participating post-editors might also contribute to the crosslinguistic variability.
Comparing the translation and post-editing task according to the pause time ranking confirmed the previous finding that the rank distributions correlate for the two tasks. The source segments increasing the pause times where exactly those that also increased the overall processing time.
In addition to the pause time analysis, the translators' subjective feedback was evaluated. The comments mainly focused on technical aspects of the post-editing task. The mentioned problems, such as terminology problems or tag reordering, were confirmed by the technical analysis. However, the translators strongly criticized the frequent occurrence of surface errors whereas the even higher number of deep errors was considered less problematic.

In the analysis, temporal, technical and cognitive results have been presented independently. In the next section they are discussed collectively to understand the more general tendencies and they are linked to the previous studies and the practical field.

5. General Discussion

This chapter discusses the findings of the crosslinguistic analysis. In the analysis, the results were presented separately according to the three different perspectives. In the discussion, these discrete findings from the temporal, technical and cognitive analysis are brought together and interpreted in combination in order to provide a more coherent picture of the post-editing process. First, the measures established for the analysis of the three different aspects are summarized and discussed. These measures provided the basis for the crosslinguistic analysis and the possibility to find answers for two research questions. The first one concerns the crosslinguistic perspective of the analysis. In section 5.2., the individual findings of each measure are summarized by focusing on properties that characterize the post-editing process across all four target languages. The second research focus is on the differences between post-editing and translation. The findings concerning this topic are summarized and discussed in section 5.3. The final section 5.4. discusses proposals how the crosslinguistic findings can be used for the practical improvement of the post-editing process.

5.1. Established Measures

In this thesis I have taken a novel approach by combining post-editing analyses in four different target languages. The post-editing data was analyzed under temporal, technical and cognitive aspects. In order to account for the research purposes and also respect the practical restrictions of the data, new measures were established. These measures rely on previous experiments about post-editing, but they have been adapted to fit the specific research questions addressed here.

The measure for the temporal ranking of segments is routed in the very common productivity measure data throughput in words per hour. In the performed analysis, the focus was not directly on the productivity of the post-editors, but on the properties of the segments that caused increased temporal post-editing effort. Therefore, the productivity measure is reversed to indicate the temporal complexity of the segment. This measure was applied on each segment instead of summing up the whole data. Calculating the processing duration (in milliseconds) for each segment made it possible to compare and rank the segments and assess the post-editing difficulty they pose on the editor. The duration was normalized by the number of words of each segments, so that long and short segments could be compared. The segments were then ranked according to this proportional processing time with the longest processing times being ranked highest. The source segments that were ranked high (<=50) in all languages are categorized as *temporally challenging.* This measure turned out to be particularly useful, because it managed to reconcile the subjective differences of the processing durations.

The processing time was normalized by the length of the machine translation segment.

In traditional translation measures, the segment length is computed on the basis of the source segment. A translator works on this source segment and transforms it, a post-editor in contrast only considers the source as content reference and works on the raw machine translation draft. Thus, for post-editing it is not very reasonable to use the source segment as the denominator. The same source content can be expressed in a different number of words depending on the target language. Considering only the source segment does not account for this crosslinguistic difference. The source segment also does not reflect the machine translation quality. The draft might be shorter than the source segment because the machine translation engine had omitted some of the content words. The post-editor then manually has to insert these missing words. In this analysis, the processing time was normalized only by the shorter length of the machine translation (like in example 38), resulting in a higher value. Considering this aspect, penalizing insufficient machine translations in that way is reasonable because the editing effort for the translator increases. On the other hand, machine translations that include unnecessary words are favoured by this measure. The post-editors have to delete wrong or irrelevant words from the machine translation draft (like in example 39), but the processing time is normalized by the longer length of the machine translation output, and is thus lower. The additional editing effort is not sufficiently captured in this case, but using the source segment length instead cannot generally account for the editing effort, either.

38. a) Source: Displays or hides the selected raster image.
    b) MT: Blendet das ausgewählte Rasterbild.
    c) Post-edited: Blendet das ausgewählte Rasterbild **ein oder aus**.

39. a) Source: On the Modify panel and click the mirror tool and select the two:
    b) MT: Nella **barra** pannello **comandi di** modifica **e** fare clic sullo strumento **Copia** speculare e selezionare i due:
    c) Post-edited: Nel pannello Modifica fare clic sullo strumento Specchio e selezionare i due:

In general, longer segments are slightly favoured by normalizing the processing time, independent of the exact choice of the denominator. Plitt and Masselot [53] already argued that there exists a minimal time that is spent on the translation of any segment. It comprises an orientation phase and the navigation within the text. This minimal time affects shorter segments stronger because it is apportioned to a smaller number of words. Hence, the smaller proportional processing times for longer segments properly reflect the reality.

Using words as the basic elements of a segment is a widely accepted strategy in translation studies, though there are some known weaknesses related to it. The number of words in a segment expressing the same content can vary strongly from one language to another. German, for example, can combine noun constructions into one single compound, agglutinating languages like Turkish can even express complex grammatical relations by only adding affixes to a word. Furthermore, function words such as articles or auxiliary verbs carry less importance in a segment and receive less attention during linguistic processing [56]. This difference is not captured by simply counting the words. Some preliminary test experiments for this analysis have also been conducted using the measure of milliseconds per character. The results have been comparable to those using words as the basic elements. This indicates, that the choice of the underlying unit has only a minor influence on the general tendencies. In western European languages, words are the most intuitive segmentation of a sentence, and therefore have been chosen as the basic elements for this analysis.

For the technical analysis, a completely new annotation scheme was developed. The scheme is inspired by the categories used in the LISA categorization scheme [41] that focuses on the evaluation of machine translation errors, but the motivation is different. The machine translation errors can only indicate the flaws of the output, they do not capture the post-editing strategies. The scheme developed for the current analysis annotates the actual technical corrections the post-editors perform, independent of the machine translation errors. A previous approach to the analysis of post-editing categories by Groves and Schmidtke [24] used the analysis trees of the machine translation system for an automatic comparison with the post-edited sentence. Not all machine translation systems provide analysis trees, therefore the scheme used in this analysis is designed for the manual annotation of post-edits. It consists of eleven categories, *Insertion*, *Deletion*, *Retranslation*, *Change of POS*, *Translate UNK*, *Detranslation*, *Reordering*, *Agreement*, *Recase*, *Formatting* and *Orthography.* The first six categories are perceived as deep edits, the last four categories count as surface edits. *Reordering* edits cannot clearly be included into one of these two groups. The matching edit category and the corresponding POS-tag of the changed element are stored during the annotation in order to distinguish between phrasal edits and edits on word level. Furthermore, the POS-tags can be used for more detailed evaluations of the post-edits. Source segments requiring three or more edits are categorized as *technically challenging.*
The annotation scheme was created for the analysis of the Autodesk data, but it can also be applied to other data sets. Related research settings for post-editing might vary according to the text types, the research purpose and the machine translation system in use. The extension to other text types should generally be feasible. However, it might be possible that the increased use of figurative language (for example, in fiction texts) might require a more elaborate distinction of stylistic post-edits. The paradigm of minimal post-editing that was addressed by the post-editing guidelines for this data excludes

purely stylistic edits, hence they were not considered in the annotation scheme. If the research emphasis slightly changes, the scheme might also require some alternation. The categorization generally covers all errors, but the categories might either be too wide (insertion, deletion, retranslation are all language errors) or too narrow (formatting also covers punctuation) depending on the research purpose. The annotation scheme can also properly cover the necessary edits for data from other machine translation systems, but the error distribution might be very different. As described in section 2.2., different systems produce different types of errors. A rule-based system, for example, would probably provoke a bigger amount of retranslation edits, but might reduce the formatting effort because the structure of the segment is retained.

In the current data, the annotation had to be performed by comparing the machine translation draft and the post-edited segments without any insights into the intermediate steps. This makes it impossible to detect temporary correction proposals, that had directly been deleted by the post-editor and are not visible in the final version. Only the corrections that are considered relevant by the post-editor for the modification of the machine translation output are getting annotated. This is a reasonable approach for the technical effort, but additional knowledge about the provisional attempts could have provided insights into the cognitive processes. This could be achieved by employing keylogging software or observation techniques.

The temporal ranking and the technical annotation can also be examined in combination in order to detect which edits take longer than others. Comparing the edit type distribution of the fifty slowest and the fifty fastest processed segments reveals that deep edits and phrasal edits are more time-consuming than surface edits.

The cognitive analysis relies on two indicators, pause times and subjective feedback. For the analysis the pause times where normalized by the length of the segment resulting in the proportional pause time in milliseconds per word. Similar to the temporal analysis, the segments were then ranked according to this pause time measure. The limitations related to the pause time measure have already been discussed in section 4.3. As described there, the pause time actually comprises the whole non-typing time and therefore does not only cover cognitive decision processes, but also other procedures like mouse movements. Thus, additional information about the amount and the duration of individual pauses that could help to identify cognitively challenging elements is not available.

The cognitive ranking was also combined with the technical annotation. The results of this combination revealed that the temporal and the cognitive ranking did not differ considerably. The segments with a long overall processing duration and the segments with a long pause time are almost identical. This supports the assumption that the pause time as measured in this data is not a sufficient indicator for cognitive processing. The cognitive difficulties could not be separated from the temporal difficulties.

The data had primarily been conducted for a commercial analysis, therefore cognitive factors were of minor interest in the original data collection. For a deeper cognitive analysis, more elaborate measures, for example, a combination of keylogging software with eye-tracking measures would be useful. In addition, for the detection of negative translatability indicators, the prior manipulation of the data would allow to confirm research hypotheses more directly. A more experimental setting makes it possible to include segments with and without presumable crosslinguistic negative indicators and directly compare the post-editing effort. In this case, the data was collected in a very realistic setting. Instead of explicitly examining only a set of intentionally designed segments, the whole amount of translated segments was analyzed. The segments causing increased post-editing effort were identified and the crosslinguistic negative translatability indicators could be determined backwards. This helps to abstract from preliminary expectations and initial definitions of negative translatability indicators. The combination of the cognitive ranking and the technical annotation showed, that segments which need phrasal modifications cause increased pause times. Phrasal edits change the structure of the segment on a deeper level, so the post-editor often needs to assess several alternatives. Thus, segments requiring phrasal edits are categorized as *cognitively challenging.*

In addition to the objective measures of the productivity test, personal feedback was collected from the participants. These comments are a very good indicator for the subjective experience during the post-editing task. The available feedback did not follow a specific protocol, but was collected very freely. Therefore the comments of the individual editors are hard to compare. More concrete results could probably be obtained by customizing the data collection for a specific research hypothesis and using a more restricted form (e.g. a questionnaire). Participants could be directly asked to comment on certain aspects of the post-editing process like cognitively challenging constructions.

The developed measures allowed to examine the post-editing process under temporal, technical and cognitive aspects in order to answer two research questions.
(1) Which properties of the source text increase the post-editing effort crosslinguistically?
(2) Are translation and post-editing effort negatively influenced by the same source segments?
The findings regarding these two questions will be discussed in the following sections.

5.2. Crosslinguistic similarities

In this thesis, post-editing data from four different target languages, namely, Italian, Spanish, French and German, have been analyzed. The research emphasis was on detecting properties of the post-editing process that hold crosslinguistically independent of the specific language pair. In general, the temporal and cognitive analysis revealed that the crosslinguistic similarities were smaller than expected. All segments were ranked according to their temporal and cognitive complexity. The distribution of these ranks varies significantly across languages. The intersection of those segments being ranked high (<=50) in all languages is relatively small, it consists of only five segments for both rankings. Previous studies by O'Brien [49] and Vasconcellos [68] assessed the post-editing effort by categorizing the difficulty of the source segment. However, this does not seem to be the only influencing factor in the current data. The crosslinguistic differences suggest, that the target language and the relation between the source and the target should also be considered in order to assess the temporal and cognitive effort. Nevertheless, the individual differences of the participants might also have contributed to the crosslinguistic variability. In future studies, a bigger sample of participants from each language can help to balance these individual characteristics and confirm whether the observation of crosslinguistic differences persists.

The technical annotation, in contrast, revealed a relatively comparable distribution of edit types across languages. *Insertion*, *Deletion*, *Retranslation* and *Recase* are very frequent edits in all languages and *Orthography*, *Detranslation* and *Translate UNK* occur only rarely. Some language-specific exceptions were detected such as the smaller amount of required casing operations in German and the higher number of vocabulary gaps in Italian, but in general the technical effort was similar for all languages. This indicates that the technical effort is mainly determined by the source language and the machine translation system. Most systems have predictable problems with certain source language phenomena that require specific edits to be corrected. The temporal and cognitive complexity of this correction in contrast might differ depending on the target language.

The combination of the technical annotation and the temporal ranking revealed three general tendencies that hold across languages. A higher number of necessary edit types generally increases the processing time. Deep edits and in particular phrasal edits have the biggest influence on the temporal effort. The combination of the cognitive ranking and the technical annotation confirmed these tendencies also for the cognitive effort.

In the previous section, the criteria for *temporally*, *technically* and *cognitively challenging* segments have been summarized. The source segments that met these criteria in all four languages were examined in order to detect common source properties that caused the increased post-editing effort. Under the temporal aspect, short noun compounds were clearly identified as crosslinguistic negative translatability indicators that increase the processing time. These expressions are usually domain-specific terms that have to be translated cautiously under consideration from glossaries and terminology references.

On the technical level, an intersection of 32 segments caused three or more edits in all four languages. From these segments, four source segment properties could be derived that increase the technical effort. Long segments, segments containing tags, segments containing you-references or imperatives, as well as segments with technical descriptions, often have ambiguous structures that are not translated properly by the machine translation engine. Some of these properties have already been considered as negative translatability indicators in previous research, but only for specific language pairs (e.g. [6]), [49]). This analysis has confirmed that these properties are not only related to a bad translatability, as a consequence they actually increased the technical post-editing effort.

On the cognitive level, segments causing phrasal edits are categorized as challenging. Only seven segments matched this criterion in all languages. The exclusivity of this criterion is surprising, as about one hundred phrasal edits occur in each language. The seven remaining segments are very diverse, hence it is difficult to derive a property that is generalizable to all of them. Long segments containing subclauses or reductions are likely to cause phrasal post-edits in all languages, but it cannot be clearly derived. Subclauses and reductions are known to cause ambiguities that are difficult to resolve ([1]), so the assumption that these properties are responsible for the increased cognitive effort seems reasonable.

The few detected crosslinguistic similarities show that results derived for a specific language pair do not necessarily hold for other languages. The data allows to individually determine the challenging segments for each language. In this thesis only the crosslinguistically generalizable features were of interest. For further research, it would also be interesting to analyze the specific features of certain language pairs. For example, if only the pair English-German had been examined, casing (67 *Recase* edits) would probably not have been identified as a relevant problem. For French as the target language, in contrast, this error causes almost one-fourth of all edits (200 *Recase* edits). The present data always used English as the source language and the target languages were all Western European. English, German, French, Italian and Spanish are the languages that are most commonly used in the European Union [13] and thus frequently need to be translated. However, the presented approach should be easily expandable to other language pairs. Translation pairs including Asian or African languages might reveal very different post-editing problems as these languages differ structurally from European languages.

5.3. Post-editing vs Translation

Both post-editing and translation transform a source segment into a suitable segment in the target language, but the contribution of the human performance on the task is different. The second aspect of the crosslinguistic analysis addressed these differences by comparing the temporal, technical and cognitive effort for the two tasks. It has been shown, that the two processes indeed differ technically, but on the temporal and cognitive level, the processing effort correlates for the two tasks. The source segments increasing the temporal and cognitive processing overlap for all languages, except for French. Previous experimental comparisons of human translation and post-editing focused on the final results. Guerberof [25] and Flournoy and Duran [20] compared the translator productivity for each task and Fiederer and O'Brien [19] analyzed quality aspects of the translation and the post-editing product. The influence of the translatability of source segments on the post-editing task has been examined by O'Brien [47] and Vasconcellos [68], but they did not compare their findings to the corresponding translation effort for the same segments. In this analysis, the two tasks have been compared under temporal, technical and cognitive aspects.

On the temporal level, it is important to note that translation generally takes longer than post-editing. The mean typing time and the pause time are significantly higher in the translation than in the post-editing task. The processing time was therefore normalized and the segments of subset B were ranked according to this normalized time to enable the comparison of the two tasks. The ranking distribution revealed that the temporally challenging segments correlated for the two tasks for German, Italian and Spanish, but not for French. The two processes are very different on the surface, but focusing on the relative effort shows that the segments increasing the processing effort overlap. Short segments containing noun compounds caused processing difficulties for post-editors as well as for translators. This indicates that the detected terminology problems are not only associated with the employment of machine translation, they are a general problem for translation tasks. Structurally challenging constructions in contrast only caused higher effort for translators, not for post-editors.

The working conditions of the two activities differ especially on the technical level. Post-editing is a correction task of an already existing text in order to transform the draft into the appropriate translation of the source text. For the post-editor, the correction of surface edits constitutes a substantial part of the task. Translators work on a deeper level of language, as they only receive the source text and create the target text themselves. In this process, corrections mainly occur only during revision and usually concern accidental omissions or unconscious errors instead of systematic flaws.

These technical differences also have influence on the cognitive challenges related to the task. Translators need to be proficient in both languages and actively have to create the translation. Post-editing resembles more a correction task, thus passive knowledge of the two involved languages might be sufficient. The challenge for the post-editor is more the exhaustive detection of errors instead of the creative mediation of language. Though the requirements for the two tasks differ, the source segments increasing the pause times correlated even stronger than for the temporal effort. As the pause times have not been a sufficient indicator for cognitive effort in this analysis, further measures should be applied before generalizing these findings. It would be interesting to compare the segments causing increased cognitive effort in more detail in order to understand how the challenges in the translation and the post-editing task differ. Understanding these differences would help to better prepare translators for the post-editing task.

Post-editing is usually perceived as just another operation field for translators. It is important to understand that the two processes are related, but the challenges for the post-editing activity differ and take the translator to get used to. Supporting translators to adapt to this new working condition will help to increase the acceptance of machine translation technology.

5.4. Practical implications

One benefit from the findings of the analysis is a better understanding of the problems post-editors face during post-editing than the previous literature has provided. Therefore, the current results can be used to improve the post-editing working process. Post-editors have to correct particularly many surface errors like casing or formatting which are neither temporally nor cognitively challenging. However, technically, they play a major role and the translators strongly emphasized them in the subjective feedback. In practice, the surface errors are those that are easier to automate and thus do not necessarily require manual correction. They occur systematically and can therefore be predicted. The surface errors require almost no linguistic knowledge, but can be minimized automatically. The deep errors differ crosslinguistically because they occur on a deeper language level, they can only be reduced by improving the machine translation quality, which is a slowly advancing process. Alternatively, facilitating the post-editing process by providing technological means and by automatizing the correction of surface errors is a simple, but still promising approach to improve the post-editing working process. Combining the findings of the analysis and the translators' proposals results in possible improvements that are relatively easy to realize. They are described in section 5.4.1.

The results also showed that post-editors spend a lot of time on solving terminology errors. The use of terminology references is facilitated in a computer-assisted translation environment. For the future it would also be useful to enable the machine translation system to learn from the post-editing corrections to minimize the post-editing effort. A first step towards this goal can be to add the human translation of words unknown to the machine translation system to the phrase table. This possibility is described in section 5.4.2.

The results have shown that translators and post-editors face different challenges. Thus, it is important to prepare post-editors for these changes in the working process and sensitize them to the new task. This can be realized in a post-editor training which is proposed in section 5.4.3.

5.4.1. Improving the workbench

The workbench introduced in section 3.2 was designed especially for testing purposes. Therefore, it was kept very simple and did not provide even basic support features like a spell-checker and a "Search & Replace" function. The participants strongly criticized the lack of these features. In production, the Moses machine translation decoder is integrated into a computer-assisted translation environment (Plitt 19.03.2010, personal communication). The machine translation matches are inserted into the target draft the same way as translation memory matches. So, the translators are used to the procedure, they only have to be aware of the different technological source. Even though most computer-assisted translation environments already come along with a possible plug-in for machine translation, it is still rarely used. In contrast to the test workbench, this working environment fulfills most of the translators' wishes for additional features. More context can be displayed and a concordance function keeps track of the correct use of terminology. Standard text edit functions like "Copy & Paste", insertion, search, spell and grammar checking, etc., are also incorporated. There exists the possibility to also display translation memory matches. Several approaches of how to combine machine translation and translation memory matches are discussed in section 2.2. The simplest solution to this combination is to only work with translation memory matches above a certain threshold and use machine translation otherwise. It is recommended to set this threshold very high as Guerberof [25] found that the modification of 80-90% translation memory matches already leads to a loss in quality and productivity compared to the post-editing of machine translations. In addition to the basic, already included features, some more operations could be introduced.

Automatic check of formatting and casing. Formatting and *Casing* are among the most frequent editing operations. In most cases, the editors changed the raw machine translation back to the source format. Even though the formatting of a segment should generally be maintained during the translation process, many machine translation systems have not yet paid attention to this issue. Integrating the preservation of formatting properties requires a change in the decoding process. The Moses system that was used for the current data was developed for research purposes, which might be the reason why formatting was not yet considered an important issue. The results showed that in practice, the lack of formatting preservation burdens the post-editors with immense additional correction effort. A possible interim solution would be to check the formatting and casing properties after the actual machine translation process. The Moses decoder can output additional alignment information like in the example below taken from the Moses tutorial [36].[15]

[15] In the original example, casing was ignored, this is corrected here to avoid confusion.

Segment 40 is the source sentence, segment 41 is the machine translation and segment 42 is the enriched output.

40. Das ist ein kleines Haus

41. This is a small house

42. This is |0-1| a |2-2| small |3-3| house |4-4|

The numbers indicate how the sentence had been segmented into phrases and to which source phrase the translation corresponds. The phrase "This is" for example is a translation of the phrase spanning from word 0 (= "Das") to word 1 (="ist") in the original segment. This allows to directly adjust the formatting of "This is" to match the formatting of "Das ist". The alignment information is particularly important, if the word order in the target segment differs from the word order in the source segment. This format comparison of the phrases could be done automatically. For casing, the comparison would probably be even easier as only one binary feature (uppercase/lowercase) has to be checked for each phrase.[16] Once this comparison is established, it might also be possible to account for the correct ordering of tags. Automating the formatting, casing and tag reordering errors would reduce the editing operations by about one third.

Confidence Scores. In the subjective feedback, the translators remarked that the matching level of the translated segment is not given as it is for fuzzy matches. This insecurity about the quality of the translated segment is also reflected in the occasionally very long processing time for segments that do not require a modification at all. Post-editors probably spend a considerably long time on deciding whether a translation is correct or not. If the machine translation comes along with a confidence score, this decision time could probably be reduced. Confidence scores indicate whether the system's translation is estimated to be a good translation. The post-editor would spend less time on searching for errors on a segment with a high confidence score. The problem is that the "matching level" cannot be calculated for machine translations as easily as for translation memory matches. Translation memory scores indicate which portion of the translated segment has already been translated before. Machine translation engines are designed to explicitly combine good phrasal translations to new translations. This ability to create translations for unseen sentences is exactly the strength of machine translation systems compared to translation memories. They adapt faster to new domains, because only the terminology needs to be learned. A good confidence score would have to consider these strengths and combine the values of the phrasal translations into one overall confidence score. Different machine translation systems have already implemented a measure that assesses the

[16] This case comparison of source and target should not be applied to case-sensitive languages like German. It is only reasonable if source and target language follow similar casing conventions.

probability of their translations (e.g. probabilistic synchronous tree-substitution grammar [70]). However, this measures cannot be generalized to other systems as they depend on the individual architecture. Blatz et al. [7] discuss machine learning approaches to confidence estimations, but they have not yet been thoroughly tested with applications. Most of the approaches rely on the use of reference translations which are not available if the actual task is translation. Further research is necessary to discover a way to accommodate the strengths of different machine translation approaches into one representative confidence score. This would not only facilitate the post-editing task, but also enable a better comparison of different machine translation systems without the need of reference translations.

Colour-Coding. As indicated before, translation memory matches and machine translations have to be treated differently. Therefore, it is important to mark them with different colours, if they occur in combination. Additionally, aspects of the translation that need specific attention from the post-editor could be highlighted with colours. Punctuation, for example, required thorough revision in the analyzed data sample. It is also possible to coordinate the machine translation output with a terminology database and highlight the source words that are contained in the terminology. The translators had mentioned in the subjective feedback that the correct terminology was not always met by the machine translation. The high number of retranslations in the technical annotation and the high temporal effort spend on complex noun compounds confirms this assumption. The colour-coding of terminology entries can remind the post-editor to properly check the translation of the corresponding word. The crosslinguistic analysis revealed that some post-editing characteristics are specific for certain languages. These language-specific characteristics could also be colour-coded when working on the corresponding language pair. In a German translation for example, the source verb could be highlighted to remind the post-editor to check whether the correspondent target verb is present in the translation.

Preprocessing. One goal of the analysis was to find crosslinguistically challenging elements that increase the post-editing effort. These elements could have received a special treatment to improve the machine translation quality. However, the results showed, that the nature of the "difficult" elements depends on the perspective of the analysis, the measure of the effort, the target language and probably even the specific post-editor. Importantly, it was found that only one source text property has been clearly identified to cause problems crosslinguistically. Proper nouns and menu items often lack context information and are therefore difficult to evaluate for the post-editors. These elements are often fixed translations. One possible preprocessing step would be to assure that the translations for menu items and proper nouns are determined in advance and are contained in the terminology. The terminology colour highlighting mentioned before can then remind the post-editor to properly check the word.

5.4.2. Automatic Modification of phrase table

Post-editing corrections generate new high quality parallel data that could be fed back to the machine translation engine. The system might even be able to learn from these corrections and improve the translation quality. This requires elaborate machine learning algorithms that work directly in the architecture of the machine translation system. However, a relatively simple interim solution focusing only on word pairs could already increase the vocabulary coverage of the system. Moses transfers source words that it has never seen before directly to the output assuming that they are proper nouns. The post-editor has to manually translate the unknown words, these corrections were categorized as *Translate UNK* in the technical annotation. These cases were relatively rare in the current analysis, because a large training sample was used. If less training material from the same domain is available, these UNK-words increase rapidly. Moses can return a list of the unknown words after the translation process. This list makes it possible to mark the UNK-words in the source and the machine translation draft. Once the post-editor corrects them, he keeps the marking around for the new translation. Thus, new term pairs can be automatically collected by combining the marked source word (the original UNK-term) and the marked target word (the correct translation of the UNK term). In example 43 the unknown word is the conjugated verb "Relaxes", it is transferred directly to the machine translation output (segment 44), only the casing is removed. The post-editor corrects the unknown word into the proper Spanish translation "Libera" (segment 45), but keeps the unknown marking (bold in this example) around the correction. The new term pair "Relaxes-Libera" can then easily be collected.

43. Source: **Relaxes** constraints

44. Raw MT: **relaxes** restricciones

45. Post-edited: **Libera** restricciones

Adding new term pairs to the phrase table usually requires a complex recalculation of the translation probabilities. In the case of these UNK-words fortunately it is assured, that there does not yet exist a phrase table entry for the corresponding source word. Thus, the pair can be added to the phrase table and the translation probability can safely be set to 1.[17]

46. Relaxes ||| Libera ||| 1.0

[17] In the phrase table other parameters might also have to be set, this depends on the specific Moses configuration. However, for an individual term pair this setting is rather simple, the parameters can take the default values.

It is not recommended to also add the reverse term pair, because it is not guaranteed that there does not yet exist an entry for "Libera". The inclusion of translations for previously unknown words is an easy example of how the reuse of post-editing corrections can improve subsequent machine translations. Once this method is established, it can also help to improve the terminology problems that caused high temporal effort. Once the translation of a complex noun compound is corrected, further occurrences of this term could be dynamically corrected. In order to learn even more complex structures from the post-editing data, advanced machine learning algorithms would be necessary. Adding phrase pairs for already existing source terms requires a more complex restructuring of the phrase table. It would be important to guarantee that the human correction is ranked higher than the machine proposals. Yet, the translation of a term might change depending on the context it occurs in. Thus, the human correction should not completely overwrite the alternative phrase pairs.

5.4.3. Training for Post-Editors

The results confirmed the assumption that post-editing is very different to translation. A translator's education does usually not yet include the acquisition of post-editing strategies. Thus, it makes sense to provide a post-editing training so that the post-editors can be sensitized to the characteristic properties of the task. If a post-editor knows more about the features specific to the machine translation system and the characteristics of a certain language pair, he will be able to detect the errors of the machine translation output considerably faster. Machine translation errors are produced systematically, and with a good knowledge of the system they can sometimes even be predicted. Post-editor training also presents the possibility to discuss the realization of the post-editing guidelines and the quality expectations of the output. The post-editing strategies might vary according to the purpose of the final translation (as described in section 2.3.). Explaining the expectations and guidelines to the post-editors in detail would increase the overall quality of the work and guarantee consistent translations. O'Brien [50] and Vasconcellos [68] already described the importance of post-editor training. O'Brien proposed different topics for the course content including theoretical and technical skills relevant to the post-editing technology. She also discussed whether post-editing needs to be performed by a translator. She described that the required post-editing skills only partially overlap with translator skills, in some aspects they are even contradictory. Fluency in source and target language and a positive attitude towards language technology can be a sufficient qualification for a post-editor, an extensive translator education might not be necessary. Nevertheless, the demand for linguistic skills in at least two languages makes translators the best candidates for a successful post-editor training.

6. Conclusions

Post-editing is a relatively new translation process in which translators correct machine translation output to guarantee a correct and stylistically adequate outcome. The process is structurally different from standard translation.

In this thesis the crosslinguistically generalizable properties of the post-editing process have been examined under temporal, technical and cognitive aspects using suitable, in part newly established methods.

It has been found that the technical effort is crosslinguistically comparable. The distribution of editing operations is similar across languages, apart from some exceptions. Long segments, segments containing tags, technical instructions and detailed descriptions cause increased technical effort in all languages. The temporal and the cognitive effort differ more strongly across languages. The target language seems to have an important influence on the temporal and cognitive processing time required for post-editing a segment. Short segments consisting of only one complex noun compound were the only property that was found to be temporally challenging in all languages. This property also caused increased temporal effort in the translation task.

Post-editing and translation are technically very different processes. However, the segments causing increased temporal and cognitive effort correlated for all languages except for French. This shows that the activities are related, but are realized differently. Thus, the challenges for post-editors are not necessarily the same as for translators. Post-editors should be sensitized to these differences in order to facilitate the task.

Surface errors had to be corrected relatively often in the analyzed data set. These corrections have been categorized as particularly cumbersome by the translators. As these errors usually occur systematically and predictably, it is possible to automate their correction in order to facilitate the post-editing process and increase the acceptance of machine translation technology.

References

[1] T. Aikawa, L. Schwartz, R. King, M. Corston-Oliver, and C. Lozano. Impact of controlled language on translation quality and post-editing in a statistical machine translation environment. In *Proc. MT Summit XI*, pages 10–14, 2007.

[2] J. Allen. Post-editing. *Computers and Translation: a Translator's Guide*, pages 297–317, 2003.

[3] J.A. Alonso and G. Thurmair. The Comprendium Translator System. In *Proceedings of the Ninth Machine Translation Summit, New Orleans, USA*, 2003.

[4] F. Alves. *Triangulating translation: perspectives in process oriented research.* John Benjamins Pub. Co., Amsterdam, 2003.

[5] S. Banerjee and A. Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. *Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, page 65, 2005.

[6] A. Bernth and C. Gdaniec. MTranslatability. *Machine Translation*, 16(3):175–218, 2001.

[7] J. Blatz, E. Fitzgerald, G. Foster, S. Gandrabur, C. Goutte, A. Kulesza, A. Sanchis, and N. Ueffing. Confidence estimation for machine translation. In *Proceedings of the 20th international conference on Computational Linguistics*, page 315. Association for Computational Linguistics, 2004.

[8] L. Bowker. *Computer-aided translation technology: a practical introduction.* Univ of Ottawa Pr, 2002.

[9] L. Bowker. Productivity vs Quality? A pilot study on the impact of translation memory systems. *Localisation Focus*, 4(1):13–20, 2005.

[10] C. Bruckner and M. Plitt. Evaluating the operational benefit of using machine translation output as translation memory input. In *MT Summit VIII*, pages 18–22. Citeseer, 2001.

[11] M. Carl, A.L. Jakobsen, and K.T.H. Jensen. Modelling Human Translator Behaviour with User-Activity Data. In *Proc. 12th EAMT Conference*, 2008.

[12] Y. Chen, M. Jellinghaus, A. Eisele, Y. Zhang, S. Hunsicker, S. Theison, C. Federmann, and H. Uszkoreit. Combining multi-engine translations with Moses. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 42–46. Association for Computational Linguistics, 2009.

[13] European Commission. Europeans and their Languages. *Eurobarometer Special*, 243, 2006.

[14] J. DeCamp. What is Missing in User-Centric MT? In *Proc. MT Summit XII*, 2009.

[15] J. Doyon, C. Doran, C.D. Means, and D. Parr. Automated machine translation improvement through post-editing techniques: Analyst and translator experiments. In *Proceedings of the Eighth Conference of the Association for Machine Translation in the Americas*, pages 346–353, October 2008.

[16] L. Dugast, J. Senellart, and P. Koehn. Statistical post-editing on SYSTRAN's rule-based translation system. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 220–223. Association for Computational Linguistics, 2007.

[17] T. Ehara. Rule Based Machine Translation Combined with Statistical Post Editor for Japanese to English Patent Translation. In *Proceedings of the Eleventh Machine Translation Summit Workshop on Patent Translation*, pages 13–18, 2007.

[18] A. Eisele, C. Federmann, H. Uszkoreit, H. Saint-Amand, M. Kay, M. Jellinghaus, S. Hunsicker, T. Herrmann, and Y. Chen. Hybrid machine translation architectures within and beyond the EuroMatrix project. In *Proceedings of the 12th annual conference of the European Association for Machine Translation*, pages 27–34, 2008.

[19] R. Fiederer and S. O'Brien. Quality and Machine Translation: A realistic objective? *The Journal of Specialised Translation*, 11:52–74, 2009.

[20] R. Flournoy and C. Duran. Machine Translation and Document Localization at Adobe: From Pilote to Production. In *Proceedings of MT Summit XII*, 2009.

[21] I. García. Research on translation tools. *Translation Research Projects*, 2:18–27, 2009.

[22] R. Green. The MT errors which cause most trouble to posteditors. *Lawson (1986)*, pages 101–104, 1982.

[23] L.A. Griffiths. Translation of Idiomatic Expressions. 2002.

[24] D. Groves, C. Wicklow, and D. Schmidtke. Identification and Analysis of Post-Editing Patterns for MT. In *Proc. MT Summit XII*, 2009.

[25] A. Guerberof. Productivity and quality in MT post-editing. In *Proc. MT Summit XII*, 2009.

[26] R. Guzmán. Manual MT Post-editing. *Translation Journal*, 11, 2007.

[27] G. Hansen. Zeit und Qualität im Übersetzungsprozess. *Copenhagen studies in language*, 27:29–54, 2002.

[28] W.O. Huijsen. Controlled language–an introduction. In *Proceedings of CLAW*, pages 1–15, 1998.

[29] M.J. Hunt. Figures of merit for assessing connected-word recognisers. *Speech Communication*, 9(4):329–336, 1990.

[30] J. Hutchins. Current commercial machine translation systems and computer-based translation tools: system types and their uses. *International Journal of Translation*, 17(1-2):5–38, 2005.

[31] P. Isabelle, C. Goutte, and M. Simard. Domain adaptation of MT systems through automatic post-editing. *Proc. of MTSummit XI*, pages 255–261, 2007.

[32] A.L. Jakobsen. Logging time delay in translation. *LSP Texts and the Process of Translation. Copenhagen Working Papers in LSP*, pages 73–101, 1998.

[33] A.L. Jakobsen. Orientation, segmentation, and revision in translation. *Empirical Translation Studies: process and product. Copenhagen Studies in Language Series*, 27:191–204, 2002.

[34] A.L. Jakobsen and L. Schou. Translog documentation. *Probing the Process in Translation: Methods and Results*, pages 151–186, 1999.

[35] R.M. Kaplan and J. Bresnan. Lexical functional grammar. *The mental representation of grammatical relations*, pages 173–281, 1982.

[36] P. Koehn. *Moses: Statistical Machine Translation System, User Manual and Code Guide.*

[37] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. Moses: Open source toolkit for statistical machine translation. In *Annual Meeting of the Association for Computational Linguistics*, Prague, Czech Republic, June 2007.

[38] H.P. Krings and G.S. Koby. *Repairing texts: empirical investigations of machine translation post-editing processes.* Kent State Univ Pr, 2001.

[39] B. Lavorel. Experience in English-French post-editing. *Lawson (1986)*, pages 105–109, 1982.

[40] V.I. Levenshtein. Binary codes capable of correcting spurious insertions and deletions of ones. *Problems of Information Transmission*, 1(1):8–17, 1965.

[41] LISA - Homepage of the Localisation Industry Standards Association. QA-model. http://www.lisa.org/LISA-QA-Model-3-1.124.0.html, accessed June 2010.

[42] N. Loorbach, J. Karreman, and M. Steehouder. The Effects of Adding Motivational Elements to User Instructions. *IEEE International Professional Communication Conference*, 2007.

[43] A. Lopez. Statistical machine translation. *ACM Computing Surveys (CSUR)*, 40(3):8, 2008.

[44] M.P. Macdonald. Can a Manual entertain? *Intercom*, 48(6):14–17, 2001.

[45] T. Mitamura. Controlled language for multilingual machine translation. In *Proceedings of machine translation summit VII*, pages 13–17. Citeseer, 1999.

[46] S. Neumann, A. Pagano, F. Alves, P. Pyykkönen, and I. da Silva. Targeting (de)metaphorization: Process-based insights. In *European Systemic Functional Linguistics Conference and Workshop*, Koper, Slovenia, July 9-12, 2010.

[47] S. O'Brien. Methodologies for measuring the correlations between post-editing effort and machine translatability. *Machine Translation*, 19(1):37–58, 2005.

[48] S. O'Brien. Eye-tracking and translation memory matches. *Perspectives: Studies in Translatology*, 14(3):185–205, 2006.

[49] S. O'Brien. Pauses as Indicators of Cognitive Effort in Post-editing Machine Translation Output. *Across Languages and Cultures*, 7(1):1–21, 2006.

[50] S. O'Brien and D. Glasnevin. Teaching Post-Editing: A Proposal for Course Content. In *6th EAMT Workshop Teaching Machine Translation*, pages 99–106, 2002.

[51] F.J. Och and H. Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003.

[52] K. Papineni, S. Roukos, T. Ward, and W.J. Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.

[53] M. Plitt and F. Masselot. A Productivity Test of Statistical Machine Translation Post-Editing in a Typical Localisation Context. *Prague Bulletin of Mathematical Linguistics*, 93(-1):7–16, 2010.

[54] C.J. Pollard and I.A. Sag. *Head-driven phrase structure grammar.* University of Chicago Press, 1994.

[55] M. Popovic and H. Ney. POS-based word reorderings for statistical machine translation. In *Proc. of the Fifth Int. Conf. on Language Resources and Evaluation (LREC)*, 2006.

[56] K. Rayner. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, 124(3):372–422, 1998.

[57] U. Reinke. Überlegungen zu einer engeren Verzahnung von Terminologiedatenbanken, Translation Memories und Textkorpora. *LDV-Forum*, 1999.

[58] U. Reuther. Two in one – Can it work? Readability and translatability by means of controlled language. *Controlled language translation, EAMT-CLAW*, 3:15–17, 2003.

[59] S. Sharmin, O. Špakov, K.J. Räihä, and A. Lykke. Where and for how long do translators look at the screen while translating? *Copenhagen Studies in Language*, 36:31–51, 2008.

[60] M. Simard and P. Isabelle. Phrase-based Machine Translation in a Computer-assisted Translation Environment. In *Proceedings of MT Summit XII*, 2009.

[61] M. Simard, N. Ueffing, P. Isabelle, and R. Kuhn. Rule-based translation with statistical phrase-based post-editing. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 203–206. Association for Computational Linguistics, 2007.

[62] S. Theison. Optimizing rule-based machine translation output with the help of statistical methods. Master's thesis, Saarland University, 2007.

[63] G. Thurmair. Comparing rule-based and statistical MT output. In *Workshop on the amazing utility of parallel and comparable corpora, LREC.* Citeseer, 2004.

[64] G. Thurmair. Hybrid architectures for machine translation systems. *Language Resources and Evaluation*, 39(1):91–108, 2005.

[65] C. Tillmann and H. Ney. Word reordering and a dynamic programming beam search algorithm for statistical machine translation. *Computational Linguistics*, 29(1):97–133, 2003.

[66] S. Tirkkonen-Condit and R. Jääskeläinen. *Tapping and mapping the processes of translation and interpreting: outlooks on empirical research.* John Benjamins Publishing Co, 2000.

[67] N. Underwood and B. Jongejan. Translatability checker: a tool to help decide whether to use MT. In *MT Summit VIII*, pages 18–22, 2001.

[68] M. Vasconcellos. Post-editing on-screen: machine translation from Spanish into English. *Translating and the Computer*, 8:133–146, 1986.

[69] M. Vasconcellos. A comparison of MT post-editing and traditional revision. In *Proceedings of the 28th annual conference of the American translators association (pp. 409Á416). Medford, NJ: Lerned Information*, 1987.

[70] M. Zhang, H. Jiang, A. Aw, H. Li, C.L. Tan, and S. Li. A tree sequence alignment-based tree-to-tree translation model. In *Proceedings of ACL*. Citeseer, 2008.

[71] A. Zollmann, A. Venugopal, F. Och, and J. Ponte. A systematic comparison of phrase-based, hierarchical and syntax-augmented statistical MT. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 1145–1152. Association for Computational Linguistics, 2008.

Appendices

A. Crosslinguistic intersection of temporally challenging segments

1. Minimum command
2. EXPORTPAGESETUP
3. License timeoutall
4. Polyline subobjects
5. License Borrowing Content Reference

B. Crosslinguistic intersection of technically challenging segments

1. Alternatively, the entire expression, sqrt (Length * Width / PI), could have been assigned to the {11941} Radius {11942} dimensional constraint, defined in a user variable, or some other combination.

2. In the {1418} plan.dwg {1419} file, ensure that the Elevation and Floor Plan layout tab is active.

3. {3229} After revising the original content in the publishing software, the designer republishes an updated DWF file, a new sheet set, or model, to begin the digital design workflow again.

4. On the ribbon, click Home tab {113} Modeling panel {114} Solid Creation drop-down {115} Revolve.

5. A full list of available functions is documented in the {13009} AutoCAD User's Guide {13010} Help topic, {13011} Constrain a Design with Formulas and Equations {13012}.

6. The intent of this tutorial is not to teach you how to draw lines and work with blocks, but rather to introduce the new AutoCAD 2009 interface.

7. Multiple filter values for the same filter category can be specified on the same line separated by a space or on separate lines.

8. The Sum column now displays the {4234} icon denoting a formula column.

9. The Non-Uniformly Scaled Blocks dialog box appears.

10. The radius of the circle changes to the radius set by you.

11. Click once inside the cell, and enter {4224} Pipes {4225} as the Display Name.

12. You have helped Viola to create annotative multileaders.

13. Some of the drawings that you work with will contain design requirements enforced within the drawing itself through the use of constraints.

14. Click {10453} Insert tab {10454} Linking & Extraction panel {10455} Extract Data {10456}.

15. When you do not need an underlay in the current drawing session, you can improve performance by temporarily unloading it.

16. They mark control locations on an object and are powerful editing tools.

17. Click {10465} Insert tab {10466} Linking & Extraction panel {10467} Extract Data {10468}.

18. {1433} Audience: {1434} AutoCAD users who want to work with the new AutoCAD 2009 interface

19. If your tags disappear, select the hatches and use the Send to Back tool on the Modify menu or right click, Draw Order>Send to Back.

20. If a report is filtered on user==User1 and user==User2, the resulting report contains usage of features by either User1 or User2.

21. Opens the {9883} Export to DWF/PDF Options palette {9884} where you can change DWF file settings such as file location, password-protection, and layer information.

22. {10952} Annotate tab {10953} Dimensions panel {10954} Baseline {10955}

23. Select to automatically scale the selection to fit the area on the paper size specified earlier.

24. Since a drawing file is normally compressed, the final size of a saved drawing file on disk will vary based on the size and number of objects in a drawing.

25. {8991} Home tab {8992} Layers panel {8993} Layer State drop-down {8994} Manage Layer States. {8995}

26. For example, the width in the illustration is constrained by the diameter constraint, {11816} dia1 {11817}, and the linear constraint, {11818} d1 {11819}.

27. At the prompt, Specify rotation angle of text <0>, press Enter to accept the default text rotation angle.

28. Toolbars organize commands and controls on small dockable windows.

29. While a workspace primarily provides toolbars, menus, ribbon tabs, and palettes, you can also use a workspace to control user interface elements for the application and drawing windows.

30. Click the rectangle tool to add a 3.5ẍ 1.25b̈earing plate to the right end of the top chord.

31. Note the location where the Network License Manager is installed and then uninstall the Network License Manager by entering the standard Linux commands, for example, {2896} rm {2897}.

32. In order to choose a language for an individual product, you first must click the Select Language for Individual Products check box, then select the language from the drop-down list.

C. Crosslinguistic intersection of cognitively challenging segments

1. Alternatively, the entire expression, sqrt (Length * Width / PI), could have been assigned to the {11941} Radius {11942} dimensional constraint, defined in a user variable, or some other combination.
2. In the {1418} plan.dwg {1419} file, ensure that the Elevation and Floor Plan layout tab is active.
3. Click inside the table to select it and to display the just-in-time (JIT) Table toolbar.
4. The Non-Uniformly Scaled Blocks dialog box appears.
5. If the New Features dialog box appears, select Maybe later and OK to close it.
6. After you click the Configure button, the following dialog boxes and options are displayed:
7. Create a link to cost estimate data stored in a spreadsheet