The Role of Compounds in Human vs. Machine Translation Quality

Kristýna Neumannová kristyna.neumannova@gmail.com
Ondřej Bojar bojar@ufal.mff.cuni.cz
Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics, Prague, Czech Republic

Abstract

We focus on the production of German compounds in English-to-German manual and automatic translation. On the example of WMT21 news translation test set, we observe that even the best MT systems produce much fewer compounds compared to three independent manual translations. Despite this striking difference, we observe that this insufficiency is not apparent in manual evaluation methods that target the overall translation quality (DA and MQM). Simple automatic methods like BLEU somewhat surprisingly provide a better indication of this quality aspect. Our manual analysis of system outputs, including our freshly trained Transformer models, confirms that current deep neural systems operating at the level of subword units are capable of constructing novel words, including novel compounds. This effect however cannot be measured using static dictionaries of compounds such as GermaNet. German compounds thus pose an interesting challenge for future development of MT systems.

1 Introduction

Assessing the quality of machine translation is a challenging task regularly tackled, e.g., in the manual evaluation of WMT translation task (Akhbardeh et al., 2021; Kocmi et al., 2022) or in WMT metrics task (Freitag et al., 2021, 2022). Various evaluation methods have been developed for this purpose. Manual evaluation in WMT has evolved from fluency and adequacy (Koehn and Monz, 2006) to direct assessment (DA, Graham et al., 2015) or MQM (Burchardt, 2013). Automatic evaluation is on the move from string matching techniques like BLEU (Papineni et al., 2002) or chrF (Popović, 2015) to embedding-based methods like COMET (Rei et al., 2020) or Prism (Thompson and Post, 2020). None of these approaches is particularly sensitive to specific subtle phenomena such as the presence or absence of compound words, a particular grammatical construction that is frequent in German. This paper focuses on German compounds, and how they occur in human and machine translations from English.

German has a highly productive word formation system mainly through compounding and derivation, especially for nouns (Barz, 2016, p. 2388). In this paper, we study German nominal compounds, which mostly consist of two constituents that are either complex or simple stems. The compounds in German are right-headed which means that the second element determines the morphosyntactic properties of the formed word. Additionally, semantically empty elements, called linking elements, can be added to the first stem of the compound (Barz, 2016, p. 2390).

Using compounds instead of multi-word expressions is a soft phenomenon related to text style, which can affect the perceived quality of the text. Native speakers regularly form new

Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track, pages 248–260
September 4–8, 2023, Macau SAR, China.

compound words to fulfill the needs requested by a particular dialogue or discourse situation. We believe that machine translation systems, operating on subword units, are able to produce complex words like humans, even if they were not included in the training data.

We know that splitting and determining German compounds is a complex task. Therefore, we relied on a list of compounds extracted from the German adaptation of WordNet called GermaNet (Henrich and Hinrichs, 2011). Operating on a closed list of compounds may provide an advantage for the analysis. Considering that the use of compounds is a stylistic matter, the exact list provides us with the possibility to group the observations of the phenomenon.

In the paper, we study several aspects of the data and models concerning the production of German nominal compounds.

2 Related Work

Most of the previous work on MT dealing with German compounds was done in the "classical" statistical machine translation (SMT). We found only a few papers, see below, about German compounds in neural machine translation (NMT), almost all of which were published before the introduction of the Transformer model (Vaswani et al., 2017), the current state of the art. Our work focuses on the production of German compounds in Transformer models, a topic that has not been adequately studied yet.

2.1 Compounds in SMT

The most common approaches to SMT operated on whole words. Therefore, they did not handle morphologically rich or compounding languages very well and dedicated methods were needed for processing compounds (by splitting them) and producing compounds (by merging them from pieces).

One of the first empirical methods for handling compounds was introduced by Koehn and Knight (2003), splitting compounds into parts that had been separately observed in the training data. The frequency of the compound constituents in the training data was the main criterion for the split.

Henrich and Hinrichs (2011) used an adapted version of a German morphological analyzer SMOR (Schmid et al., 2004) to improve the German compound splitting algorithm for determining the constituents of compounds. They combined an updated SMOR with other splitters, such as a pattern-matching-based splitter that considers all potential modifiers and heads, along with linking elements. This approach extracted a list of nominal German compounds from the German word net called GermaNet. As mentioned, we use this list for our analysis.

Sugisaki and Tuggener (2018) introduced an unsupervised method for compound splitting based on the idea of morpheme productivity, distinguishing between free morphemes (can stand alone as words) and bound morphemes within a word (appear only as parts of words). They computed the ratio between the counts of bound and free morphemes and selected a splitting with the lowest one i.e., preferring words consisting primarily of otherwise free morphemes.

Daiber et al. (2015) utilize vector representations of compounds and their parts to identify which word is likely a compound (its embedding is not far from the vector calculated from its parts).

Popović et al. (2006) focused on both German-English and English-German translation. For English-German, they split all compounds, trained the SMT system to produce split compounds and merged them in a post-processing step based on corpus statistics of compounds and their parts.

Stymne (2009) built upon Popović et al. (2006), adding a method based on a special token indicating the need to merge, and a method based on POS. These methods were evaluated in two ways: the overall translation quality and the performance of merging algorithms (the number,

type and quality of merges). It was shown that merging strategies could improve SMT quality; however, none of the investigated algorithms reached the number of compounds in the human-translated reference. The follow-up work (Stymne and Cancedda, 2011) additionally viewed the task as sequence labelling: words were labelled as to whether they should be joined or not.

Cap et al. (2014) synthesized new compounds by merging word parts based on their frequencies. Evaluation using BLEU did not show significant improvements which they sought for and validated compounds manually. Their method generated 100 more compounds (750 in total) than the baseline Moses decoder Koehn et al. (2007). Many of the generated compounds were correct translations of the source text even if they were not all confirmed by the reference translation.

2.1.1 Compounds in NMT

Neural MT reached the quality of SMT only after subword units such as Byte Pair Encoding (BPE, Sennrich et al., 2016) were invented. Splitting long words into smaller units in principle allows it to process as well as produce compounds in pieces without any dedicated focus. Weller-Di Marco and Fraser (2020) nevertheless tried explicit compound splitting as a preprocessing step, building upon Weller-Di Marco (2017) and Koehn and Knight (2003) but no significant improvement was observed.

Huck et al. (2017) investigated word segmentation strategies that incorporate more linguistic knowledge than the widely used BPE. One of the described strategies involved compound splitting and provided top-down segmentation that considers the frequency of the components, in contrast to BPE, which operates bottom-up. Compound splitting combined with suffix splitting improved BPE word segmentation in English-German translation, as evaluated by the BLEU score.

Macháček et al. (2018) examine linguistically-motivated or agnostic splits in German-to-Czech translation but observe no benefits from the motivated ones.

3 Experimental Setup

3.1 Data

In this section, we present the data that was used to analyse the presence or absence of German compounds in English-German translations, as well as the fixed dataset that was used to train our Transformer model. The compounds included in the systems' outputs and in the training data were identified based on a fixed list of compounds extracted from GermaNet.

3.1.1 GermaNet

GermaNet is a German word net that preserves the database format and structure of Princeton WordNet 1.5. Its central representation concept is the synset that groups synonyms of a given topic, such as *Streichholz* and *Zündholz* (matches for starting a fire). The word net captures semantic relations between the synsets and synonyms in them (Kunze and Lemnitzer, 2007). The authors distinguished two types of relations: lexical, such as synonymy and antonymy, and conceptual, like hyponymy, hypernymy, and others.

Henrich and Hinrichs (2011) presented a compound splitter to add semantic relations between compound constituents to GermaNet. For our analysis, we used only the list of nominal German compounds extracted from GermaNet (version v17.0, last updated in June 2022). The list contains 115,563 nominal German compounds with information on how they are split into two parts: the modifier and the head. The first part modifies the meaning of the second part, which carries the morphosyntactic features of the entire word (Barz, 2016, p. 2390). Compounds with more than two constituents can be recursively split by finding the split of its components in the GermaNet list.

3.1.2 WMT21

We used a dataset provided by the Sixth Conference on Machine Translation (WMT21, Akhbardeh et al., 2021) and tested our hypotheses on the outputs of systems submitted to the conference. Our own Transformer model was trained using the provided set of parallel training data and then tested on the Newstest2021 test set. The seven training parallel corpora were the same as those used for constrained systems submitted to WMT21. The constrained systems did not use any additional data except for the given corpora for training.

The news test set comprises around 1,000 sentences for all languages (1,002 for en-de). The authors of the test set guaranteed that the sentences were originally from the source language and then translated into the target language. Professional translation agencies performed the reference translations. Considering that English-German is a highly attractive language pair, it received special attention. A different translation agency provided a second reference, labelled "B"; however, it was found to be a post-edited version of one of the submitted systems, so it was discarded from the conference. The third reference translation was sponsored by Microsoft, labelled "C". The metric task organizers (Freitag et al., 2021) then provided a fourth reference, labelled "D".

3.2 Tools

Prior to identifying compounds in the outputs, we had to lemmatize the text. We used the UDPipe 2 (Straka, 2018) lemmatization method. In a small manual examination, we found that the pre-trained German GSD model[1] from the 2.10 version of Universal Dependencies models[2] is the best option for lemmatization of complex compounds.

Additionally, we used some minor tools during our analysis. For word segmentation, we used the subword-nmt (Sennrich et al., 2016) implementation to learn and apply BPE.[3] For estimating the overall translation quality of the outputs, we used the SacreBLEU (Post, 2018) implementation[4] of the BLEU metric.

3.3 Training of Vanilla Transformer

We selected FAIRSEQ (Ott et al., 2019) as the framework for training and evaluating Transformers. FAIRSEQ is an open-source tool used for sequence modelling. It allows researchers to train and evaluate their custom models for text-generating tasks such as translation, language modelling and summarization. It is written in PyTorch and designed to run on multiple GPUs.

We set aside 10% of the data for validation, as suggested by the translation example from FAIRSEQ.[5] Therefore, only 90% of the data was used for training. We trained several variations of the Transformer model. The modifications mainly concerned the creation of the subword dictionary, as summarized in Table 1.

We trained the models using the default FAIRSEQ Transformer configuration containing 6 decoder and 6 encoder layers, each with eight-headed attention. The setup differed from the default configuration in the following ways. The parameters were inspired by EdinSaar's submission to WMT21 (Tchistiakova et al., 2021). We operated on batches of a maximum size of 4,096 tokens. We used the Adam optimizer with setting $\beta_1 = 0.9$, $\beta_2 = 0.98$, and $\epsilon = 1e-9$. The dropout was set to 0.01. We utilized the GELU activation function. The learning rate was set to $3e-4$ and scheduled with an *inverse_sqrt* scheduler. We set 16,000 warmup updates with an initial learning rate of $1e-7$. The criterion for training was label-smoothed cross-

[1] https://universaldependencies.org/treebanks/de_gsd/index.html

[2] https://ufal.mff.cuni.cz/udpipe/2/models#universal_dependencies_210_models

[3] https://github.com/rsennrich/subword-nmt

[4] https://github.com/mjpost/sacreBLEU

[5] https://github.com/facebookresearch/fairseq/tree/main/examples/translation

system	seed	type of dictionary	size of dictionary
T40k	1	joint	40,000
T2x40k	1	separated	2 x 40,000
T10k	1	joint	10,000
T2x40k-2	1,000	separated	2 x 40,000

Table 1: Training setups of our Transformer model

entropy. The models were trained on a heterogeneous grid server that contains Quadro RTX 5000, GeForce GTX 1080 Ti, RTX A4000, and GeForce RTX 3090 cards. We utilized 8 GPU cards across several weeks to train the models.

4 Compounds in MT Outputs

In our analysis, we primarily rely on compounds that are contained in the GermaNet list and search for them in WMT21 translations. We compare the counts of compounds found in reference translations and state-of-the-art system outputs. We present counts of compounds and sentences containing at least one compound for each reference and output translation separately. We also report the number of compounds and sentences with compounds confirmed in one or more of the reference translations. The results are sorted by the decreasing number of found compounds and listed in Table 2.

Table 2 shows that human reference translations contain more compounds than any other MT system outputs. The best reference regarding the compound number is the reference "C", with 955 compounds found in 593 sentences. That is over 100 compounds more than in the best MT system. The source text for all the translations comprised 1,002 sentences, so more than half of them led to the generation of some compound in the best reference translation. Considering all sentences where at least one human translator used a compound, we get 756 sentences with 995 different compounds. For all translations, we have 898 out of 1,002 sentences where at least one compound occurred.

Considering only the number of produced compounds, the best MT system is the constrained system *Nemo*, with 842 compound occurrences in 559 sentences (see Table 2). 87% of the compounds are approved by references. Unconstrained systems that employ extra training data are expected to have better results than constrained systems. However, two constrained systems, *Nemo* and *UF*, each produced more compounds than any of the unconstrained systems. The worst system, *ICL*, contained 138 fewer compounds than the best MT system and 251 fewer compounds than the best human translation.

It is important to note that the same concept can be translated using various compounds, so even when the MT output contains a correct compound, it need not be confirmed by the reference. We mitigate this issue by considering four different human translations instead of only one, and also by reporting the number of sentences in which any compound appeared.

4.1 Novel Compounds

MT models operating on subword units have the potential to generate unseen words in their output. We first examined the number of compounds from GermaNet that were produced by systems but were not present in the training data. We found that there were no newly created compounds from GermaNet in the outputs of the constrained system. We expected this subset to be very small or empty, so it was not surprising.

We also looked at whether there were any compounds from GermaNet that were not present in the training data. We found that the training data did not include approximately 3.5% (4,200)

system	# compounds	in refs	# sents	in refs
ref-C	955		593	
ref-D	946		591	
ref-A	901		566	
ref-B	878		569	
C-Nemo	842	735	559	511
C-UF	802	710	532	487
UC-metricsystem2	801	670	533	476
UC-Online-B	798	705	532	484
UC-Facebook-AI	796	735	533	511
C-eTranslation	794	696	530	486
UC-VolcTrans-GLAT	792	756	533	521
UC-Online-W	791	741	533	515
UC-metricsystem1	790	698	530	486
UC-metricsystem3	787	641	518	475
UC-metricsystem5	783	674	531	480
C-WeChat-AI	783	707	527	493
UC-VolcTrans-AT	782	678	531	480
UC-Online-Y	776	658	522	464
UC-happypoet	770	668	526	473
UC-metricsystem4	769	685	515	475
C-Manifold	768	666	514	460
UC-Online-A	767	685	520	478
C-nuclear_trans	762	656	514	466
C-HuaweiTSC	761	673	516	473
C-UEdin	758	666	513	466
UC-Online-G	754	648	516	464
C-P3AI	740	655	505	467
C-BUPT_rush	731	627	495	443
C-ICL	704	595	485	426

Table 2: Compounds appearance in English-German translations in WMT 21 (counts of all appearances of compounds and counts of sentences with compounds plus its subsets approved by reference translations).

of the compounds from GermaNet. This set of compounds presents the upper bound to our observations: we are curious if the systems can produce compounds not seen in their training data, but our diagnosis method (the GermaNet list) offers only 4,200 compounds that could be noticed – and we have no idea if they are relevant to the test text.

Therefore, we decided to explore the subset of compounds produced by constrained systems but that were not present in the training data or the GermaNet list. However, there is no direct way to accomplish this. We collected all words that were not seen in the training data; note that we considered all words here, and manually verified which of them are compounds, see below.

Determining whether a word is a valid or conceivable German compound is not easy. We can consider all compounds produced by native speakers as proper German words. To identify valid novel words, we searched large monolingual corpora, such as Araneum Germanicum Maius (Benko, 2014) or the DWDS dictionary (Klein and Geyken, 2010). To include com-

pounds used in German articles or web pages, we used Google search.

The constrained WMT21 systems produced a total of 304 unique new words that started with a capital letter, indicating that they were possible nouns. Approximately half of them were found by Google anywhere on the Internet. During the analysis, we discovered various groups of words. Some words were of foreign origins, such as the English verb *MACED* (capitalized because it was so in the source text), human names like *Shaquia* and *Bhadauria*, and geographic locations like *Mambourin*. Regarding compounds, we discovered an example of a joint English phrase, *Speakupfordemocracy*, and many German compounds. Out of 304 novel nouns, we manually determined 229 of them as compounds. The exact number of identified compounds and foreign words for each constrained system is displayed in Table 3 below.

We examined the German compounds and discovered many of them were made up of meaningful constituents but were neither included in the training corpus nor found by Google. Naturally, they were also not found in DWDS. Below, we list several instances of this phenomenon. Most of the examples make sense as two separate words, and combining them into a compound is possible (Example 1). We also provide examples of more complex words produced by the systems that do not have any known sense (see Example 2). Their two constituents can form proper German words (Examples 2d and 2e), but their concatenation is not known as a German compound. Finally, there are also examples that cannot be clearly divided into just two parts (for instance, 2b or 2c were formed from three meaning-bearing parts).

The systems also produced compounds that existed and were found by Google but were not contained in DWDS or Araneum Germanicum Maius. The examples of these rare words we found during the analysis are listed in Example 3. These words were also produced by humans in some texts or articles but did not belong to a common vocabulary. In total, 103 of 229 novel compounds were found by Google. This analysis provides several examples of the productivity of NMT models in terms of compounds. We examined these examples further and searched for them in a bigger German corpus, namely in *Deutsche Referenzkorpus* (DeReKo).[6] The DeReKo corpus revealed that beside all compounds from Example 3, Examples 1a and 1c can also be considered as existing compounds.

(1) Words not seen in DWDS or Araneum, made from known constituents
   a. *Kondolenzbotschaft* (a condolence message)
   b. *Gladiatorenmodus* (the mode of a gladiator)
   c. *Quarantäneentscheidung* (the decision on quarantine)

(2) Very complex words not seen in DWDS or Araneum, made from known constituents
   a. *Sanktionsüberwachungsteam* (a team for observing sanctions)
   b. *Gefangenenfreistellungsprogramm* (a program for releasing prisoners)
   c. *Passagierlokalisierungsformular* (a form for localizing travellers)
   d. *Notfallgesundheitsdirektorin* (a female director for emergency health issues)
   e. *Telekommunikationsnetzausrüstung* (equipment for telecommunication networks)

(3) Rare compounds found by Google but not seen in DWDS or Araneum
   a. *Flughafenvertrag* (airport contract)
   b. *Pandemiekrise* (pandemic crisis)
   c. *Kartoffelwurzeln* (potato root)
   d. *Republikanerkollege* (a Republican colleague)
   e. *Amateurfehler* (a layman's error)

[6] https://www.ids-mannheim.de/digspra/kl/projekte/korpora

system	# nouns	n. in ref	# comp.	c. in ref	# foreign
C-Manifold	106	52	69	22	34
C-HuaweiTSC	102	57	58	24	36
C-UF	101	58	60	24	36
C-WeChat-AI	95	54	51	19	35
C-UEdin	93	56	49	20	37
C-eTranslation	92	56	55	23	32
C-Nemo	87	51	44	17	38
C-nuclear_trans	87	47	44	13	35
C-P3AI	86	45	49	15	32
C-ICL	82	47	41	15	35
C-BUPT_rush	81	43	40	11	34

Table 3: Categories of novel words (nouns, out of which some were classified as compounds and some as foreign nouns) produced by constrained systems according to our manual analysis. We also report how many of them were confirmed by the reference ("in ref").

After discovering many newly produced compounds in systems' outputs, we also explored words produced by human translators in the references that were not contained in the training data in order to compare them. We are aware of the fact that comparing the vocabulary of human translations to training corpora might not be ideal for demonstrating productivity regarding composition. However, we can consider the huge training corpora as a sample of common vocabulary knowledge.

We detected several novel compounds from our examples also in the reference translations: The compounds *Kondolenzbotshaft* and *Gladiatorenmodus* (Examples 1a and 1b above) were found in references B and D, while references A and C contained a modification of the second compound, *Gladiatormodus*. Two of the complex compounds that seemed to have no established sense were also created by humans, namely the word *Sanktionsüberwachungsteam* (Example 2a) in references B and C and *Passagierlokalisierungsformular* (Example 2c) in references A, B, and C. We found three of the listed rare compounds (Examples 3) in the references – *Flughafenvertrag* (in references A, C, and D), *Pandemiekrise* (in references B and C) and *Kartoffelwurzeln* (in all references). We can assume that these words were created correctly and reflect the discourse situation of the source test text. Particular phrases in the source text encouraged the translators to create these compounds. However, we can not easily decide the correctness of the other novel words.

After providing a manual analysis and listing some examples, we grouped the observations together. Table 3 displays the number of novel nouns created by constrained MT systems, their cooccurrence with reference translations and their distribution into categories. We distinguished three categories: compounds, foreign words or names, and others, such as web domain names or meaningless words. Only the first two categories are listed in the table. We also counted how many of the novel compounds were also present in the reference translations. In most of the constrained systems, more than a half of novel nouns appeared to be compounds, as shown in Table 3.

To conclude, the MT systems are, same as humans, capable of generating novel words, although it did not seem so when relying on a fixed list of compounds. At the same time, the number of compounds in the translations is still higher for human translators than for the MT systems when we count both novel words and compounds found by GermaNet.

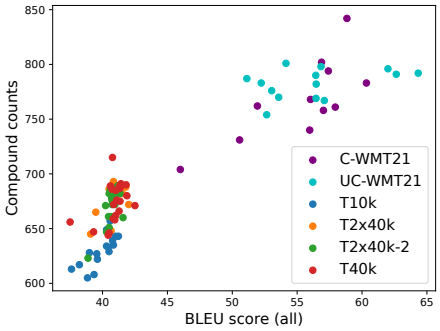

Figure 1: Comparison of BLEU scores (against 3 references) to the number of produced compounds for WMT21 systems and our systems.

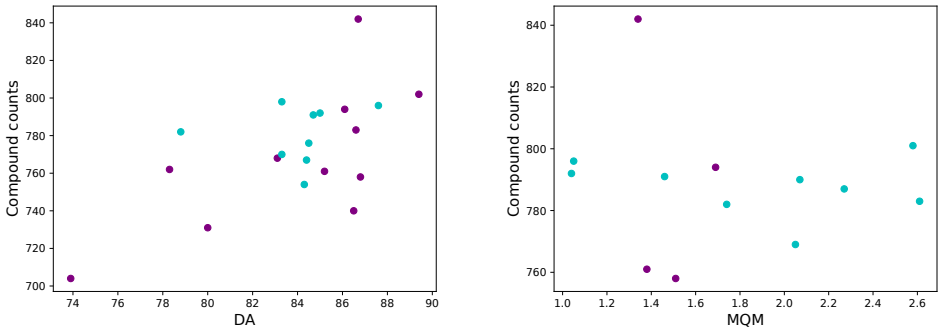

Figure 2: Comparison of human evaluation to the number of produced compounds for WMT21 systems. Legend same as in Figure 1.

4.2 Compounds vs. Overall Quality

We calculated BLEU scores for WMT21 systems to compare their overall translation quality with the number of produced compounds from GermaNet.

We visualised the relationship between both scores for all the constrained MT systems, including four versions of our Transformer, as shown in Figure 1. The graph showed the correlation between the overall quality of translations measured by BLEU and the number of generated compounds. The dependency shows an almost linear pattern. The Pearson correlation coefficient was 0.75 for constrained WMT21 systems, 0.41 for unconstrained, and 0.59 for all WMT21 systems combined. Thus, overall quality serves as a good indicator of *relative* performance in terms of compounds, although it does not reflect the human level.

To compare the number of produced compounds with human evaluation (DA and MQM), we presented the correlation in Figure 2. The Pearson correlation coefficient for DA and the compound number was 0.69 for constrained WMT21 systems, 0.17 for unconstrained, and 0.60 for all WMT21 systems combined. Regarding MQM and the compound number, the Pearson correlation coefficients were -0.24 for constrained WMT21 systems, -0.19 for unconstrained and -0.10 for all WMT21 systems combined.

In summary, these results indicate that the relationship between the number of produced

compounds and human evaluation varies depending on the evaluation metric and the type of system used (constrained vs. unconstrained). BLEU score seems to reflect the presence or absence of compounds slightly better than DA and substantially better than MQM. Nonetheless, our study highlights the potential of using the number of produced compounds as an additional metric to evaluate the quality of machine translation systems.

5 Conclusion

We examined the production of German compounds in Transformer models in English-to-German MT. Our analysis revealed that reference translations consistently contain more compounds than MT systems. We confirmed that Transformers have the ability to generate new words including compounds but evaluating compound production using closed lists or existing general manual evaluation methods (DA, MQM) is not effective. This opens space for further exploration of compound production as well as their evaluation.

Acknowledgements

This research was partially supported by the grant 19-26934X (NEUREM3) of the Czech Science Foundation.

References

Akhbardeh, F., Arkhangorodsky, A., Biesialska, M., Bojar, O., Chatterjee, R., Chaudhary, V., Costa-jussa, M. R., España-Bonet, C., Fan, A., Federmann, C., Freitag, M., Graham, Y., Grundkiewicz, R., Haddow, B., Harter, L., Heafield, K., Homan, C., Huck, M., Amponsah-Kaakyire, K., Kasai, J., Khashabi, D., Knight, K., Kocmi, T., Koehn, P., Lourie, N., Monz, C., Morishita, M., Nagata, M., Nagesh, A., Nakazawa, T., Negri, M., Pal, S., Tapo, A. A., Turchi, M., Vydrin, V., and Zampieri, M. (2021). Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online. Association for Computational Linguistics.

Barz, I. (2016). German. In Müller, P. O., Ohnheiser, I., Olsen, S., and Rainer, F., editors, *Word-Formation. An International Handbook of the Languages of Europe*, volume 4, pages 2387–2410. Mouton de Gruyter, Berlin.

Benko, V. (2014). Aranea: Yet another family of (comparable) web corpora. In Sojka, P., Horák, A., Kopeček, I., and Pala, K., editors, *Text, Speech and Dialogue*, pages 247–256, Cham. Springer International Publishing.

Burchardt, A. (2013). Multidimensional quality metrics: a flexible system for assessing translation quality. In *Proceedings of Translating and the Computer 35*, London, UK. Aslib.

Cap, F., Fraser, A., Weller, M., and Cahill, A. (2014). How to produce unseen teddy bears: Improved morphological processing of compounds in SMT. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 579–587, Gothenburg, Sweden. Association for Computational Linguistics.

Daiber, J., Quiroz, L., Wechsler, R., and Frank, S. (2015). Splitting compounds by semantic analogy. In *Proceedings of the 1st Deep Machine Translation Workshop*, pages 20–28, Praha, Czechia. ÚFAL MFF UK.

Freitag, M., Rei, R., Mathur, N., Lo, C.-k., Stewart, C., Avramidis, E., Kocmi, T., Foster, G., Lavie, A., and Martins, A. F. T. (2022). Results of wmt22 metrics shared task: Stop using bleu – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation*, pages 46–68, Abu Dhabi. Association for Computational Linguistics.

Freitag, M., Rei, R., Mathur, N., Lo, C.-k., Stewart, C., Foster, G., Lavie, A., and Bojar, O. (2021). Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774, Online. Association for Computational Linguistics.

Graham, Y., Baldwin, T., and Mathur, N. (2015). Accurate evaluation of segment-level machine translation metrics. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1183–1191, Denver, Colorado. Association for Computational Linguistics.

Henrich, V. and Hinrichs, E. (2011). Determining immediate constituents of compounds in GermaNet. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 420–426, Hissar, Bulgaria. Association for Computational Linguistics.

Huck, M., Riess, S., and Fraser, A. (2017). Target-side word segmentation strategies for neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 56–67, Copenhagen, Denmark. Association for Computational Linguistics.

Klein, W. and Geyken, A. (2010). Das digitale wörterbuch der deutschen sprache (dwds). In *Lexicographica: International annual for lexicography*, pages 79–96. De Gruyter.

Kocmi, T., Bawden, R., Bojar, O., Dvorkovich, A., Federmann, C., Fishel, M., Gowda, T., Graham, Y., Grundkiewicz, R., Haddow, B., Knowles, R., Koehn, P., Monz, C., Morishita, M., Nagata, M., Nakazawa, T., Novák, M., Popel, M., Popović, M., and Shmatova, M. (2022). Findings of the 2022 conference on machine translation (wmt22). In *Proceedings of the Seventh Conference on Machine Translation*, pages 1–45, Abu Dhabi. Association for Computational Linguistics.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Koehn, P. and Knight, K. (2003). Empirical methods for compound splitting. In *10th Conference of the European Chapter of the Association for Computational Linguistics*, Budapest, Hungary. Association for Computational Linguistics.

Koehn, P. and Monz, C. (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. Association for Computational Linguistics.

Kunze, C. and Lemnitzer, L. (2007). *Computerlexikographie: Eine Einführung*. Narr Francke Attempto Verlag.

Macháček, D., Vidra, J., and Bojar, O. (2018). Morphological and language-agnostic word segmentation for nmt. In *Proceedings of the 21st International Conference on Text, Speech and Dialogue—TSD 2018*, pages 277–284, Cham, Switzerland. Springer-Verlag.

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., and Auli, M. (2019). fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Popović, M. (2015). chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Popović, M., Stein, D., and Ney, H. (2006). Statistical machine translation of german compound words. In Salakoski, T., Ginter, F., Pyysalo, S., and Pahikkala, T., editors, *Advances in Natural Language Processing*, pages 616–624, Berlin, Heidelberg. Springer Berlin Heidelberg.

Post, M. (2018). A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Rei, R., Stewart, C., Farinha, A. C., and Lavie, A. (2020). COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Schmid, H., Fitschen, A., and Heid, U. (2004). SMOR: A German computational morphology covering derivation, composition and inflection. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Straka, M. (2018). UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.

Stymne, S. (2009). A comparison of merging strategies for translation of German compounds. In *Proceedings of the Student Research Workshop at EACL 2009*, pages 61–69, Athens, Greece. Association for Computational Linguistics.

Stymne, S. and Cancedda, N. (2011). Productive generation of compound words in statistical machine translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 250–260, Edinburgh, Scotland. Association for Computational Linguistics.

Sugisaki, K. and Tuggener, D. (2018). German compound splitting using the compound productivity of morphemes. In Barbaresi, A., Biber, H., Neubarth, F., and Osswald, R., editors, *14th Conference on Natural Language Processing - KONVENS 2018*, pages 141–147. Austrian Academy of Sciences Press. 14th Conference on Natural Language Processing (KONVENS 2018), Vienna, Austria, 19-21 September 2018.

Tchistiakova, S., Alabi, J., Dutta Chowdhury, K., Dutta, S., and Ruiter, D. (2021). EdinSaar@WMT21: North-Germanic low-resource multilingual NMT. In *Proceedings of the Sixth Conference on Machine Translation*, pages 368–375, Online. Association for Computational Linguistics.

Thompson, B. and Post, M. (2020). Automatic machine translation evaluation in many languages via zero-shot paraphrasing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 90–121, Online. Association for Computational Linguistics.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *CoRR*, abs/1706.03762.

Weller-Di Marco, M. (2017). Simple compound splitting for German. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 161–166, Valencia, Spain. Association for Computational Linguistics.

Weller-Di Marco, M. and Fraser, A. (2020). Modeling word formation in English–German neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4227–4232, Online. Association for Computational Linguistics.