Terminology in Neural Machine Translation: A Case Study of the Canadian Hansard

Rebecca Knowles **Samuel Larkin** **Marc Tessier** **Michel Simard**
National Research Council Canada
1200 Montreal Road, Ottawa, Ontario, CANADA K1A 0R6
FirstName.LastName@nrc.gc.ca

Abstract

Incorporating terminology into a neural machine translation (NMT) system is a feature of interest for many users of machine translation. In this case study of English–French Canadian Parliamentary text, we examine the performance of standard NMT systems at handling terminology and consider the tradeoffs between potential performance improvements and the efforts required to maintain terminological resources specifically for NMT.

1 Introduction

Incorporating data from a specialized or particular lexicon is a commonly-desired property of neural machine translation (NMT) systems used in computer-aided translation settings. A number of approaches have been proposed for this, including modifications to decoding, training systems for special behavior, and training with external lexicons. Results vary, highlighting the fact that they navigate a difficult compromise between imposing specific lexical choices on the decoder, and interfering as little as possible with its behavior (Yvon and Rauf, 2020). Parallel to this, lexical resources developed by terminologists and translators are not necessarily designed and formatted with NMT requirements in mind, and not all terms they contain naturally lend themselves to incorporation: for example, it may be difficult to process terms with many morphological variants or terms whose translation depends on the context. Extracting these resources' content for NMT and maintaining the two resources in sync may pose practical challenges. In light of this, it is reasonable to ask when, how, and whether it is worth implementing these methods in a real-life, practical setting.

Here, we use the scenario of Canadian Parliamentary translation as a case study to examine questions about terminology and machine translation performance. The data we use consists of transcriptions and translations of speech in the Canadian House of Commons (the proceedings, or Hansard), with most speech originally in English (then translated to French), a much smaller part spoken in French (then translated to English), and a very small fraction in other languages. Parliamentary translators have access to a document that provides guidance on terminology, from which we have manually extracted word and phrase pairs.

We are interested in the following questions:

- In this *specific* case, should we attempt to explicitly handle terminology in our NMT systems? If so, how?
- More generally, in which scenarios does it make sense to incorporate terminology into an NMT system? What tradeoffs might researchers and users want to consider?

With this data, we begin by examining just how "usable" the terminology actually is for NMT incorporation, and how consistently it is used in human translations. We then compare how an NMT system (without any special terminology handling) performs on these terms, through both automatic and manual evaluations.

In our analysis, we highlight the following considerations for researchers and users of NMT interested in handling terminology:

- How is the terminology bank formatted?

- How frequent is the terminology in the text?
- How consistently is it used by translators?
- How does an unaugmented NMT system perform?

In this particular use case, we find that the terminology bank is appropriately designed for translator use rather than optimized for machine translation, the terms are relatively infrequent in the corpus, there is a mix of how consistent the term translations should be (even in high-quality human translations), and the NMT system performs reasonably well on the terms that are most unambiguous. For these reasons, there would be a relatively high cost in terms of human time (to produce and keep current an additional machine-readable version of the term bank) to handle terminology for a relatively small amount of potential improvement. Depending on translator preferences and how much of a pain point terminology errors are, there may be appropriate alternatives, such as flagging potential terminology errors (though these also come with their own costs). We also discuss how the relative costs and payoffs may differ in other settings.

2 Data

2.1 Fixed Terms

Parliamentary translators maintain a pair of internal documents called the *Aide-mémoire du service des débats*, intended for those translating into French, and *Aide-mémoire for the House of Commons*, for those translating into English. Both documents contain a wealth of information regarding structural, orthographic and typographical conventions, common translation problems, etc. In particular, they each contain an alphabetical list of terms and phrases of interest for translators. In practice, the English *Aide-mémoire* is relatively small, with only 275 terminological entries, and so for this study, we focus on the French document, in its April 28, 2021 version. From this Microsoft Word document, we manually extracted 1162 term entries, which we annotated for usability in computer-assisted translation. We identified 605 (52%) as being "directly usable": these are entries of the form (X, Y), where X is a unique source-language term, Y is its prescribed translation in the target language, both of which can be matched in running text with minimal processing (see Section 3). In all that follows, we call these *fixed terms*. The top section of Table 1 shows examples of such entries. Of the remaining entries, 235 would require further processing for matching, such as accounting for morphological variations or disambiguating context, and 322 are monolingual, i.e. they only specify either the source or the target term, along with a full-text explanation (middle and bottom sections of Table 1, respectively).

Of course, the *Aide-mémoire* documents do not contain all the terminology there is in the Hansard. The number of topics that are addressed in parliament is huge, and parliamentary translators routinely need to consult other resources, such as the *TERMIUM Plus*[1] term bank (Bernier-Colborne et al., 2017), bilingual concordances, such as *TransSearch*[2] (Bourdaillet et al., 2010) and various internal resources.

In all that follows, we use only entries from the French *Aide-mémoire* that were identified as "directly usable". We refer to this set of entries as the *English–French Parliamentary Fixed Terms*, which we abbreviate $\mathbf{PFT_{ef}}$.[3]

2.2 Bitext

We use the XML-formatted version (original version as used by translators) of data from Sessions 39-1 to 43-2 of the Canadian Hansard (House of Commons),[4] crawled from the web.[5] All data is automatically segmented into sentences and aligned using NLTK tools (Bird and Loper, 2004).

We use this data to build NMT systems, as well as to test performance on fixed terms. The three most recent debates (120, 121, & 122) from Session 43-2, we use for evaluation. From these, we set aside 2000 randomly sampled lines for MT validation and testing; the remaining 10093 lines, which we refer to as **FT-test**, we use for evaluating the handling of terminology.

All the remaining debates are used as training data for NMT systems (see Table 2). We trained Transformer models (Vaswani et al., 2017) using Sockeye (Hieber et al., 2018) version 2.3.14, with the following modifications to default settings: we set gradient clipping to *absolute*, maximum sentence length to 200 tokens, checkpoint intervals to

[1] https://www.btb.termiumplus.gc.ca

[2] http://tsrali.com/

[3] We plan to release the PFT_{ef}, test data, and code at https://github.com/nrc-cnrc/PFT-ef-EAMT23

[4] In Session 43-2, we use data from debates 001 to 122.

[5] https://www.ourcommons.ca/documentviewer/en/house/latest/hansard

Source term	Target term	Comment
Fixed terms:		
airspace	espace aérien	
dudeplomacy	diplocopinage	
human trafficking	traite des personnes	
Require processing:		
intelligence (agency)	(organisme de) renseignement	Optional parts in parentheses.
bundle the votes	regrouper les votes	Morphological variants of the verb.
business plan	plan d'entreprise/d'activités	Depending on if it applies to a company vs. a government.
Informational (monolingual):		
	Alliés, les	with a capital "A" in the context of World Wars I & II.
	bien-être social	do not use; use *aide sociale* or *assistance sociale* (welfare)
ordinary Canadians		try to vary: *les Canadiens*, *la population*, *tout un chacun*...

Table 1: Example entries from the *Aide-mémoire du service des débats*. (Comments are ours.)

1000, we use batches of ∼8192 tokens/words, a shared vocabulary for source and target, we optimize for BLEU and perform validation on a fixed set of 1000 sentences.

Corpus	EN–FR	FR–EN	Total
Train	4,152,732	1,415,330	5,679,055
FT-test	7235	2692	10,093

Table 2: Corpus size (lines), with language direction. The two directions (EN–FR and FR–EN) do not sum to the total because we exclude certain pieces of boilerplate text for which translation direction is not specified.

3 Analysis

We begin by examining the frequency with which the terms of the PFT_{ef} appear in the text of the Hansard. To handle issues of tokenization, we begin with raw/detokenized text and use NLTK's `word_tokenize` (Bird and Loper, 2004) to tokenize the PFT_{ef} terms, the Hansard source and reference, and the (detokenized) MT output. Prior to tokenization, we perform apostrophe standardization,[6] though this impacts only a small number of segments. In this analysis, we restrict ourselves to the data where the human translation direction matches the machine translation direction.

There are 605 unique English terms in the PFT_{ef} and 600 unique French terms (599 after apostrophe standardization). The PFT_{ef} is directional and intended for English to French translation, so it is unambiguous in the English to French direction, and has some minor ambiguities in the French to English direction. This means that the most appropriate analysis is in the English–French direction, though we still include some analyses in the French–English direction (with caveats) in Table 3.[7] In most cases, a sentence contains only one instance of a particular term, making it easy to compute whether the term's translation appears on the target side or not. In the cases where a term appears more than once in the source, we do not perform alignment, but compute a clipped count: if the term appears n times in the source, we check how many times its translation appears in the target, giving credit only up to n (i.e., if it appeared $n + 1$ times, we neither penalize nor reward the extra instance). In all cases, the set of terms appearing in FT-test are a subset of those in train. Some initial observations are as follows: both the percentage of terms and the percentage where the source term's translation appears in the corresponding reference are lower for the (less-appropriate) FR–EN direction; we do not examine this in depth. Looking at the machine translation percentages as compared to the reference percentages, we find that the MT produces PFT_{ef} target terms more often than the reference does, although the gap is not particularly large.

Figure 1 shows the distribution of PFT_{ef} term occurrences in EN–FR training data. Five appear more than 10,000 times: climate change (14586), liberal party (16537), first nations (26702), conservatives (53883), and budget (67943).

We focus our attention on the English–French portion of the *FT-test* data set. Of the 7235 English text segments in the sample, 595 (8.2%) contain at least one (lowercase) match to one of the PFT_{ef} terms. As some segments contain more than one source term, there are a total of 694 instances of source terms in that data set. For 594 of

[6] Converting three different characters to one standard.

[7] In the case of the ambiguous French–English pairs, we used the final entry as the corresponding term.

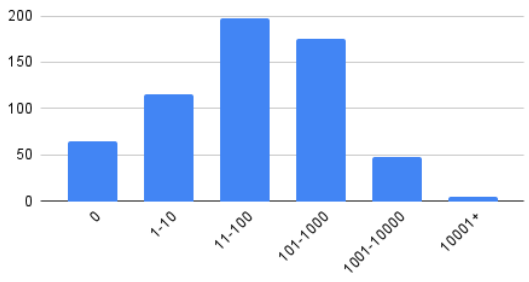

Figure 1: Distribution of PFT_{ef} source-side term occurrences in the EN–FR training data.

Corpus	% of terms	Src. #	% ref.	% MT
EN–FR				
Train	89.4% (541)	376680	78.0%	-
FT-test	13.2% (80)	694	85.4%	87.2%
FR–EN				
Train	66.3% (397)	120845	81.0%	-
FT-test	5.0% (30)	230	75.7%	76.1%

Table 3: PFT_{ef} term matches in corpora. The **% of terms** column shows what percentage of the full set of unique source side terms appeared in the corpus (type count in parentheses). The **Src. #** column shows raw match counts. The **% ref.** column shows the percentage of instances that had both a source term on the source side and its translation from the PFT_{ef} in the target reference (counts clipped; extra instances in the target side are neither penalized nor rewarded); the **% MT** column shows the same but for MT output.

these instances, we find that the reference translation uses the corresponding French term from the PFT_{ef}.[8] Looking at the remaining 100 term instances, i.e. those for which the reference translation does not contain the prescribed target term, we quickly identify that 6 correspond to alignment errors: as explained in Section 2, our corpus was segmented and aligned automatically; this process occasionally produces errors, in the form of badly segmented and misaligned segments. We discard the offending segments and their translations (both reference and MT) for the rest of this analysis. This leaves us with 94 (13.6%) occurrences of PFT_{ef} terms for which the reference translations do not use the corresponding target term.

We perform a similar analysis on the machine translations of the EN–FR *FT-test* data set. We find 602 (87.2%) translations that contain the prescribed French term, versus 88 (12.8%) that don't.

There are various reasons why a prescribed term might not appear in a translation, including a (human or machine) translation error.[9] However, in many cases, a missing term does *not* imply an error. For example, a translation might have been formulated in such a way that the entity or notion to which the term refers is referred to with a paraphrase or a pronoun in the translation. In other cases, the context may render the term redundant or superfluous. Sometimes a term occurrence is actually part of a larger term within which it should be translated differently; for example, while the prescribed French translation for *climate change* is *changement climatique*, the official French name for the "Intergovernmental Panel on *Climate Change*" (IPCC) is "Groupe d'experts intergouvernemental sur l'*évolution du climat*" (GIEC).

To better understand how humans and MT behave with regard to PFT_{ef} terms, we manually annotated a subset of the *FT-test* data set. We collected all *FT-test* segments that matched one or more source terms from the PFT_{ef}, but for which either the reference or the machine translation did not contain at least one occurrence of the prescribed translation for each matching source term. In all, there are 123 such source segments, each with two translations: 28 for which the reference translation uses the prescribed term but the MT doesn't; 40 for which the MT uses the prescribed term but the reference translation doesn't; and 55 for which both translations are missing a prescribed term. In order to get a better balance between the translations that use the prescribed terms and those that don't, we added 49 segments, randomly selected from *FT-test* that match both source and target terms. In all, our annotation set contains 172 distinct segments, containing 185 source term matches.

For each of the 185 term instances, the reference and the machine translations were analyzed to determine whether the matched source term was correctly translated in the context. The question that annotators were asked was: "Is the term highlighted in the Source rendered correctly in the

[8] Again, when a segment contains multiple matches of the given source term, we verify that the reference translation contains at least as many occurrences of the corresponding target term as of the source term.

[9] It is worth noting that parliamentary translators are not always to blame for terminology errors found in the reference translations. In some cases, the Hansard will contain excerpts from pre-existing documents, for which an official translation already exists. Translators are not permitted to fix errors in these pre-existing translations. In other situations, the fault may lie with the speaker in the House of Commons which may have used an incorrect or inexact term; it is then the translator's duty to attempt to fix this, by translating the speaker's *intent* rather than their words.

Translation?"[10] A first-pass annotation was performed by two of the authors.[11] Each annotator assigned one of three tags to each translation: *Correct*, *Incorrect*, or *Unsure*. The two annotators then jointly produced consensus labels by reconciling their differences together.

All term translations with a *Unsure* label that remained after consensus were then submitted to a second-pass annotation (43 of the 370 translations). This second pass was done through individual interviews with three volunteer translators from the parliamentary service.[12] From these judgments, we assign the majority label.[13]

	Reference		**MT**	
target term appears:	yes	no	yes	no
Translation is:				
Correct	91	78	97	58
Incorrect	0	16	0	30

Table 4: Manual annotation of reference and machine translations for instances of PFT_{ef} source terms. We provide separate counts for translations that use the corresponding PFT_{ef} target term and those that don't.

Table 4 reports overall counts of *Correct* vs. *Incorrect* translations, for reference and machine translations, with and without the prescribed translated term. When the target term was used in the translation, the translation of the source term was always judged to be correct: this was true for both reference ($91/91$) and machine ($97/97$) translations. We find that reference translations that don't use the prescribed term are still overwhelmingly judged positively by annotators: only 16 of 94 such reference translations (17%) were labelled as incorrect. In contrast, 30 of the 88 machine translations (34%) not using the prescribed target term were judged to be incorrect.

[10]The original question was formulated in French as: "*Le terme « X » dans la Source est-il rendu adéquatement dans la Traduction? (Oui/Non)*" with *X* replaced by the actual term.

[11]The annotation of the 49 segments in which both translations contained the target term was performed by a single annotator.

[12]This process conforms to the recommendations of our institution's Research Ethics Board, who were consulted regarding this work.

[13]In practice, there were 67 *Unsure* translations. But 24 of these were deemed similar enough to another example that it was possible to derive their labels from second-pass annotations once these were completed.

4 Related Work

We now briefly discuss a number of approaches that have been applied to the problem of handling fixed terms, including modifications to decoding, training systems for special behavior, and training with external lexicons. For a much more extensive review of approaches to lexicons and terminology resources in NMT, see Yvon and Rauf (2020). These approaches can be applied independently or combined, and each has various strengths and weaknesses. Decoding modifications, such as lexically constrained decoding (Hokamp and Liu, 2017; Post and Vilar, 2018) typically come with strong guarantees (i.e., that the desired term will appear in the output), do not require the lexicon to be known in advance, and do not necessarily require any modification to training procedures. Downsides to these include that they may be overly strict (e.g., failing to inflect forms) and that forcing low probability output can harm overall translation quality ("reference aversion"). There is also no guarantee that the tokens are in the correct location, are produced by translating the correct source token, or are not concatenated with adjacent tokens. Hasler et al. (2018) seek to improve terminology placement in constrained decoding by incorporating alignment (via attention) to tie the relevant source tokens to the desired target token output. Susanto et al. (2020) modify the beam search procedure to enforce translation of words (as specified in XML-style input) or to perform look-ahead to ensure they are generated.

Training for special behavior, through placeholders (Post et al., 2019) or factors (Dinu et al., 2019) does not require a fixed lexicon in advance, but it does not offer the same strong guarantees of producing fixed terms. However, it sometimes successfully results in correctly inflected terms. Bergmanis and Pinnis (2021) expand on Dinu et al. (2019), specifically with the goal of better handling morphological variants.

If a lexicon is fixed in advance, it can be incorporated into NMT training (Arthur et al., 2016; Nguyen and Chiang, 2018), though this does not hold strong guarantees of lexicon production and does not generalize to new lexicon entries in the future. Exel et al. (2020) compare the approaches in Dinu et al. (2019) with constrained decoding, and find that in their use case, this training for specific behavior "offers a good trade-off for terminology enforcement in a production setting." They also

note that even baseline systems had fairly high performance on translation quality, though term translation did lag behind the specialized systems.

5 Discussion and Conclusions

Our analysis shows that MT is twice as likely as humans to commit terminology errors in the Hansard, for terms in the PFT_{ef}: when the MT system does not produce the target term, its term translation is incorrect approximately 34% of the time, as compared to 17% reference translations in the same scenario (see Table 4). This is not surprising, and clearly, MT researchers still have work to do. It is, however, useful to put in perspective the numbers that lead to this conclusion. Our tests were conducted on a set of 7235 English segments. Of these, less than 10 percent (694) matched any of the 605 terms of the PFT_{ef}. In the cases where they did match a term, the MT produced the prescribed translation in over 85% of its translations. We did not manually validate the quality of all these translations, but evidence suggests that it is very unlikely that any of these contains errors relative to the PFT_{ef} terms (no doubt, they contain other types of errors). Even when the translation does not use the prescribed term, two-thirds of machine translations are adequate with regard to PFT_{ef} terminology. In the end, we estimate that the MT makes terminology errors in approximately one out of every 250 Hansard segments (0.4%).

In this work we focused on the kinds of "fixed" fixed terms that could be most easily incorporated into lexicon-based approaches to NMT fixed term augmentation. In our setting, this meant excluding close to half the terms from the translators' term bank (48%). In particular, we excluded terms that would almost always require significant inflection (e.g., verbs), though some approaches to handling fixed terms are capable of handling morphological variation and future work may wish to broaden the use of terms to more fully capture the kinds of term banks used by translators, as argued by Bergmanis and Pinnis (2021). Unlike prior work that has dealt with fixed terms by enforcing terminology in the test sets (Alam et al., 2021), we leave the parallel text as it is, but also examine cases where, even within our more constrained setting, fixed terms are not "fixed" in the strictest sense. We observed situations where they are fluently replaced with pronouns (to avoid repetition), where the term is translated differently as part of a larger phrase, and other such sources of variation. We note that this may be particular to this corpus and term bank; a corpus that is heavy on highly-technical terminology (e.g., chemistry, medicine) might have a greater proportion of terms that are *truly* fixed. Thus we encourage researchers and users to check how "fixed" the terminology is in real text, even if only at the shallow automatic level.

In light of this, and in a scenario such as ours, it seems reasonable to ask whether it is worth implementing any of the methods outlined in Section 4. To cover only the terms we analyzed here, most approaches would be suitable. However, we note that the *Aide-mémoire* documents are periodically updated, which would require retraining in the case of approaches that require a known and fixed terminology in advance.[14] Even though the NMT system made twice as many terminology errors than the reference text did (when the target term was not produced), its term translations were still judged to be adequate the majority of the time. This raises the question: if we enforced term translation, what would happen in those sentences? Would the result be just as good, or might it produce less-fluent translations? As we did not perform manual evaluation of quality beyond the terms, this is not a question that our current data can answer.

One simple alternative to consider is to automatically flag to the translator's attention those translations (human or machine) that do not match the PFT_{ef} term when the source segment does. However, it should be noted that this too has a cost, not so much in software development, but in maintenance of the lexical resources, which must then be encoded in machine-readable format. This may include expanding morphological variants, as well as keeping the machine-readable term bank up-to-date. This would need to be weighed against the time spent correcting machine translation errors, as well as the potential inconvenience or trust loss due to flags that are false positives. The time needed to correct MT errors should be weighed against the time needed to maintain and update the resource specifically as a tool for the NMT system.

[14] Note here that we are not putting in question the *Aide-mémoire* documents themselves. As pointed out earlier, these documents are rich in information. They serve an invaluable role for parliamentary translators in documenting terminological decisions and for training newcomer translators to the service. Importantly, they are designed for translator use, allowing for information about context and ambiguity that is often skimmed over in work on "fixed" terms.

Acknowledgements

We wish to thank the parliamentary translation team and all our partners at the Canadian government's Translation Bureau. Without their help and expertise, this work would not have been possible.

References

Alam, Md Mahfuz Ibn, Ivana Kvapilíková, Antonios Anastasopoulos, Laurent Besacier, Georgiana Dinu, Marcello Federico, Matthias Gallé, Kweonwoo Jung, Philipp Koehn, and Vassilina Nikoulina. 2021. Findings of the WMT shared task on machine translation using terminologies. In *Proceedings of the Sixth Conference on Machine Translation*, pages 652–663, Online, November. Association for Computational Linguistics.

Arthur, Philip, Graham Neubig, and Satoshi Nakamura. 2016. Incorporating discrete translation lexicons into neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1557–1567, Austin, Texas, November. Association for Computational Linguistics.

Bergmanis, Toms and Mārcis Pinnis. 2021. Facilitating terminology translation with target lemma annotations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3105–3111, Online, April. Association for Computational Linguistics.

Bernier-Colborne, Gabriel, Caroline Barrière, and Pierre André Ménard. 2017. Fine-grained domain classification of text using TERMIUM plus. In *Proceedings of Language, Ontology, Terminology and Knowledge Structures Workshop (LOTKS 2017)*, Montpellier, France, September. Association for Computational Linguistics.

Bird, Steven and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain, July. Association for Computational Linguistics.

Bourdaillet, Julien, Stéphane Huet, Philippe Langlais, and Guy Lapalme. 2010. Transsearch: from a bilingual concordancer to a translation finder. *Machine Translation*, 24(3-4):241.

Dinu, Georgiana, Prashant Mathur, Marcello Federico, and Yaser Al-Onaizan. 2019. Training neural machine translation to apply terminology constraints. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3063–3068, Florence, Italy, July. Association for Computational Linguistics.

Exel, Miriam, Bianka Buschbeck, Lauritz Brandt, and Simona Doneva. 2020. Terminology-constrained neural machine translation at SAP. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 271–280, Lisboa, Portugal, November. European Association for Machine Translation.

Hasler, Eva, Adrià de Gispert, Gonzalo Iglesias, and Bill Byrne. 2018. Neural machine translation decoding with terminology constraints. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 506–512, New Orleans, Louisiana, June. Association for Computational Linguistics.

Hieber, Felix, Tobias Domhan, Michael Denkowski, David Vilar, Artem Sokolov, Ann Clifton, and Matt Post. 2018. The sockeye neural machine translation toolkit at AMTA 2018. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 200–207, Boston, MA, March. Association for Machine Translation in the Americas.

Hokamp, Chris and Qun Liu. 2017. Lexically constrained decoding for sequence generation using grid beam search. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1535–1546, Vancouver, Canada, July. Association for Computational Linguistics.

Nguyen, Toan and David Chiang. 2018. Improving lexical choice in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 334–343, New Orleans, Louisiana, June. Association for Computational Linguistics.

Post, Matt and David Vilar. 2018. Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1314–1324, New Orleans, Louisiana, June. Association for Computational Linguistics.

Post, Matt, Shuoyang Ding, Marianna Martindale, and Winston Wu. 2019. An exploration of placeholding in neural machine translation. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 182–192, Dublin, Ireland, August. European Association for Machine Translation.

Susanto, Raymond Hendy, Shamil Chollampatt, and Liling Tan. 2020. Lexically constrained neural machine translation with Levenshtein transformer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3536–3543, Online, July. Association for Computational Linguistics.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Guyon, I., U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Yvon, François and Sadaf Abdul Rauf. 2020. Utilisation de ressources lexicales et terminologiques en traduction neuronale. Technical report, LIMSI-CNRS.