Relations between comprehensibility and adequacy errors in machine translation output

Maja Popović
ADAPT Centre, School of Computing
Dublin City University, Ireland
maja.popovic@adaptcentre.ie

Abstract

This work presents a detailed analysis of translation errors perceived by readers as comprehensibility and/or adequacy issues. The main finding is that good comprehensibility, similarly to good fluency, can mask a number of adequacy errors. Of all major adequacy errors, 30% were fully comprehensible, thus fully misleading the reader to accept the incorrect information. Another 25% of major adequacy errors were perceived as almost comprehensible, thus being potentially misleading. Also, a vast majority of omissions (about 70%) is hidden by comprehensibility. Further analysis of misleading translations revealed that the most frequent error types are ambiguity, mistranslation, noun phrase error, word-by-word translation, untranslated word, subject-verb agreement, and spelling error in the source text. However, none of these error types appears exclusively in misleading translations, but are also frequent in fully incorrect (incomprehensible inadequate) and discarded correct (incomprehensible adequate) translations. Deeper analysis is needed to potentially detect underlying phenomena specifically related to misleading translations.

1 Introduction

While automatic evaluation metrics are very important and invaluable tools for rapid development of machine translation (MT) systems, they are only a substitution for human assessment of translation quality. Various methods have been proposed and used for the human evaluation of MT quality by assigning overall scores to MT outputs, such as (ALPAC, 1966; White et al., 1994; Koehn and Monz, 2006; Callison-Burch et al., 2007; Roturier and Bensadoun, 2011; Graham et al., 2013; Barrault et al., 2019), and all of them rely on at least one of the three translation quality criteria: comprehensibility (comprehension, intelligibility), adequacy (fidelity, semantic accuracy), and fluency (grammaticality). Comprehensibility reflects the degree to which a translated text can be understood, adequacy reflects the degree to which the translation conveys the meaning of the original text in the source language, and fluency reflect the grammar of the translated text. The raters are usually asked to assign an overall score for the given translation criterion. In order to get more details about translation performance, error classification and analysis emerged in the field of MT (Vilar et al., 2006; Lommel et al., 2014; Klubička et al., 2018; Van Brussel et al., 2018).

However, there is less work dealing with human perception of MT quality and errors. For statistical phrase-based MT systems (SMT), Kirchhoff et al. (2014) and Federico et al. (2014) were identifying error types which are mostly disliked by readers. In the last five years, systems based on artificial neural networks (NMT) have become the new state of the art. Several evaluation studies, such as (Castilho et al., 2017; Klubička et al., 2018; Van Brussel et al., 2018) reported that these systems are able to produce more fluent and readable translations, but that they are still sufferring from adequacy issues. In addition, many participants mentioned that good fluency of NMT outputs makes it more difficult to spot adequacy errors such as omissions or mistranslations. Such "fluently inadequate" errors may mislead readers into trusting the content based on fluency alone, especially when surrounded by fluent and adequate parts of a text (Martindale and Carpuat, 2018). Automatic identification of such errors for both SMT and NMT systems has been investigated in (Martindale et al., 2019) and it is confirmed that these errors appear much more often in NMT system.

To the best of our knowlegde, comprehensibility, while being a very important translation quality factor, has not been investigated in depth yet. It

Proceedings of the 24th Conference on Computational Natural Language Learning, pages 256–264

https://doi.org/10.18653/v1/P17

should be stressed that comprehensibility is very different from fluency – a fluent text can be incomprehensible (for example "Colorless green ideas sleep furiously."), and vice versa (for example "All these experiment was carry out this year.").

Our main research questions are:

RQ1 Are there "comprehensible inadequate" translations which are misleading human readers so that they fully trust the MT output despide adequacy errors?

In other words: how many adequacy errots are perceived as comprehensible?

RQ2 If the answer to the RQ1 is "yes", which types of translation errors are mainly related to these translations?

As a first step, a group of evaluators annotated problematic parts of the given machine translated text. They were not asked to assign any error labels, only to mark the parts of the text which they perceived as problematic for the given translation criterion. They first annotated all comprehensibility issues, and after about two weeks, all adequacy issues. For each criterion, they were asked to distinguish major and minor issues. We then analysed all major issues in order to examine relations between comprehensibility and adequacy and identify error types.

The analysis was carried out on English user reviews (as a case of "mid-way" genre between formal and informal written language) translated into Croatian and Serbian (as a case of mid-size less-resourced morphologically rich European languages).

It is worth noting that the aim of this work is not to compare MT systems, nor to estimate their overall performance for the given language pairs and domain in order to potentially improve them. The aim of this work is to explore relations between two aspects of human perception of translation quality.

2 Related work

Lot of research on MT evaluation deals with classification and analysis of MT errors, for example (Vilar et al., 2006; Farrús et al., 2010; Stymne and Ahrenberg, 2012; Lommel et al., 2014; Klubička et al., 2018). Few papers deal with human perception of these errors, but neither of them defines precisely which criterion is the translation quality based on.

Kirchhoff et al. (2014) uses conjoint analysis to investigate user preferences for error types of SMT systems. First, the errors in MT outputs were annotated, and then MT outputs with different error types were given to the crowd evaluators. They were asked to choose the MT output which they like best and to give the reason for their preference. One of the findings is that the frequencies of error types are not related to the user preferences. The most dispreferred error type was word order error, although it was the least frequent one. It was followed by word sense errors (ambiguity), then morphological errors (most frequent ones), whereas errors in function words were the most tolerable.

A similar study on SMT outputs based on linear mixed-effects models is described in (Federico et al., 2014), aiming to estimate the impact of different translation errors to the overall translation quality. For each MT output, experts were asked to assign a score on a 5-point scale while other experts annotated the errors. The results confirmed that the frequency of errors of a given type does not correlate with human preferences. Another finding is that omissions and mistranslations have the highest impact on the overall translation quality. In addition, it is observed that certain combinations of errors have less impact than each of those error types ocurring in isolation.

In the last few years, with the emergence of NMT systems which generate much more fluent and readable outputs but still are prone to adequacy errors, some studies have concentrated on investigating adequacy and fluency errors. Martindale and Carpuat (2018) carried out a survey to determine how users respond to good translations compared to translations that are either adequate but not fluent, or fluent but not adequate. This study showed that users strongly disliked disfluent translations, but were much less bothered with adequacy errors. Therefore, it was concluded that fluent translations with adequacy errors can mislead the reader to trust an incorrect meaning. Automatic identification of these misleading "fluently inadequate" translations using source text, reference human translation and MT output was proposed in (Martindale et al., 2019), and the main finding was that NMT systems generate more misleading translations than SMT systems. However, the question about how many adequacy errors are actually hidden by fluency remained open.

To the best of our knowledge, the relation be-

tween adequacy and comprehensibility has not been investigated yet. Comprehensibility, similarly to fluency, has an immediate effect on the reader, while adequacy problems can be perceived only if the reader has access to the source text or to a correct translation to find out that the meaning is wrong. This means that comprehensibility may have the same misleading effect making the reader accept an incorrect information. On the other hand, because comprehensibility is different than fluency (fluent sentences can be incomprehensible and vice versa), the effects might be different.

3 Data set

Our analysis has been carried out on written user-generated content, namely user reviews. Two types of publicly available user reviews written in English have been analysed: IMDb movie reviews[1] (Maas et al., 2011) and Amazon product reviews[2] (McAuley et al., 2015). A set of those user reviews was translated into Croatian and Serbian, two closely related mid-size less-resourced morphologically rich European languages. The reviews were translated[3] by three on-line systems: Google Translate[4], Bing[5] and Amazon translate[6]. The analysed text consists of a mixture of MT outputs from the three systems including 222 translated reviews consisting of about 1500 sentences (segments) and 19837 untokenised words in total.

This text was then given to the annotators to mark comprehensibility and adequacy issues, and the process is described in details in the next section.

The annotated text is publicly available under the Creative Commons CC-BY licence.[7]

3.1 Annotating comprehensibility and adequacy issues

As mentioned in Introduction, comprehensibility reflects the degree to which a translated text can be understood, and adequacy reflects the degree to which the translation conveys the meaning of the original text in the source language. Comprehension should be assessed without access to the original text in the source language (or a correct translation), while the original text (or a correct translation) is mandatory for adequacy. Therefore, each annotator first completed the annotation of comprehension issues while reading only the translation. After completing (usually after about two weeks), they annotated adequacy issues by comparing the translation with the original source text. For each criterion, the annotators were asked to distinguish two levels of issues: major issues and minor issues. While for this particular study we are interested only in major issues, we did not want any errors to remain unannotated. The following guidelines were given to the annotators:

Comprehensibility:

- mark all parts of the text (single words, small or long phrases, or entire sentences) which are not understandable (it does not make sense, it is not clear what it is about, etc.) as major issues;
- mark all parts of the text (again: words, phrases or sentences) which seem understandable but contain grammatical or stylistic errors as minor issues;
- if it seems that something is missing, add "XXX" to the corresponding position.

Adequacy:

- mark all parts of the translation (single words, small or long phrases, or entire sentences) which have different meaning than the original English text as major issues;
- mark all parts of the translation (again: words, phrases or sentences) which do not actually change the meaning of the source text, but contain sub-optimal lexical choices or grammar errors as minor issues;
- if some parts of the original English text are missing in the translation, add "XXX" to the corresponding position in the translation;
- if there are any errors in the source language[8] (spelling or grammar errors, etc.), mark its

[1] https://ai.stanford.edu/~amaas/data/sentiment/

[2] http://jmcauley.ucsd.edu/data/amazon/

[3] at the end of January 2020

[4] https://translate.google.com/

[5] https://www.bing.com/translator

[6] https://aws.amazon.com/translate/

[7] https://github.com/m-popovic/QRev-annotations/tree/master/initial-analysis

[8] Detailed instructions for errors in the source text are particularly relevant for evaluating user generated content.

translation as major or minor issue if it does not correspond to the *intented* English word even though it is a correct translation of the erroneous English word.

The annotators were seeing the entire reviews during the process, not only isolated segments or blocks of 2-3 segments. In this way, it was ensured that the annotators were able to spot any context-dependent issues.

We wanted the texts to be annotated by a reliable group of readers which is neither too homogeneous as a group of professional translators nor too heterogeneous as crowd evaluators. Therefore, the annotation was performed by computational linguistics researchers and students, fluent in the source language and native speakers of the target language. They had different backgrounds, coming from technical studies, translation studies as well as from humanities.

Because the annotators were not asked to perform any fine-grained categorisation, the inter-annotator agreement was high – annotators assigned identical issue tags to more than 70% of words. More details about the annotation process can be found in (Popović, 2020).

4 Analysis of comprehensibility and adequacy issues

Table 1 presents overall percentages[9] of words perceived as issues, separately for each of the two translation criteria. In total (including both target languages and all three MT systems), 9.5% words in the text were perceived as incomprehensible, and the meaning of 9.9% words was changed in the translation process. As for minor issues, 13.5% words were perceived as slightly difficult to understand, and 12.8% were not translated in the optimal way.

It can be noted that the overall amounts of comprehensibility and adequacy issues are similar. However, it does not necessarily mean that the majority of words is perceived both as incomprehensible and inadequate. Therefore, we examined major comprehensibility and adequacy issues in depth.

[9] raw counts divided by the total number of words in the text including those without issues and the omission marks "XXX"

quality aspect	grade	raw count	%words
comprehension	major	1887	9.5
	minor	2673	13.5
adequacy	major	1963	9.9
	minor	2539	12.8

Table 1: Raw counts and percentages of words (normalised by the total number of words, including those without issues and the omission marks "XXX") perceived as problematic for comprehensibility and adequacy.

4.1 Relations between different types of issues

In order to determine presence or absence of misleading translations, we explored the following cases of different relations between comprehensibility and adequacy errors:

- only major adequacy issue A_{maj}
  (comprehensible inadequate translation)
  – incorrect information is accepted –

  The meaning of the original text is changed but the translation is readable and comprehensible. The reader feels comfortable with the text and does not notice any problem, thus accepting the incorrect meaning.

- A_{maj}+C_{min} – major adequacy and minor comprehension issues
  (almost comprehensible inadequate translation)
  – incorrect information can be accepted –

  The meaning of the original text is changed, and the reader finds this incorrect meanining slightly difficult to understand. The reader is therefore susceptible to accept this incorrect meaning.

- A_{maj}+C_{maj} – both major issues
  (incomprehensible inadequate translation)
  – incorrect information is discarded –

  The meaning of the original text is changed, and the reader is not able to understand this changed meaning. The reader clearly notices that there is something wrong with the text.

- C_{maj}+A_{min} – major comprehension and minor adequacy issues
  (incomprehensible almost adequate translation)
  – almost correct information is discarded –

  The meaning of the original text is basically conveyed to the translation, only not in an

issue types	affected words		
	raw count	%words	%A_{maj}
only A_{maj}	588	2.96	30.0
A_{maj}+C_{min}	490	2.47	24.9
A_{maj}+C_{maj}	885	4.46	45.1
C_{maj}+A_{min}	342	1.72	
only C_{maj}	660	3.33	
C_{min}+A_{min}	1254	6.32	
only C_{min}	929	4.68	
only A_{min}	943	4.75	

Table 2: Raw counts and percentages of words (normalised by the total number of words, including those without issues and the omission marks "XXX") of all combinations of perceived issue types. For cases involving major adequacy issues, the percentages normalised by the total number of major adequacy issues are shown, too, in order to estimate the portion of hidden adequacy issues. For the sake of completenes, the numbers are presented for minor issues, too, although they were not further analysed in this work.

optimal way, but the reader cannot understand it. The reader is thus missing some correct information.

- only major comprehension issue C_{maj} (incomprehensible adequate translation)
  – correct information is discarded –

  The meaning of the original text is correctly conveyed to the translation, but the reader cannot understand it. The reader is therefore not able to get the fully correct information.

Table 2 presents raw counts and percentages of words perceived in the described ways. For the sake of completeness, the numbers for minor issue types are shown as well. The numbers are generally in line with the findings of the previous work (Kirchhoff et al., 2014; Federico et al., 2014) regarding lack of correlation between the error frequency and perception of severity – in our texts, the frequencies of words perceived only as minor issues are higher than the frequencies of words perceived as major issues.

As already mentioned, minor issues were not further analysed in this work, because by definition they were not perceived as essential: either the meaning was preserved although not conveyed in the best way, or the translation was slightly difficult to understand, or both.

Misleading translations Table 2 shows that about 3% of words in the translated text are misleading, and 2.5% are potentially misleading. This means that of every 100 words in the translation, 3 are fully accepted by the reader although their meaning is not correct, and 2 can be potentially accepted. Furthermore, from all major adequacy errors in the text, only 45.5% are incomprehensible. About 30% of adequacy errors are fully hidden so that the reader does not notice any problem, and about 25% are partially hidden because the reader is not fully sure that s/he understands the text, but s/he is very susceptible to accept the meaning.

All in all, the portion of misleading translations is not negligible, so we continued our analysis by trying to identify error types associated with such translations. Also, we wanted to explore whether there are error types related (almost) exclusively to them.

4.2 Error types

For each (group of) word(s) perceived as compreensibility or adequacy major issue, we assigned an error type. The error types were not predefined by any particular error typology, but identified while looking into the text. It is worth noting that many error types were identified, but most of them are ocurring rarely in the text. Also, for some of the annotated words no particular error type could be defined, which is probably an effect of annotators' personal preferences. The most frequent error types perceived as misleading translations can be defined as follows:

- **ambiguity**

  The obtained translation for the given word is in principle correct, but not in the given context (word sense error).

- **mistranslation**

  The obtained translation for the given word is incorrect.

- **noun phrase**

  An English sequence consisting of a head noun and additional nouns and adjectives is incorrectly translated. Formation rules for Serbian and Croatian are rather different than for English and there is often no unique solution. The examples in the table below represent two English noun collocations and their reference translations into Serbian and Croatian together with the corresponding English glosses. This

type of issue is relevant for many Slavic languages.

language	NP1	NP2
en	grill cover	chocolate cake
sr/hr	poklopac za roštilj	čokoladni kolač
gloss	cover for grill	'chocolaty' cake

- **spelling error in source**

  A word in the original text in the source language has spelling errors which result in incorrect translation. This type of issue is especially relevant for user-generated content.

- **subject-verb agreement**

  A verb inflection in the translation denoting person does not correspond to the subject.

- **untranslated**

  A word in the source language is simply copied to the translated text.

- **word-by-word translation**

  A sequence of source words is translated as single words – the translation choice of each word looks random, both lexically and morphologically, without taking into account any context.

Table 3 shows these error types and their percentages for misleading translations. These error types are the certainly "dangerous" because they can easily mislead the reader to accept incorrect information. However, the very same error types are often perceived as fully incorrect (incomprehensible inadequate), too. Furthermore, they (except of untranslated words) even often lead to discarding correct information (incomprehensible adequate). Further in-depth analysis is needed to determine whether there are some underlying phenomena related exclusively to the misleading translations.

Five examples of different perceptions of ambiguity errors, noun phrase errors and word-by-word translations are presented in Table 4. All sentences except 3) have misleading parts (fully misleading marked as red and potentially misleading as green). In the sentences 1) and 2) there is only one misleading ambiguous word. The incorrectly chosen variants of these words are fully comprehensible so that without the source text, the reader was not able to figure out that the information is not correct. On the other hand, the ambiguous word in the sentence 3), together with the noun phrase, is perceived as both incomprehensible and inadequate (marked as violet). Sentences 4) and 5) illustrate how different parts of a phrase translated word-by-word are perceived in different ways: violet denotes fully incorrect, red denotes misleading, and cyan denotes discarding almost correct translation. It might be worth noting that all sentences are perfectly fluent except the sentence 3) which is very disfluent.

Propagation effect Table 3 also shows that there is a strong effect of *propagation* for comprehensibility – many correct words are perceived as incomprehensible because of errors in surrounding words. In many cases, the reader finds the whole sentence incomprehensible. An example of propagation can be seen in Table 5. All words in bold are correct, but all were perceived as major comprehensibility issues due to different types of errors in surrounding words: a red misleading omission, a fully incorrect violet word, and an incomprehensible group of almost correct cyan words. It should be mentioned that for some adequacy errors, annotators also marked one or two neighbouring words which were not really incorrect, but that happened very rarely.

Omissions Since several studies reported that the omissions are generally problematic to spot without access to the source text, we compared the frequencies of omissions percieved only as comprehensibility issue, only as adequacy issue, and as both (regardless of the severity grade).

Table 6 confirms the previous findings: a vast majority of omissions (71.5%) was perceived only as adequacy error. Only 9% of actual omissions were also perceived as comprehensibility issues. Apart from this, 19% of omissions were perceived as exclusively comprehensibility issues and are not related to anything actually omitted from the source text. The most probable reason is the influence of other surrounding errors, but further analysis is needed to better understand this effect.

5 Conclusions

This work presents the results of a detailed analysis of translation errors perceived by readers as major comprehensibility and/or major adequacy issues. The main finding is that good comprehensibility, similarly to good fluency, can mask a number of adequacy errors. Of all major adequacy errors, 30% were fully comprehensible, thus fully misleading the reader to accept the incorrect information. An-

incorrect information is:			discarded information is:	
accepted only A_{maj}	potentially accepted A_{maj}+C_{min}	discarded A_{maj}+C_{maj}	almost correct C_{maj}+A_{min}	correct only C_{maj}
ambiguity 24.0	ambiguity 24.8	ambiguity 26.7	ambiguity 16.4	*propagation 33.3*
mistranslation 6.0	mistranslation 8.9	mistranslation 8.2	noun phrase 8.4	ambiguity 14.4
word-by-word 5.0	noun phrase 8.4	noun phrase 6.9	mistranslation 5.8	noun phrase 7.0
noun phrase 4.4	untranslated 8.2	untranslated 6.7	{noun case 5.8}	word-by-word 4.1
source spelling 3.8	word-by-word 5.4	word-by-word 5.5	word-by-word 5.8	mistranslation 3.7
subject-verb 3.8	(subject-verb 3.2)	(subject-verb 4.5)	subject-verb 4.9	(subject-verb 2.0)
untranslated 3.8	(source spelling 2.4)	(source spelling 3.5)	{POS ambiguity 3.5}	(source spelling 1.2)

Table 3: The most frequent error types perceived as a particular issue combination. The numbers represent percentages of the error type perceived as the issue combination – 24.0% of all comprehensible inadequate translations (accepted incorrect information) are ambiguity errors, 6.0% are mistranslations, etc. Parentheses indicate that the error type was not in the top list for the given issue combination, but it is presented for comparison because it is in the top list for misleading traslations.

other 25% of major adequacy errors were perceived as almost comprehensible, thus being potentially misleading. In addition, a vast majority of omissions (about 70%) is hidden by comprehensibility.

Further analysis of those misleading translations was carried out in order to find out which types of translation errors are perceived in this way. Ambiguous words, mistranslations, noun phrases, untranslated words, word-by-word translations, subject-verb agreement and spelling errors in the original text were identified as the most frequent error types in misleading translations. Although noun phrase problems are typical for Slavic languages and errors in the source text are typical for user generated content, the rest of the error types is rather general. However, none of these error types is exclusively related to misleading translations, but are also frequent in fully incorrect (incomprehensible inadequate) and discarded correct (incomprehensible adequate) translations. Deeper analysis is needed to potentially detect underlying phenomena specifically related to misleading translations.

Apart from the obvious directions for future work such as analysing more texts and including more language pairs and domains, the presented analysis can be expanded in the following directions: including fluency in the analysis, including all minor issues in the analysis, further analysis of omissions, and investigating co-ocurrences of different error types. Another experiment could include monolingual annotators for comprehensibility in order to completely eliminate potential influence of knowlegde of the source language.

Acknowledgments

This research is being conducted with the financial support of the European Association for Machine Translation (EAMT) under its programme "2019 Sponsorship of Activities" at the ADAPT Research Centre at Dublin City University. The ADAPT SFI Centre for Digital Media Technology is funded by Science Foundation Ireland through the SFI Research Centres Programme and is co-funded under the European Regional Development Fund (ERDF) through Grant 13/RC/2106.

We would like to thank all the evaluators for providing us with annotations and feedback.

References

ALPAC. 1966. Language and machines. Computers in translation and linguistics.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

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

Sheila Castilho, Joss Moorkens, Federico Gaspari, Iacer Calixto, John Tinsley, and Andy Way. 2017. Is neural machine translation the new state of the art? *The Prague Bulletin of Mathematical Linguistics*, 108(1):109 – 120.

1)	A_{maj} (misleading)	
	source	Extremely uncomfortable
	MT	Izuzetno **neprijatno AMB**
	gloss	Extremely awkward
2)	A_{maj} (misleading)	
	source	The special effects with the mummy's ghost
	MT	Specijalni efekti s duhom **mame AMB**
	gloss	The special effects with the mother's ghost
3)	A_{maj}+C_{maj} (fully incorrect)	
	source	Best readily available food coloring
	MT	Najbolje **lako AMB obojenje hrane NP**
	gloss	Best easy coloring of_food
4)	A_{maj} (misleading), A_{maj}+C_{min} (potentially misleading)	
	source	they fit easily under my snow pants and they don't show through
	MT	lako se **uklapaju AMB** ispod mojih snežnih pantalona i ne **prolaze WBW kroz njih WBW**
	gloss	they concord easily under my snow pants and not they_go through them
5)	A_{maj} (misleading), A_{maj}+C_{maj} (fully incorrect), A_{min}+C_{maj} (almost correct discarded)	
	source	No matter how much care I used in throwing it
	MT	Bez obzira koliko **briga WBW sam koristio WBW u bacanju WBW**
	gloss	No matter how_many cares I used in the throwing

Table 4: Examples of ambiguity errors (AMB), noun phrase errors (NP) and word-by-word translations (WBW) perceived as comprehensible inadequate (red), almost comprehensible inadequate (green), incomprehensible almost adequate (cyan) and incomprehensible inadequate translations (violet).

source	For the kind of shipping they want it would be reasonable to expect a better presentation.
MT	**Za vrstu dostave** XXX želе da bi bilo **razumno očekivati bolju prezentaciju**.
gloss	**For the kind of shipping** {which} they_want that would be **reasonable to expect better presentation.**

Table 5: Example of propagation effect for major comprehensibility issues. All words in bold are correct, but perceived as incomprehensible due to different types of errors in surrounding words (presented in colour).

percentage of omissions perceived:	
only as comprehensibility issues	19.2
only as adequacy issues	71.5
as both	9.3

Table 6: Percentage of omissions perceived only as comprehensibility issues, only as adequacy issues, and as both types of issues.

Mireia Farrús, Marta Ruiz Costa-Jussà, José B. Mario, and José Adrián R. Fonollosa. 2010. Linguistic-based evaluation criteria to identify statistical machine translation errors. In *Proceedings of the 14th Annual Conference of the European Association for Machine Translation (EAMT 2010)*, St. Raphaël, France.

Marcello Federico, Matteo Negri, Luisa Bentivogli, and Marco Turchi. 2014. Assessing the impact of translation errors on machine translation quality with mixed-effects models. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, Doha, Qatar.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

Katrin Kirchhoff, Daniel Capurro, and Anne M. Turner. 2014. A conjoint analysis framework for evaluating user preferences in machine translation. *Machine Translation*, 28(1):117.

Filip Klubička, Antonio Toral, and Víctor M. Sánchez-Cartagena. 2018. Quantitative Fine-grained Human Evaluation of Machine Translation Systems: A Case Study on English to Croatian. *Machine Translation*, 32(3):195–215.

Philipp Koehn and Christof Monz. 2006. Manual and automatic evaluation of machine translation between European languages. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 102–121, New York City.

Arle Lommel, Aljoscha Burchardt, Maja Popović, Kim Harris, Eleftherios Avramidis, and Hans Uszkoreit. 2014. Using a new analytic measure for the annotation and analysis of MT errors on real data. In *Proceedings of the 17th Annual Conference of the European Association for Machine Translation (EAMT 2014)*, pages 165–172.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning Word Vectors for Sentiment Analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics and Human Language Technologies (ACL-HLT 2011)*, pages 142–150, Portland, Oregon, USA.

Marianna Martindale and Marine Carpuat. 2018. Fluency over adequacy: A pilot study in measuring user trust in imperfect MT. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (AMTA 2018)*, pages 13–25, Boston, MA.

Marianna Martindale, Marine Carpuat, Kevin Duh, and Paul McNamee. 2019. Identifying fluently inadequate output in neural and statistical machine translation. In *Proceedings of Machine Translation Summit XVII*, pages 233–243, Dublin, Ireland.

Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton van den Hengel. 2015. Image-Based Recommendations on Styles and Substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2015)*, pages 43–52, Santiago, Chile.

Maja Popović. 2020. Informative manual evaluation of machine translation output. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING 2020)*, Online.

Johann Roturier and Anthony Bensadoun. 2011. Evaluation of MT Systems to Translate User Generated Content. In *Proceedings of the MT Summit XIII*, Xiamen, China.

Sara Stymne and Lars Ahrenberg. 2012. On the practice of error analysis for machine translation evaluation. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, pages 1785–1790, Istanbul, Turkey.

Laura Van Brussel, Arda Tezcan, and Lieve Macken. 2018. A fine-grained error analysis of NMT, SMT and RBMT output for English-to-Dutch. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan.

David Vilar, Jia Xu, Luis Fernando D'Haro, and Hermann Ney. 2006. Error analysis of statistical machine translation output. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

John White, Theresa O'Connell, and Francis O'Mara. 1994. The ARPA MT evaluation methodologies: evolution, lessons, and future approaches. In *Proceedings of the 1994 Conference, Association for Machine Translation in the Americas*, pages 193–205.