# Exploring the effects of vocabulary size in neural machine translation: Galician as a target language

**Daniel Bardanca Outeirinho**[1] and **Pablo Gamallo**[1] and **Iria de-Dios-Flores**[1 2] and **José Ramom Pichel Campos**[1]
[1]Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS)
Universidade de Santiago de Compostela
[2] Department of Translation and Language Sciences, Universitat Pompeu Fabra
{danielbardanca.outeirino, pablo.gamallo, iria.dedios, jramon.pichel}@usc.gal

## Abstract

We present a systematic analysis of the influence of vocabulary size on the performance of Neural Machine Translation (NMT) models, with a particular focus on Galician language models (Basque-Galician, Catalan-Galician, and English-Galician). The study encompasses an exploration of varying vocabulary sizes employing the Byte Pair Encoding (BPE) subword segmentation methodology, with a particular emphasis on BLEU scores. Our results reveal a consistent preference for smaller BPE models. This preference persists across different scales of training data. The study underscores the importance of vocabulary size in NMT, providing insights for languages with varying data volumes.

## 1 Introduction

This research is part of an initiative dedicated to the advancement of linguistic technologies specifically designed for the Galician language (de Dios-Flores et al., 2022). Before the beginning of this initiative, Galician Machine Translation (MT) systems were rule-based (e.g. Apertium (Forcada et al., 2011)), thus one of the objectives of this initiative is to bring Galician up to speed on MT technology by spearheading the development of NMT models between Galician and other strategic languages (Ortega et al., 2022). These include English, and the remaining official languages of the Kingdom of Spain: Basque, Catalan, and Spanish.

While the ultimate goal of our project is the creation of open multilingual models with other strategic languages, such as Portuguese (European variant), our initial focus has been on crafting bilingual models for the target language pairs. This allows us to have greater control over the quality of the parallel corpora, which contain original and synthetic data, as well as over the optimal size of the vocabulary built with the tokenization models. The aim of this paper is precisely to study and identify the most appropriate vocabulary size for training and inference within a given language pair and specific training corpus. Specifically, we investigate what is the most optimal vocabulary size as a function of the size of the parallel training corpus, taking into account that there are substantial divergences in the sizes of the training corpora for the language pairs under consideration. For instance, the Galician-Basque corpus is much smaller than the Galician-English corpus.

The main contribution of this work lies in the development of experiments that substantiate the trends identified in the few existing studies focused on exploring the optimal vocabulary size in NMT. The remainder of this paper is organized as follows: in Section 2, we discuss the challenges posed by the Zipfian distribution in NMT and the BPE approach. Section 3 describes the experiments we performed, including the language pairs and the range of vocabulary sizes tested. Section 4 discusses the results observed across all models and varied data sizes, highlighting the significance of vocabulary size in NMT when training bilingual models for languages with diverse data volumes.

## 2 Related work: vocabulary size in NMT

The words present in natural language models tend to follow a Zipfian distribution, where a word's rank is roughly inversely proportional to its frequency within any given natural language corpus. As a result, a small number of words are highly frequent, while the majority fall into the tail end of low or very low frequencies. This Zipfian distribution produces at least two challenges for any NMT system (and NLP systems in general). On one hand, the input sequence often contains many words that were not learned previously during training. On the other, the word distribution is unbalanced, potentially creating biases towards frequent patterns and severely degrading performance (Johnson and Khoshgoftaar, 2019).

To address these two issues, a subword vocabulary is employed, entailing the decomposition of word types into smaller components. The most popular approach is known as Byte Pair Encoding (BPE) (Sennrich et al., 2016). BPE fundamentally allows the breakdown of infrequent words into more common subwords. Translation is a technique that inherently requires an open vocabulary. Therefore, the utilization of subword models to address issues related to unbalanced word distribution is a prevalent practice in NMT. By employing BPE to encode rare and unknown words as sequences of subword units and choosing the appropriate level of subword segmentation, we can enhance translation performance (Kudo, 2018). Since the appearance of this algorithm, it has become standard practise to incorporate word segmentation approaches relying on BPE when developing NMT models. It is a very effective algorithm, but the reasons for this effectiveness are not well understood (Galle, 2019).

Subword models can prove especially advantageous for languages with limited linguistic resources, as the availability of parallel corpora is scarce and limited in size. Consequently, a significant portion of the vocabulary is absent from these datasets. Previous work showed that reducing the number of BPE merge operations resulted in substantial improvements, reaching a decrease of 5 points of BLEU (Sennrich and Zhang, 2019) when tested on RNN models. Lankford et al. (2021) achieved significantly different results by altering the vocabulary sizes of several small English-Irish Transformer models trained on the same parallel corpus. The authors observed that the best results were achieved with a BPE model optimized to produce a small subword vocabulary of 16k tokens. It is important to note that although BLEU scores provide a useful metric for evaluating machine translation performance, no single metric can perfectly evaluate the quality of machine-translated text. Therefore, a combination of BLEU scores with other metrics such as COMET (Rei et al., 2020), and human evaluation are necessary to fully understand the limitations of a model.

Furthermore, Gowda and May (2020) analyze the effect of various vocabulary sizes on NMT performance on several language pairs with different corpora sizes. Their experiments revealed that a large vocabulary with more than 30K tokens is unlikely to produce optimal results unless the parallel corpora is large. On small (30K tokens) to medium (1.3M tokens) corpora sizes, a small vocabulary of less than 10K tokens is sufficient.

Following the experimental strategy of Gowda and May (2020), our primary goal in this short paper is to determine the optimal BPE vocabulary size for different sizes of training parallel corpora between Galician and Catalan, Basque and English. Our findings are then compared with those of Gowda and May (2020), who conducted similar research on four different language pairs: English-German, German-English, English-Hindi, and English-Lithuanian. Notably, the importance of considering vocabulary sizes in language modeling enterprises go beyond NMT. For instance, similar effects to those observed in NMT are related to those studies focusing on how to transfer vocabulary from the pre-trained model to the fine-tuned model (e.g. Samenko et al. (2021) and Bostrom and Durrett (2020)). In these studies the vocabulary size is a relevant element that needs to be considered when training a fine-tuned model, similarly to how it also influences the quality of translation models.

## 3 Experiments

To conduct the study proposed in this work, we performed two distinct experiments involving the following three language pairs: Basque-Galician (eu-gl), Catalan-Galician(ca-gl), and English-Galician(en-gl). Given that the parallel corpora available for these pairs vary in size, we were able to analyze the impact of vocabulary size at various scales: small (eu-gl), medium (eu-gl, ca-gl), and large (en-gl).

| Model | Size |
|---|---|
| eu-gl aut | 400k |
| eu-gl aut+sint | 3.5M |
| ca-gl | 3.5M |
| en-gl | 30M |

Table 1: Size of the parallel corpus for each model

Table 1 offers a numerical representation of each scale. The eu-gl pair was tested on two models trained with different datasets: small(400k) and medium (3.5M), whereas ca-gl and en-gl were always trained on the same dataset of 3.5M and 30M lines respectively. This is because original data for eu-gl i.e. data that was originally written by hu-

mans in these languages, was scarce compared to the other two language pairs. In order to compensate for this disparity and improve the quality of the translation model, a new dataset with synthetic data was developed. These new data were the result of combining the Portuguese-Galician (pt-gl) module of Apertium (Forcada et al., 2011) and transliterating text written in Portuguese orthography to the local Galician spelling as described in(Ortega et al., 2022). It is also important to note that the linguistic distance between the source languages (i.e. Catalan, English, Basque) and the target (Galician) varies considerably. All models utilized in the development of this paper are publicly available on our GitHub repository [1].

**Experiment 1:** The first experiment involved training new models with vocabularies ranging from 1k to 50k. Both source and target vocabularies were kept separate. The BLEU scores obtained are the result of evaluating all language pairs on the FLORES-200 dataset (Team et al., 2022). Experiments involving vocabularies higher than 50k were not included because they did not alter the analysis and conclusions presented in the next section. All models for this experiment were based on a transformer architecture with 6 layers, 8 attention heads, and 512 hidden vector size.

**Experiment 2:** In the second experiment, we created new BPE models for each language based on a fixed corpus size of 400k tokens, which matched the size of our smallest parallel corpus. These models were not used to train the models presented in Experiment 1. We wanted to examine how the BPE models segmented the words into subwords and how that affected the translation quality. Our goal was to find out if there was a direct link between BLEU scores and subword ratio, which we define as the result of dividing the number of words in a text by the total amount of subwords generated by the BPE algorithm.

## 4 Results and discussion

**Experiment 1:** Figure 1 shows the BLEU scores for the translation pairs (y-axis) using models with different vocabulary sizes (x-axes). The trends observed indicates a preference for smaller BPE models. It seems that models with a vocabulary size exceeding 40,000 yield inferior results compared

[1] https://github.com/proxectonos/propor2024_vocabulary

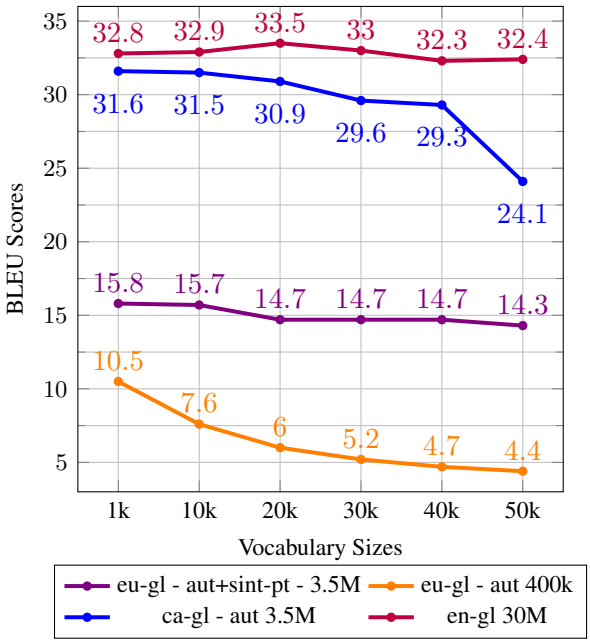


Figure 1: BLEU scores for the for translation pairs using models with different vocabulary sizes

to those with smaller vocabularies. This trend remains consistent across all models, regardless of the volume of training data and language. Interestingly, the preference for smaller BPE models becomes more pronounced as the size of the training data decreases. For instance, a compact eu-gl model (400k) paired with a BPE model trained with a 1k vocabulary size yields a BLEU score that is twice as high as that of a model trained with the same dataset but a vocabulary size of 30k. Both intermediate (3.5M) and large models (30M) continue to perform better with fewer than 30k types. However, larger datasets do not exhibit as significant a variation in performance between 1k and 40k tokens. While intermediate-sized models for eu-gl and ca-gl still performed optimally at 1k, the difference in BLEU score between 1k and 10k is marginal, at only 0.1, compared to a difference of 2.9 BLEU in the smallest model. Moreover, in the case of ca-gl there is a significant performance drop with 50k models, a trend not observed in the other two language pairs tested. This raises the question of whether linguistic proximity between Catalan and Galician could be playing a role here. These findings are generally in agreement with Gowda and May (2020), where small vocabulary sizes perform the best, and the smaller the training data, the earlier the score peaks. However, while what they labeled as big datasets (4.5M sentences) performed better at 48k vocabulary size, we have

found that our similarly sized (3.5M sentences) still performed optimally with smaller vocabulary sizes. Even our largest model trained on a significantly bigger dataset of 30M sentences preferred much lower sizes, performing its best with a 20k token configuration.

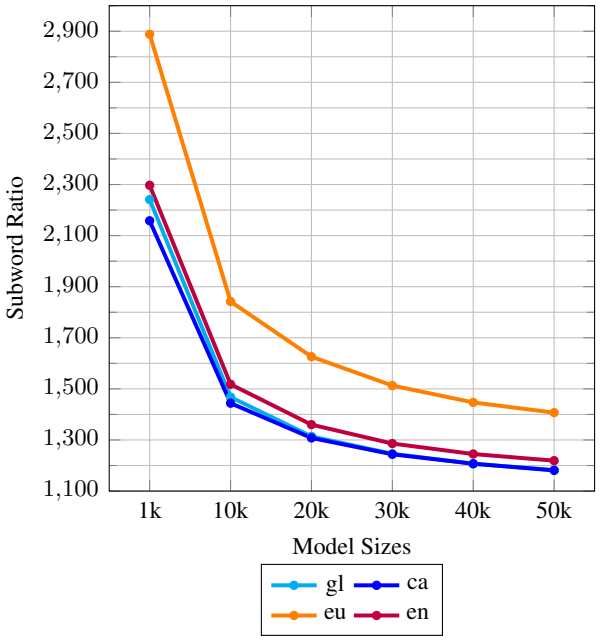


Figure 2: Subword ratio for Galician, Basque, Catalan, and English

**Experiment 2:** Figure 2 shows the evolution of the *subword ratio* in all languages used during Experiment 1 as the vocabulary size increases. We find that, as expected, BPE models produced a greater number of subword divisions the smaller the model is. Out of the four languages, Basque, which is a morphologically rich non-Indo-European agglutinative language, stands out for always producing more subdivisions than the three Indo-European languages represented. The subword ratio shows that there is a clear difference between a morphologically agglutinative language (with many more word divisions) and non-agglutinative languages. By contrast, no striking differences are observed between clearly inflectional languages, such as Galician and Catalan (Romance languages) and English, with a more limited inflection.

Finer subdivision, however, is not directly linked to higher BLEU scores. From the observations depicted in the two figures, it seems that smaller vocabulary sizes tend to result in more word subdivisions, which improves the granularity and detail of new models when dealing with small training data, but when dealing with larger datasets, the importance of a small or big vocabulary (which always result in a lower subword ratio) seems to be overridden by the sheer size of the input data.

## 5 Conclusion

We presented a systematic analysis of the influence of vocabulary size on the performance of NMT models. When juxtaposing the findings from Experiments 1 and 2, it becomes apparent that models with reduced vocabulary sizes not only lead to an increased number of word subdivisions but also tend to produce superior BLEU scores. This implies that a reduction in vocabulary size could potentially enhance both the detail of the models and the quality of their translations. Nevertheless, it is crucial to take into account the unique attributes of each dataset and language, such as proximity between source and target languages, data size, and the morphology of each language, when determining the most suitable vocabulary size.

Overall, our results align with the general recommendation by Gowda and May (2020) to prefer small rather than large vocabulary sizes. This holds especially true for us when dealing with small datasets (less than 1.5M), which seem to benefit from extremely small vocabulary sizes (1k). We concur with this observation. Nevertheless, our findings question the necessity of expanding the vocabulary beyond 20k when training models for Galician. Regarding vocabulary sizes, it becomes evident that small vocabularies should consistently be considered as the initial choice for new models.

## Acknowledgements

This publication was produced within the framework of the Nós Project, which is funded by the Spanish Ministry of Economic Affairs and Digital Transformation and by the Recovery, Transformation, and Resilience Plan - Funded by the European Union - NextGenerationEU, with reference 2022/TL22/00215336, and by the Junta de Galicia through the collaboration agreements signed in with the University of Santiago de Compostela in 2021 and 2022.

Additionally, the authors of this article received funding from MCIN/AEI/10.13039/501100011033 and the European Union Next Generation EU/PRTR (TED2021-130295B-C33), the Galician Government (ERDF 2014-2020: Call ED431G 2019/04, and ED431F 2021/01), by

MCIN/AEI/10.13039/501100011033 (grants with references PID2021-128811OA-I00 and TED2021-130295B-C33, the latter also funded by the European Union Next Generation EU/PRTR), and a Juan de la Cierva Grant (JDC2022-049433-I) funded by MCIN/AEI/10.13039/501100011033 and the European Union Next Generation EU/PRTR.

We are grateful to CESGA (Centro de Supercomputación de Galicia) for allowing us access to their infrastructure to carry out the experiments.

## References

Kaj Bostrom and Greg Durrett. 2020. Byte pair encoding is suboptimal for language model pretraining. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4617–4624, Online. Association for Computational Linguistics.

Iria de Dios-Flores, Carmen Magariños, Adina Ioana Vladu, John E. Ortega, José Ramom Pichel, Marcos García, Pablo Gamallo, Elisa Fernández Rei, Alberto Bugarín-Diz, Manuel González González, Senén Barro, and Xosé Luis Regueira. 2022. The nós project: Opening routes for the Galician language in the field of language technologies. In *Proceedings of the Workshop Towards Digital Language Equality within the 13th Language Resources and Evaluation Conference*, pages 52–61, Marseille, France. European Language Resources Association.

M. L. Forcada, M. Ginestí-Rosell, J. Nordfalk, J. O'Regan, S. Ortiz-Rojas, J. A. Pérez-Ortiz, F. Sánchez-Martínez, G. Ramírez-Sánchez, and F. M. Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*, 25:127–144.

Matthias Galle. 2019. Investigating the effectiveness of bpe: The power of shorter sequences. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 1375–1381, Hong Kong, China. Association for Computational Linguistics.

Thamme Gowda and Jonathan May. 2020. Finding the optimal vocabulary size for neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3955–3964, Online. Association for Computational Linguistics.

Justin M. Johnson and Taghi M. Khoshgoftaar. 2019. Survey on deep learning with class imbalance. *Journal of Big Data*, 6:1–54.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Seamus Lankford, Haithem Alfi, and Andy Way. 2021. Transformers for low-resource languages: Is féidir linn! In *Proceedings of Machine Translation Summit XVIII: Research Track*, pages 48–60, Virtual. Association for Machine Translation in the Americas.

John Ortega, Iria de Dios-Flores, José Ramom Pichel, and Pablo Gamallo. 2022. A neural machine translation system for galician from transliterated portuguese text. In *SEPLN-PD 2022. Annual Conference of the Spanish Association for Natural Language Processing 2022: Projects and Demonstrations,* , pages 92–95.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 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.

Igor Samenko, Alexey Tikhonov, Borislav Kozlovskii, and Ivan P. Yamshchikov. 2021. Fine-tuning transformers: Vocabulary transfer. *CoRR*, abs/2112.14569.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 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.

Rico Sennrich and Biao Zhang. 2019. Revisiting low-resource neural machine translation: A case study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 211–221, Florence, Italy. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.