Submitted 09/2019; Published 12/2019

NMT's wonderland where people turn into rabbits. A study on the comprehensibility of newly invented words in NMT output

Lieve Macken LIEVE.MACKEN@UGENT.BE
Laura Van Brussel LAURAVBRUSSEL@GMAIL.COM
Joke Daems JOKE.DAEMS@UGENT.BE

LT^3*, Language and Translation Technology Team, Ghent University*

Abstract

Machine translation (MT) quality has improved enormously since the arrival of neural machine translation (NMT). The most noticeable improvement compared to statistical MT systems is the increased grammaticality and fluency of the produced MT output. At the lexical level, the quality of NMT systems is less promising. New types of lexical mistakes appear in NMT output, such as the occurrence of non-existing words, i.e. words that are not part of the vocabulary of the target language and were thus invented by the NMT system. For MT use cases in which readers only have access to the MT output without the source text, such non-existing words can affect comprehension as the intended source meaning may not be recovered. To investigate if and to what extent non-existing words in English-to-Dutch NMT output impair comprehension, an experiment was set up in SurveyMonkey. Eighty-six participants were given 15 non-existing words (5 single words and 10 noun compounds) and were either asked to describe the meaning of these words or to select the correct meaning from a predefined list. The words were presented either in isolation or in sentence context. Participants were asked to indicate how confident they were about their answer. Results show that non-existing words indeed impair comprehension as in 60% of the cases the participants gave a wrong answer. Sentence context had a positive impact and made it easier for the participants to determine the meaning of the non-existing word. Participants were also more confident about their answer when the words were presented in sentence context.

1. Introduction

There is a large consensus in academia and in the translation industry that with the arrival of neural machine translation (NMT), the quality of machine translations has improved significantly. Quality improvements, compared with (the previous state-of-the-art) statistical machine translation (SMT) engines, have been demonstrated for a wide variety of language pairs and text types using both human and automatic evaluation methods (Bentivogli et al. 2016, Burchardt et al. 2017, Toral and Sánchez-Cartagena 2017, Klubička et al. 2018, Shterionov et al. 2018, Jia et al. 2019, Daems and Macken 2019). The most noticeable improvement is the increased grammaticality and fluency of the produced MT output. For English (EN) into Dutch (NL), the language pair we focus on in this article, Van Brussel et al. (2018) carried out a large-scale error annotation on 665 sentences translated by both Google's SMT and Google's NMT systems, using the fine-grained SCATE error taxonomy (Tezcan et al. 2017), which distinguishes between the well-known quality aspects of fluency (assessing the well-formedness of the target language) and accuracy (the transfer of source content, sometimes called adequacy). In general, the total number of errors in the output of the NMT engine halved in comparison to the SMT version, and the NMT system contained even less than half of the number of fluency errors, which corroborates other research findings. However, Van Brussel et al. (2018) also demonstrated that the overall better fluency scores of NMT could be mainly attributed to the fact that the NMT output contained much less grammatical errors than the SMT system (250 vs. 932) and hardly any spelling mistakes (95 vs. 253). On the other hand, the NMT system contained more lexical errors (358 vs. 235), of which 54 (or 15%) were non-existing words. In this paper we focus on this category of non-existing (or fictional) words. The problem of fictional words was also identified by Lesznyák (2019) in the context of post-editing NMT at the European Commission.

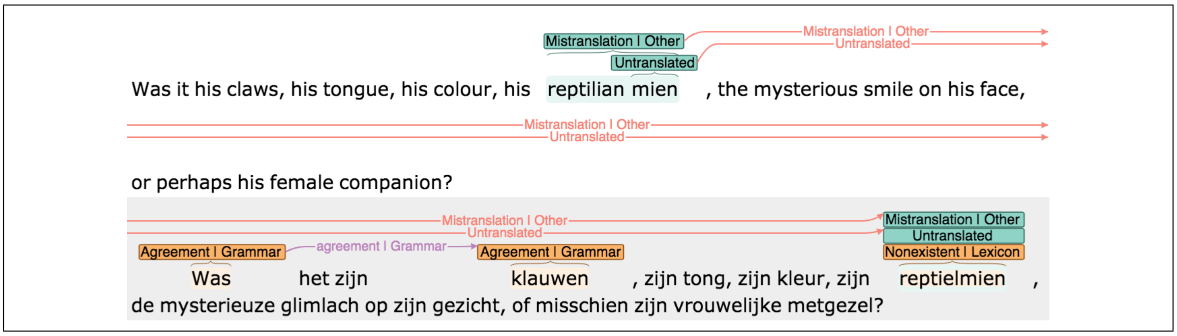
We define non-existing words as words that are not part of the vocabulary of the target language and were thus invented by the NMT system. When translating from English into Dutch, we spotted different types of non-existing words in the output of NMT systems, e.g. *guzzelen* as translation for *guzzle* (NL: *opschrokken, zwelgen*), *familiekonijn* (literally *family rabbit*) as translation for *family rabbi* (NL: *rabbijn van de familie*), *plattelandsbroeder* (literally *country brother*) as translation for *country squire* (NL: *landjonker*) and *diplomaatgenoot* (literally *diplomat companion*) as translation for *diplomat husband* (NL: *diplomaat* en *echtgenoot*). There are several reasons why an NMT system creates new non-existing words. One reason is that, although NMT systems have made huge progress in many areas, they sometimes still generate a too literal translation for different types of multiword expressions such as compounds, e.g. *mijlpaalonderzoek* for *landmark study* (NL: *belangrijk onderzoek*) and idiomatic expressions, e.g. *de zweephand hebben* for *to have the whip hand* (NL: *de macht hebben over, beheersen*). Another reason, specific for NMT systems, is that they operate at sub-word level to reduce vocabulary size. Sennrich et al. (2016) came up with a word segmentation technique inspired by byte pair encoding, a technique used for data compression, which iteratively merges the most frequent character n-grams into a single symbol. This technique can be applied separately on the source and target vocabulary or jointly on the union of the source and target vocabularies. The main advantage of operating at sub-word level, apart from vocabulary reduction, is that the system can translate unseen words in the training data and improves the translation of rare words. In SMT, different techniques were applied to handle unknown words: morphologically complex words such as compounds were segmented into their component parts and other unknown words (or out-of-vocabulary words) were often just copied to the target. In NMT, these frequent merged character n-grams often correspond to linguistically motivated units, such as compound parts or affixes, but this is not always the case. In some cases things go wrong and the NMT system generates a translation for an orthographically very similar word (e.g. *rabbit* instead of *rabbi* or *scared* instead of *scarred*) or just combines translated n-grams into a very plausible target word (e.g. *guzzelen*).

With the arrival of NMT, the quality of MT has improved to the extent that MT is becoming an attractive solution to deal with the increased need for translated content, which is reflected in new use cases such as business-to-consumer e-commerce and user-generated content (see Levin et al. (2017) for the Booking.com example). As a result, in the near future, readers will be more often confronted with 'raw' (unedited) MT output. It is therefore important to understand MT quality and its impact on end-users (Castilho and O'Brien 2017). Martindale and Carpuat (2018) studied user trust in imperfect MT and their study suggests that fluency problems affect user trust more strongly than adequacy issues. Scarton and Specia (2016) studied the comprehensibility of different RBMT and SMT systems using reading comprehension tests, and used a set of human translations as a control. In their experiments participants did not achieve higher scores when reading the human translated texts. Macken and Ghyselen (2018) followed up on this work and compared the results of reading comprehension tests based on NMT output with human translations. Their results were in line with the findings of Scarton and Specia (2016) as the reading comprehension tests did not reveal major differences between the human and machine translated texts, although participants quoted certain passages that hindered comprehension. The NMT mistakes that bothered readers most were grammatical problems, repetitions, inconsistent translations and unidiomatic constructions. Castilho and Guerberof Arenas (2018) set up a small-scale pilot study involving six participants to compare reading comprehension of SMT and NMT. They reported shorter task completion times and higher satisfaction levels for NMT than SMT. Their results for cognitive load were inconclusive due to the low number of participants.

The reading comprehension experiments cited above provide valuable information about the quality of entire machine translated documents. But they cannot provide information about the impact of specific error types on comprehension. This pilot study focuses on a very specific problem in NMT output and proposes a methodology to investigate if and to what extent non-existing words in English-to-Dutch NMT output impair comprehension. The remainder of this article is organized as follows: in Section 2 we outline the methodology used in this study, in Section 3 we present the results and Section 4 concludes the paper and discusses future work.

2. Methodology

In the framework of the ArisToCAT-project[1], translation errors have been manually annotated in English-Dutch MT output of Google Translate and DeepL during the summer of 2017 and 2018 using the fine-grained error taxonomy of Tezcan et al. (2017). The source texts were collected from various existing corpora such as the Dutch Parallel Corpus (Macken et al. 2011), Dundee (Kennedy 2003), and Provo (Luke and Christianson 2018). To facilitate the annotation process we used WebAnno[2](Yimam et al. 2013). An example of an annotated sentence pair according to the SCATE error taxonomy is given in Figure 1.


Was it his claws, his tongue, his colour, his reptilian mien , the mysterious smile on his face,

or perhaps his female companion?

Was het zijn klauwen , zijn tong, zijn kleur, zijn reptielmien ,

de mysterieuze glimlach op zijn gezicht, of misschien zijn vrouwelijke metgezel?

Figure 1: Example of annotated sentence pair according to the SCATE error taxonomy in WebAnno

We manually analysed 175 sentences containing non-existing words, and found that two thirds of the cases were noun compounds and one third single words.

Within the noun compounds we distinguished the following three subclasses:

- Noun compounds that consist only of existing words, e.g. *baard+juffrouw* (EN: *bearded lady*; NL: *bebaarde vrouw*; 53% of all non-existing words)
- Noun compounds that contain one untranslated element, e.g. *zelfmoord+bomber* (EN: *suicide bomber*; NL: *zelfmoordterrorist*; 10% of all non-existing words)
- Noun compounds that contain one non-existing element, e.g. *dinosaurus+cuke* (EN: *dinosaur puke*; NL: *braaksel van dinosaurussen*; 12% of all non-existing words)

Within the single words we checked whether the translation resembled the correct Dutch translation or the original English word. To make that decision we looked at the suggestions that were provided by the Dutch and English spell checker of Microsoft Word. The following three subclasses were distinguished:

- Single words that resemble the correct Dutch word, e.g. *giecheert* ∼ *giechelt* (EN: *giggles*; NL: *giechelen*; 11% of all non-existing words)
- Single words that resemble the original English word, e.g. *plaintief* ∼ *plaintive* (EN: *plaintive*; NL: *klagend*; 4% of all non-existing words)

[1] `https://research.flw.ugent.be/projects/aristocat`
[2] `https://webanno.github.io/webanno/`

- Single words that do not resemble the correct Dutch or original English word, e.g. *zwaande* (EN: *swanky*; NL: *chic*; 10% of all non-existing words)

To test the comprehensibility of the invented words, we set up an experiment in SurveyMonkey. Participants were given 15 non-existing words (10 compounds and 5 single words). They were told that they would see 15 weird words for which they either had to describe or to select the correct meaning, but they were not aware that the words were created by an NMT system. The words were presented in three different conditions and we made sure that each participant saw each word only once. In the first condition (condition A) we showed the non-existing word in isolation and the participants were asked to give a synonym or describe the meaning of the presented word; in the second condition (condition B), the non-existing word was presented in sentence context and participants had to select the correct answer from a predetermined list of possible answers; in the third condition (condition C) the non-existing word was also presented in sentence context and the participants were again asked to give a synonym or describe the meaning of the presented word. After each word the participants were asked to indicate their confidence on a 4-point scale ranging from *Not sure at all* to *Very sure*. Examples of the three conditions are given in the Appendix. In total, 86 participants filled in the survey and we obtained 1151 valid observations.[3] We tested 40 different non-existing words (26 compounds and 14 single words). For each non-existing word we collected between 27 and 30 observations.

The participants were all native speakers of Dutch living in Flanders. The median age of the participants was 29 years (range 18-67). All answers were scored as either correct or wrong by one of the authors on the basis of the original source sentence. The collected data was first analysed using exploratory graphs in Microsoft Excel and afterwards analysed using generalised linear mixed models and cumulative link mixed models in R (version 3.6.1.), using the packages lme4 (Bates et al. 2015) and ordinal (Christensen 2019), with lmerTest (Kuznetsova et al. 2017) to obtain significance scores.

3. Results

In the analysis section, we first look at the variable 'correctness' and then at the variable 'confidence'. For each variable we first discuss the exploratory graphs and then results of the mixed models.

3.1 Correctness

If we look at the percentage of correct answers in the different conditions in Figure 2, we can conclude that, in general, the participants had difficulties to understand the words the NMT systems had invented. In total, 60% of the answers were wrong.

Sentence context helps to determine the correct meaning as the percentage of wrong answers is higher (77%) when the words were presented in isolation (Condition A), compared to 59% in Condition C. But even in the "easiest" condition (Condition B), in which the words were presented in sentence context and the participants had to select the correct answer from a predefined list, 44% of the answers were incorrect.

If we look at the percentage of correct answers per subtype in Figure 3, we see that for single words the non-existing words that resemble a Dutch word were the easiest to understand, with 70% of correct answers. For compounds, the words containing a non-existing element are the hardest to understand (28% of correct answers), whereas compounds containing an untranslated element were the easiest (63% correct answers). The results for the single words however should be interpreted with caution as the data set used for single words is relatively small.

[3] A number of test words were removed as some answers in the multiple choice condition were found to be too ambiguous.

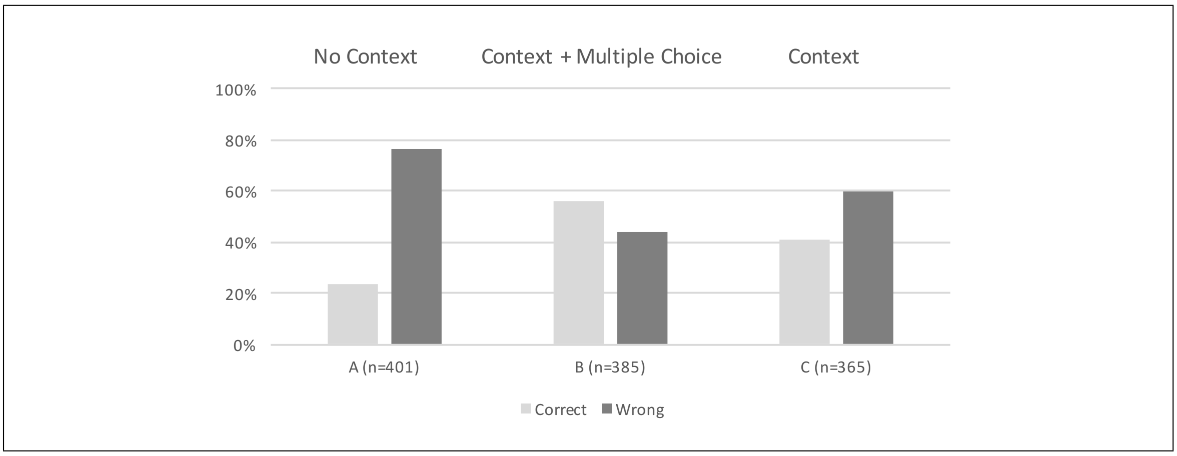


Figure 2: Percentage of correct (light grey) and wrong (dark grey) answers per condition

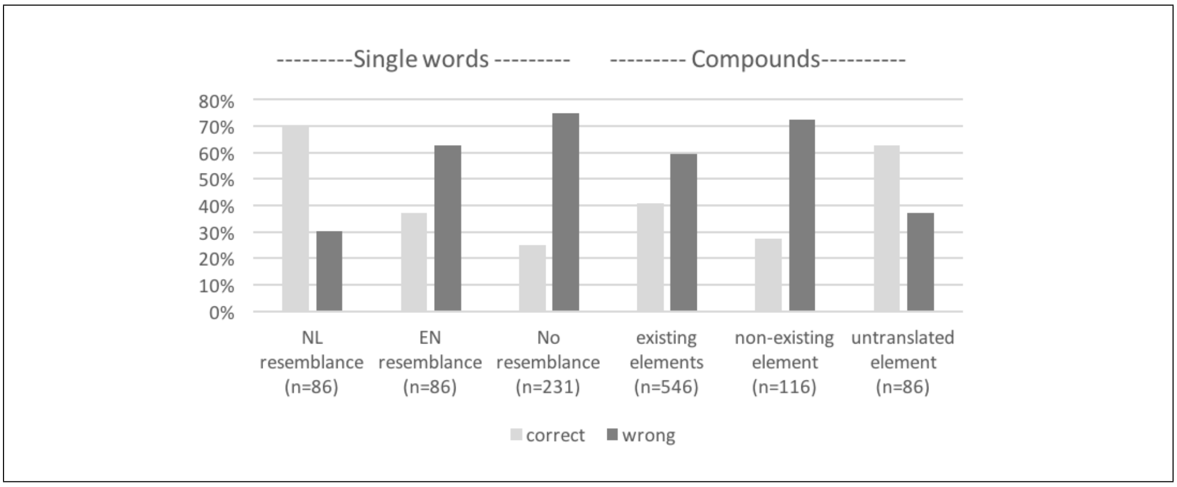


Figure 3: Percentage of correct (light grey) and wrong (dark grey) answers per (sub)type

Apart from the exploratory graphs we also built a binomial generalised linear mixed effects model in R. The dependent variable was 'correctness', the predictor variables we tested were condition, item subtype, and confidence, as well as any combination of these predictors with and without interaction. To account for individual differences across participants (some participants might be better at the task than others) and item-inherent characteristics (some items might be easier to interpret than others), we added both participant code and items as random effects. The models with interaction effects did not converge, so we opted for a reduced model without interaction. Including random slopes made the model unidentifiable, so these were not part of the final model. The model with the best fit (according to comparison of AIC values[4] between the null model without predictor variables and the different models with just one, both, or all three predictor variables) was the model that included condition plus confidence. The model confirmed that Condition B is more often correct than condition A (effect size = 2.08; $p < 0.001$), and that condition C is more often correct than condition A (effect size = 1.19; $p < 0.001$). In addition, correctness was found to be influenced by a higher degree of confidence (effect size = 0.45, $p < 0.001$). The effects can be seen in the effects plots in Figure 4.

[4] AIC or Akaike Information Criterion (Akaike 1974) estimates the amount of information loss of a model. In model comparison, the model with the lowest AIC value is generally the better model.

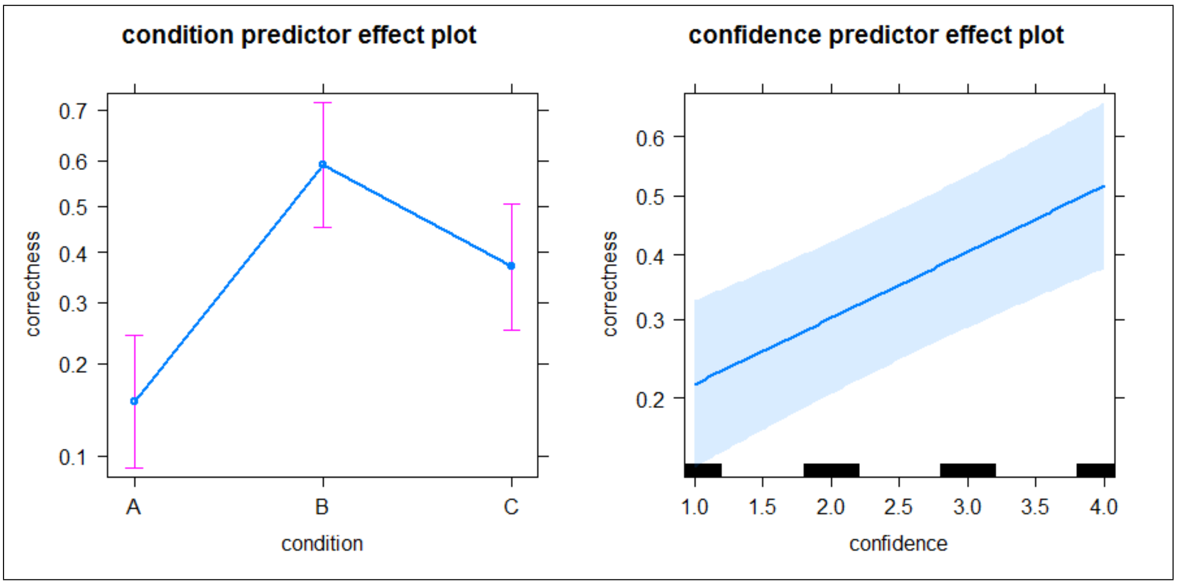


Figure 4: Effect of 'condition' and 'confidence' on correctness

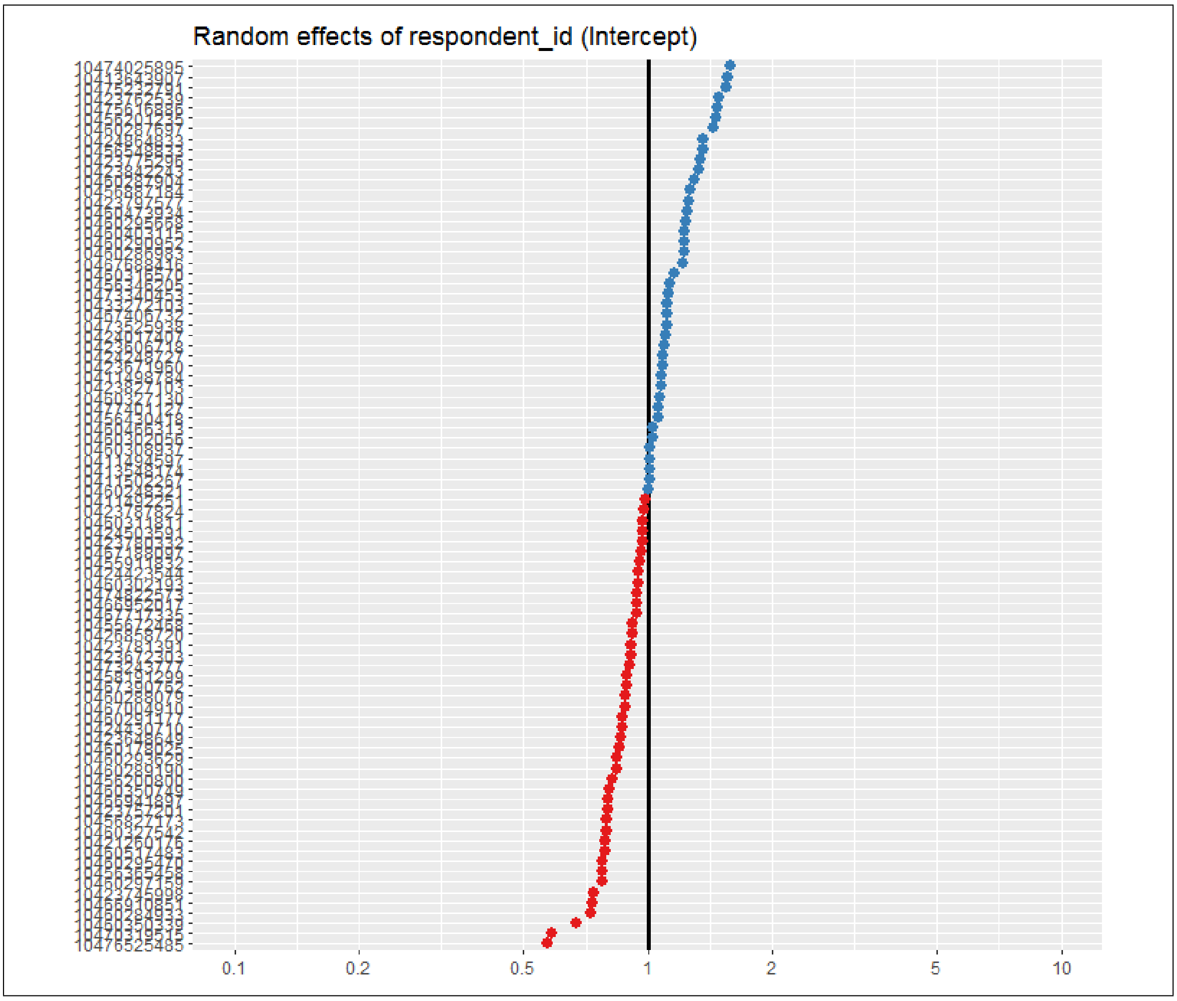


Figure 5: Random effects of participants for correctness

Random effects showed not so much variation across participants (see Figure 5), but, as can be seen in Figure 6, random effects did show variation across items, with items *douchegordijnmast* (EN: *shower curtain pole*), *mijlpaalonderzoek* (EN: *landmark study*) and *aardlingen* (EN: *earthlings*) being the easiest words to understand and *bosoevers* (EN: *woodland banks*) and *krabbenbomen* (EN: *crab trees*) the most difficult.

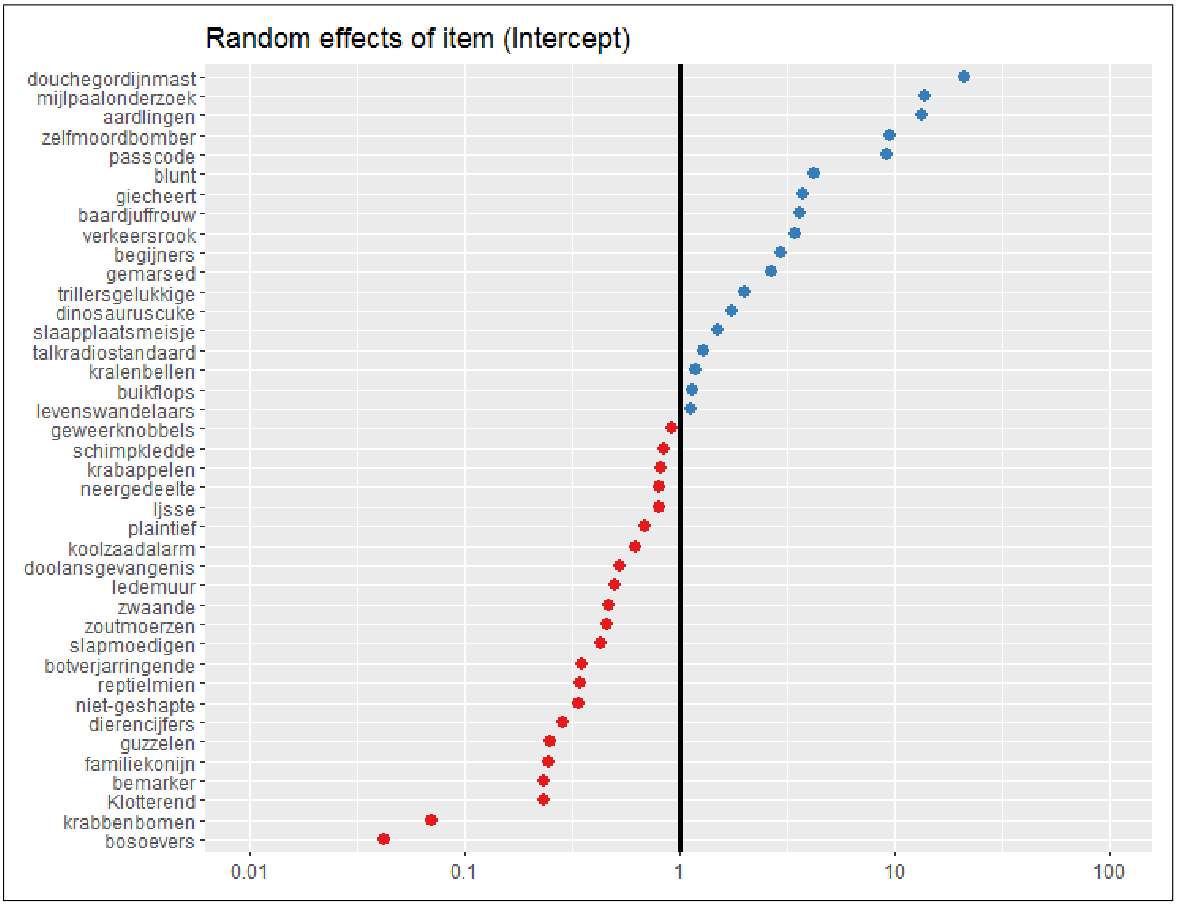


Figure 6: Random effects of items for correctness

3.2 Confidence

As mentioned in section 2, participants had to indicate how confident they were about their answer after each question on a 4-point scale and select one of the following options: *Not sure at all*, *not so sure*, *a little bit sure* and *very sure*. When we look at the distribution of the confidence scores for all correct and wrong answers separately in Figure 7, we see higher confidence scores for correct answers, suggesting that participants were more confident about correct answers than about wrong answers. Still, in 13% of all wrong answers, participants were *very sure* about their answer.

Condition has some impact on the confidence scores as participants used the label *very sure* more often in condition C (in 36% of the correct answers and 17% of the wrong answers) than in the other conditions as can be seen in Figure 8 and participants use the label *not sure at all* most when the words were presented in isolation (Condition A) and when they gave a wrong answer (45%).

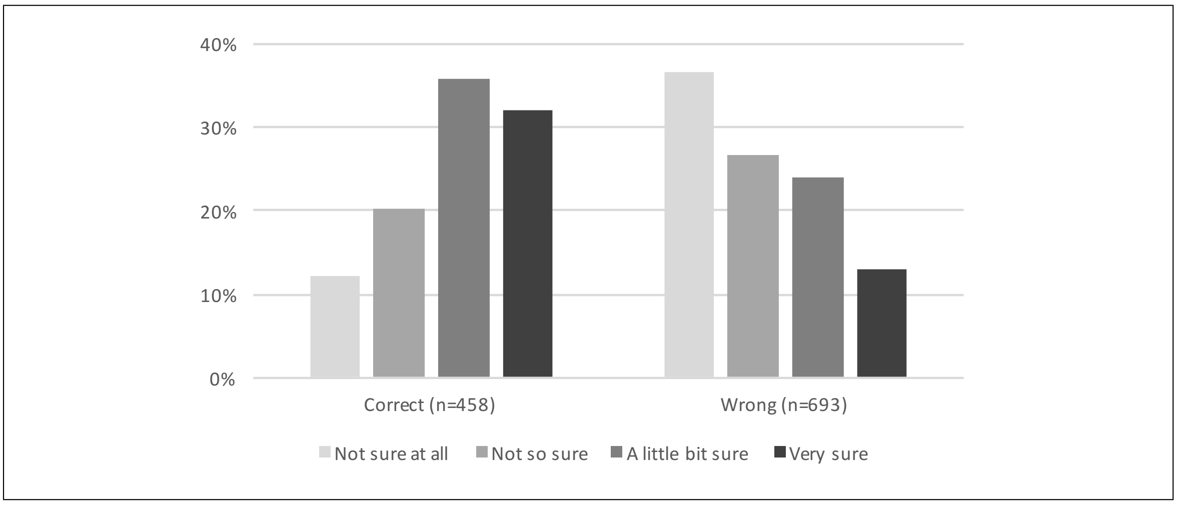


Figure 7: Distribution of different confidence scores per type of answer (correct/wrong)

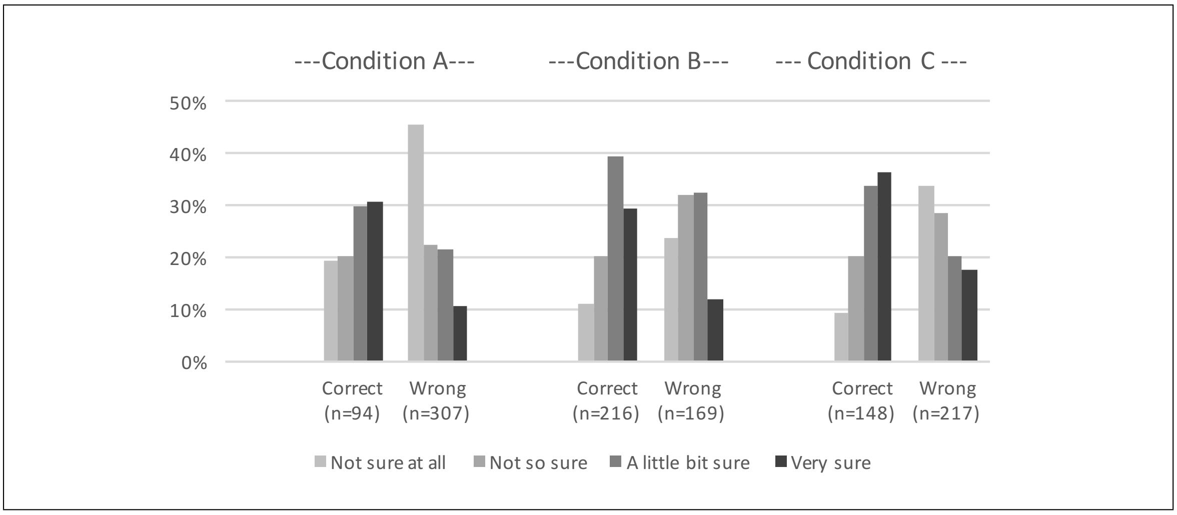


Figure 8: Distribution of different confidence scores per condition and type of answer (correct/wrong)

As 'confidence' is an ordinal variable, we built a cumulative link mixed model in R with 'confidence' as response variable. Predictor variables that were tested were condition, item subtype, correctness, and the different combinations of those three (with and without interaction effects). To account for individual differences across participants and item-inherent characteristics, we again added both participant code and items as random effects. As with the first model, including interaction effects or random slopes made the model unidentifiable, so we worked with a reduced model without interaction effect or random slopes. The model with the best fit (according to comparison of AIC values between the null model without predictor variables and the different models with just one, both, or three predictor variables) was the model that included condition and correctness as predictors.

Significant findings in the model were higher confidence in condition B (effect size = 0.98; $p < 0.001$) and C (effect size = 0.74; $p < 0.001$) and an increase in confidence when the answer was correct (effect size = 0.84, $p < 0.001$). These effects are visualised in the effect plot in Figure 9.

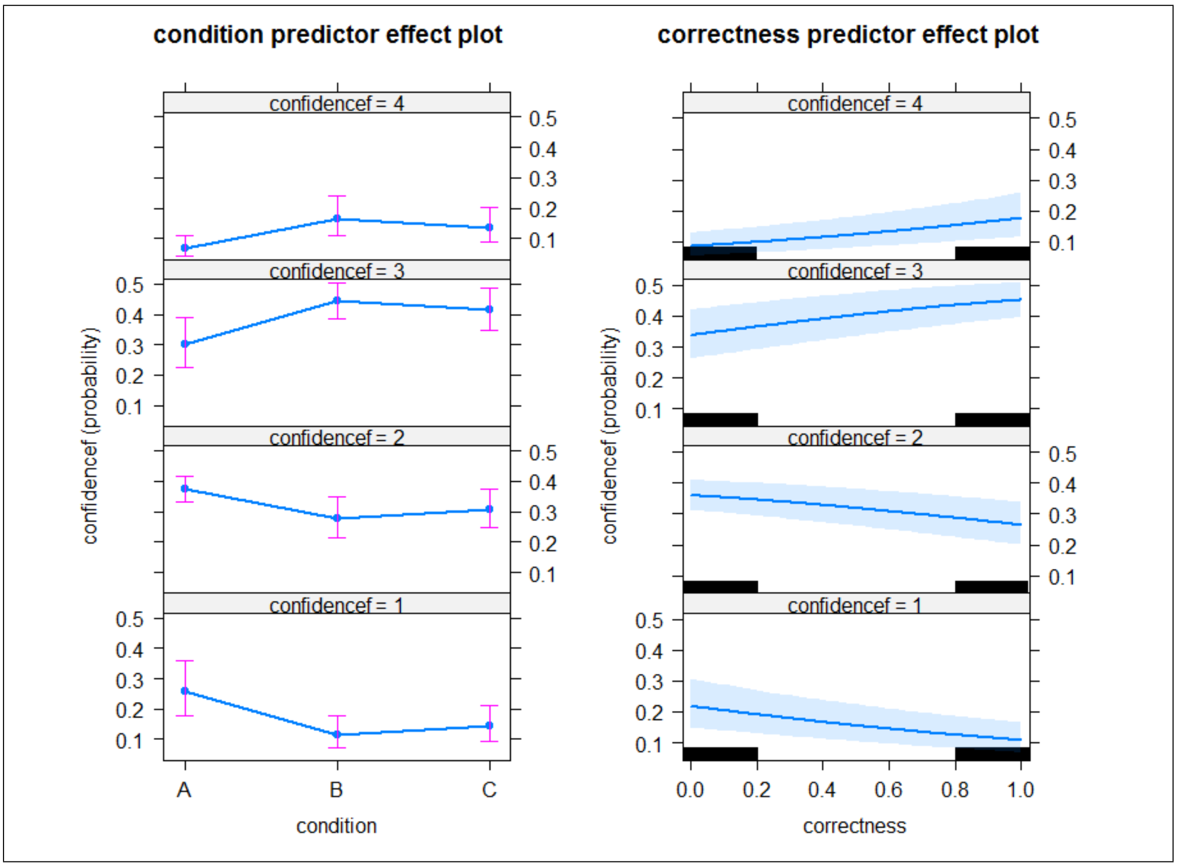


Figure 9: Effect of condition and correctness on confidence

Random effects show much more variability between individual participants than was the case for ‘correctness’ (see Figure 10), which makes sense, as ‘confidence’ is intuitively a subjective factor. As was the case for ‘correctness’, random effects (see Figure 11) showed variation across items, with *giecheert* (EN: *giggles*), *doolansgevangenis* (EN: Maze prison) and *zwaande* (EN: *swanky*) being the items participants were least confident about and items *mijlpaalonderzoek* (EN: *landmark study*),*aardlingen* (EN: *earthlings*) and *blunt* (EN: *bluntly*) the items participants were most confident about.

When comparing the random effect plots of items for ‘correctness’ and ‘confidence’, we notice that, in general, participants made a correct estimation of their answers: in most cases, confidence is higher for items that were correct more often, and lower for items that were incorrect more often. Cases of overestimation are, in our opinion, most problematic from a comprehensibility point of view, as readers are unaware of their misinterpretation of the intended meaning. There are 8 such cases in the data, of which ’bosoevers’ and ’niet-geshapte’ are the most severe, as participants were very confident of their answers while these items were often answered incorrectly. While less dangerous than overestimation, cases of underestimation can also impair comprehensibility or readability, as readers are not sure of the intended meaning, despite guessing it correctly. The most severe case of underestimation in the data is ’giecheert’, which is the word participants were least confident about, although it was answered correctly more often than many other items.

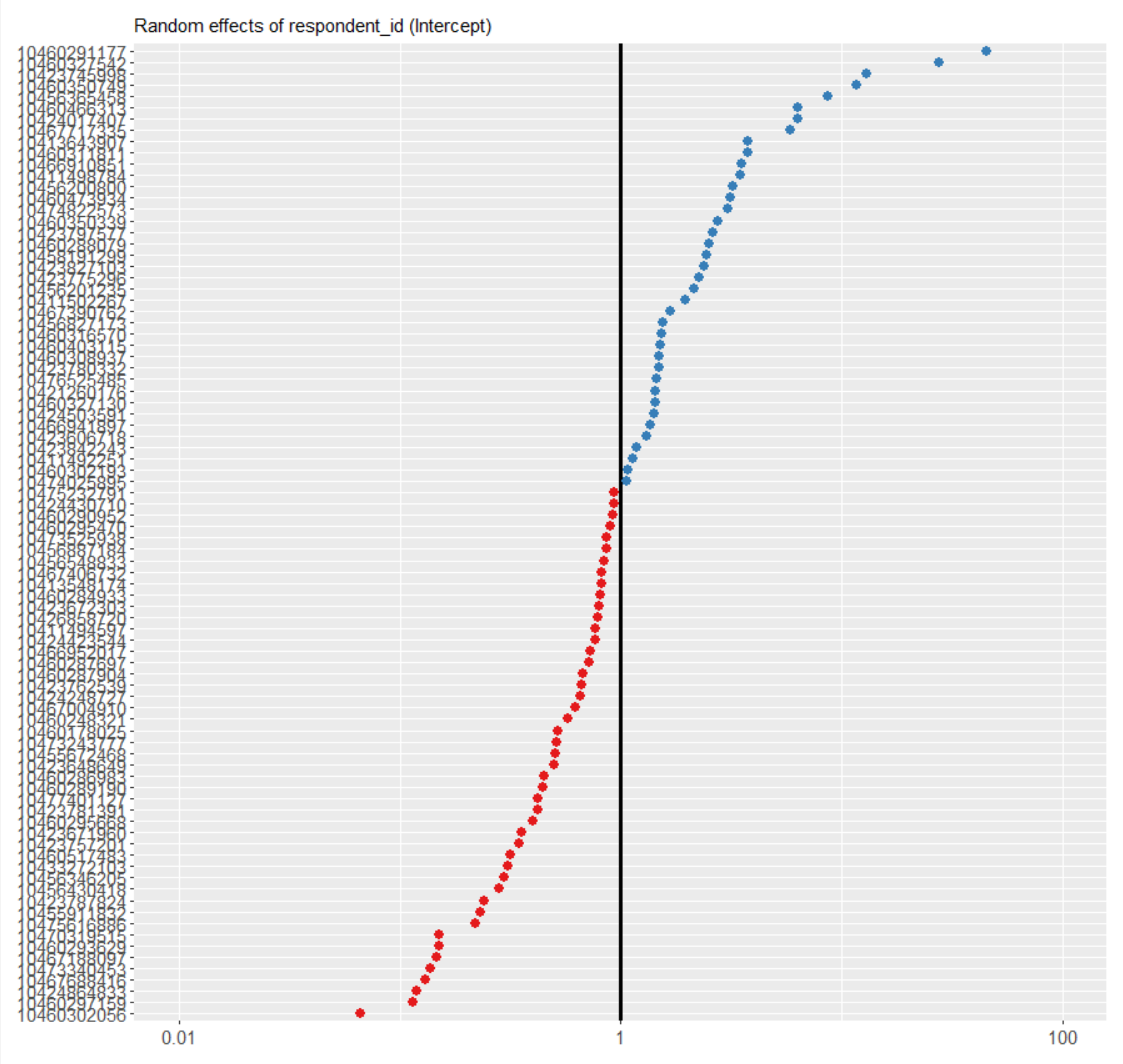


Figure 10: Random effects of participants for confidence

4. Conclusion

In this article we investigated the problem of non-existing words in NMT output and their impact on comprehension. In previous studies reading comprehension tests did not reveal major differences between human and machine translated texts (Scarton and Specia 2016, Macken and Ghyselen 2018), but both studies assessed comprehension on text level. The results of our pilot study show that non-existing words indeed impair comprehension as in 60% of the cases the participants gave a wrong answer. Sentence context helps to determine the correct meaning of non-existing words as the percentage of wrong answers is higher in Condition A (77%) compared to Condition C (59%).

Research on user confidence in NMT output is scarce. Martindale and Carpuat (2018) found that fluency problems affect user trust more strongly than adequacy problems, but they did not distinguish between different types of adequacy and fluency issues. In this study, we focus on one specific lexical problem, namely non-existing words invented by the NMT system. We show that context has a positive impact on how confident participants were about their answer. When the

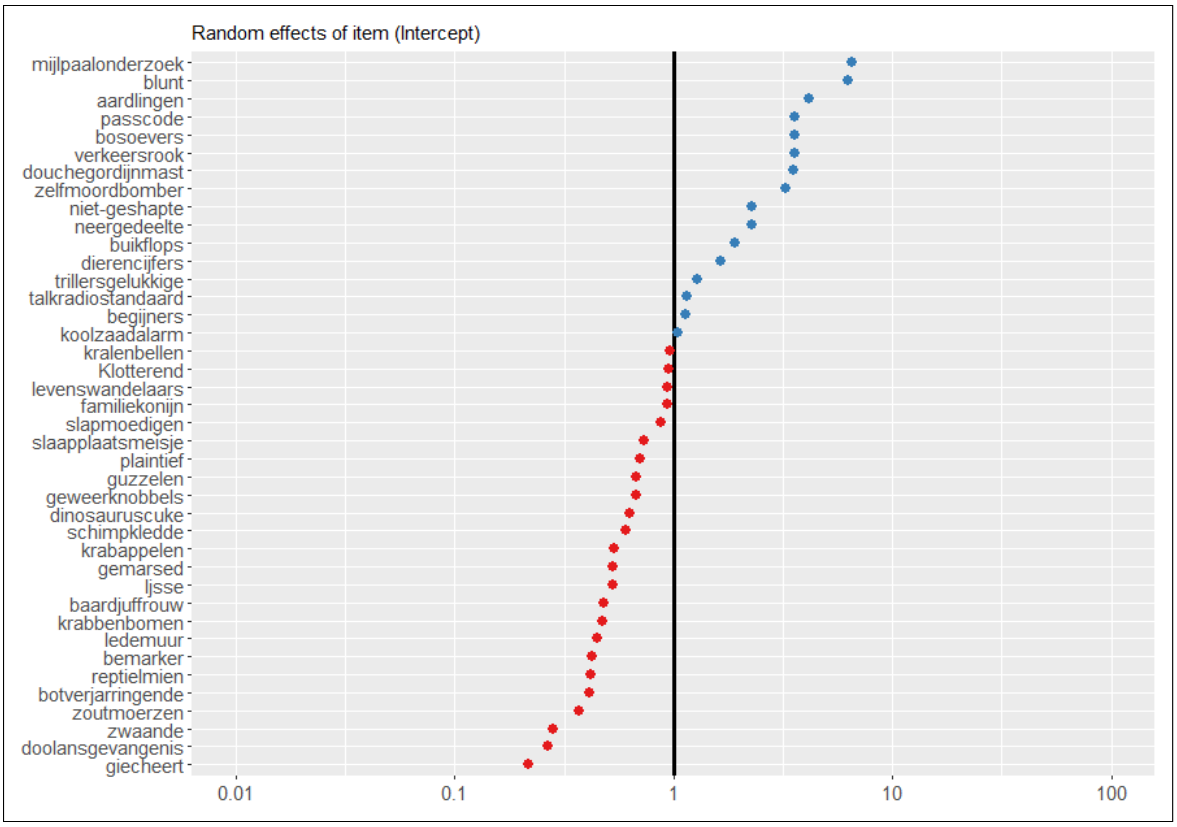


Figure 11: Random effects of items for confidence

words were presented in sentence context they either gave a higher confidence score when the answer was correct or a lower confidence when the answer was wrong.

We distinguished between compounds and single words and identified different subtypes within each category. Unfortunately, the dataset was not large enough to include subtype as a meaningful predictor in any of the statistical models. From the dataset we can tentatively suggest that participants are less confident about items containing non-existing elements, and that items containing untranslated elements are often answered correctly. However, a larger dataset with more respondents would be needed to statistically substantiate these trends.

This pilot study has some important limitations. It is a small-scale study in which only 40 non-existing words were tested. Due to the experimental set-up the context was artificially limited to one sentence, which may have influenced the results. The machine translation output was generated with third-party MT systems (Google Translate and DeepL) in 2017 and 2018, and as machine translation technology advances very quickly, we cannot be certain whether these systems still generate the same output.

In future work we will focus on more error types and study the reading behaviour of participants in a more natural setting, using more advanced techniques. We will collect and analyse eye movements of participants reading Dutch machine-translated text to investigate the impact of different categories of MT errors (syntactic versus semantic, function words versus content words, short-distance versus long-distance triggers of errors) on the underlying comprehension process.

5. Acknowledgments

This study is part of the ArisToCAT project (Assessing The Comprehensibility of Automatic Translations), which is a four-year research project (2017-2020) funded by the Research Foundation – Flanders (FWO) – grant number G.0064.17N.

References

Akaike, Hirotugu (1974), A new look at the statistical model identification, *Selected Papers of Hirotugu Akaike*, Springer, pp. 215–222.

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker (2015), Fitting linear mixed-effects models using lme4, *Journal of Statistical Software* **67** (1), pp. 1–48.

Bentivogli, Luisa, Arianna Bisazza, Mauro Cettolo, and Marcello Federico (2016), Neural versus Phrase-Based Machine Translation Quality: a Case Study, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Austin, Texas, pp. 257–267.

Burchardt, Aljoscha, Vivien Macketanz, Jon Dehdari, Georg Heigold, Jan-Thorsten Peter, and Philip Williams (2017), A Linguistic Evaluation of Rule-Based, Phrase-Based, and Neural MT Engines, *The Prague Bulletin of Mathematical Linguistics* **108**, pp. 159–170.

Castilho, Sheila and Ana Guerberof Arenas (2018), Reading comprehension of machine translation output: What makes for a better read, *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*, Association for Computational Linguistics, Universitat d'Alacant, Alacant, Spain, pp. 79–88.

Castilho, Sheila and Sharon O'Brien (2017), Acceptability of machine-translated content: a multi-language evalutation by translators and end-users, *Linguistica Anvtverpiensia: New Series: Themes in Translation studies* **16**, pp. 120–136.

Christensen, R. H. B. (2019), ordinal-regression models for ordinal data. R package version 2019.4-25. http://www.cran.r-project.org/package=ordinal/.

Daems, Joke and Lieve Macken (2019), Interactive adaptive SMT versus interactive adaptive NMT: a user experience evaluation, *Machine Translation* **33** (1), pp. 117–134.

Jia, Yanfang, Michael Carl, and Xiangling Wang (2019), Post-editing neural machine translation versus phrase-based machine translation for English–Chinese, *Machine Translation* **33** (1), pp. 9–29.

Kennedy, Alan (2003), *The Dundee Corpus*, School of Psychology, The University of Dundee.

Klubička, Filip, 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), pp. 195–215.

Kuznetsova, Alexandra, Per B. Brockhoff, and Rune Haubo Bojesen Christensen (2017), lmertest package: Tests in linear mixed effects models, *Journal of Statistical Software* **82** (13), pp. 1–26.

Lesznyák, Ágnes (2019), Hungarian translators' perceptions of neural machine translation in the European commission, *Proceedings of Machine Translation Summit XVII Volume 2: Translator, Project and User Tracks*, European Association for Machine Translation, Dublin, Ireland, pp. 16–22. https://www.aclweb.org/anthology/W19-6703.

Levin, Pavel, Nishikant Dhanuka, and Maxim Khalilov (2017), Machine translation at booking.com: Journey and lessons learned., *Proceedings of the 20th Conference of the European Association for Machine Translation,*, Vol. User Studies and Project/Product Descriptions, EAMT, Prague, Czech Republic, pp. 81–86.

Luke, Steven G. and Kiel Christianson (2018), The Provo Corpus: A large eye-tracking corpus with predictability norms, *Behavior Research Methods* **50** (2), pp. 826–833.

Macken, Lieve and Iris Ghyselen (2018), Measuring Comprehension and User Perception of Neural Machine Translated Texts: A Pilot Study, *Proceedings of the 40th Conference Translating and the Computer*, AsLing, London, UK, pp. 120–126.

Macken, Lieve, Orphée De Clercq, and Hans Paulussen (2011), Dutch parallel corpus: a balanced copyright-cleared parallel corpus, *META* **56** (2), pp. 374–390.

Martindale, Marianna J. and Marine Carpuat (2018), Fluency over adequacy: a pilot study in measuring user trust in imperfect MT, *Proceedings of AMTA 2018*, Vol. 1: MT Reseacrh Track, Boston, pp. 13–25.

Scarton, Carolina and Lucia Specia (2016), A reading comprehension corpus for machine translation evaluation, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, European Language Resources Association (ELRA).

Sennrich, Rico, Barry Haddow, and Alexandra Birch (2016), Neural machine translation of rare words with subword units, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Berlin, Germany, pp. 1715–1725.

Shterionov, Dimitar, Riccardo Superbo, Pat Nagle, Laura Casanellas, Tony O'Dowd, and Andy Way (2018), Human versus automatic quality evaluation of NMT and PBSMT, *Machine Translation* **32** (3), pp. 217–235.

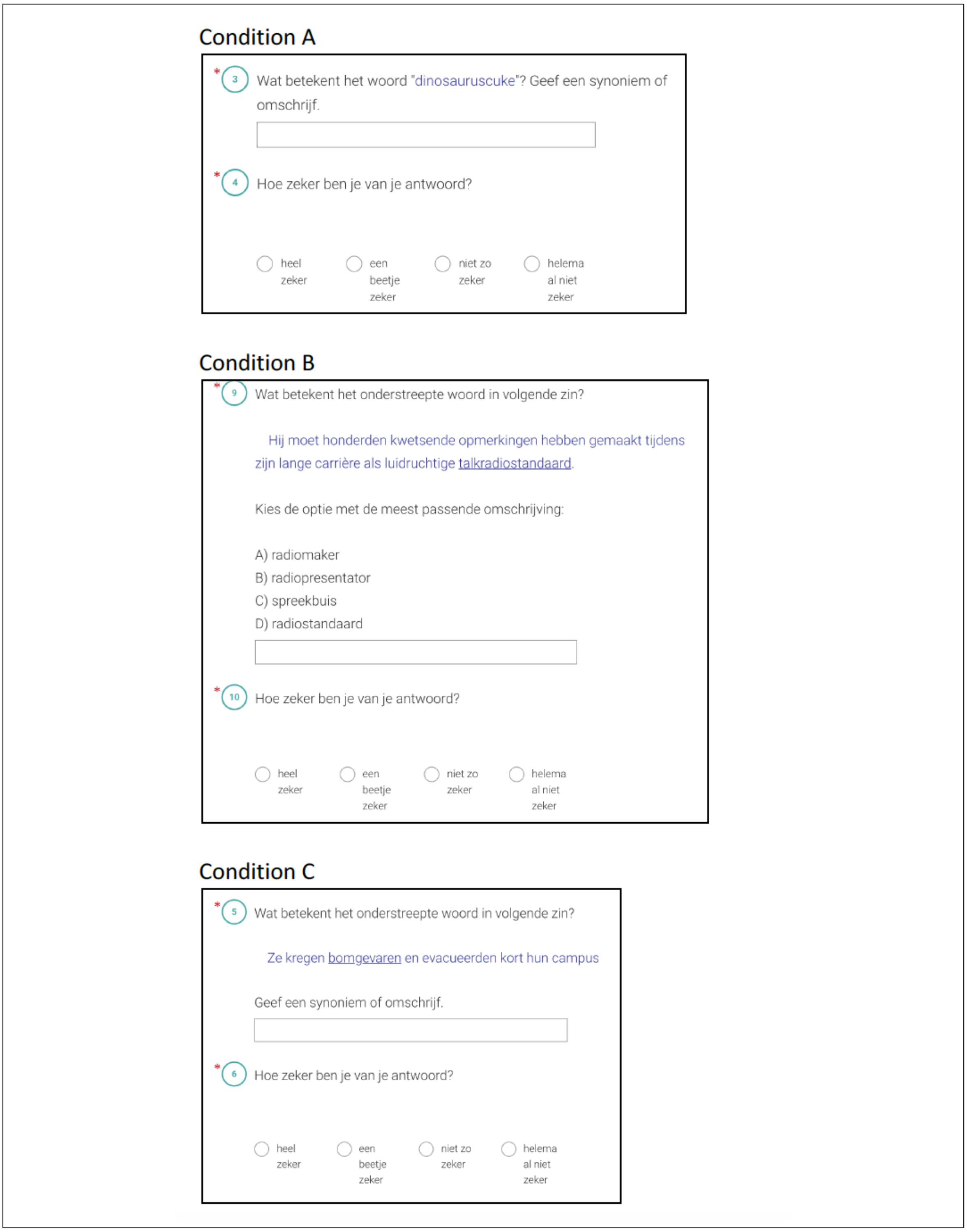
Tezcan, Arda, Veronique Hoste, and Lieve Macken (2017), SCATE taxonomy and corpus of machine translation errors, *in* Pastor, Gloria Corpas and Isabel Durán-Muñoz, editors, *Trends in E-tools and resources for translators and interpreters*, Vol. 45 of *Approaches to Translation Studies*, Brill | Rodopi, pp. 219–244.

Toral, Antonio and Víctor M. Sánchez-Cartagena (2017), A multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, Association for Computational Linguistics, Valencia, Spain, pp. 1063–1073.

Van Brussel, Laura, Arda Tezcan, and Lieve Macken (2018), A fine-grained error analysis of NMT, SMT and RBMT output for English-to-Dutch, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, European Language Resources Association (ELRA), Miyazaki, Japan, pp. 3799–3804.

Yimam, Seid Muhie, Iryna Gurevych, Richard Eckart de Castilho, and Chris Biemann (2013), WebAnno: A flexible, web-based and visually supported system for distributed annotations, *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Association for Computational Linguistics, Sofia, Bulgaria, pp. 1–6. https://www.aclweb.org/anthology/P13-4001.

6. Appendix



Condition A

* 3 Wat betekent het woord "dinosauruscuke"? Geef een synoniem of omschrijf.

* 4 Hoe zeker ben je van je antwoord?

○ heel zeker ○ een beetje zeker ○ niet zo zeker ○ helemaal niet zeker

Condition B

* 9 Wat betekent het onderstreepte woord in volgende zin?

Hij moet honderden kwetsende opmerkingen hebben gemaakt tijdens zijn lange carrière als luidruchtige talkradiostandaard.

Kies de optie met de meest passende omschrijving:

A) radiomaker
B) radiopresentator
C) spreekbuis
D) radiostandaard

* 10 Hoe zeker ben je van je antwoord?

○ heel zeker ○ een beetje zeker ○ niet zo zeker ○ helemaal niet zeker

Condition C

* 5 Wat betekent het onderstreepte woord in volgende zin?

Ze kregen bomgevaren en evacueerden kort hun campus

Geef een synoniem of omschrijf.

* 6 Hoe zeker ben je van je antwoord?

○ heel zeker ○ een beetje zeker ○ niet zo zeker ○ helemaal niet zeker

Figure 12: Screenshots of questions in the three conditions in SurveyMonkey