UNIVERSITY OF ZAGREB
FACULTY OF HUMANITIES AND SOCIAL SCIENCES
Department of Information and Communication Sciences
Department of English

MASTER THESIS

Fine-grained Human Evaluation of an English to Croatian Hybrid Machine Translation System

Filip Klubička

Mentors: *Nikola Ljubešić and Mateusz Milan Stanojević*

Zagreb, July 2017.

My sincerest thanks to Gema-Ramírez Sánchez for babysitting me in Spain and showing me how to laugh in the face of seemingly insurmountable issues, and countless thanks to Antonio Toral, Víctor Sánchez-Cartagena and all other members of the Abu-MaTran team for their hard work, cooperation and immense practical help with the most stubborn of systems; I cherish the friendship and memory of working with them. I am also very much indebted to Maja Popović for her expert advice on how to approach the annotation task, and to Denis Kranjčić for actually approaching the annotation task. Of course, this thesis would not have been written were it not for my ever-watchful emotional support network, consisting of my family and good friends (including, but not limited to, TSONTS, DDFIT, JDI and Dekameron), and for that I owe them my thanks. On the other hand, this thesis would have certainly been completed much earlier were it not for the crazy cult embodied in the Academic choir Concordia discors, and I would not have it any other way; thank you for being an inextricable part of my life. Finally, I bestow my sincerest gratitude upon my mentors: to Mateusz Milan Stanojević for wholeheartedly supporting and reinforcing my enchantment with linguistics, and to Nikola Ljubešić for the countless opportunities bestowed upon me, for believing in me in spite of the occasional missed deadline, and for showing me that there is magic in automation.

CONTENTS

1. Introduction

Translation has always been a valiant human endeavour, and machine translation presents itself as a tool to make that task easier. However, as there is yet no optimal way to build a system that would be able to account for all the nuance and complexities that accompany the task of translation, automatic translation can still be quite problematic to do. Especially considering that very often the difficulty of creating a satisfactory machine translation system rises in proportion with the distance between the languages in the language pair being translated. In other words, the more different the languages are from each other, the more of a challenge it becomes to build these systems, regardless of the approach one takes to do it.

Thus, when it comes to a language pair such as English and Croatian, two languages that belong to different language groups and, one being analytic and the other synthetic, are really quite different from one another, approaching the task of building a machine translation system can be daunting. A possible way to tackle the problem of suiting the system to such a distant language pair is to make the system more complex. In theory, implementing specific methods to address some of the particularities of translation between the languages involved should raise the quality of the system.

Following a similar train of thought, one of the many products of the Abu-MaTran project[1] was a hybrid machine translation system between English and Croatian, which was built with the aim of exploring ways to improve the quality of translations in this particular language pair; and, according to automatic metrics, the quality did indeed improve. However, standard approaches to evaluating the quality of a machine translation system are, more often than not, quantitative, and not qualitative. So when applied to this particular system, they do not provide answers to questions such as: "Does the new approach improve a specific linguistic aspect of the output?", "How is the resulting text different from what a standard statistical setup would produce?", "What specifically does it mean for the text that the scores have improved?, "How important are the observed differences?" and ultimately "How reliable are those metrics in general?" Even though these questions are interesting from a purely scientific point of view, answering them is also important because such insights can guide

[1] More about the project and its efforts, to which the author of this thesis has contributed as an early-stage researcher, can be found on the following URL: `https://www.abumatran.eu`

future development - knowing what happens with the translations on this level allows developers to better tailor their systems to the languages and data they are working with, and it can better guide the development of new systems, giving rise to more advanced translation methods.

It is exactly for these reasons that this thesis aims to adapt and apply an evaluation methodology that could adequately investigate the impact that a particular hybrid machine translation model has on solving the complexities of translation from English to Croatian. The main objective of the thesis is to show that using a factored translation model truly does affect the grammatical quality of the language produced by a machine translation system. In addition, this thesis delves beyond just that and pinpoints exactly which aspects of the translation are improved, and which might need more work in the future.

The rest of the thesis is structured as follows: Chapter 2 provides an overview of the extensive theoretical background underpinning this thesis, Chapter 3 describes the layout of the performed experiments, while Chapter 4 describes the annotation process and the results of the analyses. Finally, the thesis is rounded off with a discussion and a concluding chapter.

2. Theoretical background

Machine translation, abbreviated as MT, is a sub-field of computational linguistics that investigates the use of software to translate text or speech from one language to another. The field of machine translation, as it is known today, has begun its development in the 1950s, with Warren Weaver's *Memorandum on Translation* (1949), where he discusses the possibility of using digital computers to translate documents between natural human languages. Due to its vast potential, as well as its underestimated limitations, it has since had several ups and downs in popularity. Or, as Goutte et al. (2009, 1) describe it, "a long history of ambitious goals and unfulfilled promises". However, they point out how the recent surge in computers' processing power has made MT more accessible and usable, while the ever-growing need for content localisation due to globalisation and the internet has launched the demand for MT to unprecedented heights. Subsequently, the field of MT research is stronger than ever, and is constantly inventing novel ways to improve the machine translation paradigm. As a result of this development, there are currently many different approaches to doing MT.

2.1. Approaches to machine translation

Translation is in itself a complex and multi-layered task, even for humans, so it is no surprise that very different approaches to MT have been developed through the years. These include rule based (transfer, interlingua, dictionary based), statistical, hybrid, example-based and, most recently, neural MT. Given that it is the aim of this thesis to compare two of these approaches - statistical and hybrid MT - it is essential to elaborate on them in more detail.

2.1.1. The rule-based paradigm

In order to understand the hybrid paradigm, one must understand both elements that comprise it, and the rule-based machine translation paradigm makes one of them.

Often abbreviated as RBMT, or called the "Classical Approach" to MT, it is based on explicit linguistic information about source and target languages, which is essentially retrieved from unilingual, bilingual or multilingual dictionaries, and grammars covering the main semantic, morphological, and syntactic regularities of each language respectively. Given some

input sentences (in a source language), an RBMT system generates output sentences (in a target language) on the basis of morphological, syntactic, and semantic analysis of both the source and the target languages involved in a concrete translation task (Okpor, 2014, 160).

In itself, this approach has many advantages, according to Okpor (2014, 161):

- No bilingual texts are required. This makes it possible to create translation systems for languages that have no texts in common, or even no digitised data whatsoever.
- The approach is domain independent. Rules are usually written in a domain independent manner, so the vast majority of rules will "just work" in every domain, and only a few specific cases per domain may need rules written for them.
- In theory, RBMT is open to the possibility to use targeted rules to correct any error that the system might produce, even if the trigger case is extremely rare. This is in contrast to statistical systems, where their capabilities are limited to what they can learn from the training data
- It offers total control. Because all rules are hand-written, one can easily debug a rule based system to see exactly where a given error enters the system, and why.
- Once developed, they have high reusability value. Because RBMT systems are generally built from a strong source language analysis that is fed to a transfer step and target language generator, the source language analysis and target language generation parts can be shared between multiple translation systems, requiring only the transfer step to be specialised. Additionally, source language analysis for one language can be reused to bootstrap a closely related language analysis.

However, as pointed out by Okpor (2014, 161), this approach is not without shortcomings:

- There is an insufficient amount of really good, comprehensive dictionaries, while building new ones is expensive.
- RBMT systems require the manual development of linguistic rules, which can be costly, and which often do not generalise to other languages.
- In bigger RBMT systems, it is hard to deal with idiomatic expressions, ambiguity, and rule interactions (having many specific rules can negatively impact unrelated cases).
- The systems usually fail to adapt to new domains. Although RBMT systems do provide a mechanism to create new rules and extend or adapt the lexicon, changes are usually very costly and the results, frequently, do not pay off.

So generally, it can be said that RBMT systems are not too flexible, and their plateau, though nonexistant in theory, is in practice relatively low, while development costs are high. In practice, their best use nowadays is for MT between closely related languages, especially

smaller languages that might be under-resourced. This is being done for many language pairs, including Croatian and Serbian (Klubička et al., 2016), for which the open-source Apertium platform (Forcada et al., 2010) has been utilised, as it enables free and open development of RBMT systems. However, an Apertium system (or most RBMT systems for that matter) would not be feasible for a more distant language pair, such as English and Croatian, as the languages are simply too different. To address all the specificities of translating between these languages, the RBMT system would have to be far more complex than what Apertium can offer, while the number of rules and regularities that would need to be manually encoded is just too high to make the endeavour worthwhile. Thus, current trends indicate that a statistical approach is preferable in such a case.

2.1.2. The statistical paradigm

Statistical machine translation, abbreviated as SMT, is a machine translation paradigm where translations are generated on the basis of statistical models, the parameters of which are derived from the analysis of bilingual text corpora.

The general setting of SMT, according to Goutte et al. (2009, 2), is to learn how to translate from a large corpus of pairs of equivalent source and target sentences. This is typically a machine learning framework: there is an input (the source sentence), an output (the target sentence), and a model trying to produce the correct output for each given input. This is done according to the probability distribution that a string in the target language is the translation of a string in the source language.

The problem of modelling the probability distribution has been approached in a number of ways - SMT models were initially word based (Models 1-5 from IBM Hidden Markov model (Vogel et al., 1996) and Model 6 (Och and Ney, 2003)), but significant advances were made with the introduction of phrase based models (Koehn et al., 2003). There has also been work that incorporates syntax or quasi-syntactic structures (Chiang, 2005) into the model.

The baseline system evaluated in this thesis is a phrase-based statistical model, so focus will be directed towards this particular approach, which is also the currently dominant approach in statistical MT. Similar to word-based models, its attractiveness comes from separating the task of finding the highest probability of translation into two subtasks. It creates a translation model - the probability that the source string is the translation of the target string - and a language model - the probability of seeing that target language string. Finding the best translation is done by picking the one that gives the highest probability using log-linear models[1].

[1]Log-linear models are mathematical models that take the form of a function whose logarithm is a linear combination of the parameters of the model. For more details refer to Christensen (2002), as well as `https://en.wikipedia.org/wiki/Log-linear_model`

In phrase-based translation, the aim is to reduce the restrictions of word-based translation by translating whole sequences of words, where the lengths may differ. The sequences of words are called blocks or phrases, but typically do not consist of linguistic phrases, but rather phrasemes found using statistical methods from corpora. It has been shown that restricting the phrases to linguistic phrases (e.g. syntactic categories) decreases the quality of translation (Chiang, 2005).

The most popular and widely-used implementation of PBMT is the Moses system (Koehn et al., 2007), a complete out-of-the-box open-source translation system for academic research. It consists of all the components needed to preprocess data, train language models and translation models. It also contains tools for tuning these models and evaluating the resulting translations using the BLEU score[2].

The statistical approach is often presented in contrast with rule-based approaches, as stated by Koehn (2010), and indeed, its benefits over the rule-based approach are quite numerous:

- More efficient use of human and data resources
- There are many parallel corpora in machine-readable format and there is even more monolingual data.
- Generally, SMT systems are not tailored to any specific pair of languages.
- More fluent translations owing to use of a language model

However, as the other approaches, it too has its shortcomings:

- Corpus creation can be costly, depending on availability and quality of data.
- Specific errors are difficult to predict and fix.
- Results may have superficial fluency that masks translation problems.
- Statistical machine translation usually works less well for language pairs with significantly different word order.
- The benefits obtained for translation between Western European languages are not representative of results for other language pairs, owing to smaller training corpora and greater grammatical differences.

So generally, one can say that given enough data, SMT works well enough for a native speaker of one language to get the approximate meaning of what is written by the other native speaker, but just like with RBMT, there is a plateau. The real difficulty is getting enough data to either train a good enough general model, or obtain enough in-domain data to focus on translating a specific domain. But it is still a question whether such systems would be able to handle all the specificities of the languages in question; specificities that

[2]BLEU score explained in detail in Section 2.2

might be easier solved with a rule-based approach. Which prompts the question of the ever-present cost-benefit tradeoff to rear its head - even though the large multilingual corpus of data needed for statistical methods to work is not necessary for the grammar-based methods, such grammar methods need a skilled linguist to put in a lot of time and work to carefully design the grammar that they use.

This is where hybrid models come in, with the underlying idea of combining the best of both worlds.

2.1.3. Hybrid models

As Goutte et al. (2009, 3) state, although the early SMT models essentially ignored linguistic aspects, a number of efforts have attempted to reintroduce linguistic considerations into either the translation models or the language models, which led to the development of hybrid MT models.

Hybrid MT is a method of machine translation that is characterised by the use of multiple machine translation approaches within a single machine translation system. The motivation for developing hybrid machine translation systems stems from the failure of any single technique to achieve a satisfactory level of accuracy, and hybrid systems have indeed been successful in improving the accuracy of the translations.

There are several approaches to hybridisation, as laid out by Hutchins (2007). There is the multi-engine approach, which involves running multiple machine translation systems in parallel, and then generating the final output by combining the output of all the sub-systems. Most commonly, these systems use statistical and rule-based translation subsystems, but according to Hutchins (2007, 14), other combinations have also been explored.

Then, there is statistical rule generation, an approach that involves using statistical data to generate lexical and syntactic translation rules. The input is then processed with these rules as if it were a rule-based translator. This approach attempts to avoid the difficult and time-consuming task of creating a set of comprehensive, fine-grained linguistic rules by extracting those rules from a training corpus (Hutchins, 2007, 14).

Another approach is the multi-pass approach, which involves serially processing the input multiple times. The most common technique used in multi-pass machine translation systems is to pre-process the input with a rule-based machine translation system. The output of the rule-based pre-processor is passed to a statistical machine translation system, which produces the final output (Hovy, 1996).

But other than these, there are ways to more tightly integrate explicit linguistic information into the translation model. One such way calls upon factored models, and this is the approach used by the other system evaluated in this thesis.

Koehn and Hoang (2007) introduced factored translation models, which are an extension

of phrase-based models, where source words are enriched with linguistic annotation (e.g., lemmas, parts of speech, morphological tags). Separate distributions then model translation from source lemmas to target lemmas, and from source parts of speech and morphology to their target equivalent. A deterministic morphological generator, finally, combines target lemmas and morphological information to reconstruct target surface forms (i.e., actual words). This is illustrated in Figure 2.1.

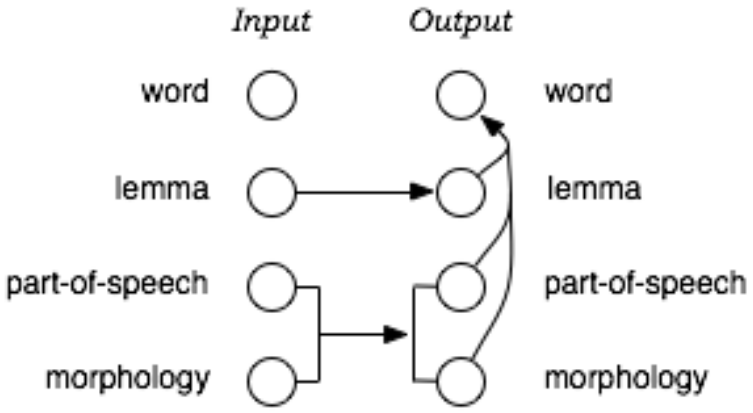


Figure 2.1: Graphic representation of the factored model

Such an approach is of great assistance when translating from English into other languages, as this task requires solving problems that are negligible the other way around. Goutte et al. go on to elaborate that morphology, for instance, is very simple in English compared to most other languages, where verbs can have tens of alternative forms according to mood, tense, etc.; nouns can have different forms for nominative, accusative, dative, and so on. Dictionaries for such languages tend to be much larger (empirical linguists speak of a lower token/type ratio), and reliable statistics are harder to gather. Moreover, when translating from a morphologically poor language (e.g., English) into a morphologically rich one (e.g., Croatian), purely word- or phrase-based models can have a hard time, since generating the appropriate morphology might require rather sophisticated forms of analysis on the source, and n-gram-based language models can only go so far. Meanwhile, factored models help in this regard, as they do not only transfer surface forms, but take with them any kind of additional annotations.

When it comes to implementations of the model, Moses offers the option to use factored models, which makes it freely available to use and do research with[3].

[3]Installation and usage guidelines, as well as the graph presented in Figure 2.1, can be found on the following link: `http://www.statmt.org/moses/?n=Moses.FactoredModels`

2.2. Machine Translation Evaluation

Regardless of the approach, one must be able to judge the quality of the output a translation system produces.

There is a myriad of ways to perform it, and as Goutte et al. (2009, 3) say, entire books have been devoted to discussing what makes a translation a good translation, while "relevant factors [for evaluation] range from whether translation should convey emotion as well as meaning, to more down-to-earth questions like the intended use of the translation itself."

They go on to say that, if one were to restrict one's attention to machine translation, "there are at least three different tasks which require a quantitative measure of quality:

1. assessing whether the output of an MT system can be useful for a specific application (absolute evaluation)

2. (a) comparing systems with one another, or similarly (b) assessing the impact of changes inside a system (relative evaluation)

3. in the case of systems based on learning, providing a loss function to guide parameter tuning"

Goutte et al. give an overview of different evaluation methods in their book, saying that MT evaluation has almost become a field of study in and of itself (this being reflected in the many ways one can perform it), but thus far no single method can really be crowned as the best. A relevant paper on this topic by Han and Wong (2016) offers an extensive overview of the current state-of-the-art methods.

Evaluation methods can essentially be divided into two groups: automatic and manual (human) evaluation (Han and Wong, 2016, 1). Depending on the task, it can be more or less useful or practical to require human intervention in the evaluation process. As Goutte et al. (2009, 3) point out, "on one hand, humans can rely on extensive language and world knowledge, and their judgment of quality tends to be more accurate than any automatic measure. On the other hand, human judgments tend to be highly subjective, and have been shown to vary considerably between different judges, and even between different evaluations produced by the same judge at different times."

Whatever one's position is concerning the relative merits of human and automatic measures, there are often contexts where requiring human evaluation is simply impractical because it is either too expensive or time-consuming. In such contexts fully automatic measures are necessary.

2.2.1. Automatic evaluation

A good automatic measure should above all correlate well with the quality of a translation as it is perceived by human readers. The ranking of different systems given by such a measure (on a given sample from a given distribution) can then be reliably used as a proxy for the ranking humans would produce.

In many cases new measures are justified in terms of correlation with human judgment. Many of the measures that will be briefly described below can reach Pearson correlation coefficients[4] in the 90% region on the task of ranking systems using only a few hundred translated sentences.

For this reason most automatic measures actually evaluate something different, sometimes called human likeness. For each source sentence in a test set, a reference translation produced by a human is made available, and the measure assesses how similar the translation proposed by a system is to the reference translation. Ideally, one would like to measure how similar the meaning of the proposed translation is to the meaning of the reference translation: an ideal measure should be invariant with respect to sentence transformations that leave meaning unchanged (paraphrases). One source sentence can have many perfectly valid translations. However, most measures compare sentences based on superficial features which can be extracted very reliably, such as the presence or absence of n-grams in the references compared to the proposed translation. As a consequence, these measures are far from being invariant with respect to paraphrasing. In order to compensate for this problem, at least in part, most measures allow considering more than one reference translation. This has the effect of improving the correlation with human judgment, although it imposes on the evaluator the additional burden of providing multiple reference translations.

Some of the most popular and widespread automatic evaluation metrics are based on the Levenshtein distance metric (Levenshtein, 1966). This is a string metric for measuring the difference between two sequences. Informally, the Levenshtein distance between two words is the minimum number of single-character edits (insertions, deletions or substitutions) required to change one word into the other, It may sometimes also be referred to as edit distance.

The following list explains some of the more popular metrics used today.

– **Word error rate (WER)** (Niessen et al., 2000) is the sum of insertions, deletions, and substitutions normalised by the length of the reference sentence. A slight variant

[4]The Pearson Correlation coefficient measures the strength and direction of linear relationships between pairs of continuous variables. By extension, the Pearson Correlation evaluates whether there is statistical evidence for a linear relationship among the same pairs of variables in the population. For more details, refer to: `https://en.wikipedia.org/wiki/Pearson_correlation_coefficient` and `http://libguides.library.kent.edu/SPSS/PearsonCorr`

(WERg) normalises this value by the length of the Levenshtein path, i.e., the sum of insertions, deletions, substitutions, and matches: this ensures that the measure is between zero (when the produced sentence is identical to the reference) and one (when the candidate must be entirely deleted, and all words in the reference must be inserted).

- **Position-independent word error rate (PER)** (Tillmann et al., 1997) is a variant that does not take into account the relative position of words: it simply computes the size of the intersection of the bags of words of the candidate and the reference, seen as multi-sets, and normalises it by the size of the bag of words of the reference.
- **Translation edit rate (TER)** (Snover et al., 2006) is another variant which, Similar to WER, counts the minimal number of insertions, deletions, and substitutions, but unlike WER it introduces a further unit-cost operation, called a "shift," which moves a whole substring from one place to another in the sentence.
- **Bilingual evaluation understudy (BLEU)** (Papineni et al., 2002) differs from the previous measures, insomuch that it does not compute edit distance between the candidate translation and the reference, but is based on n-grams. More specifically, it is based on clipped n-gram precision, i.e., the fraction of n-grams in a set of translated sentences that can be found in the respective references. The BLEU score is expressed as the geometric mean of clipped n-gram precisions for different n-gram lengths (usually from one to four), multiplied by a factor (brevity penalty) that penalises producing short sentences containing only highly reliable portions of the translation.
- **National Institute of Standards and Technology measure (NIST)** (Doddington, 2002) is very similar to BLEU - it is the arithmetic mean of clipped n-gram precisions for different n-gram lengths, also multiplied by a (different) brevity penalty. Additionally, when computing the NIST score, n-grams are weighted according to their frequency, so that less frequent (and thus more informative) n-grams are given more weight.

There are many other approaches, according to Goutte et al. (2009, 7-8) such as the general text matcher (GTM) measure (Melamed et al., 2003), the ROUGE-L and ROUGE-W methods (Lin, 2004), the BLANC (Lita et al., 2005) and METEOR (Banerjee and Lavie, 2005) measures, which are all focused on recall. Then there are measures that use syntax, proposed by Liu and Gildea (2005) and Giménez and Màrquez (2007), and meta-measures like ORANGE (Lin and Och, 2004) and IQMT (Giménez and Amigó, 2006), which combine and evaluate existing measures to obtain more reliable scores.

However, going into all of them in more detail would be beyond the scope of this thesis. At this point it is more beneficial to move the discussion towards manual evaluation methods.

2.2.2. Human evaluation

Even when it comes to human evaluation, the task of judging whether a translation is 'good' or 'correct' is a difficult one. As already stated, agreement between judges is often very low. This is related to the fact that there is always more than one acceptable translation, in combination with the high subjectivity of evaluators, who often even disagree with themselves, given some amount of time.

Having clearly defined definitions of what constitutes an acceptable translation, and a robust framework to allow for the least amount of ambiguity, can somewhat help remedy the problem.

An overview of methods is laid out by Han and Wong (2016), who divide the approaches into traditional and advanced ones.

Traditional human assessment:

- Defining the concepts of **Intelligibility and Fidelity** (Carroll, 1996) and using them to evaluate MT was one of the earliest human assessment methods. The requirement that a translation be intelligible means that a translation should, as much as possible, read like normal, well-edited prose and be readily understandable in the same way that such a sentence would be understandable if originally composed in the translation language. The requirement that a translation is of high fidelity or accuracy requires that the translation should, as little as possible, twist, distort, or controvert the meaning intended by the original.
- Church and Hovy (1993) propose that evaluators look at three aspects: **Fluency, Adequacy and Comprehension**. Adequacy scores are assigned by the evaluator, who looks fragments in the translation (a fragment is defined as being delimited by syntactic constituents and containing sufficient information), and judges their adequacy on a scale from 1-to-5. The adequacy score is computed by averaging the judgments over all of the decisions in the translation set.

  The fluency evaluation determines whether the sentence is well-formed and fluent in context, and is compiled in the same manner as for adequacy, except that the evaluator makes intuitive judgments on a sentence by sentence basis for each translation. The evaluators are asked to determine whether the translation is "good English" without being provided a reference translation.

  Comprehension can also be called “Informativeness”, its objective being to measure a system’s ability to produce a translation that conveys all the necessary information from the source (based on the reference set of expert translations).
- There is further development in the traditional assessment domain, and most is based on redefining and refining the notions of fluency, accuracy and adequacy. For ex-

ample, Bangalore et al. (2000) conduct research in which they define several kinds of accuracy, including simple string accuracy, generation string accuracy, and two corresponding tree-based accuracies.

Meanwhile, Specia et al. (2011) conduct a study of MT adequacy and assign to it 4 scores, ranging from highly adequate (the translation faithfully conveys the content of the input sentence), through fairly adequate (the translation generally conveys the meaning of the input sentence, with word order or tense/voice/number issues and untranslated words), poorly adequate (the content of the input sentence is not adequately conveyed by the translation) to completely inadequate (the content of the input sentence is not conveyed at all by the translation).

Additionally, the "Linguistics Data Consortium" (LDC) developed two five-points scales representing fluency and adequacy, where adequacy indicates how much of the meaning expressed in the reference translation is also expressed in a hypothesis translation; whereas fluency indicators involve both grammatical correctness and idiomatic word choices. Sccording to Han and Wong (2016), this has become the widely used methodology when manually evaluating MT.

However, there are more approaches to evaluating MT systems and their outputs, beyond specifying categories such as adequacy, fluency, etc. Han and Wong (2016) cite these as advanced human assessment methods:

- **Task-oriented approaches** measure MT systems in light of the tasks for which their output might be used. For example, Voss and Tate (2006) introduce the task-based MT output evaluation by extracting who/when/where type elements from translations. They later extend this work into event understanding. Nießen et al. (2000), on the other hand, developed a task-oriented evaluation methodology for translation from Japanese to English, seeking to associate the output used in the evaluation with a scale of language-dependent tasks, such as scanning, sorting, and topic identification. They develop an MT proficiency metric with a corpus of multiple variants which are usable as a set of controlled samples for user judgments. The principal steps include identifying the user-performed text-handling tasks, discovering the order of text-handling task tolerance, analysing the linguistic and non-linguistic translation problems in the corpus used in determining task tolerance, and developing a set of source language patterns which correspond to diagnostic target phenomena.
- King et al. (2003) add some **extended criteria** to a large range of manual evaluation methods for MT systems themselves, not just their output. So in addition to accuracy, they include suitability (to the particular context in which the system is to be used), interoperability, reliability, usability, efficiency, maintainability and portability.

– **Utilising post-editing** can also be a way of evaluating translation quality. By comparing translations from scratch and the post-edited result of an automatic translation, one can evaluate a system's performance. However, this type of evaluation is quite time consuming and very reliant on the skills of the translator/posteditor. An interesting example of a metric that is designed in such a manner is the human-targeted translation edit rate (HTER) (Snover et al., 2006), based on the number of editing steps between an automatic translation and a reference translation. While WER and TER only consider a pre-defined set of references, and compare candidates to them, in computing HTER a human is instructed to perform the minimal number of operations to turn the candidate translation into a grammatical and fluent sentence that conveys the same meaning as the references. HTER is then defined as the number of editing steps divided by the number of words in the acceptable translation.

– **Segment ranking** is a way of manual evaluation without explicitly assigning values to a system's output, but rather performing a relative comparison of several systems. Judges are asked to provide a complete ranking over all the candidate translations of the same source segment (Callison-Burch et al., 2011, 2012). For example, five systems can be randomly selected for the judges to rank. Each time, the source segment and the reference translation are presented to the judges together with the candidate translations of five systems. The judges will rank the systems from 1 to 5, allowing for tied scores. The collected pairwise rankings can be used to assign a score to each participated system to reflect the quality of the automatic translations.

Another, more linguistically inclined approach to manual MT evaluation that Han and Wong do not go into, consists of assessing the quality of an MT system by performing error analysis to analyse the linguistic quality of the output texts. According to Spillner (1991, abstract), erros are information, so it is no surprise that in the fields of second language learning, speech-language pathology, and even sociolinguistics, error analysis is deemed to be an essential methodological tool for tasks such as diagnosis and evaluation of the language acquisition process, investigation of the language of speakers with and without language disorders, and the study of native speaker reactions to errors made by non-native speakers (Bussmann et al., 2006, 378). By analogy, error analysis also plays an important role in the task of both human and machine translation - if, for example, an error can show that a language learner does not understand a grammatical rule, such as subject-verb agreement, then it can say the same thing about the output of an MT system. This is the reasoning that underpins the utilisation of the error analysis approach in the research leading up to this thesis. The details of the approach used in the thesis will be elaborated on in Section 4.1.

3. Experimental setup

This section describes the MT systems and the datasets used in the experiments. In short, two systems have been developed - a classic PBMT system, and a factored PBMT system. Both systems were trained on the same parallel data, hence only the underlying architecture can affect the output, and not the training data.

3.1. Data

A set of publicly available English–Croatian parallel corpora was considered for use as training data. It was comprised of the DGT Translation Memory[1], HrEnWaC[2], JRC Acquis[3], OpenSubtitles 2013[4], SETIMES[5] and TED talks[6]. All these corpora were concatenated and had cross-entropy based data selection performed on them[7].

Once the data was ranked, the highest ranked 25% sentence pairs were kept inside the corpus. As a result, the training corpus contains 4,786,516 sentences.

In addition, training a PBMT system also requires monolingual data for language modelling. To this end, the target side of the aforementioned parallel corpora was concatenated with hrWaC, the Croatian web corpus (Ljubešić and Klubička, 2014).

A development set was constructed from the first 1,000 sentences of the English test set used at the WMT12 news translation task[8], translated by a professional translator into Croatian. Similarly, the test set consists of the first 1,000 sentences of the English test set of the WMT13 translation task[9], which was again manually translated into Croatian. The properties of the test set are elaborated on in Section 3.3.

[1] https://ec.europa.eu/jrc/en/language-technologies/dgt

[2] https://www.clarin.si/repository/xmlui/handle/11356/1058

[3] http://tinyurl.com/CroatianAcquis

[4] http://opus.lingfil.uu.se/OpenSubtitles2013.php

[5] http://nlp.ffzg.hr/resources/corpora/setimes/

[6] http://opus.lingfil.uu.se/TedTalks.php

[7] This means that the language model training data was selected by training an in-domain language model and then using it to score text segments from out-of-domain data sources, selecting segments based on a score cutoff. For more details, refer to (Moore and Lewis, 2010).

[8] http://www.statmt.org/wmt12/translation-task.html

[9] http://www.statmt.org/wmt13/translation-task.html

3.2. MT systems

A PBMT and a factored PBMT system were trained on the data described in Section 3.1.

The PBMT system was built with Moses v3.0[10]. In addition to the default models, hierarchical reordering (Galley and Manning, 2008), an operation sequence model (Durrani et al., 2011) and a bilingual neural language model (Devlin et al., 2014) were employed. Pirinen et al. (2015) provide a detailed description of this system in the third project deliverable, but most of its relevant components have been mentioned or described throughout this thesis.

The factored PBMT system, is described in detail by Sánchez-Cartagena et al. (2016). As explained in Subsection 2.1.3, factored models can break down the translation of words in the translation of different factors (surface forms, lemmas, lexical categories, morphosyntactic information, etc.). Among the different ways these factors can be combined, Sánchez-Cartagena et al.'s model produces a surface form factor and a morphosyntactic description (MSD) factor for each word in the output, and uses two different language models, one operating on surface forms and another one on MSDs. Skadiņš et al. (2010) report that this setup is effective (and efficient in terms of decoding time) when the target language is highly inflected but the source language is not, since it helps the decoder to produce grammatically correct phrases that have not been observed in the training corpus. When building the factored PBMT system, two aspects were considered :

- Order of the MSD language model. The order of surface-form based language models is usually set to 5. As the number of different MSDs is several orders of magnitude lower than the number of different surface forms, a greater order can also be considered.
- Corpora tagging algorithm. In order to obtain the MSD factor of the target language side of the parallel corpus and the target language monolingual corpus, a part-of-speech (PoS) tagger is needed. For these purposes, the state of the art CRF PoS tagger for Croatian (Ljubešić et al., 2016) was used, which achieves an accuracy of 92.5%.

3.3. The test set

The test used in this evaluation is an asset developed quite early in the Abu-MaTran project, and consists of 1000 sentences that were manually translated by the author of this thesis. The sentences were taken from a test set published within the WMT13 workshop, which consisted of article text extracted from various online news publications. It is available in a number of European languages but not Croatian. Thus, part of the data (48 articles) was

[10] `https://github.com/moses-smt/mosesdecoder/tree/RELEASE-3.0`

translated into Croatian for the purpose of experiments and evaluation of systems built within the Abu-MaTran project.

Not much metadata about this test set has been published by the workshop organisers. However, a detailed look at its content reveals that the articles comprising the corpus indeed belong exclusively to the news domain, but the topic coverage is relatively wide, ranging from political across economical and medical to scientific.

The categories detected in the dataset are, in descending order, politics, culture, science, lifestyle, economy, health, technology and sports, as well as an "other" category for unspecific news articles (the two were loosely about religious tourism and icy roads during winter, respectively). It should also be noted that 4 of the articles are in fact interviews (1 of each belonging to politics, culture, lifestyle and sports). A detailed visualisation of the data is presented in Figure 3.1, and from the graph it is obvious that politics is a topic that dominates the discourse, comprising ~35% of the texts.

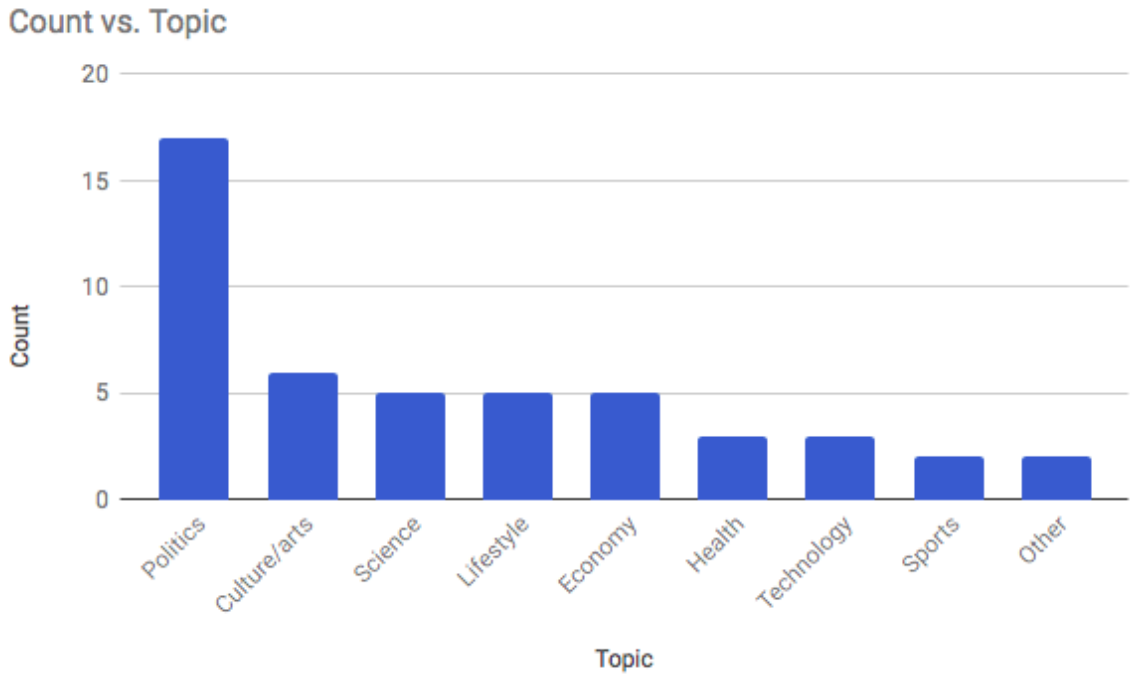


Figure 3.1: The distributions of topics across the test set

It is important to consider this aspect of the dataset, because even though the training data was of mixed origins and could be considered general-domain data, the test data belongs exclusively to the news-domain, and is dominated by a specific topic. This can have implications for the output, as the register and topics are highly specialised. Certainly, in order to make a definitive claim about the existence of domain bias, additional experiments are needed that tackle this particular question. Even so, it is quite possible that the results of the evaluation might be significantly different if the systems were presented with a different

test set, one coming from either a different domain, or containing a mix of domains (similar to what is found in the training data).

3.4. Automatic evaluation

As is common practice, the MT systems were evaluated automatically, in order to easily obtain an idea of their performance. Evaluation was done using the BLEU and TER metrics, as described in Section **??**. The scores obtained from this evaluation are presented in Table 3.1.

System	BLEU	TER
PBMT	0.2544	0.6081
Factored PBMT	0.2700	0.5963

Table 3.1: Automatic evaluation (BLEU and TER scores) of the two MT systems.

As the table suggests, the factored PBMT model leads to a substantial improvement upon pure PBMT, resulting in a 6% relative increase in terms of BLEU score, with a similar trend in TER score improvement (a higher BLEU score means implies a better performance, whereas TER indicates improved performance via a lower score).

4. Error Annotation

This chapter describes the motivation for conducting manual error analysis, describes the framework and overall annotation process, and presents the results.

The fact that Croatian is a synthetic language, and English is analytic, gives rise to specific translation issues between these two languages. Croatian is rich in inflection, has a rather free word order and other similar phenomena that English does not. For example, grammatical categories that do not exist in English, like gender or case, may be particularly hard to generate reliably in a Croatian translation. The factored PBMT system was built with the goal to directly address such issues, specifically agreement issues.

Indeed, as shown in Section 3.4, automatic evaluation shows significant improvement for the factored system over the PBMT system. However, as is the nature of automatic metrics, the automatic scoring methods do not indicate whether any of the linguistic problems mentioned earlier have been addressed by the systems. The question of whether the linguistic quality, or rather, grammaticality of the output is improved is not answered by automatic evaluation. Are cases and gender handled better? Is there better agreement? Is the fluency of the translation higher?

In order to provide answers to these research questions, a thorough comparison of the two systems is necessary. This can be done by systematically analysing their outputs via manual error analysis. In this way one can obtain a more complete picture of what is happening in the translation, while still quantifying the results. This, in turn, can provide pointers on where to act to obtain further improvements in the future.

4.1. Multidimensional Quality Metrics

After considering different ways of performing this task, using an established framework for error annotation seemed like a good approach. A suitable candidate for this was the Multidimensional Quality Metrics (MQM) framework, developed within the QTLaunchpad project[1]. This is a framework for describing and defining custom translation quality metrics. It provides a flexible vocabulary of quality issue types and a mechanism for applying them

[1] http://www.qt21.eu/mqm-definition/definition-2015-06-16.html

to generate quality scores. It does not impose a single metric for all uses, but rather provides a comprehensive catalog of quality issue types, with standardised names and definitions, that can be used to describe particular metrics for specific tasks.

The main reason the MQM framework was chosen for this task was the flexibility of the issue types and their granularity — it provides a reliable methodology for quality assessment, that still allows for the picking and choosing of specific error tags to use.

4.2. The tagset

The MQM guidelines propose a great variety of tags on several annotation layers[2]. However, the full tagset, which is provided in the appendix of this thesis, is too comprehensive to be viable for any annotation task, so the process begins with choosing the tags to use in accordance to the research question. It is good practice to start off with the core tagset, a default set of evaluation metrics (i.e. error categories) proposed by the MQM guidelines, as seen in Figure 4.1.

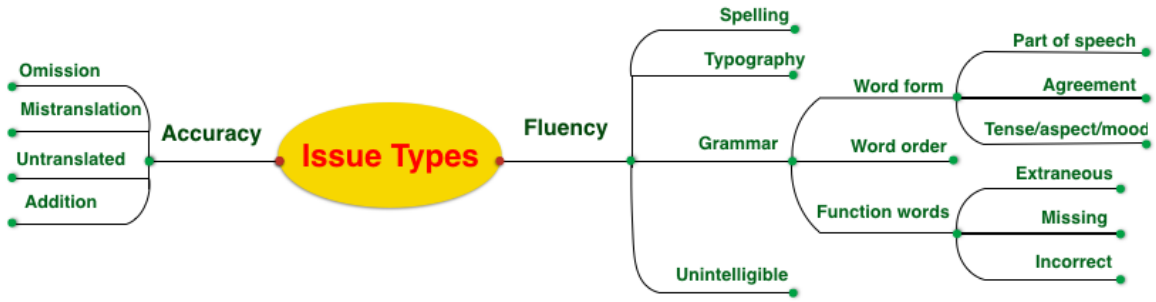


Figure 4.1: Visual representation of the core error categories proposed by the MQM guidelines

However, given the morphological complexity of Croatian and the level at which interventions have been made in the factored system, one can argue that these core categories are not detailed enough, or rather, do not allow for an analysis of the specific phenomena that could be of interest. Some such categories, like specific *Agreement* types, are not present in the core tagset, while some errors, like *Typography*, are outside the scope of what this research is examining. So a modified set of tags was created by rearranging the hierarchy, adding new tags and removing ones that are of little consequence. This new tagset can tentatively be called the Slavic Tagset, as this expansion allows for identification of grammatical errors which are commonly shared by Slavic languages, but are omitted from the core tagset. This tagset is outlined in Figure 4.2.

[2]The full taxonomy of error tags can be found on the following URL:
`http://www.qt21.eu/mqm-definition/issues-list-2015-12-30.html`

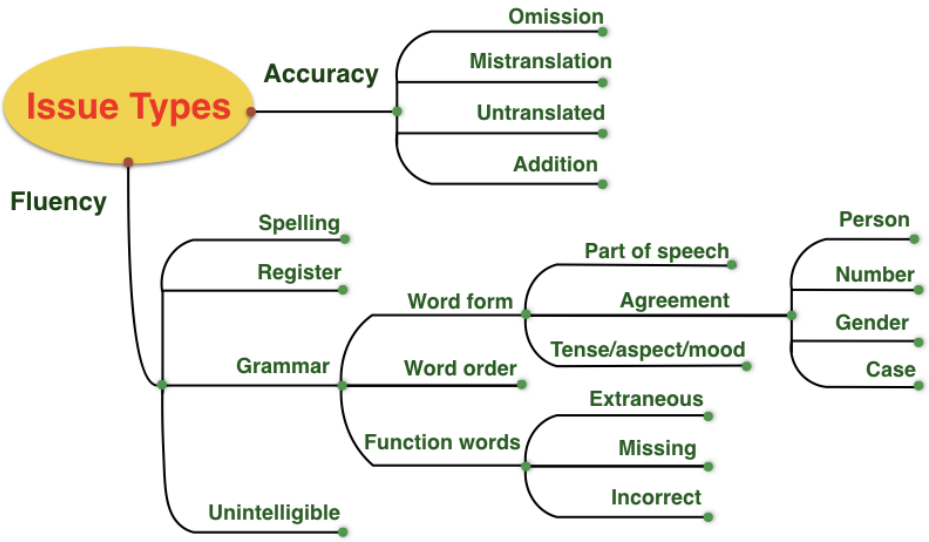


Figure 4.2: The Slavic tagset, a modified version of the MQM core tagset

As can be seen from the graphic, there is a hierarchy in place, a sort of error taxonomy. Errors are divided into two main groups, or levels - **Accuracy** and **Fluency**. As stated in the MQM usage guidelines[3], *Accuracy* addresses the extent to which the target text accurately renders the meaning of the source text. For example, if a translated text tells the user to push a button when the source tells the user not to push it, there is an accuracy issue. *Fluency*, on the other hand, relates to the monolingual qualities of the source or target text, relative to agreed-upon specifications, but independent of relationship between source and target. In other words, fluency issues can be assessed without regard to whether the text is a translation or not. For example, a spelling error or a problem with register remain issues regardless of whether the text is translated.

It has to be said that at first look this distinction might seem obvious and clear-cut, but in practice it is anything but. Very often examples can seem like they belong into either category, so it is up to the annotators' judgement to decide which level is a better fit, and then being consistent in following through on the decisions made regarding dubious examples.

Each of the top nodes in the taxonomy has numerous child-nodes and branches. As such, Accuracy branches out out into errors of Mistranslation, Omission, Addition and Untranslated.

The **Mistranslation** category describes issues that arise when the content on the target side of the translation does not accurately represent the content on the source side. This is one such error type that might cause trouble for annotators, as it can seemingly overlap with the *Fluency* branch: According to the guidelines, only one error should be tagged, and *Accuracy* trumps *Fluency* if the required information is present in the source text.

[3] `http://www.qt21.eu/downloads/MQM-usage-guidelines.pdf`

Source:	For example, **websites** provide...
Correct:	Na primjer, **internetske stranice** pružaju...
Translation:	Na primjer, **internetska stranica** pružaju...

Table 4.1: Example of a grammatical *Mistranslation* error.

An example of this is shown in Table 4.1, where the only actual error is the translation of 'website' in the singular rather than the plural, which is explicitly encoded via the -s morpheme in the source text. However, this error then causes a subject-verb agreement error, where the translated subject is singular, but the verb has been correctly translated in plural. This example should, according to the guidelines, be classified only as *Mistranslation*, even though it also shows problems with agreement. If the subject had been translated properly (as the plural), the subject-verb agreement problem would be resolved, so in this case only 'internetsku stranicu' should be tagged as a *Mistranslation*.

As for the other, less ambiguous nodes in the *Accuracy* branch, there is:

- **Omission**, to be used when content is missing from the translation that is present in the source. It should be reserved for those cases where content present in the source and essential to its meaning is not found in the target text.
- **Addition**, where the target text includes text not present in the source.
- **Untranslated**, to be tagged when there is content that should have been translated, but has not been. It should be noted that, if a term is passed through untranslated, it should be classified as *Untranslated* rather than as *Mistranslation*.

On the other side, the *Fluency* branch has many more nodes and sub-branches, a total of 4 different levels.

At the first level, it branches into *Spelling*, *Register*, *Unintelligible* and *Grammar*.

- **Spelling** is to be used if there are issues related to spelling of words, including capitalisation. When it comes to PBMT, such errors most often crop up due to noisy data.
- **Register** was not present in the core tagset, but can be found in the extended set. It was included in the Slavic tagset because preliminary insights into the data showed a potential usefulness for annotating a breach of standardness, which has indeed cropped up a couple of times the systems' outputs. It does not happen often, but sometimes a synonym for a word can be used, one that is a correct translation in a very general sense, but is actually sub-standard and would not normally be found in that sentence or that particular context. E.g. She was the first woman in space. > Bila je prva ženska u svemiru.

– **Unintelligible** indicates a major breakdown in fluency. This category should be used sparingly, for cases where further analysis is too uncertain to be useful. If an issue is categorised as *Unintelligible*, no further categorisation is required. *Unintelligible* can refer to texts where a significant number of issues combine to create a text for which no further determination of error type can be made or where the relationship of target to source is entirely unclear. For example:

Source:	... he is doomed to come across unconscientious people, who, pretending to help, will force upon him **a "ticket" to terrible, cramped barracks**...
Correct:	... suđen mu je susret s nesavjesnim ljudima koji će mu, pravećи se da pomažu, dati **"kartu" za grozne, pretrpane barake**...
Translation:	... on je osuđen na to da nesavjesni ljudi, koji, pretvarajući se da pomažu, natjerat će ga **na "kartu" užasna, skučenim vojarni**...

Table 4.2: Example of an *Unintelligible* error.

– **Grammar** errors denote issues related to the grammar or syntax of the text, other than spelling and orthography. Given that this is a parent branch, it ought to be used only if none of its subtypes accurately describe the issue.

Considering that grammaticality of translations is in the focus of this research, it is quite useful that *Grammar* is branched out further, as already proposed by the core tagset, to the second level.

– **Function words** is a tag that signifies when linguistic function words such as prepositions, particles, and pronouns are used incorrectly. It branches into three additional specific nodes.

– **Word order** tag denotes text where the word order is incorrect.

– **Word form** tags should be used when the wrong form of a word is used. Subtypes should be used when possible.

Both *Word form* and *Function words* have their own children at the third level of the hierarchy. When it comes to *Function words*, the selection is simple:

– **Incorrect** indicated when an incorrect function word is used.

– **Extraneous** indicates whether an unneeded function word is present.

– **Missing** indicates when a needed function word is missing

By comparison, the *Word form* family is rather more complex at the third level:

– **Tense/aspect/mood** refers to inappropriate use of verbal forms.

- **Part of speech**, an error that occurs when a word is the wrong part of speech. Though it does not appear often in translations to Croatian, the tag was still included for completeness. An English example would be: "Read these instructions careful" instead of "Read these instructions carefully."
- **Agreement** error in the core set covers issues where two or more words do not agree with respect to grammatical features such as case. Given that the Slavic tagset includes these categories as children, the parent *Agreement* tag is used when one phrase has more than one *Agreement* error (this is elaborated on in Section 4.4).

As explained in Subsection 2.1.3, factored models take grammatical categories such as number, gender, case and person with them into translation. In order to accurately analyse the impact of such a transfer, it was necessary to add these categories to the taxonomy, which allows for more fine-grained linguistic insight. Thus, the following *Agreement* categories were added as children to the lowest level of the hierarchy:

- **Number** indicates disagreement in number (e.g. "Čovjek hodaju").
- **Gender** indicates disagreement in gender (e.g. "crveni mačka").
- **Case** indicates disagreement in case (e.g. "pored stolicu")
- **Person** indicates disagreement in person (e.g. "Oni smo gledali")

Once the taxonomy is defined, it is hardcoded as a simple XML file that indicates all the levels and relationships in the hierarchy. This file, together with the test data, is then uploaded to the MQM annotation tool and the annotation can begin.

4.3. The translate5 tool

Translate5[4] was used in order to carry out the annotations. It is a web-based tool that implements annotations of MT outputs using hierarchical taxonomies, as is the case of MQM.

Translate5 has a fairly intuitive interface. The data is presented in a table where each row represents a sentence, and each column a version of that sentence - source, reference, translation 1, translation 2. All these elements are flexible and malleable - there need not be a reference translation, one could evaluate only 1 MT system, or even a text that is not even MT-related. A screenshot of the whole interface can be found on the following page.

[4] `http://www.translate5.net/`

translate5

Authenticated user: Filip Klubicka
Login name: fklubicka
Task: m3-fm_enhr_case_rnd

Logout | Tasks | My settings | English

Files

Segment list and editor

Editor modes | Hide tags | Short tag view | Full tag view | Reset sorting / filtering | QM Subsegment Statistics

Nr.	source	reference	mt_out1	mt_out2	Comments
1	A shuttle bus brings people to it from the foot of the mountain.	Direktni bus koji kreće s podnožja planine prevozi ljude do vrha.	[5 Shuttle bus 5] dovodi ljude [2 na to 2] [22 iz 22] podnožja planine.	[5 Shuttle bus 5] dovodi ljude [2 na to 2], [22 iz 22] podnožja planine.	
2	Now, everybody is divided into micro societies, it's hard to be liked by everyone.	Sada su svi podijeljeni u mikrodruštva, teško je sviđati se svima.	Sada, svi se [14 dijeli 14] na mikro [16 društava 16], teško je [7 da se svidimo 7] drugima.	Sada, svi su podijeljeni u mikro društva, teško je [7 da se svidimo 7] drugima.	mt_out1 hrase agr *Filip Klubicka 2016-09-0* ... (1 more comment)
3	Right now, even if I need a few knuckledusters, I get them through someone I trust.	Trenutno, ako trebam par boksera, nabavit ću ih preko nekog kome vjerujem.	Sada, čak i ako trebam nekoliko [5 knuckledusters 5], dobivam ih kroz [16 neko 16] [3 3] vjerujem.	Sada, čak i ako trebam nekoliko [5 knuckledusters 5], dobivam ih kroz nekoga [3 3] vjerujem.	mt_out1 phrase ag *Filip Klubicka 2016-09-0*
4	Once this number has been found, the other nodes can easily check that it is the right one.	Jednom kad je taj broj pronađen, ostali čvorovi lako mogu provjeriti radi li se o pravoj transakciji.	Kad ovaj broj [18 je 18] pronađen, drugi čvorovi mogu lako provjeriti da je ispravna.	Nakon što je ovaj broj [4 je 4] pronađen, ostali čvorovi mogu lako provjeriti da je ispravna.	
5	Because, in the opinion of a number of people working in palliative care, great moments occur at the very heart of such regression.	Jer, prema mišljenju ljudi koji rade u palijativnoj skrbi, veliki trenutci dešavaju se u samom srcu takve regresije.	Jer, po mišljenju velikog broja ljudi koji rade u [16 palijativne 16] skrbi, veliki trenuci se [7 odvijaju 7] u samom srcu takve regresije.	Jer, po mišljenju velikog broja ljudi koji rade u palijativnoj skrbi, veliki trenuci se javljaju u samom srcu takve regresije.	mt_out1 phrase ag *Filip Klubicka 2016-09-0*
6	But we'll have something to say against some very good European teams.	Ali imat ćemo nešto za reći protiv nekih vrlo dobrih europskih timova.	Ali ćemo imati nešto za reći protiv [2 neke vrlo dobre europske 2] momčadi.	Ali ćemo imati nešto za reći protiv vrlo dobrih europskih timova.	
7	The Ministry of the Interior does not put arms from the illegal market back into circulation	Ministarstvo unutarnjih poslova ne vraća oružje s crnog tržišta natrag u cirkulaciju	Ministarstvo unutarnjih poslova ne stavi [2 ruke 2] [22 na 22] ilegalnom tržištu natrag u promet	Ministarstvo unutarnjih poslova ne stavi [2 ruke 2] od nelegalnog tržišta natrag u promet	
8	People used to get together in flocks, Bohemians liked one thing, the simple people, something else.	Ljudi su se prije nalazili u jatima, boemima se sviđala jedna stvar, jednostavni ljudi, nešto drugačije.	Ljudi [2 koriste 2] da se okupe u jatima, [2 primilo 2] je volio jednu stvar, jednostavni ljudi, nešto drugo.	Ljudi [2 koji se koriste 2] kako bi se [2 zajedno 2] u jatima, [4 gomila 4] boema sviđala jedna stvar, jednostavni ljudi, nešto drugo.	
9	How do you explain this progression?	Kako objašnjavate taj napredak?	Kako [7 objašnjavaš 7] ovo [2 napredovanje 2]?	Kako [7 objašnjavaš 7] ovo [2 napredovanje 2]?	
10	If a migrant does not understand the language, says Sebelev with certainty, he is doomed to come across unconscientious people, who, pretending to help, will force upon him a "ticket" to terrible, cramped	"Ako migrant ne razumije jezik", sa sigurnošću kaže Sebelev, "suđen mu je susret s nesavjesnim ljudima koji će mu, praveći se da pomažu, dati "kartu" za grozne, pretrpane barake u kojima će pored	Ako [3 3] još ne [14 razumiju 14] jezik, [18 kaže [4 da 4] sa sigurnošću Sebelev 18], on je osuđen [3 3] da nesavjesni ljudi, koji, pretvarajući se da [17 [14	Ako doseljenik ne [14 razumiju 14] jezik, [18 kaže Sebelev sa sigurnošću 18], on je osuđen da [2 se preko 2] nesavjesni ljudi, koji [4 su 4], pretvarajući se [2 u pomoć	mt_out2 sentence vp) *Filip Klubicka 2016-09-0* ... (3 more comments)

Segment meta data

Terminology
Keine Terminologie vorhanden!

QM Subsegments
add QM Subsegment
Severity: Critical
Comment:

QM
- OK
- Minor errors
- Must be reworked

Status
- Status 1
- Status 2
- Status 3
- Nicht gesetzt

Comments for the current segment

Translate5 enables annotation of text in the form of spans i.e. every annotation has a beginning and an ending in the text. This is done by simply selecting the desired text span with the cursor and then selecting the appropriate annotation from the drop-down menu.

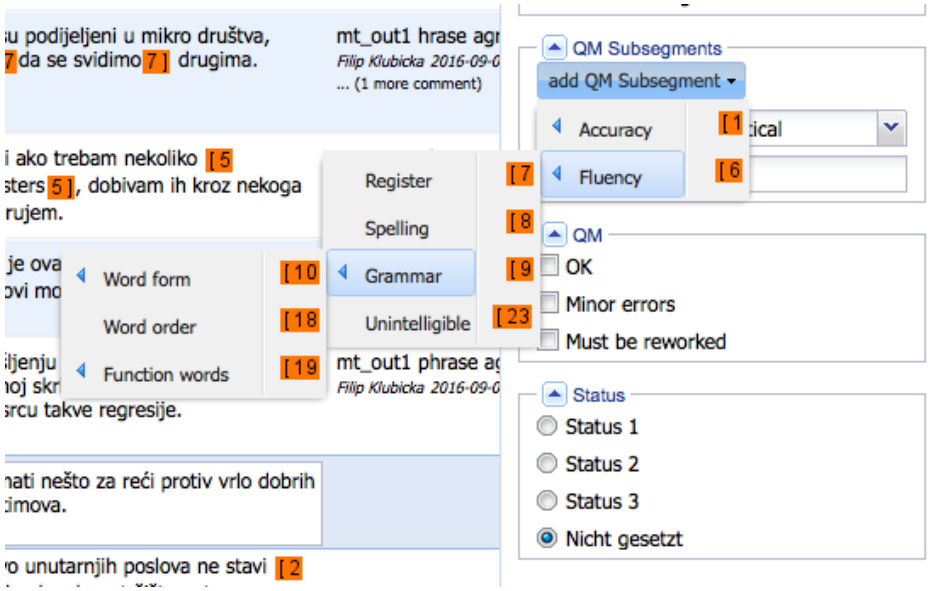


Figure 4.3: The drop-down menu that presents a choice of tags from the error hierarchy

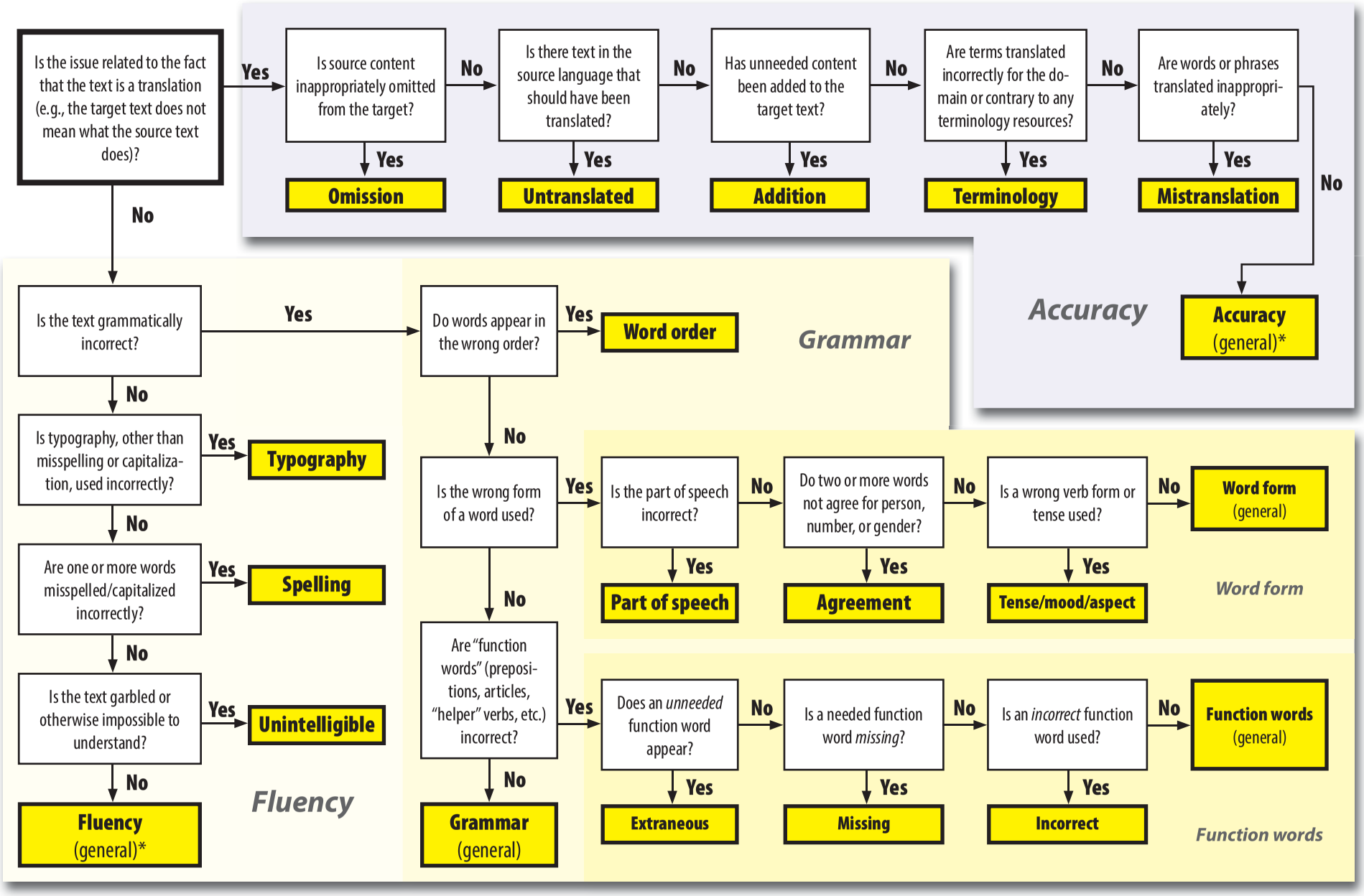
Even though translate5 does what it is intended to do, it has a number of bugs that require some attention, as well as a lack of some features that one might expect from such a tool. Most notably, the system offers no tools that allow for any sort of statistical analysis of the annotated data, except for a raw error count per dataset. Some more detailed and informative automatic analysis and the ability to compare the performance of different annotators on the same dataset would be quite welcome.

4.4. Annotation setup

Two annotators, who both had prior experience with MQM as well as the same background (an MA in English linguistics and information science), were familiarised with the MQM framework and the translate5 system. They were presented with both systems' outputs at the same time, given a choice in which order to annotate, but did not know which outputs belonged to which system, effectively performing blind annotation. Annotation was performed in accordance to the official MQM annotation guidelines, which offer detailed and unambiguous instructions for annotation within the MQM framework, and even provide a flowchart/decision tree to help the annotation process. This chart is presented on the following page.

MQM Annotation Decision Tree

Note: For any question, if the answer is unclear, select "No"



* Please describe any **Fluency (general)** or **Accuracy (general)** issues using the Notes feature.

However, some additional instructions and clarifications on how to annotate cases specific to the new tag set were needed. The following decisions were made regarding specific cases that do not appear in the official MQM guidelines:

- If something has been omitted in the translation, annotate the empty space between the tokens where the omitted text would be found had it not been omitted.
- If there is an agreement error within a noun phrase, include the whole affected phrase in the span of the annotation.
- If, however, there is sentence agreement (e.g. between a noun and a verb), mark as the span the whole subject and verb. In practice, this means that one error tag will span tokens that do not necessarily have an agreement error attached to them. On the other hand, there will not be any superfluous *Agreement* annotations overall. (This discrepancy is accounted for in the downstream.)
- However, if several or all grammatical categories (gender, case, number) are causing an agreement error, then tag this as a blanket Agreement error, so as to avoid tagging the same word/phrase multiple times.

The data that was thus annotated consists of 100 random sentences collected from the test set introduced in Section 3.3. These sentences were translated by both MT systems, and were then annotated by both annotators (i.e. each system translated the same 100 sentences, each annotator annotated the 200 translated sentences, making a total of 400 annotated sentences.)

The annotators were presented with the source text, a reference translation and both unannotated MT outputs at the same time. Once the sentences were annotated, the annotation data was extracted and inter-annotator agreement was calculated, the output was analysed to see what the number of error tags can say about the performance of each system.

4.5. Inter-Annotator Agreement

Though carefully thought out and developed, the MQM metrics, and more broadly MT evaluation in general, are notorious for resulting in low inter-annotator agreement scores. This is attested by the body of work that has addressed this issue, most notably Lommel et al. (2014), who worked specifically on MQM, and Callison-Burch et al. (2007), who investigated several tasks. This is why it is important to also check to what extent the annotators agree on the task at hand and whether this is consistent with other work done with MQM so far.

Once the data was annotated, observed agreement was approximated on the level of sentence, and inter-annotator agreement was calculated using the Cohen's Kappa (κ) metric

(Cohen, 1960), which does not only take into account the observed agreement (Pr(a)), but also accounts for chance agreement (Pr(e)), as seen in equation 4.1.

$$\kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)} \tag{4.1}$$

Agreement was calculated on the annotations of every system separately, as well as on a concatenation of annotations, in order to both see whether there are differences in agreement across systems, as well as to get an idea of overall agreement between annotators. Additionally, Cohen's κ was also calculated for every error type separately. Detailed results can be found in Table 4.3 and Appendix B.

Error type	**PBMT**	**Factored**	**Concatenated**
Accuracy			
Mistranslation	0.51	0.48	0.50
Omission	0.34	0.39	0.36
Addition	0.50	0.54	0.52
Untranslated	**0.86**	**0.86**	**0.86**
Fluency			
Unintelligible	0.39	0.32	0.35
Register	0.37	0.20	0.29
Word order	0.56	0.33	0.46
Function words			
Extraneous	0.56	0.32	0.44
Incorrect	0.37	0.18	0.27
Missing	0.00	0.49	0.40
Tense/aspect/mood	0.44	0.36	0.40
Agreement	0.24	0.41	0.32
Number	0.53	0.55	0.54
Gender	0.46	0.59	0.53
Case	0.53	0.49	0.53
All errors	**0.56**	**0.49**	**0.53**

Table 4.3: Inter-annotator agreement (Cohen's κ values) for the MQM evaluation task. The highest scores for any individual system as well as the overall score are shown in bold.

Generally, the table shows that the annotators agree better on evaluations of the PBMT system than on the factored system, and that the overall agreement scores are relatively low, the average total κ being approximately 0.53. Furthermore, the κ scores are relatively

consistent across all error types, mostly ranging between 0.35 and 0.55. According to Cohen, such figures constitute moderate agreement.

There is one obvious outlier, however — the Untranslated category. Agreement on this error is extremely high when compared to other error types. This does make sense, as untranslated text is quite an unambiguous and easily detectable phenomenon, so high agreement among annotators would be expected.

The fact that the average agreement scores fall under the 'moderate agreement' category is also expected, given the complexity of both the problem and the annotation schema. However, these scores are in fact much higher than what has been reported in similar work, most notably by Lommel et al. (2014), who achieve far lower MQM annotation κ scores, ranging between 0.25 and 0.34. Though the relatively high scores obtained on this task can certainly be called a success, the comparison with Lommel et al. should be taken with a grain of salt, as these calculations are just an approximation compared to theirs. Lommel et al.'s setup was considerably different - they calculated agreement on the token level, while here it was done at the sentence level.

The calculations are approached differently here in order to attempt to account for some of the problems that come with span-level annotation. As Lommel et al. (2014, 4) point out, a "fundamental issue that the QTLaunchPad annotation encountered was disagreement about the precise scope of errors". In other words, though annotators can agree that a sentence contains the same issue, they might disagree on the span that the issue covers. An example is shown in Table 4.4 (annotations marked in bold).

Source:	Trakhtenberg was the presenter of many programs before Hali-Gali times.
Annotator_1:	Bio **je** voditelj **Trakhtenberg** brojnih programa Hali-Gali prije puta.
Annotator_2:	**Bio je voditelj Trakhtenberg** brojnih programa Hali-Gali prije puta.

Table 4.4: Example of annotator disagreement on error span on the example of a *Word order* error.

This case shows that annotators can agree on the nature and categorization of issues, yet still disagree on their precise span-level location. Even though they are instructed to mark minimal spans, i.e., spans that cover only the issue in question, they frequently disagree as to what the scope of these issues is. Lommel et al. (2014, 4) hypothesize that this may be due to the fact that the two reviewers perceive the issue differently, and so see different spans as cognitively relevant. In some instances this disagreement may reflect differing ideas about optimal solutions, while in others the problem may have more to do with perceptual units in the text.

In cases where annotators disagree on the span of the annotation, even Lommel et al. are uncertain on how best to assess IAA. Thus, building on their work and exploring a sentence-

level approach is a direction worth pursuing, as there seems to be no optimal solution, as both the sentence- and token-level approach come with certain drawbacks. However, to dispel any doubt in the reliability of the annotators' judgements on the task at hand, further analysis of the results shows that both annotators' annotations point to comparable conclusions, both when considered separately and together. This is elaborated on in Section 4.6.

4.6. Results of MQM evaluation

4.6.1. Raw annotations

Table 4.5 on the following page presents the sum of raw annotations for every error type, for each system and both annotators, as exported directly from the translate5 system.

By examining the table one can easily detect that both annotators have judged that the PBMT system contains more errors than the factored PBMT system (317 and 264 errors in PBMT, whereas 276 and 199 errors in factored PBMT). This trend is consistent across most fine-grained error categories as well.

However, even though simply counting the errors can provide a rough idea of which system performs better, one could claim that this approach does not allow for a proper quantification of the quality of the outputs, and thus cannot adequately represent the findings. Certainly, based on data from Table 4.5 the claim can be made that the PBMT system produces less errors in general, or less errors of a specific type, but given that the outputs are different, as is the number of tokens in each translation, there is an obvious need to somehow normalise the data.

4.6.2. Normalised annotations

There seems to be no related work on how to approach normalisation of MQM results, as all the papers simply count the number of MQM tags and stop there. Even though it does makes sense to count the absolute number of errors, as this is what one wants to minimise in a system's output, such an approach is still a bit problematic; if for no other reason, then due to the fact that such a setup does not allow for any kind of statistical significance testing, especially a statistical comparison of differences between two or more systems. Furthermore, the two systems may output sentences of different lengths, which is indeed the case in the data explored here: in the 100 annotated sentences, the phrase-based system produced an average of 18.99 tokens per sentence, whereas the factored system averaged on 18.89. Though not too striking a difference in this particular case, these reasons led to the decision to approach the normalisation task on the token-level.

	PBMT		Factored PBMT	
Error type	**Anno_1**	**Anno_2**	**Anno_1**	**Anno_2**
Accuracy	103	125	92	93
Mistranslation	78	80	64	64
Omission	13	22	11	12
Addition	4	14	10	8
Untranslated	8	9	7	9
Fluency	214	139	184	106
Unintelligible	2	3	2	4
Register	13	6	12	4
Spelling	0	2	0	4
Grammar	197	128	170	94
Word Order	26	16	25	8
Function words	22	10	25	6
Extraneous	4	3	4	2
Incorrect	16	7	18	3
Missing	2	0	3	1
Word Form	149	102	119	80
Part of Speech	10	2	11	4
Tense...	30	23	27	17
Agreement	109	76	80	58
Number	17	12	12	10
Gender	15	9	18	12
Case	71	40	38	23
Total error count	**317**	**264**	**276**	**199**

Table 4.5: Raw annotation data from PBMT and Factored systems: number of error for each error type, per annotator.

Instead of counting just the error tags produced by each annotator, the tokens that these errors are assigned to are counted - tokens that do and tokens that do not have an error annotation. However, some error categories are a bit specific for such a simple approach and require special consideration while counting:

- *Omission*: given that an omission error was never assigned to a token, but to an empty space, and yet it still needs to be counted somehow, it has been assumed that 1 token was omitted for every omission error, and so every omission error was given one phantom token to latch on to, in order to perform the calculations.
- *Agreement*: given that an agreement error can either span a single phrase or a whole sentence (more on this in Subsection 4.7), it is not always the case that what is annotated should also be counted when normalising. In the event of sentence-level (dis)agreement, not all the words in the span are actually part of the agreement error. Thus the decision had to be made on how to automatically count the number of tokens in such cases, the options being to either ignore the problem and count everything in the span, or assume that one agreement error most often spans two tokens, and count each agreement error as such, no matter the span of the annotation. Both approaches have been tested by creating two datasets differing only in the way that *Agreement* errors are counted, and then performing the analysis described below on both. A comparison of the results has shown that both approaches point towards the same broad conclusions. However, the assumption that one agreement error spans two tokens yielded more tidy results, so it has been presented in the body of the thesis. The alternative approach can be found in the Appendix.
- Tokens tagged with another category are counted as the per their annotation span.

Once these counts have been made and the sums are divided by the total number of tokens, they give a more concrete idea of the ratio of tokens with and without errors.

The results of such analysis again show that the factored PBMT system has a smaller error ratio. This is further backed up by a chi-squared (χ^2) statistical significance test, which shows that the difference in the total number of tokens with errors is statistically significant, with an average p value lower than 0.0001.

Furthermore, the extracted data allows for insight into which error types are the ones making a significant impact on this result. So the same measurements are repeated, but instead of performing them on all error types combined, they were performed for each specific error category. Where values were too small for Pearson's test[5] to handle, Fisher's test[6] for statistical significance was performed instead. The combined results of the calculations and

[5] `http://vassarstats.net/newcs.html`

[6] `https://www.graphpad.com/quickcalcs/contingency1/`

	PBMT		Factored PBMT				
Error type	**No error**	**Error**	**No error**	**Error**	χ^2	p	ϕ
Accuracy	3467	369	3525	291	9.65	**0.0019**	0.0355
Mistranslation	3547	289	3586	230	6.87	**0.0088**	0.03
Omission	3801	35	3793	23	2.44	0.1183	0.0179
Addition	3814	22	3797	19	0.21	0.6468	0.0052
Untranslated	3813	23	3797	19	0.36	0.5485	0.0069
Fluency	3195	641	3298	518	14.64	**0.0001**	0.0437
Unintelligible	3790	46	3769	47	0.02	0.8875	0.0016
Register	3810	26	3794	22	0.31	0.5777	0.0064
Spelling	3833	3	3812	4	-	0.7257	-
Grammar	3270	566	3371	445	15.97	**<0.0001**	0.0457
Word order	3752	84	3752	64	2.65	0.1035	0.0186
Function words	3801	35	3780	36	0.02	0.8875	0.0016
Extraneous	3829	7	3810	6	0.07	0.7913	0.003
Incorrect	3810	26	3790	26	0	1	0
Missing	3834	2	3812	4	-	0.4518	-
Word form	3389	447	3471	345	14.06	**0.0002**	0.0429
Part of speech	3822	14	3800	16	0.14	0.7083	0.0043
Tense...	3775	61	3765	51	0.85	0.3566	0.0105
Agreement	3466	370	3540	276	14.41	**0.0001**	0.0434
Number	3778	58	3772	44	1.87	0.1715	0.0156
Gender	3788	48	3756	60	1.42	0.2334	0.0136
Case	3614	222	3694	122	29.89	**<0.0001**	0.0625
Total errors	2826	1010	3007	809	27.77	**<0.0001**	0.0602

Table 4.6: Processed annotation data from both annotators concatenated: each system's total number of tokens with and without errors, with statistical significance test results (chi-squared (χ^2), p-value and effect size ϕ). Statistical significance is marked with bold where p-value is <0.05.

transformations are presented in Table 4.6 on the previous page. Separate counts for each individual annotator can be found in the Appendix.

Several things can be concluded from this table. Firstly, when looking simply at the grand total of tokens with and without errors, the difference between the systems is statistically significant by a wide margin. In other words, the factored system has significantly fewer errors than the pure PBMT system. The overall error rate is in this case reduced by 20%.

A separate analysis of specific error types that contribute to this score reveals that only some of the error categories are significantly different between the two systems. In the table, these values are marked in bold - the difference in number of errors is statistically significant when it comes to errors in the categories of general *Accuracy*, and more specifically *Mistranslation*, as well as errors in the categories of general *Fluency*, *Grammar*, *Word form* and *Agreement*, as well as most specifically, *Agreement in Case*.

The final point is most interesting, as this was one of the most intriguing questions at the beginning: Is there a way to better handle agreement when translating to Croatian? It can now be confidently said that the factored system produces significantly less agreement errors overall, and given the specific agreement types, the system handles agreement in case significantly better.

However, one should also note the effect size (ϕ coefficient) of these measurements. The ϕ coefficient is a measure that reports the size of the impact an independent variable has on a dependent variable. In this case, the p-value indicates whether there is a statistically significant difference, and if there is, the ϕ coefficient indicates how big or strong the difference is. The ϕ coefficient is often reported together with the p-value to provide a more complete picture of the effects. In this case, wherever the difference is statistically significant, the ϕ coefficient is very small. This is consistent with the relatively small increase in BLEU score that was observed during automatic testing.

Such low ϕ coefficients might indicate that these differences, though meaningful, are quite minor. Even so, the differences are undeniably statistically significant, and the factored model is a step closer to solving some of the specific issues of translating from English to Croatian. The resulting system produces language that is a little more fluent and a little more grammatical, which, among other things, can be of help when it comes to the task of post-editing.

4.7. Syntactic annotation of agreement

Given that the most significant gains have been obtained on agreement, it would be interesting to look at agreement more in depth, not just at the level of morphology (gender, number, case), but also at the level of syntax. Especially considering that two syntactically different

types of agreement have been subsumed under the MQM Agreement tags - local, short-distance agreement (or phrase agreement), which concerns agreement of elements within a phrase; and long-distance agreement (or sentence agreement), which concerns agreement of elements in the sentence, that have much wider spans and are further apart. For example, local agreement would be agreement between an adjective and a noun, or between a preposition and the following noun, while sentence agreement would be agreement between a verb and a noun. Here are some examples of agreement errors at these levels:

Phrase disagreement:	Veliki broj ljudi radi **u palijativne skrbi**.
Sentence disagreement:	Stalna antikorupcijska **jedinica**, koja se bori protiv svakog oblika korupcije, **nastao** je 2011. godine.

Table 4.7: Example of different types of agreement errors.

This distinction is not only important linguistically, but also from a technical perspective - given that PBMT works with text segments that are situated close together, it would be interesting to see whether this is reflected in the agreement improvements as well.

Thus, an additional layer of annotation was conducted, outside the framework of MQM. Each MQM agreement error was categorised as being either phrase or sentence agreement. Additionally, the POS of elements participating in the error was marked as well, in order to gain insight into what might be better handled. The results of this categorisation are presented in Table 4.8.

	Phrase agreement			**Sentence agreement**	
Phrase	**PBMT**	**Factored**	Phrase	**PBMT**	**Factored**
PP+NP	20	12	NP+VP	21	16
ADJ+N	13	7	NP+NP	2	2
NP+CNJ(+NP)	3	4	CNJ+VP	1	2
N+N	4	4	VP+VP	1	1
Total	40	27	**Total**	25	21

Table 4.8: Breakdown and categorisation of agreement errors found in the annotated data.

As the table suggests, the factored PBMT model leads to a substantial improvement upon pure PBMT, as the number of phrase agreement errors has been reduced from 40 to 27, which results in a 30% relative reduction in errors. When looking at sentence agreement, there is also an observed reduction in errors, but it is far smaller - 25 to 21, only 16% less. It is interesting to note that considerable gains have been observed in prepositional phrases and

noun phrases that contain a premodifying adjective, as well as improvement in long-distance agreement between nouns and verbs.

Of course, this analysis comes with the same issue encountered when counting MQM errors. Knowing that the factored model produces less agreement errors overall, it is no surprise that it produces less of both phrase and sentence agreement errors. Thus, to determine whether these differences are statistically significant overall, a chi-squared (χ^2) statistical significance test was performed on the data. Calculated through a 2x2 contingency table, where the rows contained counts for each agreement type (Phrase/Sentence), while columns contained counts for agreement errors in each MT system (PBMT/Factored). The null-hypothesis in this setting is that there is no link between the MT system and the frequency of a specific agreement type that it produces (i.e. that no matter which system is employed, both phrase and sentence agreement errors are relatively similar). The resulting p-value turns out to be 0.6987, revealing no overall statistical significance. In other words, it cannot be claimed that the factored system produces less phrase agreement errors than the PBMT system does.

However, given that these data points are quite small, this analysis would certainly benefit if the number of analysed sentences, examples and errors were larger. Perhaps a larger sample size might be indicative of a different trend in error reduction, which could more conclusively shed light on how different models impact these types of agreement errors.

5. Discussion

As elaborated on in Section 4.6, there is no doubt that the factored system produces fewer errors overall, and specifically fewer *Mistranslation* and fewer *Agreement* errors, as well as fewer errors in the corresponding parent categories (like *Accuracy*, *Word form*, *Grammar* and *Fluency*). Even though the effect sizes are quite small, it is still interesting to note the largest effect size belongs to *Agreement_Case*, which further gives credence to the hypothesis that the factored system handles *Agreement* better than PBMT does. This might lead one to the conclusion that the factored system is superior to a standard PBMT system. However, the situation is decidedly not so clear cut.

Because, generally, both systems are still very similar, as they both have an SMT basis and have both been trained and tested on the same data. The only difference is that the factored model also takes into account explicit linguistic information. This theoretically gives it an advantage, as it is what undoubtedly causes the drop in the number of errors produced, compared to the pure PBMT system. The factored model is also extremely interesting in its own right, as it is able to infer regularities from the data on its own, without being given explicit rules as would be the case in an RBMT system - where an RBMT system would be literally instructed that, for example, "nouns and their premodifying adjectives have to agree in gender, number and case", the factored system was not instructed in any such explicit way, but is able to produce such constructions all the same, doing so more reliably than the PBMT system. However, training a factored model requires more work and resources compared to the PBMT system - an additional language model needs to be trained, and a PoS tagger for the target language needs to be available. And though these components are responsible for the improved results, they also make the process of building the system more expensive, while, in this case, providing relatively minor improvements.

Thus, once again, the tradeoff between cost and benefit is something that needs to be considered when deciding on an approach. The factored model is certainly less expensive and less time-consuming when compared to RBMT, as it does not require nearly as many work hours to implement (assuming that the parallel corpora needed to train the model exist and are more or less readily available), and when compared to PBMT, the factored model outperforms, if only so slightly. Still, whether the improvement in performance is worth the

additional work also depends on the ultimate goal of the developer. If investment of time and resources is not an issue, and the goal is to get the best performance possible, the factored model is probably a good choice. If time and resources are scarce, a PBMT model would suffice. If the goal, for some reason, is to fix a specific grammatical issue in the output, it might be more beneficial to perform some rule-based post-processing on the PBMT output and fix the specific errors with a targeted rule. This is assuming that the issue in question is not too complex for a single rule to handle; if it is, and a factored system is available off-the-shelf with enough data to cover the phenomenon in question, it might still be the preferred choice.

Whatever the case may be, the bottom line is that there is no simple answer - there is no 'one best model' that will solve everyone's problems, and the choice of a system ultimately depends on many factors. That is, however, not the question this thesis aims to answer; its aim is, rather, to shed some more light on two particular models and highlight their strengths and weaknesses. To that end, it shows that using factored models can have an impact on the linguistic quality of the output.

However, this study is not without its limitations. Or rather, one need be careful when drawing conclusions and generalising the results. Because many variables in this pipeline could have an effect on the result. First of all, the training data that the systems are trained on can always be improved (e.g. enlarged, cleaned, trimmed, specialised, generalised etc.). Another way to strengthen the arguments of this thesis would be to run the same procedure on multiple case studies; that is, either splitting the training data into subgroups, or gathering more data and training and testing on it separately. Having consistent results over multiple case studies would show that the conclusions arrived at in this thesis are sound.

Furthermore, the test data can also have a profound impact on the results, as already discussed in Section 3.3. Aligning it better with the type of training data, or just picking a different domain to test on is something to be considered. Building on that, the manual evaluation was performed only on 100 sentences from the test set. Even though this is enough to perform statistical analyses and draw some conclusions, such analyses always benefit from having more data points (as also mentioned in Appendix B). Annotating all 1000 sentences available in the test set, though expensive, would certainly yield more conclusive results. It would also allow for annotator adjudication - checking in with the annotators every 100 sentences and discussing their differences and possible misunderstandings, or outlying issues, in order to smooth out the annotation process and raise inter-annotator agreement.

Additionally, it would also be beneficial to employ more than 2 annotators. A higher number of annotators necessarily means more reliable annotation data. In addition, it allows for more detailed analysis of inter-annotator agreement, which might reveal outliers, i.e. under- or over-performing annotators. This, in turn, can foster discussion and a better

understanding of the problem at hand.

Finally, all these results are still restricted to the English-Croatian language pair. Given the similarities of languages belonging to the same groups, the results would probably hold for a language pair such as English-Slovene (or another Slavic language), but the impact of factored versus phrase-based models on a language pair such as French-Japanese cannot be predicted based on the findings in this research.

The ultimate claim that can be made is that between these two English to Croatian MT systems, that are trained on the same general domain data, and tested on the same domain-specific data, according to annotations of two annotators, the factored system is the one that performs better. However, if any of the variables in that statement were to change, the same result, though possible, cannot be expected without further testing. More experiments are needed to warrant drawing general conclusions.

6. Conclusion

There are several contributions that this thesis makes to the corpus of scientific literature and MT research.

The error taxonomy that was developed for this research, while only used for the English-to-Croatian language direction, should be applicable for the analysis of MT errors for any language direction going from a Germanic language to a Slavic one, as it takes into account grammatical properties specific to these languages.

Additionally, this thesis describes an approach to statistically analysing and interpreting the results of MQM error annotation that goes beyond simple counting of error tags.

Furthermore, fine-grained manual evaluation performed for the purpose of this thesis has provided answers to several questions, one of which was the main drive behind the development of the factored system - is there a way to better handle agreement when translating to Croatian? Given that the factored system produces sentences with far less errors overall and significantly less agreement errors, and generally language that is more fluent and more grammatical, a confident claim can now be made that factored models result in significantly less agreement errors compared to pure PBMT. Specifically, they produce less agreement in case, and the bulk of error reduction happens when translating local agreement.

There are many possible lines of future work. The methodology can always be applied to another language pair (e.g. English-Czech), or to the same language pairs but a different MT model. For example, in addition to comparing English to Croatian PBMT and factored system, an NTM system could be added to the comparison. Additionally, a bigger dataset could be annotated, with more annotators and a more controlled IAA analysis (such as evaluating the annotators in batches, consider the results, adjust the annotation process if needed and continue with the annotation).

Another interesting line of work could be a further adaptation of the tagset. In its current version, it has proved to be informative when comparing PBMT to factored PBMT, and it has shown many specificities of the translation. However, the only other error category with a significant reduction in errors is *Mistranslation*. This might be in part due to its broad definition, as, according to the MQM guidelines, it covers both lexical selection and, less intuitively, translation of grammatical properties. For example, if "cats[pl.]"

is translated as "mačka[sg.]", this is to be tagged as *Mistranslation*, in spite of the correct lexical choice. This makes it quite a vague category, so if one would wish to perform an even more nuanced linguistic error analysis, adding additional layers to the *Accuracy* branch (such as *Mistranslation-Gender*, *Mistranslation-Case*, *Mistranslation-Number*, *Mistranslation-Person*) seems like a promising direction to follow.

Furthermore, the question remains as to how important these results are overall when it comes to the effect they have in the usefulness of the translation in a particular application. Certainly, the reduction in *Fluency* errors results in more fluent texts which is of help when it comes to the task of post-editing, but more nuanced experiments are needed to confidently claim how impactful these differences are in practice. Thus, performing post-editing experiments on the outputs of both systems could prove to be an interesting extrinsic way of evaluating the systems' performance and the differences between them.

BIBLIOGRAPHY

Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005.

Srinivas Bangalore, Owen Rambow, and Steve Whittaker. Evaluation metrics for generation. In *Proceedings of the first international conference on Natural language generation-Volume 14*, pages 1–8. Association for Computational Linguistics, 2000.

H. Bussmann, K. Kazzazi, and G. Trauth. *Routledge Dictionary of Language and Linguistics*. Taylor & Francis, 2006. ISBN 9781134630387. URL `https://books.google.nl/books?id=O0-9Iw0Qh6EC`.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. (meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158. Association for Computational Linguistics, 2007.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Omar F Zaidan. Findings of the 2011 workshop on statistical machine translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 22–64. Association for Computational Linguistics, 2011.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. Findings of the 2012 workshop on statistical machine translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, WMT '12, pages 10–51, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics. URL `http://dl.acm.org/citation.cfm?id=2393015.2393018`.

John B Carroll. An experiment in evaluating the quality of translations. *Mechanical Translation*, 1996.

David Chiang. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*,

ACL '05, pages 263–270, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics. doi: 10.3115/1219840.1219873. URL https://doi.org/10.3115/1219840.1219873.

Ronald Christensen. *Plane answers to complex questions. The theory of linear models.* New York, NY: Springer, 3 edition, 2002.

Kenneth W Church and Eduard H Hovy. Good applications for crummy machine translation. *Machine Translation*, 8(4):239–258, 1993.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960. doi: 10.1177/001316446002000104. URL http://dx.doi.org/10.1177/001316446002000104.

Jacob Devlin, Rabih Zbib, Zhongqiang Huang, Thomas Lamar, Richard Schwartz, and John Makhoul. Fast and Robust Neural Network Joint Models for Statistical Machine Translation. In *Proceedings of ACL*, pages 1370–1380, 2014.

George Doddington. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, pages 138–145. Morgan Kaufmann Publishers Inc., 2002.

Nadir Durrani, Helmut Schmid, and Alexander Fraser. A joint sequence translation model with integrated reordering. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1045–1054. Association for Computational Linguistics, 2011.

Mikel L. Forcada, Mireia Ginestí Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Gema Ramírez-Sánchez, Felipe Sánchez-Martínez, and Francis M. Tyers. Apertium: a free/open-source platform for rule-based machine translation platform. *Machine Translation*, 2010.

Michel Galley and Christopher D Manning. A simple and effective hierarchical phrase reordering model. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 848–856. Association for Computational Linguistics, 2008.

Jesús Giménez and Enrique Amigó. Iqmt: A framework for automatic machine translation evaluation. 2006.

Jesús Giménez and Lluís Màrquez. Linguistic features for automatic evaluation of heterogenous mt systems. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 256–264. Association for Computational Linguistics, 2007.

Cyril Goutte, Nicola Cancedda, Marc Dymetman, and George Foster. *Learning machine translation*. Neural information processing series. MIT Press, 2009. ISBN 0262072971,9780262072977,9780262255097.

Aaron Li-Feng Han and Derek Fai Wong. Machine translation evaluation: A survey. *CoRR*, abs/1605.04515, 2016. URL http://arxiv.org/abs/1605.04515.

Eduard H Hovy. Deepening wisdom or compromised principles? - the hybridization of statistical and symbolic mt systems. *IEEE Expert*, 11(2):16–18, 1996.

John Hutchins. Machine translation: A concise history. *Journal of Translation Studies*, 2007.

Margaret King, Andrei Popescu-Belis, and Eduard Hovy. Femti: creating and using a framework for mt evaluation. In *Proceedings of MT Summit IX, New Orleans, LA*, 2003.

Filip Klubička, Gema Ramírez-Sánchez, and Nikola Ljubešić. Collaborative development of a rule-based machine translator between croatian and serbian. *Baltic Journal of Modern Computing*, 4(2):354, 2016.

Philipp Koehn. *Statistical Machine Translation*. Cambridge University Press, New York, NY, USA, 1st edition, 2010. ISBN 0521874157, 9780521874151.

Philipp Koehn and Hieu Hoang. Factored translation models. In *EMNLP-CoNLL*, pages 868–876, 2007.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 48–54, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics. doi: 10.3115/1073445.1073462. URL http://dx.doi.org/10.3115/1073445.1073462.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=1557769.1557821.

VI Levenshtein. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10:707, 1966.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop*, volume 8. Barcelona, Spain, 2004.

Chin-Yew Lin and Franz Josef Och. Orange: a method for evaluating automatic evaluation metrics for machine translation. In *Proceedings of the 20th international conference on Computational Linguistics*, page 501. Association for Computational Linguistics, 2004.

Lucian Vlad Lita, Monica Rogati, and Alon Lavie. Blanc: learning evaluation metrics for mt. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 740–747. Association for Computational Linguistics, 2005.

Ding Liu and Daniel Gildea. Syntactic features for evaluation of machine translation. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 2005.

Nikola Ljubešić and Filip Klubička. {bs,hr,sr}WaC – web corpora of Bosnian, Croatian and Serbian. In *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*, pages 29–35, Gothenburg, Sweden, 2014. Association for Computational Linguistics.

Nikola Ljubešić, Filip Klubička, Željko Agić, and Ivo-Pavao Jazbec. New inflectional lexicons and training corpora for improved morphosyntactic annotation of croatian and serbian. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may 2016. European Language Resources Association (ELRA). ISBN 978-2-9517408-9-1.

Arle Richard Lommel, Maja Popovic, and Aljoscha Burchardt. Assessing inter-annotator agreement for translation error annotation. In *MTE: Workshop on Automatic and Manual Metrics for Operational Translation Evaluation*, 2014.

I. Dan Melamed, Ryan Green, and Joseph P. Turian. Precision and recall of machine translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume of the Proceedings of HLT-NAACL 2003–short papers-Volume 2*, pages 61–63. Association for Computational Linguistics, 2003.

Robert C. Moore and William Lewis. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers*, ACLShort '10, pages 220–224,

Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=1858842.1858883.

Sonja Niessen, Franz Josef Och, Gregor Leusch, Hermann Ney, et al. An evaluation tool for machine translation: Fast evaluation for mt research. In *LREC*, 2000.

Sonja Nießen, Franz Josef Och, Gregor Leusch, Hermann Ney, et al. An evaluation tool for machine translation: Fast evaluation for mt research. In *LREC*, 2000.

Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Comput. Linguist.*, 29(1):19–51, March 2003. ISSN 0891-2017. doi: 10.1162/089120103321337421. URL http://dx.doi.org/10.1162/089120103321337421.

MD Okpor. Machine translation approaches: issues and challenges. *International Journal of Computer Science Issues (IJCSI)*, 11(5):159, 2014.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL http://dx.doi.org/10.3115/1073083.1073135.

Tommi Pirinen, Raphael Rubino, Víctor Sánchez-Cartagena, and Antonio Toral. Abu-matran deliverable d4.1b mt systems for the third development cycle. Technical report, 2015.

Víctor M Sánchez-Cartagena, Nikola Ljubešić, and Filip Klubička. Dealing with data sparseness in SMT with factored models and morphological expansion: a case study on croatian. *Baltic Journal of Modern Computing*, 4(2):354, 2016.

Raivis Skadiņš, Kārlis Goba, and Valters Šics. Improving smt for baltic languages with factored models. In *Human Language Technologies: The Baltic Perspective: Proceedings of the Fourth International Conference, Baltic HLT 2010*, volume 219, page 125. IOS Press, 2010.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas*, pages 223–231, 2006.

Lucia Specia, Najeh Hajlaoui, Catalina Hallett, and Wilker Aziz. Predicting machine translation adequacy. In *Machine Translation Summit*, volume 13, pages 19–23, 2011.

B. Spillner. *Error Analysis: A comprehensive bibliography*. Library and Information Sources in Linguistics. John Benjamins Publishing Company, 1991. ISBN 9789027284792. URL https://books.google.nl/books?id=pW5CAAAAQBAJ.

Christoph Tillmann, Stephan Vogel, Hermann Ney, Arkaitz Zubiaga, and Hassan Sawaf. Accelerated dp based search for statistical translation. In *Eurospeech*, 1997.

Stephan Vogel, Hermann Ney, and Christoph Tillmann. Hmm-based word alignment in statistical translation. In *Proceedings of the 16th Conference on Computational Linguistics - Volume 2*, COLING '96, pages 836–841, Stroudsburg, PA, USA, 1996. Association for Computational Linguistics. doi: 10.3115/993268.993313. URL http://dx.doi.org/10.3115/993268.993313.

Clare R Voss and Calandra R Tate. Task-based evaluation of machine translation (mt) engines: Measuring how well people extract who, when, where-type elements in mt output. In *Proceedings of the 11th Annual Conference of the European Association for Machine Translation (EAMT-2006)*. Citeseer, 2006.

Warren Weaver. Translation. In William N. Locke and A. Donald Boothe, editors, *Machine Translation of Languages*, pages 15–23. MIT Press, Cambridge, MA, 1949. Reprinted from a memorandum written by Weaver in 1949.

Appendix A
Full MQM taxonomy and decision tree

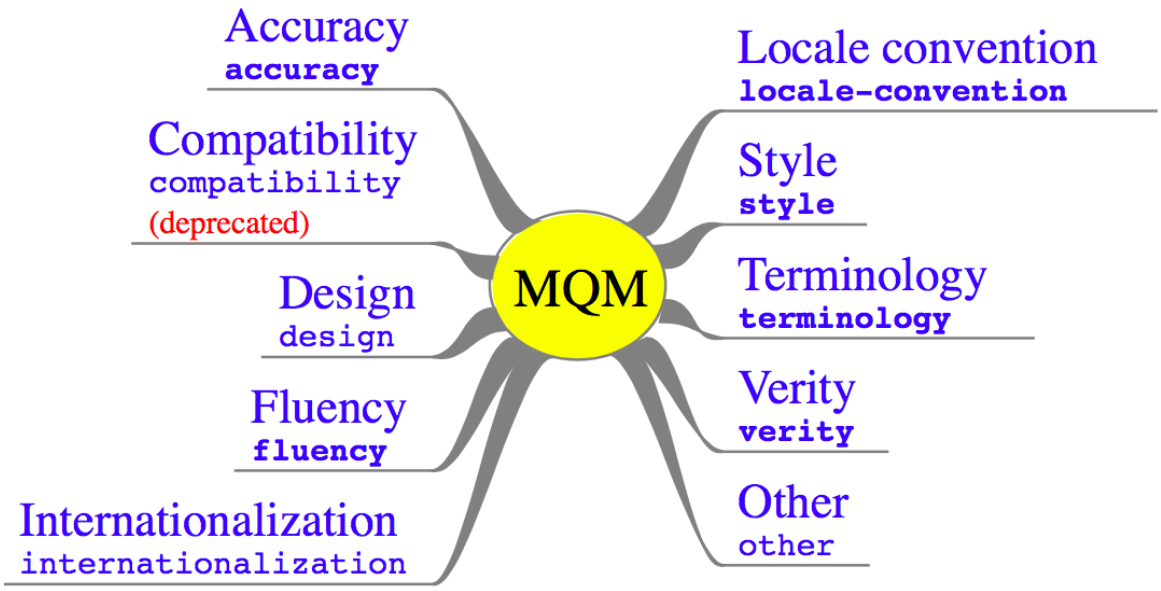


Figure A.1: Graphic of all the top-level MQM categories

Only the full Accuracy and Fluency trees will be presented here, but as can be seen in Figure A.1[1] , there are many other top level issue types like *Design*, which includes issues related to the physical presentation of text, typically in a "rich text" or "markup" environment; *Internationalisation*, which covers areas related to the preparation of the source content for subsequent translation or localisation; *Locale convention*, which relates to the formal compliance of content with locale-specific conventions, such as use of proper number formats; *Style*, which relates to what is commonly known as "Style", defined both formally (in style guides) and informally (e.g., a "light style" or an "engaging style"); *Terminology*, which relates to the use of domain- or organisation-specific terminology; and *Verity*, which relates to the suitability of content for the target locale and audience.

[1] `http://www.qt21.eu/mqm-definition/issues-list-2015-12-30.html`

Multidimensional Quality Metrics (MQM): Full Decision Tree

The Multidimensional Quality Metrics (MQM) Framework provides a hierarchical categorization of error types that occur in translated or localized products. Based on a detailed analysis of existing translation quality metrics, it provides a flexible typology of *issue types* that can be applied to analytic or holistic translation quality evaluation tasks. Although the full MQM issue tree (which, as of November 2014, contains 115 issue types categorized into five major branches) is not intended to be used in its entirety for any particular evaluation task, this overview chart presents a "decision tree" suitable for selecting an issue type from it. In practical terms, however, an individual metric would have a smaller decision tree that covers just the issues contained in that metric.

To use the decision tree start with the first question and follow the appropriate answers until a specific issue type is reached.

Start here →

General 1
Is the issue related to a difference in meaning between the source and target?
- Yes: Go to **Accuracy**
- No: Go to **General 2**

General 2
Is the issue related to the linguistic or mechanical formulation of the content
- Yes: Go to **Fluency**
- No: Go to **General 3**

General 3
Is the issue related to the appropriateness of the content for the target audience or locale (separate from whether it is translated correctly)?
- Yes: Go to **Verity**
- No: Go to **General 4**

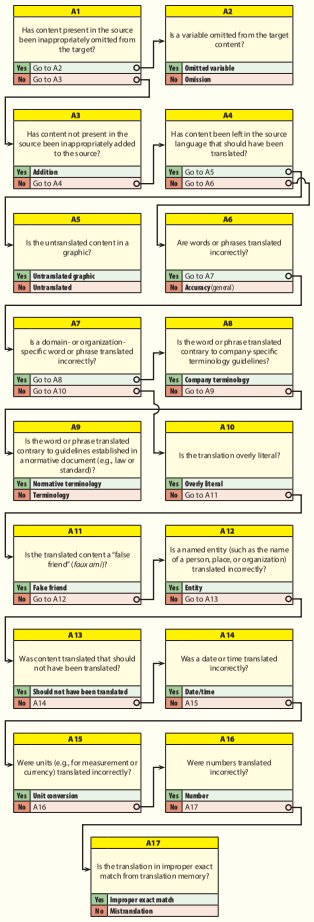
General 4
Is the issue related to the presentational/display aspects of the content?
- Yes: Go to **Design**
- No: Go to **General 5**

General 5
Is the issue related to whether or not the content was set up properly to support subsequent translation/adaptation?
- Yes: Go to **Internationalization**
- No: Go to **General 6**

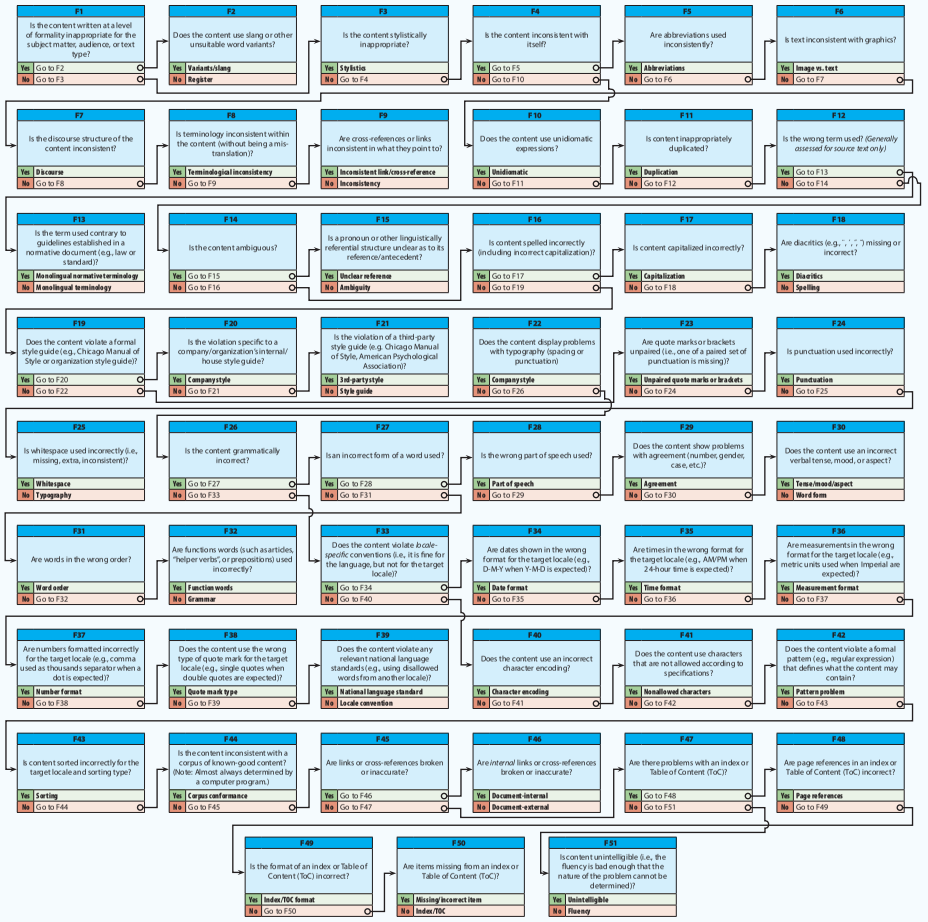
General 6
Is the issue addressed in the **Compatability** branch
- Yes: Go to **Compatability**
- No: **Other**

Note: If the issue is found in the Compatability branch it is probably not a translation product issue, but instead a process issue and would normally not be addressed in MQM. If the issue is **Other**, it may not be a translation-related issue since translation-related issues would normally fall into one of the major branches.

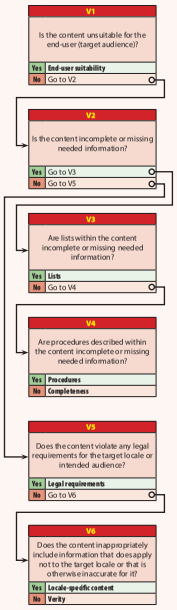
Accuracy



Fluency



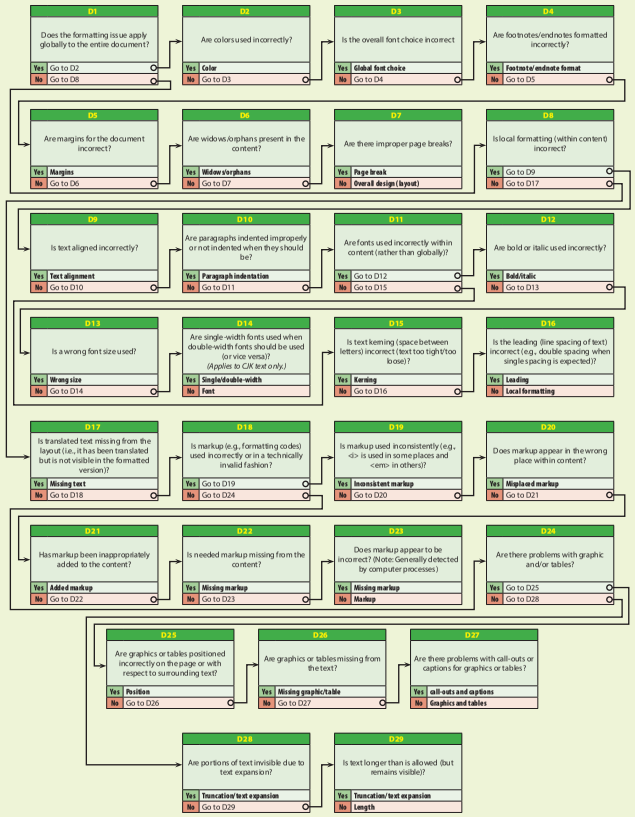
Verity



Internationalization

Subtypes of **Internationalization** are currently undefined.

Design



Compatability (deprecated) – These issues are included primarily for compatability with the LISA QA Model

Application compatibility • Bill of materials/runlist • Book-building sequence • Covers • Deadline • Delivery • Does not adhere to specifications • Embedded text • File format • Functional • Output device • Printing • Release guide • Spines • Style, publishing standards • Terminology, contextually inappropriate

To Learn More:

MQM

http://www.qt21.eu

MQM definition: http://qt21.eu/mqm-definition/

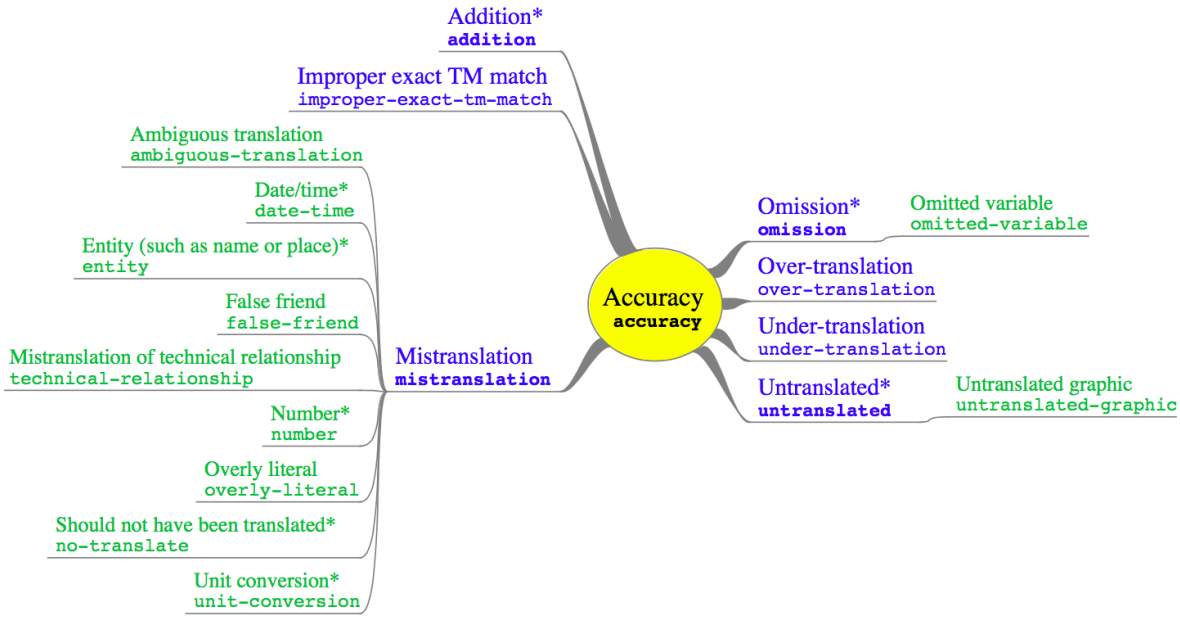


Figure A.2: Graphic representing the full branch of the *Accuracy* top-level MQM category

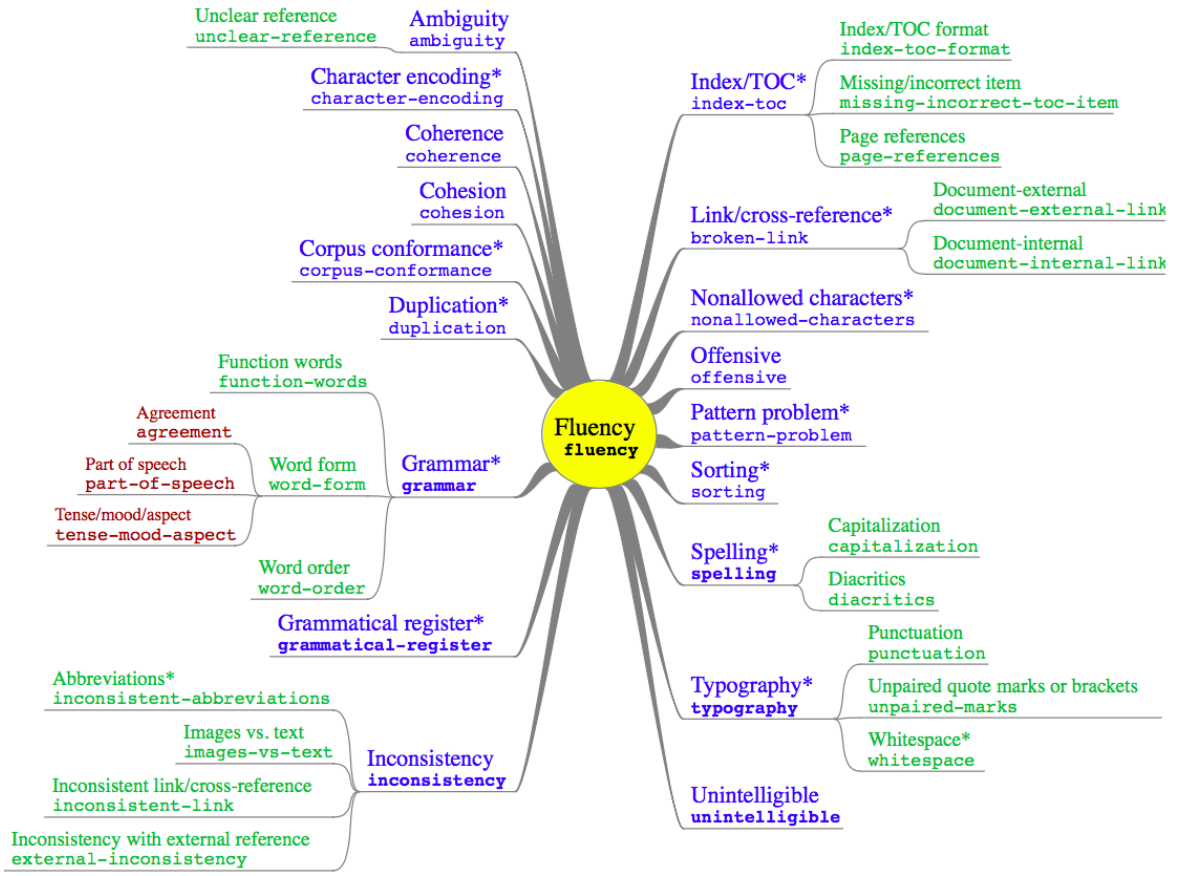


Figure A.3: Graphic representing the full branch of the *Fluency* top-level MQM category

Appendix B
Individual annotation normalisation results with alternative agreement counts

Given that there were many forks in the road towards the analysis of annotation data, several decisions had to be made on how to best analyse and present the gathered data. This section of the Appendix contains tables that present counts and calculations, both individual and alternative to Table 4.6.

Tables B.1 and B.2 show the normalised data comparison of the two systems according to Annotator_1. The difference between them is that *Agreement* errors were counted differently, as described in Subsection 4.6.2. Tables B.3 and B.4 show the normalised data comparison of the two systems according to Annotator_2. The difference between them is also the different counting of *Agreement* errors. It is interesting to note that the p-values obtained from Annotator_1's annotations are consistent in both variations - differences between the same error categories are statistically significant, regardless of how *Agreement* errors were counted. However, p-values obtained from Annotator_2's annotations are slightly less consistent - the difference between the bottom most node in the hierarchy, *Agreement_case*, is statistically significant regardless of how *Agreement* errors were counted, but all of its parents up until *Fluency* are not statistically significant in the case where *Agreement* errors are counted in accordance with the annotation span. Finally, table B.5 presents a concatenated version of the results, similar to Table 4.6, but with *Agreement* counted according to the annotation span. It highlights that, regardless of the fact that agreement tokens were counted differently, the same error categories have statistically significant differences.

This comparison between annotators and annotation count variations shows that, even if the annotators are looked at separately, they both arrive at the same two conclusions. If they are concatenated, they might enforce each other's differences: though Annotator_2's annotations did not show a statistically significant difference in *Mistranslation*, once concatenated, this it becomes a category that is also has a statistically significant difference between the two systems.

	PBMT		Factored PBMT				
Error type	**No error**	**Error**	**No error**	**Error**	χ^2	p	ϕ
Accuracy	1749	188	1772	155	3.3	0.0693	0.0292
Mistranslation	1776	161	1802	125	4.69	**0.0303**	0.0348
Omission	1924	13	1916	11	0.16	0.6892	0.0064
Addition	1932	5	1916	11	-	0.1417	-
Untranslated	1928	9	1919	8	-	1	-
Fluency	1624	313	1676	251	7.61	**0.0058**	0.0444
Unintelligible	1910	27	1902	25	0.07	0.7913	0.0043
Register	1921	16	1911	16	0	1	0
Spelling	1937	0	1927	0	-	1	-
Grammar	1667	270	1717	210	8.21	**0.0042**	0.0461
Word order	1907	30	1900	27	0.14	0.7083	0.006
Function words	1912	25	1897	30	0.49	0.4839	0.0113
Extraneous	1933	4	1923	4	-	1	-
Incorrect	1918	19	1904	23	0.41	0.522	0.0103
Missing	1935	2	1924	3	-	0.6865	-
Word form	1723	214	1775	152	11.25	**0.0008**	0.054
Part of speech	1925	12	1915	12	0	1	0
Tense...	1902	35	1897	30	0.37	0.543	0.0098
Agreement	1771	166	1818	109	12.4	**0.0004**	0.0566
Number	1915	22	1910	17	0.62	0.431	0.0127
Gender	1913	24	1907	20	0.37	0.543	0.0098
Case	1827	110	1871	56	18.06	**<0.0001**	0.0684
Total errors	1436	501	1521	406	16.91	**<0.0001**	0.0668

Table B.1: Normalised annotation data from Annotator_1 with **all error tokens**: each system's total number of tokens with and without errors, where *Agreement* errors were counted in accordance with the annotation span. Includes with statistical significance test results (chi-squared (χ^2), p-value and effect size ϕ). Statistical significance is marked with bold where p-value is <0.05.

	PBMT		Factored PBMT				
Error type	**No error**	**Error**	**No error**	**Error**	χ^2	p	ϕ
Accuracy	1749	188	1772	155	3.3	0.0693	0.0292
Mistranslation	1776	161	1802	125	4.69	**0.0303**	0.0348
Omission	1924	13	1916	11	0.16	0.6892	0.0064
Addition	1932	5	1916	11	2.29	0.1302	0.0243
Untranslated	1928	9	1919	8	-	1	-
Fluency	1574	363	1626	301	6.61	**0.0101**	0.0414
Unintelligible	1910	27	1902	25	0.07	0.7913	0.0043
Register	1921	16	1911	16	0	1	0
Spelling	1937	0	1927	0	-	1	-
Grammar	1617	320	1667	260	6.94	**0.0084**	0.0424
Word order	1907	30	1900	27	0.14	0.7083	0.006
Function words	1912	25	1897	30	0.49	0.4839	0.0113
Extraneous	1933	4	1923	4	-	1	-
Incorrect	1918	19	1904	23	0.41	0.522	0.0103
Missing	1935	2	1924	3	-	0.6865	-
Word form	1672	265	1724	203	8.98	**0.0027**	0.0482
Part of speech	1925	12	1915	12	0	1	0
Tense...	1902	35	1897	30	0.37	0.543	0.0098
Agreement	1719	218	1767	160	9.53	**0.002**	0.0497
Number	1903	34	1903	24	1.7	0.1923	0.021
Gender	1907	30	1891	36	0.59	0.4424	0.0124
Case	1795	142	1851	76	20.82	**<0.0001**	0.0734
Total errors	1386	551	1471	456	11.47	**0.0007**	0.0545

Table B.2: Normalised annotation data from Annotator_1 with ***Agreement* as two tokens**: each system's total number of tokens with and without errors, where *Agreement* errors were counted as spanning two tokens. Includes with statistical significance test results (chi-squared (χ^2), p-value and effect size ϕ). Statistical significance is marked with bold where p-value is <0.05.

	PBMT		Factored PBMT				
Accuracy	1718	181	1753	136	6.71	**0.0096**	0.0421
Mistranslation	1771	128	1784	105	2.29	0.1302	0.0246
Omission	1877	22	1877	12	2.91	0.088	0.0277
Addition	1882	17	1881	8	3.21	0.0732	0.0291
Untranslated	1885	14	1878	11	0.35	0.5541	0.0096
Fluency	1562	337	1585	304	1.84	0.175	0.022
Unintelligible	1880	19	1867	22	0.24	0.6242	0.008
Register	1889	10	1883	6	0.98	0.3222	0.0161
Spelling	1896	3	1885	4	-	0.7256	-
Grammar	1594	305	1617	272	2.03	0.1542	0.0231
Word order	1845	54	1852	37	3.16	0.0755	0.0289
Function words	1889	10	1883	6	0.98	0.3222	0.0161
Extraneous	1896	3	1887	2	-	1	-
Incorrect	1892	7	1886	3	-	0.3431	-
Missing	1899	0	1888	1	-	0.4987	-
Word form	1658	241	1660	229	0.28	0.5967	0.0086
Part of speech	1897	2	1885	4	-	0.4516	-
Tense...	1873	26	1868	21	0.51	0.4751	0.0116
Agreement	1688	211	1686	203	0.13	0.7184	0.0059
Number	1860	39	1846	43	0.22	0.639	0.0076
Gender	1859	40	1837	52	1.67	0.1963	0.021
Case	1807	92	1837	52	11.33	**0.0008**	0.0547
Total errors	1371	528	1449	440	10.13	**0.0015**	0.0517

Table B.3: Normalised annotation data from Annotator_2 with **all error tokens**: each system's total number of tokens with and without errors, where *Agreement* errors were counted in accordance with the annotation span. Includes with statistical significance test results (chi-squared (χ^2), p-value and effect size ϕ). Statistical significance is marked with bold where p-value is <0.05.

	PBMT		Factored PBMT				
Error type	**No error**	**Error**	**No error**	**Error**	χ^2	p	ϕ
Accuracy	1718	181	1753	136	6.71	**0.0096**	0.0421
Mistranslation	1771	128	1784	105	2.29	0.1302	0.0246
Omission	1877	22	1877	12	2.91	0.088	0.0277
Addition	1882	17	1881	8	3.21	0.0732	0.0291
Untranslated	1885	14	1878	11	0.35	0.5541	0.0096
Fluency	1621	278	1672	217	8.28	**0.004**	0.0468
Unintelligible	1880	19	1867	22	0.24	0.6242	0.008
Register	1889	10	1883	6	0.98	0.3222	0.0161
Spelling	1896	3	1885	4	-	0.7256	-
Grammar	1653	246	1704	185	9.38	**0.0022**	0.0498
Word order	1845	54	1852	37	3.16	0.0755	0.0289
Function words	1889	10	1883	6	0.98	0.3222	0.0161
Extraneous	1896	3	1887	2	-	1	-
Incorrect	1892	7	1886	3	-	0.3431	-
Missing	1899	0	1888	1	-	0.4987	-
Word form	1717	182	1747	142	5.17	**0.023**	0.0369
Part of speech	1897	2	1885	4	-	0.4516	-
Tense...	1873	26	1868	21	0.51	0.4751	0.0116
Agreement	1747	152	1773	116	5	**0.0253**	0.0363
Number	1875	24	1869	20	0.35	0.5541	0.0096
Gender	1881	18	1865	24	0.9	0.3428	0.0154
Case	1819	80	1843	46	9.31	**0.0023**	0.0496
Total errors	1440	459	1536	353	16.91	**<0.0001**	0.0668

Table B.4: Normalised annotation data from Annotator_2 with ***Agreement* as two tokens**: each system's total number of tokens with and without errors, where *Agreement* errors were counted as spanning two tokens. Includes with statistical significance test results (chi-squared (χ^2), p-value and effect size ϕ). Statistical significance is marked with bold where p-value is <0.05.

	PBMT		Factored PBMT				
Error type	**No error**	**Error**	**No error**	**Error**	χ^2	p	ϕ
Accuracy	3467	369	3525	291	9.65	**0.0019**	0.0355
Mistranslation	3547	289	3586	230	6.87	**0.0088**	0.03
Omission	3801	35	3793	23	2.44	0.1183	0.0179
Addition	3814	22	3797	19	0.21	0.6468	0.0052
Untranslated	3813	23	3797	19	0.36	0.5485	0.0069
Fluency	3186	650	3261	555	8.31	**0.0039**	0.033
Unintelligible	3790	46	3769	47	0.02	0.8875	0.0016
Register	3810	26	3794	22	0.31	0.5777	0.0064
Spelling	3833	3	3812	4	-	0.7257	-
Grammar	3261	575	3334	482	8.94	**0.0028**	0.0342
Word order	3752	84	3752	64	2.65	0.1035	0.0186
Function words	3801	35	3780	36	0.02	0.8875	0.0016
Extraneous	3829	7	3810	6	0.07	0.7913	0.003
Incorrect	3810	26	3790	26	0	1	0
Missing	3834	2	3812	4	-	0.4518	-
Word form	3381	455	3435	381	6.93	**0.0085**	0.0301
Part of speech	3822	14	3800	16	0.14	0.7083	0.0043
Tense...	3775	61	3765	51	0.85	0.3566	0.0105
Agreement	3459	377	3504	312	6.37	**0.0116**	0.0289
Number	3775	61	3756	60	0	1	0
Gender	3772	64	3744	72	0.52	0.4708	0.0082
Case	3634	202	3708	108	29.2	**<0.0001**	0.0618
Total errors	2807	1029	2970	846	22.41	**<0.0001**	0.0541

Table B.5: Normalised annotation data from both annotators concatenated, with **all error tokens**: each system's total number of tokens with and without errors, where *Agreement* errors were counted in accordance with the annotation span. Includes statistical significance test results (chi-squared (χ^2), p-value and effect size ϕ). Statistical significance is marked with bold where p-value is <0.05.

Fine-grained Human Evaluation of an English to Croatian Hybrid Machine Translation System

Abstract

This research compares two approaches to statistical machine translation - pure phrase-based and factored phrase-based - by performing a fine-grained manual evaluation via error annotation of the systems' outputs, and subsequently analysing the results. The error types considered in the annotation are compliant with the multidimensional quality metrics (MQM), and the annotation is performed by two annotators. Inter-annotator agreement is relatively high for such a task, and the results show that the factored system, i.e. hybrid system, performs much better, reducing the amount of errors produced by the pure phrase-based system by 20%.

Keywords: machine translation, human evaluation, PBMT, factored PBMT, MQM, inter-annotator agreement, statistical machine translation, error annotation

Detaljna evaluacija jezično oplemenjenog statističkog sustava za strojno prevodenje s engleskog na hrvatski

Sažetak

U ovom se istraživanju usporeduju dva pristpa strojnom prevodenju - klasično statističko strojno prevodenje (na temelju fraze) i hibridno strojno prevodenje (statističko s nadodanim jezičnim znanjem) - tako što se provodi detaljna manualna evaluacija preko označavanja pogrešaka u prijevodima obaju sustava, te analizom rezultata takvog označavanja. Vrste pogrešaka koje se uzimaju u obzir pri označavanju uskladene su sa multidimenzionalnim mjerama kvaliteta (MQM), a označavanje provode dva anotatora. Slaganje medu anotatorima se pokazalo relativno visokim za ovakav zadatak, a rezultati pokazuju da hibridni sustav generira manje pogredaka pri prijevodu, čak 20% manje od klasičnog statističkog sustava.

Ključne riječi: strojno prevodenje, manualna evaluacija, MQM, slaganje izmedu anotatora, statističko strojno prevodenje, označavanje pogrešaka