THE UNIVERSITY *of* EDINBURGH

Linguistic Typology for Neural Machine Translation

Arturo Oncevay

Doctor of Philosophy

Institute for Language, Cognition and Computation
School of Informatics
University of Edinburgh

2023

Abstract

The vast structural diversity of languages worldwide, compounded by the problem of scarce resources, remains a challenge for machine translation research. To address this problem, we leverage knowledge from the field of linguistic typology, which describes the structural diversity and common properties of the world's languages.

In this thesis, we investigate which variables or concepts of linguistic typology impact neural machine translation performance. First, we propose a combined language representation that encodes language variables from typology databases, specifically syntax variables, and fuses them with pre-trained multilingual language embeddings. We demonstrate the higher quality of the new language representations by assessing their performance on computational typology tasks such as typological feature prediction and phylogenetic inference. Next, we show that the combined language space can be leveraged to improve multilingual machine translation tasks and reduce negative transfer by creating clusters of multilingual models with significantly related languages, obtaining benefits across languages with different training sizes compared to robust baselines, and with the advantage of working efficiently when adding new languages to a multilingual setting.

Furthermore, we investigate the impact of typological variables associated with morphology on machine translation. Morphology examines how words are formed, a process that varies across languages and is related to subword segmentation, a critical aspect of current machine translation systems. Specifically, we demonstrate that a higher degree of morphological fusion and synthesis usually corresponds to lower translation quality. We perform this analysis at the word level and obtain consistent results at the segment level for several language pairs. Finally, we study the extreme case of high synthesis (polysynthesis) and low-resource scenarios, which are typically present in endangered languages from the Americas. We build machine translation resources for Amerindian languages, find that unsupervised segmentation methods perform comparably or better than morphologically supervised ones, and propose a less data-dependent segmentation strategy based on syllable units with promising results in our case study.

Overall, our work sheds light on the impact of linguistic typology on machine translation, specifically on the relevance of syntax and morphological variables in low-resource and structurally diverse languages.

Lay Summary

Machine translation, which allows computers to automatically translate text from one language to another, is a difficult task, especially when dealing with languages that are very different from each other or have few resources available. In this research, the author uses knowledge from the field of linguistic typology, which describes the structural diversity and common properties of the world's languages, to investigate how they impact machine translation performance for these challenging scenarios.

The author proposes a new method to represent languages that combines information from typology databases with pre-trained language embeddings, which leads to better performance on tasks related to understanding the properties of languages. The study also shows that this new method can be used to improve multilingual machine translation and reduce negative transfer caused by differences between languages. The thesis then focuses on the impact of morphology, which studies word formation processes, on machine translation. The author finds that, morphological complexity, measured in the variables of synthesis and fusion, is related to machine translation performance. Finally, the thesis evaluates the extreme case of high-synthesis and low-resource languages, such as those found in endangered languages from the Americas, by developing machine translation resources for Amerindian languages and proposing a less data-dependent segmentation approach with promising results for a case study.

Overall, this research highlights the importance of understanding the structural diversity of languages and the impact of linguistic typology on machine translation.

Acknowledgements

I would like to express my heartfelt appreciation to my mother, father, brother, and grandmother for their unwavering support and unconditional love. I am deeply grateful for the special role they have played in my journey. I also extend my gratitude to my dearest friends and extended family from Peru for their continuous encouragement throughout my years in Edinburgh.

My deepest gratitude goes to my supervisors, Lexi and Barry, for their invaluable support, guidance, and understanding. I am indebted to them for their encouragement, constructive criticism, and belief in my abilities. Without them, this thesis would not have been possible.

I would also like to express my appreciation to my peers at the School of Informatics at the University of Edinburgh, as well as to other researchers, collaborators, co-authors, and friends from around the world who have played vital roles in my academic journey. Their insightful discussions and feedback have significantly enriched my academic experience.

Furthermore, I acknowledge César and Andrés, former fellow researchers at PUCP, now friends, for introducing me to the world of research and guiding my early academic steps in Peru. Likewise, I am also grateful to my friend Roberto, whose passion for endangered languages ignited my interest in developing language technologies for understudied languages in NLP.

Lastly, I am thankful for the many friends and lovely people I have met during my time in Edinburgh. The community here made this beautiful city feel like home. I will definitely come back to visit Edinburgh again and again.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Arturo Oncevay*)

To my younger self: be proud of who you are and what you can achieve.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

While low-resource languages have historically faced challenges in benefiting from Neural Machine Translation (NMT) due to a lack of training data and resources, recent advancements in transfer learning (Zoph et al., 2016) and multilingual learning (Bapna et al., 2022) have significantly improved NMT's ability to translate many low-resource languages (Haddow et al., 2022). However, given that there are thousands of languages spoken worldwide, according to Glottolog (Nordhoff and Hammarström, 2012), only a tiny fraction of all possible language pairs can currently be translated effectively by machine translation systems. Although the lack of resources is the most pressing and widely discussed limitation (Joshi et al., 2020), there is another far less studied challenge: the vast structural diversity in languages worldwide. This presents a severe problem for developing robust natural language processing (NLP) applications for multiple languages, including machine translation (Ponti et al., 2019).

Languages exhibit significant variations in their grammar, syntax, morphology, or vocabulary (Sapir, 1921; Comrie, 1989). For example, some languages have complex case systems that determine the grammatical roles of nouns and pronouns in a sentence, while others have simpler case systems or rely more on word order or prepositions. In Spanish, for instance, a noun's gender and number can change its form, and the adjective that modifies it must also agree in gender and number. In Chinese, however, word order is much more important than inflection, and the same word can be used as a noun or a verb without changing its form. This illustrates how different languages convey meaning differently, and these challenges are particularly pronounced when working with low-resource languages. Due to the limited availability of training data

for many of these languages, it is often difficult to obtain sufficient examples to account for every possible variation.

While multilingual machine translation for dozens of languages is currently the most common approach to address the language diversity issue (Aharoni et al., 2019; Zhang et al., 2020), it is not necessarily a full solution. Rather, it is a pragmatic approach given the limited availability of data for many languages and the lack of explicit knowledge about their structural differences. Nonetheless, by leveraging data for multiple languages, massive multilingual NMT systems can offer significant benefits in terms of transfer capabilities and shared vocabulary representation space for all languages involved, particularly when translating between high-resource and low-resource languages (Haddow et al., 2022), and works particularly well for languages that are similar (Bapna et al., 2022).

However, what does language similarity mean in this context? Different corpus-based metrics and language family relationships have been applied in several approaches, such as ranking models for choosing a suitable language candidate for transfer learning (Lin et al., 2019) or clustering similar languages in smaller groups for training multilingual models (Tan et al., 2019). Nonetheless, corpus-based similarity metrics rely on the availability of data, which is severely restricted for underrepresented or endangered languages, such as those spoken in the Americas (Mager et al., 2021) or Africa (Nekoto et al., 2020). Additionally, while language family relationships provide a means of categorising languages into groups, they only offer surface-level information and split groups of contrasting sizes. For example, in two widely used multilingual datasets for machine translation, TED Talks (Qi et al., 2018) and OPUS-100 (Zhang et al., 2020), the number of Indo-European languages significantly surpasses that of other language families (33 out of 53 in TED Talks and 49 out of 100 in OPUS-100). In contrast, different language families contain only a handful of entries; for instance, TED Talks includes only three languages from the Turkic and Uralic families, and they are the language families with more members, included in the dataset, after the Indo-European group.

To fully address the challenge of language diversity, it is necessary to gain insights into the shared and distinct features across languages that are not necessarily family-related, as well as to distinguish between differences within a language family. Linguistic typology presents an opportunity to address this challenge, as it is a field that describes and studies the structural diversity, common properties, and patterns of the world's languages (Comrie, 1989). However, computational research often demands

concrete data rather than abstract concepts. Fortunately, machine-readable typology databases such as the World Atlas of Language Structure (WALS; Dryer and Haspelmath, 2013) offer a wealth of language-level variables that expert linguists in different linguistic fields have crafted. Linguistic typology databases store, for example, variables about the word order in languages, including whether a language is head-initial or head-final, which refers to whether the head of a phrase appears first or last, respectively. They also store data on morphological typology, such as the level of fusion present in a language and whether affixation or compounding are used to build new words.

These kinds of features can be used as a valuable source of knowledge to support multilingual NLP tasks, as previous surveys have highlighted (O'Horan et al., 2016; Ponti et al., 2019). For instance, Daiber et al. (2016) used WALS typological features of word order to develop a universal reordering, a relevant task in the statistical machine translation era, prior to the predominance of neural models. Such insights can be particularly beneficial in developing more accurate and effective machine translation models for underrepresented and endangered languages, which often lack sufficient training data. However, it is important to note that not all typological features or concepts might be relevant for machine translation applications, and additional research is needed to determine which features are most useful. Therefore, we raise our main research question: **which variables or concepts of linguistic typology impact neural machine translation performance?**

To address this, we need to look at the details. Table 1.1 provides examples of typological features sourced from WALS. Notably, word order features related to syntax may hold significant relevance for expanding machine translation to languages with limited resources (Daiber et al., 2016). This is because more accurate translations require a precise ordering of words in the output sequence, and different approaches have been developed to pre-order or re-order data to enhance low-resource machine translation (Zhou et al., 2019; Murthy et al., 2019), especially for translating between distant languages. Moreover, the typological features from morphology, which studies the word formation processes, may also hold relevance in this context. This is particularly important in machine translation, which relies on subword segmentation[1] for limiting the vocabulary size of a model, and needs to compound subword elements into meaningful words at the decoding step, including the ability to generate rare words

[1] A subword is a smaller textual unit than a word, often a character fragment or character sequence based on their frequency. See more details in §2.1.2 of Chapter 2.

Area	Feature	Values	#langs.
Word Order	Order of Subject (S), Order (O) and Verb (V)	SOV	564
		SVO	488
		VSO	95
		VOS	25
		OVS	11
		OSV	4
		No dominant order	189
	Order of Adjective (A) and Noun (N)	A-N	373
		N-A	879
		No dominant order	110
		Internally-headed relative clauses	5
Morphology	Inflectional Synthesis of the Verb	Categories per word: 0-1	5
		2-3	24
		4-5	52
		6-7	31
		8-9	24
		10-11	7
		12-13	2
	Fusion of Selected Inflectional Formatives	Exclusively concatenative	125
		Exclusively isolating	16
		Exclusively tonal	3
		Tonal/isolating	1
		Tonal/concatenative	2
		Ablaut/concatenative	5
		Isolating/concatenative	13
Lexicon	Finger and Hand	Identical	72
		Different	521
	Number of Basic Colour Categories	3-4	20
		4.5-5.5	26
		6-6.5	34
		7-7.5	14
		8-8.5	6
		9-10	8
		11	11

Table 1.1: Samples of features from WALS (Dryer and Haspelmath, 2013), including their area, values, and number of annotated languages per feature-value.

not seen during training (Sennrich et al., 2016b). However, while the morphological features may be helpful, the number of annotations related to them is often limited, and their division into categories instead of a continuous variable may be inadequate from a computational perspective. Conversely, features from the lexicon area, such as whether a language uses a single word for both "finger" and "hand" or different words for each, may not be as significant in enhancing machine translation performance.

Although linguistic typology-based representations offer extensive knowledge about thousands of human languages, integrating this information into NMT research lacks a clear and general approach (Ponti et al., 2019). Not all the features are grounded in a specific behaviour, and in practice, they would merely group languages together when they share a particular property. For this reason, we first investigate typological features as a group rather than individually. To this end, we focus on developing a vector space representation of languages that incorporates typological information. Previous computational approaches to WALS and other linguistic databases have extracted similar language vectors based on multiple features, with a particular emphasis on syntax, by reducing redundancy and dropping variables with highly sparse annotations (Littell et al., 2017). However, data-driven methods such as the embedding matrix of a multilingual model can also produce similar representations (Malaviya et al., 2017; Östling and Tiedemann, 2017), which may contain different and significant information (Bjerva et al., 2019b). Therefore, we pose the question: **is it possible to build a combined language representation that incorporates complementary sources of information from the typological databases and NMT-learned representations?** A vector space of languages that is aware of typological information has the potential to measure language similarity. As we previously noted, language similarity is an important factor for multilingual machine translation, as we can identify pairs of languages that are more related by different criteria beyond family relationships, and may benefit from being trained together. We then further ask: **can we leverage the combined language space to improve multilingual machine translation applications?**

Ponti et al. (2019) surveyed that previous studies that leverage typological knowledge to improve machine translation have primarily focused on syntax, especially word order features (Daiber et al., 2016; Ponti et al., 2018). Likewise, the language representation space developed by Littell et al. (2017) that we used in our previous analysis is syntax-focused. However, it is important to note that linguistic diversity extends beyond syntax to other domains, particularly morphology. For example, gender and verb conjugation patterns are morphological phenomena that exhibit significant variation

across languages. In Spanish, for instance, verbs can have up to 50 forms depending on the tense, aspect, and mood, while in English, verbs have approximately 12 forms. In contrast, verbs are not conjugated in Mandarin Chinese. Additionally, gender marking is another aspect of morphology that differs across languages. Spanish distinguishes between two grammatical genders, masculine and feminine, while Zapotec and other languages that are spoken in Mexico distinguish two more: animate, and inanimate. Similarly, Swahili has more than 20 noun classes, each with its own set of gender markers.

Given the rich morphological diversity in languages, it must be significant to study the impact of different morphological phenomena on machine translation performance, but also on subword segmentation, which is an essential prior step for the task (Amrhein and Sennrich, 2021). Although several approaches have been proposed, such as training character or segmentation-free machine translation models (Gao et al., 2020; Libovický et al., 2022) or proposing morphologically-aware subword segmentation methods (Creutz and Lagus, 2002; Ataman et al., 2017), the relationship between morphological complexity and translation performance has shown mixed results. Furthermore, none of these studies has considered typological knowledge as a source for measuring morphological complexity and diversity in different languages. More precisely, Payne (2017) characterises the morphological typology of a language through two phenomena: synthesis, which refers to the number of morphemes[2] that make up a single word, and fusion, which refers to the number of inflections and meanings that can be combined within specific morphemes. In light of these considerations, we raise the question of **whether morphological typology variables are relevant in the context of machine translation**.

After analysing the morphological variable of synthesis, we observed that a higher degree of these phenomena, known as polysynthesis, hinders the performance of machine translation systems. For this reason, we adopt a pragmatic approach and investigate the performance of machine translation systems in an extreme scenario, where highly synthetic and low-resource languages are used. Our final research question thus becomes: **how can we improve machine translation performance for such challenging languages?** To answer this question, we focus our efforts on understudied and endangered languages of the Americas, which often exhibit the aforementioned traits of high synthesis and scarce resources (Mager et al., 2018b), thereby contributing to the

[2] A morpheme is the smallest meaningful linguistic unit within a word, often carrying semantic or grammatical significance. For more details, refer to §2.3 in Chapter 2.

diversity of machine translation research. We collect and develop new datasets, analyse the impact of unsupervised and morphologically supervised segmentation methods, and propose a new segmentation approach that is less dependent on data. Through these efforts, we aim to broaden the scope of attention given to such languages and advance the field of machine translation from different angles.

1.2 Thesis Structure

This thesis demonstrates that we can leverage linguistic typology variables and knowledge to analyse and improve the quality of neural machine translation models.

In summary, we first show that we can fuse language representations extracted from a typological knowledge base, syntactic features specifically, and embeddings from multilingual machine translation models, and evaluate the encoded information from each source. Then, we show that our language representations can leverage the positive transfer of multilingual machine translation models by providing a better and more efficient measurement of language similarity. Going further, we focus on morphological typology and analyse the variables of synthesis and fusion, to study how they impact machine translation quality at the word and phrase levels for different language pairs. Finally, we test an extreme case of machine translation with languages of high synthesis, or polysynthesis, and extremely low-resource. To do so, we focus on endangered languages of the Americas by building training and evaluation resources, and we finally propose a less data-driven segmentation approach.

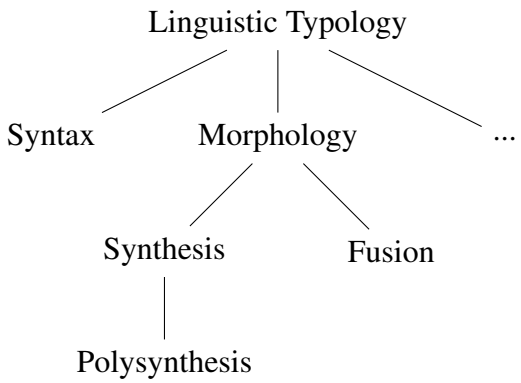


Figure 1.1: Our approach towards the impact of linguistic typology on machine translation. At the first level, we use syntactic variables crafted from typological databases such as WALS (Dryer and Haspelmath, 2013). Afterwards, we focus on Payne (2017)'s variables of morphological typology. Finally, we go deeper into polysynthesis.

Figure 1.1 shows the overview of the linguistic typology's domains that we focused on. Furthermore, a summary of the chapters that constitute this thesis is as follows:

Chapter 2 provides a theoretical framework that introduces the neural machine translation task, multilingual and transfer learning, subword segmentation, automatic evaluation metrics for machine translation, linguistic typology and morphological typology. The chapter offers a foundation for subsequent analyses in the thesis.

Chapter 3 proposes a method for measuring language differences through the use of typological features, specifically syntax, combined with pre-trained language embeddings. The chapter evaluates the effectiveness of this method in two computational typology tasks and a correlation analysis with genetic similarity, showing that the shared space of languages leads to better results.

Chapter 4 examines the usefulness of the proposed language representations in multilingual NMT, specifically in clustering related languages and ranking candidate languages for creating multilingual models. The findings reveal that the language representations perform as well as robust baselines for both low-resource and high-resource languages, and across different language families.

Chapter 5 focuses on morphology and examines whether the morphological variables of fusion and synthesis affect machine translation performance. The chapter proposes methods to quantify these variables in different languages and analyses their impact on translation quality using English-Turkish and English-Spanish as case studies. The study finds that Turkish nouns and verbs with higher synthesis and Spanish verbs with higher fusion are more challenging to translate, as supported by a human evaluation. The analysis is extended to the phrase level, with consistent results.

Chapter 6 takes a pragmatic approach and examines translation performance in extreme scenarios of high synthesis and low-resource languages, using understudied and endangered languages of the Americas as case studies. The chapter compiles and develops NMT resources and finds that morphologically-aware segmentation methods are not significantly better than standard unsupervised segmentation. Additionally, a less-data-dependent segmentation approach using syllables is proposed with effective results for our case study.

Chapter 7 concludes the thesis and presents ideas for future work.

1.3 Contributions

We make the following contributions:

- A method to compute language representations, which have higher quality than using only linguistic typology vectors or pre-trained language embeddings in computational tasks such as typological feature prediction and phylogenetic inference for languages.
- A tool, based on the aforementioned language representations, to retrieve related languages to build clusters for multilingual machine translation models that benefit low-resource language pairs and isolated languages.
- Methods that perform the first computational quantification of synthesis and fusion using standard NLP evaluation sets for Turkish and Spanish, plus experiments that analyse the relationship between the two indices and machine translation quality at word and phrase-level.
- Annotations of synthesis and fusion in machine translation evaluation sets for English–Turkish and English–Spanish language pairs.
- Human evaluation data for synthesis and fusion in machine translation experiments for English–Turkish and English–Spanish.
- A compilation and development of machine translation resources for highly synthetic and extremely low-resource languages of the Americas, including multilingual machine translation baselines.
- A new morphological segmentation dataset for Shipibo-Konibo, an endangered language spoken in the Americas
- A less data-dependent segmentation method based on syllables to support the translation into a highly synthetic and extremely low-resource language.
- Human evaluation of machine translation models for Spanish–Shipibo-Konibo.

1.3.1 Publications

The main articles that are described in this thesis are as follows:

- Oncevay et al. (2020) presents the research on language representations and their application on multilingual machine translation, which are included in Chapters 3 and 4.
- Oncevay et al. (2022a) quantifies the morphological variables of synthesis and fusion and analyses their impact on machine translation performance. This is included in Chapter 5.
- Oncevay (2021) describes the multilingual machine translation experiments for four highly synthetic and extremely low-resource languages of the Americas. This is the first part of Chapter 6.
- Oncevay et al. (2022b) introduces a syllable-based segmentation method that is less-data dependent and effective for language with high synthesis and low resource traits. This is the last part of Chapter 6.

Finally, we contributed to further research that is related to this thesis:

- In Mager et al. (2021), we contributed to a larger effort to build machine translation resources for ten indigenous languages of the Americas.
- In Mager et al. (2022), as second author, we compared several unsupervised and morphologically-supervised segmentation methods for highly synthetic and low-resource languages. My specific contribution is the development of a new morphological segmentation dataset for Shipibo-Konibo, and the execution plus statistical significance analysis of the machine translation experiments. This is partially added in Chapter 6.
- In Zariquiey et al. (2022), we argued about the importance of bridging the fields of language documentation and NLP, in order to support revitalisation efforts of endangered languages. This article further represents our motivation to contribute, with our machine translation research, to understudied languages, especially indigenous languages of the Americas.

Chapter 2

Background

This chapter briefly describes essential concepts and methods for the development of the thesis. First, we introduce the machine translation task, including neural network architectures, subword segmentation methods, transfer and multilingual learning, and automatic metrics used for evaluation. Then, we introduce linguistic-related concepts, such as the areas of linguistic typology and morphology, including morphemes and morphological segmentation. Finally, we delve into morphological typology and the degrees of synthesis and fusion.

2.1 Machine translation

In the field of NLP, Machine Translation (MT) refers to the task of automatically translating text from one language to another. The MT process consists of various stages, including text preprocessing and translation model training. Two main types of MT systems exist: rule-based and data-driven. Rule-based MT relies on manually designed grammars and dictionaries to translate text, but these approaches have limitations when handling the complexity and ambiguity of natural language. Data-driven MT, on the other hand, trains models on large parallel corpora to learn translation patterns. Recently, neural machine translation models have gained popularity due to their success in achieving state-of-the-art performance on several language pairs.

Despite significant progress in MT, the task still poses several challenges, such as dealing with low-resource languages and maintaining translation quality for rare or out-of-vocabulary words (Koehn and Knowles, 2017; Sennrich et al., 2016b). Consequently, researchers continue to explore various techniques to improve MT systems, including multilingual and transfer learning.

2.1.1 Neural machine translation

Neural Machine Translation (NMT) is a type of sequence-to-sequence (seq2seq) model (Sutskever et al., 2014), which is a deep learning-based approach for machine translation that has achieved state-of-the-art results in recent years. NMT models typically consist of an encoder and a decoder, both of which are neural networks. In a general perspective, the encoder processes the input sentence and produces a fixed-length vector, which is then used by the decoder to generate the output sentence word-by-word.

2.1.1.1 Recurrent Neural Networks

One of the early and successful architectures for NMT is based on Recurrent Neural Networks or RNNs. RNNs are a type of neural network that can process variable-length input sequences by maintaining an internal state, or "memory", that is updated at each step. The internal state captures information about the entire input sequence up to the current step, and is used to generate the output sequence (Jurafsky and Martin, 2019).

Formally, an RNN can be defined as follows:

$$h_t = \sigma(W_{ih}x_t + W_{hh}h_{t-1} + b_h) \tag{2.1}$$

where x_t is the input at time step t, h_{t-1} is the internal state at the previous time step, W_{ih} and W_{hh} are weight matrices, b_h is a bias vector, and σ is a non-linear activation function, such as the hyperbolic tangent function.

The output at each time step is then computed as:

$$y_t = \text{softmax}(W_{oh}h_t + b_o) \tag{2.2}$$

where W_{oh} and b_o are the weight matrix and bias vector for the output layer, respectively, and softmax is a normalisation function that ensures the output probabilities sum to one.

2.1.1.2 Recurrent Neural Networks with Gated Mechanisms

Although RNNs have been successful in many sequence-to-sequence tasks, including NMT, they can struggle to handle long sequences and may suffer from the vanishing gradient problem, meaning that the gradients become too small to be effective during

training. To address these issues, researchers have introduced gated network architectures, such as Long Short-Term Memory (LSTM; Hochreiter and Schmidhuber, 1997) and Gated Recurrent Unit (GRU; Cho et al., 2014).

LSTM and GRU incorporate specialised gating mechanisms that enable them to control the flow of information within the network's internal state. These mechanisms selectively retain and update information over longer sequences, mitigating the vanishing gradient issue and enhancing the model's ability to capture context.

2.1.1.3 Recurrent Neural Networks with Attention

In addition to gated networks, which aid in handling long-range dependencies, Bahdanau et al. (2015) introduced the concept of attention, further improving the alignment and translation of source and target sequences.

The attention mechanism allows the model to selectively focus on different parts of the input sequence, depending on the context of the output sequence being generated. This can improve the model's ability to capture relevant information from the input, especially for long input sequences.

Formally, the attention mechanism can be defined as follows:

$$e_{ti} = \text{score}(h_{t-1}, \tilde{h}_i) = v_a^\top \tanh(W_a h_{t-1} + U_a \tilde{h}_i) \tag{2.3}$$

where h_{t-1} is the hidden state of the decoder RNN at the previous time step, $\tilde{h}_i$ is the i-th hidden state of the encoder RNN, W_a and U_a are learnable weight matrices, and v_a is a learnable weight vector. The score function computes a scalar score for each encoder hidden state, indicating how relevant it is to the decoder's current state.

The attention scores are then used to compute a weighted sum of the encoder hidden states:

$$c_t = \sum_{i=1}^{T_x} \alpha_{ti} \tilde{h}_i \tag{2.4}$$

where α_{ti} is the attention weight assigned to the i-th hidden state, defined as:

$$\alpha_{ti} = \frac{\exp(e_{ti})}{\sum_{j=1}^{T_x} \exp(e_{tj})} \tag{2.5}$$

The context vector c_t represents a summary of the relevant parts of the input sequence at the current time step, which can be used to generate the output sequence.

2.1.1.4 Transformers

A more recent and robust architecture for NMT is the Transformer (Vaswani et al., 2017). The Transformer eliminates the need for recurrence by using a self-attention mechanism that directly connects all the positions in the input and output sequences, allowing the model to attend to different parts of the input and output at each position.

The Transformer consists of an encoder and a decoder, each composed of multiple layers of self-attention and position-wise feedforward networks. During training, the Transformer uses teacher forcing, where the decoder inputs the ground truth tokens for each time step, rather than its own predicted tokens.

The self-attention mechanism in the Transformer can be defined as follows:

$$\begin{aligned} \text{MultiHead}(Q,K,V) &= \text{Concat}(\text{head}_1,\ldots,\text{head}_h)W^O \\ \text{head}_i &= \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \\ \text{Attention}(Q,K,V) &= \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \end{aligned} \tag{2.6}$$

where Q, K, and V are query, key, and value matrices, respectively. MultiHead computes multiple sets of attention heads in parallel, which are concatenated and projected back to the model's original dimensionality with the matrix W^O. The Attention function computes the dot product between the query and key matrices, and scales the result by $\sqrt{d_k}$, where d_k is the dimensionality of the key vectors. The softmax function is then applied to obtain the attention weights, which are used to compute a weighted sum of the value matrix.

During decoding, the Transformer uses an autoregressive strategy, where the output sequence is generated one symbol at a time, conditioned on the previously generated symbols. Specifically, the output at time step t is computed as:

$$\begin{aligned} y_t &= \text{softmax}(W_{out}h_t + b_{out}) \\ h_t &= \text{TransformerDecoder}(y_{<t}, X) \end{aligned} \tag{2.7}$$

where $y_{<t}$ represents the previously generated symbols, W_{out} and b_{out} are the weight matrix and bias vector for the output layer, respectively, and TransformerDecoder is a modified Transformer that takes as input the entire source sequence X and the previously generated symbols, and outputs the internal state h_t at time step t. The decoding process continues until a special end-of-sequence symbol is generated.

2.1.2 Subword units in NMT

Another challenge that NMT faces is the handling of rare and out-of-vocabulary (OOV) words (Sennrich et al., 2016b). These words are infrequent in the training data and may not be properly learned by the model, leading to translation difficulties. To address this issue, researchers have introduced subword units, which break down words into smaller units that are more frequent and allow the model to handle rare words more effectively. These units, stripped of inherent linguistic meaning, offer a workable solution. For example, "unhappiness" could be segmented into more frequent "un", "happ" and "iness" subwords.

Subword segmentation is applied to the parallel data before training, as the subwords serve as input to the NMT model. During training, the model learns to translate between languages by processing the subword representations and generating target translations. For NMT, two of the most prominent subword segmentation methods are Byte Pair Encoding (BPE) and Unigram Language Model (UnigramLM):

BPE (Sennrich et al., 2016b) is a data compression algorithm that works by iteratively merging the most frequent pairs of bytes in a corpus until a predefined number of merge operations has been reached. This process creates a dictionary of subword units, which can then be used to encode and decode text.

UnigramLM or uniLM (Kudo, 2018) is a language model that uses a unigram distribution to segment words into subword units. The algorithm first counts the frequency of each character in the corpus, then greedily selects the most frequent character n-grams until a specified number of subword units has been reached. This process creates a vocabulary of subword units, which can then be used to segment text.

2.1.3 Transfer learning in NMT

While NMT has shown promising results, it still suffers from a significant limitation: the need for large amounts of parallel data to train high-quality models. This is particularly challenging for low-resource languages, for which there is limited parallel data available. To address this challenge, transfer learning has been explored for NMT (Zoph et al., 2016). Transfer learning involves the pretraining of a model on a large amount of source data, followed by fine-tuning on the target task with a smaller amount of data.

Concerning the factors for choosing the parent language pair, Kocmi and Bojar (2018) argued that the corpus size is more relevant than the similarity of the parent and child languages. This is consistent with Lin et al. (2019), as they investigated several metrics (corpus-based, typological, geographical) and identified that the corpus-based heuristics are more significant for ranking a candidate parent language pair for transfer learning.

2.1.4 Multilingual machine translation

Another approach to address the challenge of limited parallel data is multilingual NMT, which involves training a single model on multiple languages. Multilingual NMT has been shown to be effective in improving translation quality for low-resource languages (Johnson et al., 2017).

Multilingual NMT can be formulated as a joint optimisation problem that aims to minimise the negative log-likelihood of all available parallel data across multiple languages. Formally, the objective function can be defined as:

$$\sum_{i=1}^{n} \sum_{j=1}^{m_i} -\log p(y_{i,j}|x_{i,j}, l_i) \tag{2.8}$$

where n is the number of languages, m_i is the number of parallel sentences for language l_i, $x_{i,j}$ and $y_{i,j}$ are the source and target sentences for the j-th parallel sentence in language l_i, respectively. The objective function can be optimised using gradient-based optimisation techniques such as stochastic gradient descent.

Furthermore, Johnson et al. (2017) proposed a simple approach to enable multilingual NMT, without modifying the neural network architecture, by introducing an artificial (language identity or language-specific) token at the beginning of the input sentence. For instance, consider a scenario where the model translates from one source language to multiple target languages. In this setting, a special token like $< 2es >$ (where *es* stands for Spanish) could be used to indicate the target language. Conversely, in a many-to-one setting where a model translates from various source languages to a single target language, a language-specific token such as $< es >$ can be introduced. We will focus on the language identity token when developing our language representations in Chapter 3.

It's important to note that while these language-specific tokens can effectively handle multilingual translation, they are just one approach among several to achieve better performance. For instance, a shared vocabulary is another key aspect that enhances the

efficiency of multilingual models by sharing the model's parameters in a shared vector space representation of words. In a traditional NMT model, each word (or subword) in the source and target languages has its own separate vector space representation. In a shared vocabulary setup, words from all languages are mapped into a common representation space. This means that words with similar meanings or translations in different languages could be closer together in the shared space, allowing the model to capture cross-lingual relationships. We will focus on more strategies for improving multilingual NMT models in Chapter 4.

2.1.5 Evaluation in Machine Translation

While human evaluation remains the gold standard for assessing translation performance, it is often time-consuming and expensive. In this context, automatic metrics play a pivotal role in MT evaluation. Typically, automatic metrics analyse the degree of overlap between reference translations and system-generated translations, enabling a quantitative assessment of translation quality. These evaluation metrics provide a quicker and more scalable alternative for researchers to make informed decisions about model adjustments and other strategies.

In this thesis, we present results using three automatic evaluation metrics for MT:

Bilingual Evaluation Understudy (BLEU) (Papineni et al., 2002): BLEU measures the overlap between a machine-generated sentence and one or more reference sentences using n-gram precision. BLEU considers both precision (how many words in the generated sentence match a reference) and brevity (how concise the generated sentence is compared to references). One advantage of BLEU is its simplicity and efficiency, making it easy to compute. However, BLEU's reliance on n-gram matches might lead to high scores for sentences with surface-level similarities but lacking adequacy (semantic correctness) or fluency. Besides, BLEU doesn't account for word order and might not be well-suited for languages with flexible word order or complex syntactic structures.

Character n-gram F-score (chrF) (Popović, 2015): chrF measures the similarity between a reference sentence and a machine output based on character-level n-grams. In contrast to BLEU, chrF provides a more holistic view of translation quality by considering not only word-level but also character-level correspondences. However, a metric like chrF may not fully capture semantic and syntactic nuances in translations, and its

use of character-level n-grams might penalise sentences that are structurally similar but differ in character representation.

Crosslingual Optimized Metric for Evaluation of Translation (COMET) (Rei et al., 2020): COMET aims to address the limitations of semantic correctness in traditional metrics. COMET uses pretrained, Transformer-based multilingual language models, such as multilingual BERT (Pires et al., 2019) or XLM-ROBERTa (Conneau et al., 2020). These models are typically trained to reconstruct masked tokens by uncovering the relationship between those tokens and their neighbours. For this reason, COMET can capture semantic similarity and assess whether a translation maintains the meaning of the original. Furthermore, COMET strongly correlates with human judgement for several language pairs. Nevertheless, its reliance on pretrained models may introduce biases and limits their application for evaluating the translations of low-resource and under-represented languages.

While we use BLEU, the most widespread metric in the MT field, to assess the results of our experiments in Chapter 4, we consider that reporting different metrics can provide a more comprehensive assessment of machine-generated text. For this reason, we present outcomes measured by the chrF metric in Chapters 5 and 6, as well as the application of COMET in Chapter 5, wherein we explore the quality of machine-generated text in terms of linguistic variables (as discussed in §2.4) and low-resource and morphologically-complex languages.

2.2 Linguistic typology

Linguistic typology is the study of patterns and variability in the structure and use of human languages (Comrie, 1989). It aims to identify cross-linguistic generalisations and variations in different language features, such as word order, grammatical categories, phonology, and morphosyntax. Linguistic typology seeks to provide insights into the ways that languages differ from one another and how they reflect and shape human cognition and communication.

2.2.1 Typology databases

Linguistic typology databases are large-scale collections of cross-linguistic data that enable researchers to compare and analyse linguistic features across languages. These

databases typically contain information on various aspects of language structure, including phonetics, morphology, syntax, and semantics. The data in these databases are often coded based on standardised criteria and can be searched, filtered, and analysed using computational tools (O'Horan et al., 2016). The most relevant typological database used in NLP research is the Word Atlas of Languages Structures, which is described as follows.

2.2.1.1 Word Atlas of Languages Structures

The World Atlas of Language Structures (WALS) (Dryer and Haspelmath, 2013) is one of the most widely used linguistic typology databases. WALS contains information on over 2,600 languages, covering 143 linguistic features. The database includes interactive maps, tables, and visualisations that enable researchers to explore the distribution and diversity of linguistic features across languages and regions. WALS has been used in a wide range of linguistic research, including studies on language evolution, language contact, language acquisition, and in multilingual NLP applications (Ponti et al., 2019).

More specifically, WALS stores typological variables in the following domains: Word Order, Morphology, Lexicon, Phonology, Nominal Categories, Nominal Syntax, Verbal Categories, Simple Clauses, Complex Sentences, Sign Languages and Others. Table 1.1 shows an example of features stored in the database.

Before delving into morphological typology, we first describe the field of morphology and the morphological segmentation task in the following section.

2.3 Morphology and morphemes

In linguistics, the study of morphology and morphemes holds significant importance as it delves into the foundational elements of language structure. Morphology examines the structure and formation of words, while morphemes are the smallest meaningful units that constitute these words, playing a vital role in conveying semantic and grammatical information (Payne, 2017).

Morphemes can exist independently as free morphemes or be attached to others, resulting in bound or joint morphemes. Morphological typology investigates various processes that bind morphemes and analyses the diverse mechanisms languages employ to create and modify words (see more details in §2.4).

2.3.1 Morphological segmentation

Morphological segmentation refers to the process of dividing a word into its individual morphemes (Harris, 1951). This segmentation involves identifying and isolating prefixes, suffixes, roots, and other morphological elements within a word.

In recent NLP literature, machine learning methods are extensively applied to the task of morphological segmentation using both unsupervised and supervised approaches. In the case of unsupervised methods, algorithms do not require any labeled data to learn from. One widely used method in this category is called Morfessor.

Morfessor is a family of unsupervised segmentation algorithms, developed by Creutz and Lagus (2002). The key idea behind Morfessor is to segment words into meaningful character sequences based on the statistical properties of the language. Morfessor algorithms operate by iteratively splitting and merging the most frequent character units until a desired number is reached. This makes Morfessor suitable for morphologically rich languages where words can have a large number of inflections and derivations. Other popular methods that are based on Morfessor are **FlatCat** (Grönroos et al., 2014) and **LMVR** (Ataman et al., 2017).

One of the disadvantages of unsupervised learning methods for morpheme segmentation is that they can sometimes result in ambiguity and variability in segmentation outcomes. These methods lack access to external linguistic resources or specific language rules during segmentation, which may lead to struggles in accurately identifying certain morpheme boundaries, particularly in cases where morphemes are less transparent, context is crucial for accurate segmentation, or due to irregularities present in some languages.

In contrast, supervised morphological segmentation methods are trained on labeled data (e.g., unsegmented words as inputs and their segmented morphemes as outputs). A popular method applies Pointer Generator Networks, described as follows:

Pointer Generator Networks are a type of neural network architecture that can be used for morphological segmentation (See et al., 2017; Mager et al., 2020). These networks combine a seq2seq model with a pointer network, allowing the model to generate output sequences by copying or generating tokens from a source sequence. In the context of morphological segmentation, the source sequence is typically a sentence or text in an unsegmented form, and the output sequence is a segmented form, where

the morphemes have been separated. The seq2seq component of the model learns to encode the input sequence and generate the output sequence, while the pointer network helps the model to select appropriate morphemes from the input sequence to generate the output.

Despite outperforming unsupervised methods in morphological segmentation (see §6.5 in Chapter 6), supervised approaches rely on the availability of annotated segmented data, which is scarce for most languages and requires specialised knowledge to create.

2.4 Morphological typology

Morphological typology is a subfield of linguistics that aims to classify and compare languages based on their morphological structures and word-formation processes (Payne, 2017), such as agglutination or fusion.

In the field of NLP, researchers often label languages as agglutinative or fusional based on the explicit boundaries between morphemes. For example, Turkish is considered a highly agglutinative language (e.g. in Ataman et al. (2017)) because the morpheme boundaries tend to be clear, and each morpheme typically expresses its own lexical or grammatical meaning. In contrast, Spanish is often labelled as fusional (e.g. in Mager et al. (2018b)) because many meanings or grammatical information could be fused in a single morpheme. However, holistic types for languages fail to capture the full complexity of morphological typology. Early typological studies attempted to quantify these variables (e.g. fusion) and avoided characterising languages with a single type (Sapir, 1921; Greenberg, 1960; Comrie, 1989).

Based on this earlier research, recent work by Payne (2017) highlights the importance of the indices of synthesis and fusion. These indices provide a more pragmatic and comprehensive approach to characterising morphological typology. Before delving into their definition, we first describe more details about morphemes.

2.4.1 Synthesis

The index of synthesis is calculated as the average number of morphemes per word (Payne, 2017), as we can see in Figure 2.1. Highly isolating languages, such as Mandarin or, to a lesser extent, English, have a synthesis value that is closer to 1. In con-

trast, synthetic languages, like Turkish and Inuktitut, have higher synthesis values[1].

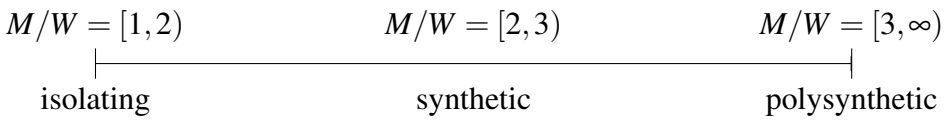


Figure 2.1: Index of Synthesis. M: morphemes, W: words. Adapted from Payne (2017).

To illustrate the case of synthesis, we present two examples from Mandarin Chinese, a Sino-Tibetan isolating language, and Asháninka, an Arawak synthetic language spoken in South America:

(1) 他再有他想要去鹿。
tā méi zài yǒu shuō tā xiǎng yào qù liè lù
3SG NEG again EXIST say 3SG want comp go hunt deer

'He did not say again that he wanted to go hunt deer.'
(Payne, 2017)

(2) *nonkotsitasanomempebentajeibetamanakero*
no-n-kotsi-t-asano-mempe-bent-a-jei-be-t-aman-ak-e-ro
1-IRR-cook-EP-INTENS-pretend-EP-PL-frust-EP-early-PRF-IRR-3F

'We will really pretend to cook for her early in the morning without success'
(Jaime Montoya Samamé, fieldnotes)

In the Mandarin Chinese example (1), provided by Payne (2017), we observe that each word corresponds to a discrete meaningful unit, as is typical in a strictly isolating language. Furthermore, the translated English gloss also demonstrates low synthesis, with a close one-to-one correspondence between words and annotated lexical or grammatical units. In contrast, the Asháninka example (2) shows a verb with a highly synthetic structure, encompassing several morphemes. This morphological complexity is most pronounced within the verb in many highly synthetic languages.

When the degree of synthesis surpasses three, some scholars argue that the language exhibits polysynthesis (as is evident in the Asháninka example). However, the

[1] Sapir (1921) classified synthesis as an index distinguishing analytic from synthetic language types. Sapir's framework associated the term "isolating" with agglutinative, fusional, and symbolic types, encompassing various techniques for encoding grammatical information. While linguistic typology literature presents varying perspectives on these indexes, we adopt Payne's definition for our experiments in Chapter 5.

boundary remains debatable, as other factors such as discourse or topic considerations can impact the index calculation (Payne, 2017).

2.4.1.1 Polysynthetic languages.

Polysynthesis is a term used to describe highly synthetic languages, in which multiple morphemes can be combined to form a single word. However, linguists have defined polysynthesis in different ways, leading to some debate about the exact criteria that a language must meet to be considered polysynthetic. For example, some definitions require that the language allows for the incorporation of nouns or other words into the complex verb morphology, as well as having agreement morphemes and pronominal affixes (Baker, 1996; Mithun, 1986).

While there are various definitions of synthesis and polysynthesis in linguistics, for the purposes of our experiments in Chapter 5, we will focus on Payne's definition, which takes into account the number of morphemes per word, a definition that is particularly useful from a pragmatic standpoint.

2.4.2 Fusion

Fusion refers to the degree to which morphemes in a language carry multiple grammatical or semantic features, with the ratio of fusional morphemes to the total number of morphemes being used as an index of this phenomenon. Figure 2.2 illustrates the concept of the index of fusion, according to Payne (2017). This index ranges from 0 to 1, with languages that are highly agglutinative (e.g., Turkish) scoring low and those that are highly fusional (e.g., Spanish) scoring high.

Figure 2.2: Index of Fusion: Number of fusional junctures (F) per all junctures (J). Adapted from Payne (2017).

For instance, we can analyse two words in Turkish and Spanish as follows:

(3) *Yaramaz-laş-tırıl-a-mıy-abilen-ler-den-miş-siniz*
useless-act-like-E-NEG-able-PL-of-doubt-2.be

'It seems that you are one of those who are incapable of being useless'
(Payne, 2017)

(4) *Conden-ó*
condemn-3.SG.PAST.IND.PFV

'Condemned (verb)'

We observe that, in the example in Turkish (3), each grammatical category or lexical meaning corresponds to a single morpheme. This is a clear example of a highly agglutinative word, where the boundaries of the morphemes are transparent. The only exception is the last morpheme '-siniz', which fuses the meaning of the 2nd person and the verb "to be." In contrast, the example in Spanish (4) demonstrates that the last morpheme '-ó' fuses five grammatical categories, including the 3rd person, singular number, past tense, indicative mood, and perfective aspect.

To compute the index of fusion, one needs to determine the number of fusional junctures (F) among all junctures (J) in a given language. A fusional juncture is a morpheme that fuses multiple grammatical or semantic features into a single unit, such as a morpheme that carries both tense and aspect. By contrast, a non-fusional juncture is a morpheme that carries a single feature or is composed of multiple morphemes that each carry only one feature, such as a separate morpheme for tense and a separate morpheme for aspect.

However, computing the fusion index can be challenging to automate. For example, Payne (2017) has suggested various types of morphemes that can potentially be considered as fusional joints, including prefixes, suffixes, infixes, circumfixes, compounding, and non-concatenative processes such as reduplication, apophony, and subtractive morphology. Additionally, there are autosegmental morphemes that may also count as fusional. Nevertheless, current automatic tools are not yet capable of identifying these cases for most languages.

2.4.3 Interaction of Synthesis and Fusion

Payne (2017) discusses how the index of synthesis interacts with the index of fusion. Highly synthetic languages have a greater number of morpheme junctures, which

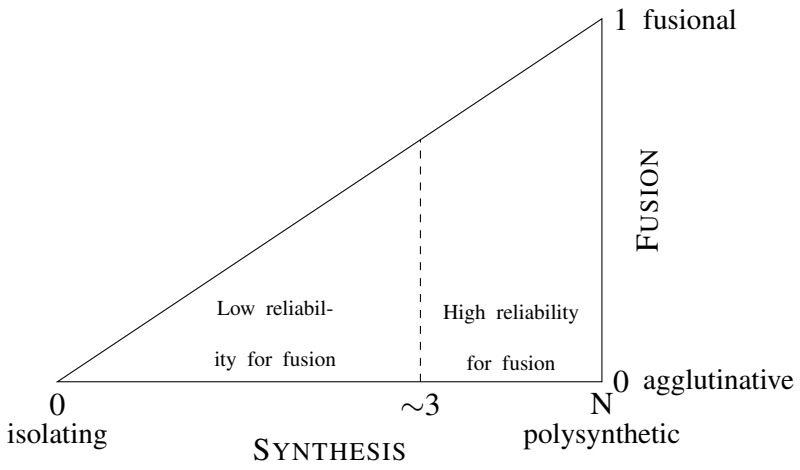


Figure 2.3: Interaction between Synthesis and Fusion: the triangle of reliability. Adapted from Payne (2017).

makes it more reliable to calculate the index of fusion. However, in highly isolating languages with no morpheme junctures, the index of fusion would always be irrelevant. This relationship between synthesis and fusion can be illustrated as a "triangle of reliability" in a two-dimensional graph, as shown in Figure 2.3.

Finally, it is important to note that the indices of synthesis and fusion are orthogonal and can be used together to characterise a language. For example, a language like Turkish is highly synthetic, with many morphemes per word, but it is also agglutinative and has a low fusion index. On the other hand, a language like Spanish is highly fusional but has a lower degree of synthesis. In our experiments (see Chapter 5), we will focus on Turkish for synthesis and Spanish for fusion, as they represent approximately extreme points along these axes and are well-studied examples in the literature (Ataman et al., 2017; Mager et al., 2018b).

Chapter 3

Multi-view Language Representations

To investigate the impact of linguistic typology variables and knowledge on machine translation, we first construct a typologically-aware vector space representation of languages by fusing language embeddings extracted from multilingual NMT models with linguistic variables of syntax from typology databases. Our ultimate objective is to determine whether these typologically-informed language representations effectively leverage multilingual NMT applications (as discussed in the following chapter, Chapter 4). However, before delving into their impact on machine translation, we first examine whether the proposed method for combining typology data and NMT embeddings effectively enhances the vector space representation of languages itself, and whether it successfully preserves both the typology knowledge and data-driven learned information in a single representation. The content of this chapter is based on the work of Oncevay et al. (2020).

3.1 Introduction

A vector space representation of languages enables the measurement of language similarity, which is a crucial factor in analysing the benefits of transfer learning and leveraging NMT models, particularly multilingual ones (Kudugunta et al., 2019; Malaviya et al., 2017; Tan et al., 2019). For example, Kudugunta et al. (2019) demonstrated how the representations learned in the encoder of a multilingual machine translation model are related to a language similarity factor. However, their approach only utilised language family relationships as the basis for measuring similarity. Likewise, Malaviya et al. (2017) learned a dense vector space of languages by training a multilingual NMT model with a language identification token at the beginning of each sentence. They

then predicted linguistic typology features using the language vectors in a classification task, showing that these vectors encode knowledge of language similarity. Additionally, Lin et al. (2019) developed ranking models to retrieve candidate languages for transfer learning applications and obtained positive results. They utilised a diverse set of similarity metrics beyond language family relationships, including corpus-based statistics and language features extracted from typological databases, such as WALS (Dryer and Haspelmath, 2013).

In this context, it is essential to understand that there are different ways to build a vector space representation of languages, and each approach can offer potential benefits to multilingual machine translation research. This leads us to ask a key question: **is it possible to build a combined language representation that incorporates complementary sources of information?** To investigate this, we focus on two types of language representations: data-driven representations learned by multilingual neural models (Malaviya et al., 2017; Östling and Tiedemann, 2017; Bjerva et al., 2019b) and linguistic variables extracted from typological databases (Littell et al., 2017; Dryer and Haspelmath, 2013).

To achieve our goal of studying the application of language representations in multilingual machine translation models (see next chapter, Chapter 4), we must focus on data-driven language representations. These are dense vectors in high-dimensional spaces that can be learned during the training processes of multilingual language modelling (Östling and Tiedemann, 2017; Bjerva et al., 2019b) or machine translation systems (Malaviya et al., 2017). The primary aim of these studies is to uncover and analyse what kind of information can be encoded from large collections of monolingual and parallel corpora about languages. Such information is valuable for transfer learning, as Lin et al. (2019) found that corpus-based metrics are one of the most significant factors in their study. However, the language diversity in the corpus-based representations is limited.

On the other hand, linguistic typology databases store information about languages that are under-represented in machine translation or NLP repositories. For example, WALS (Dryer and Haspelmath, 2013) encodes 143 language-level features for 2,679 languages. This surpasses any multilingual machine translation dataset, which typically contains only hundreds of languages, such as the OPUS-100 with 100 languages (Zhang et al., 2020). As surveyed by Ponti et al. (2019), linguistic typology variables, such as the word order features shown in Table 1.1, have the potential to enhance machine translation applications (Daiber et al., 2016; Ponti et al., 2018). However, they

are categorical or discrete features with sparse values, and the mean coverage of variables per language is barely around 14%. Therefore, combining these sparse variables with a dense vector space of languages could enrich them and be beneficial for further applications.

For these reasons, we propose to compute language representations using the best of both views, typological knowledge bases and data-learned vectors, with minimal information loss, and we name them **multi-view language representations**. Therefore, we compute a shared space of discrete and continuous features, where each source embeds specialised knowledge about language similarity. The linguistic vectors can measure typological similarity, whereas data-learned embeddings correlate with other kinds of language relationships (e.g., genetic or cognate sharing, according to Bjerva et al. (2019b)). To analyse whether each kind of information is induced in our new representations, we inspect how much typological knowledge is present by predicting features for new languages (typological feature prediction in §3.5), and we infer language phylogenies to inspect whether specific relationships are induced from the data-learned vectors (§3.6).

3.2 Related work

To build vector space representations of languages, several works have focused on extracting language features from either typological knowledge bases or data-driven tasks. On the typological side, Littell et al. (2017) developed the URIEL knowledge base and the lang2vec tool, which provide a straightforward extraction of binary features from typological, geographical, and phylogenetic databases. Using this approach, they constructed a language representation of 103 syntax binary features by one-hot-encoding categorical variables from the World Atlas of Language Structures (WALS) (Dryer and Haspelmath, 2013), adding binary features from the Syntactic Structures of World Languages (SSWL) (Collins and Kayne, 2011), and processing short descriptions on syntactic typological features from Ethnologue (Simons and Fenning, 2019). Likewise, in a series of related work, Murawaki (2015, 2017, 2018) built latent language representations by leveraging typological linguistic variables from WALS.

On the other hand, language representations are also encoded from data-driven tasks such as multilingual NMT or language modelling. For example, Östling and Tiedemann (2017) learned dense language representations using a character-based neural language modelling, where a language identification token is concatenated to char

embeddings in each time step. A similar procedure is illustrated later in Figure 3.1. Related approaches are followed by Bjerva and Augenstein (2018b); Bjerva et al. (2019b), where they further explored what kind of language relationships are represented in the learned embeddings. Language modelling has also been complemented with other linguistic-related target tasks to extract more specialised embeddings, for instance, in phonetics (Tsvetkov et al., 2016) and phonology (Bjerva and Augenstein, 2018a). In the case of NMT, Malaviya et al. (2017) trained a many-to-English model using Bible translations with a language identity token at the beginning of each input sentence. They further identified that the language vectors are robust for predicting typological features. Likewise, Tan et al. (2019) obtained NMT-learned language embeddings to cluster the languages and train multilingual NMT models.

Our goal of combining typological variables and data-learned representations is closer to Bjerva et al. (2019a). They built a generative model from typological features and use language embeddings, extracted from factored language modelling at character-level, as a prior of the model to extend the language coverage. However, their main goal is to develop a generative model to infer typological variables and complete empty typological database entries from the information in the data-learned embeddings.

3.3 Multi-view language representations

We chose **canonical correlation analysis** (CCA) as our method for obtaining a shared representation of languages for several reasons. First and foremost, our goal is to fuse parallel representations of the same language into a single shared space, and CCA is specifically designed for finding a projection of two views for a given set of data with related parallel entries. Moreover, CCA is a well-established method with a long history of successful applications in various fields, including NLP, where it has been used to match word translations from bilingual lexicons (Haghighi et al., 2008) and to derive word embeddings (Faruqui and Dyer, 2014; Dhillon et al., 2015; Osborne et al., 2016). Additionally, CCA-based methods are popular choices for inspecting representations extracted from neural models (Raghu et al., 2017), including multilingual NMT models (Kudugunta et al., 2019).

Using CCA, we can find linear combinations that maximise the correlation of the two sources in each coordinate iteratively, as described by Hardoon et al. (2004). In our case, the two sources are linguistic typological features and data-learned embeddings.

Represented by two random vectors $X \in \mathbb{R}^d$ and $Y \in \mathbb{R}^{d'}$, the views are projected in a shared space with m dimensions, by maximising their correlation in each coordinate and retaining as little redundancy as possible. CCA solves a sequence of optimisation problems for $j \in \{1..m\}$ where $a_j \in \mathbb{R}^{1\times d}$ and $b_j \in \mathbb{R}^{1\times d'}$:

$$\underset{a_j,b_j}{\operatorname{argmax}} \operatorname{corr}\left(a_j X^\top, b_j Y^\top\right) \text{ such that } \begin{cases} \operatorname{corr}\left(a_j X^\top, a_k X^\top\right) = 0, & k < j \\ \operatorname{corr}\left(b_j Y^\top, b_k Y^\top\right) = 0, & k < j \end{cases}$$

where the corr function returns the Pearson correlation between two vectors (pairwise element). We only retain j-th dimension if $\operatorname{corr}(a_j, b_j) \geq 0.5$, as we look for moderate to strong uphill correlation scores between the transformed views.

CCA considers all dimensions of the two views as equally important. This approach is particularly useful because we want to identify the shared structure between two potentially redundant sets of data: knowledge base features that are mostly one-hot-encoded, and task-learned features that inherit the high dimensionality of the embedding layer. However, few samples and sparsity can make the convergence harder, so we also consider **singular value decomposition** (SVD) as an appealing alternative to address the redundancy issue. SVD factorises the data matrix, composed of language vectors in our case, to compute the principal components and singular values. Furthermore, to deal even more with sparsity in the one-hot-encoded vectors, we adopt a truncated SVD approximation, which is also known as latent semantic analysis in the context of linear dimensionality reduction for term-count matrices (Dumais, 2004).

The two-step transformation of SVD followed by CCA is called **singular vector canonical correlation analysis** (SVCCA; Raghu et al., 2017), and it is particularly useful for understanding the representation learning throughout neural network layers. SVCCA is a powerful tool for analysing the redundancy and distinctiveness of different features, and can help us identify to what extent two representations capture the same information about language properties.

3.3.1 Representation of unseen languages

Once the CCA-based model has been trained, it can be used to generate representations of new languages that were not seen during training. Given a new language represented as a typological feature vector $\mathbf{x}$ or an embedding $\mathbf{y}$, we can obtain its CCA-based representation $\mathbf{z}$ as follows:

$$\mathbf{x}' = \mathbf{U_x}\mathbf{x} \quad \text{and} \quad \mathbf{y}' = \mathbf{U_y}\mathbf{y}$$

where $\mathbf{U_x}$ and $\mathbf{U_y}$ are the left singular vectors obtained from the SVD of the training data matrices $\mathbf{X}$ and $\mathbf{Y}$, respectively. Then, the CCA-based representation $\mathbf{z}$ is given by:

$$\mathbf{z} = \mathbf{w_x}^T\mathbf{x}' = \mathbf{w_y}^T\mathbf{y}'$$

where $\mathbf{w_x}$ and $\mathbf{w_y}$ are the projection matrices obtained during the CCA training phase. This allows us to represent unseen languages in the same vector space as the trained languages, facilitating comparison and analysis. Therefore, we can take advantage of the large language coverage of typological language vectors in contrast to data-learned language embeddings from multilingual models.

3.3.2 SVD optimisation

While performing SVD on the input views to reduce their dimensionality, we can select an appropriate threshold that determines the variance we aim to preserve from each view after the dimensionality reduction. Specifically, we consider a range of threshold values in the range of 0.5 to 1.0, with incremental steps of 0.05, and select the threshold that achieves the best performance in the specific task.

It's worth noting that setting the threshold to 1.0 bypasses SVD altogether, resulting in CCA being computed on the full feature vectors or embeddings. However, this may not always be desirable, as it can lead to overfitting or computational issues for high-dimensional input views.

In the experiments conducted in this chapter, we did not utilise a separate validation set for hyperparameter fine-tuning. This decision was driven by the specific characteristics of our dataset and research objectives. Firstly, our study involves a one-leave-out cross-validation approach for the classification task of typological features (see §3.5), where each data point serves as its own validation set during training and testing. This methodology allows us to effectively evaluate the performance of different threshold values on all available data points, providing a comprehensive assessment and insights into the model's generalisation capacity. Furthermore, the size of our language typological dataset is limited, making it challenging to reserve a portion of the data for validation without significantly reducing the available training samples.

Secondly, for the inference task of language phylogenies (see §3.6), the evaluation criteria primarily focus on structural properties and the overall accuracy of the inferred tree. The selection of the optimal threshold value for SVD dimensionality reduction is crucial for this task. Instead of using a separate validation set, we conducted an in-depth analysis of various SVD settings and selected the configuration that yielded the best results based on a transparent evaluation metric.

In summary, our primary objective is to maximise performance within the constraints of our dataset and research goals for our proposed method and all the baselines in each task.

3.4 Sources for language representations

After introducing SVCCA, we turn our attention to the sources or views from which we build our SVCCA shared space. The quality and representativeness of these sources are crucial for the effectiveness of the fused language representation.

3.4.1 Linguistic typology view

As we consistently noted, one popular source of linguistic features is WALS, which provides a comprehensive database of structural features of languages. However, working directly with WALS features can be challenging due to the categorical and redundant nature of the variables, as well as incomplete data entries for many languages (Murawaki, 2015; Bjerva et al., 2019a). As an example, two redundant word order features of WALS are "Order of Subject, Object and Verb" and "Order of Subject and Object". To overcome these challenges, we shift our attention to URIEL, a compilation of linguistic knowledge bases provided by Littell et al. (2017).

URIEL includes a Python library called lang2vec that has processed not only WALS, but also the SSWL (Collins and Kayne, 2011) database and language documentation resources from Ethnologue (Simons and Fenning, 2019). They encoded 103 linguistic typology features from the syntax domain. These features are represented as binary vectors, which we refer to as URIEL Syntax or U_S. Although many redundant features are reduced, some binary features have a strong correlation between them, such as S_SUBJECT_AFTER_VERB and S_SUBJECT_BEFORE_VERB.

In the following experiments, we will use U_S as one of the views to build our SVCCA-based shared space.

3.4.2 NMT-learned view

Following with the next view, we worked with three sets of data-learned language embeddings extracted from different multilingual NMT models. The first two are taken from previous studies, while the third one is developed by us.

Firstly, we exploit the NMT-learned embeddings obtained by Malaviya et al. (2017) using Bible translations. We look for all the entries that intersect with the language entries in U_S, summing up to 731 languages. They trained a many-to-English NMT model with a recurrent neural network (Sutskever et al., 2014). To learn the language embeddings, they added a pseudo-token identifying the source language at the beginning of every input sentence, as it is shown in Figure 3.1. After the training is finished, the vector of the language identity token is extracted from the embedding matrix. We refer to these 512-dimensional vectors as L_B.

Secondly, we use the language embeddings learned in a multilingual many-to-English NMT model by Tan et al. (2019). They trained a small Transformer model (Vaswani et al., 2017) (only two encoder and decoder layers) using 23 languages of the WIT^3 corpus (Cettolo et al., 2012). A significant difference from the first set of vectors (L_B) is the use of factors in the architecture, meaning that the embedding of every input token was concatenated with the embedded pseudo-token that identifies the source language. Figure 3.1 compares the two approaches. We refer to these embeddings as L_W, and they have only 256 dimensions.

Finally, we train a new set of embeddings using the approach of Tan et al. (2019). We trained a multilingual many-to-English Transformer model with factored inputs[1]. However, in our case, we use 53 languages of the TED corpus processed by Qi et al. (2018). For representing the language identification token, we prefer factored embeddings over initial pseudo-tokens as we identified a difference for encoding information about language similarity. This is discussed further in the next chapter (see Chapter 4, §4.6), where we applied our SVCCA-based language representations on multilingual machine translation tasks.

In the following part, we introduce the 23 and 53 languages that are part of the WIT^3 and TED corpora, which are used to train the multilingual NMT models to extract the L_W and L_T language embeddings, respectively.

[1]More details about the model and training are described in Chapter 4, §4.4.1.

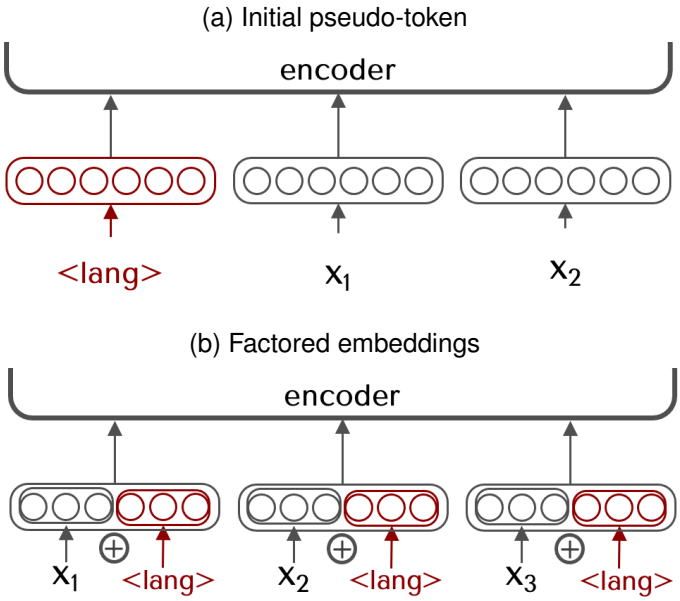


Figure 3.1: Two approaches to introduce a language identity token ($<$ lang $>$): (a) Initial pseudo-token at the beginning of the input sentence; and (b) factored embeddings concatenated to every input token. $\oplus$ is a concatenation operation.

3.4.3 Languages and parallel corpora

Both WIT^3 (Cettolo et al., 2012) and TED (Qi et al., 2018) corpora are built upon TED talk translations. The 23 languages processed by Cettolo et al. (2012) are included in the URIEL database and the U_S vectors. However, from all the languages pre-processed by Qi et al. (2018), we needed to manually correct the mapping of some language ISO codes to identify the correct language vector in URIEL:

- *zh* (*zho* , Chinese macro-language) mapped to *cmn* (Mandarin Chinese).
- *fa* (*fas* , Persian inclusive code for 11 dialects) mapped to *pes* (Western/Iranian Persian).
- *ar* (*ara* , Arabic) mapped to *arb* (Standard Arabic).

We also disregard working with artificial languages like Esperanto (eo) or variants like Brazilian Portuguese (pt-br) and Canadian French (fr-ca). Therefore, we work with 53 languages in TED.

In table 3.1, we present the list of all the languages for the TED and WIT^3 corpora, alongside the following details: ISO 693-2 code, language family and the size of the

training set in thousands of sentences. Furthermore, Figure 3.2 shows the size of each language (paired with English) for both datasets. Language are sorted in descending order by language family size (sum up of the each language size) and by their language-pair size.

We observed that, besides Indo-European sub-families, language family groups are not well-represented in either corpus, and many of these groups consist of only one language (8 in WIT^3 and 11 in TED). This emphasises the need to incorporate additional sources of information, such as typological vectors, into the mix to achieve more diverse representation of languages.

3.5 Analysis of typological knowledge

After introducing the SVCCA method and the views of language representation we are fusing together, we ask whether we are preserving or enhancing the knowledge encoded from each source in the new shared space, such as the typological knowledge encoded in U_S. For this purpose, we investigate the computational typology task of feature prediction as follows.

3.5.1 Prediction of typological features

As we explained earlier, a typological feature can be a word order specification, such as whether the adjective is predominantly placed before or after the noun (features #24 and #25 of U_S). In our task, we aim to predict syntactic features (U_S) using various combinations of multi-view language representations. To control phylogenetic relationships, we consider two validation settings: "one-language-out" and "one-language-family-out" (Bjerva et al., 2019a). In the "one-language-family-out" setting, for example, we remove an entire language family from the training set, and use it as a test set. This setting helps us evaluate the generalisation capabilities of our models across different languages and language families.

Previous work has shown that task-learned embeddings are potential candidates to predict features of a linguistic typology database (Malaviya et al., 2017). Therefore, data-learned embeddings serve as a strong baseline for our task of predicting syntactic features. However, we aim to go beyond the baseline and enhance the NMT-learned language embeddings with typological knowledge from their linguistic typology parallel view using the SVCCA method.

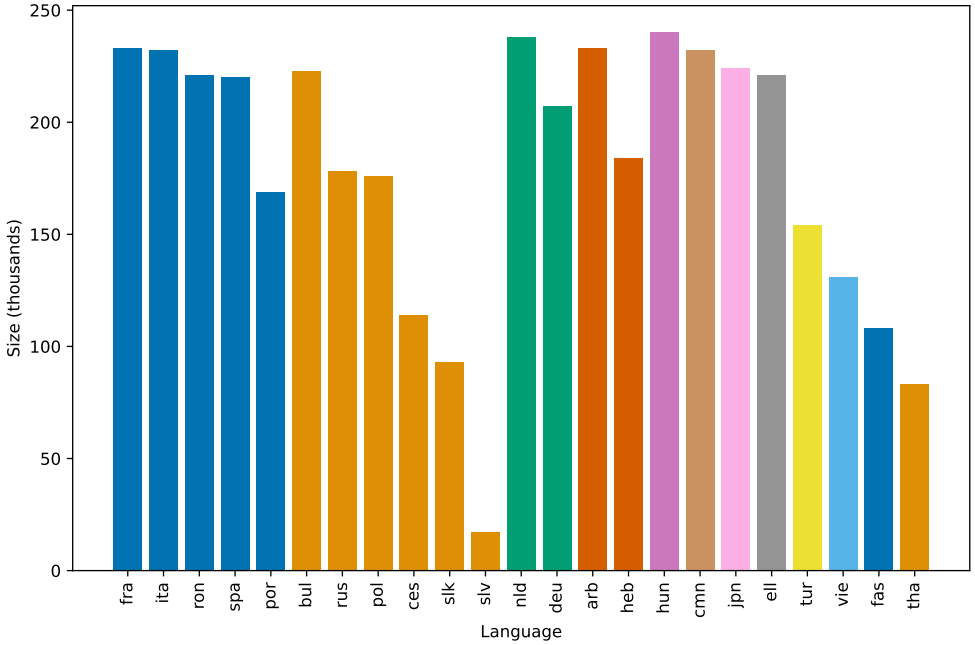
Lang. family	ISO	Language	TED	WIT^3	Lang. family	ISO	Language	TED	WIT^3
Afroasiatic	heb	Hebrew	208	184	IE/Hellenic	ell	Greek	132	221
	arb	Arabic	211	233	IE/Indo-Iranian	ben	Bengali	4	
Austroasiatic	vie	Vietnamese	169	131		urd	Urdu	5	
Austronesian	zlm	Malay	5			mar	Marathi	9	
	ind	Indonesian	85			kur	Kurdish	10	
Dravidian	tam	Tamil	6			hin	Hindi	18	
IE/Albanian	sqi	Albanian	43			fas	Persian	148	108
IE/Armenian	hye	Armenian	21		IE/Italic	glg	Galician	9	
IE/Balto-Slavic	bel	Belarusian	4			por	Portuguese	50	169
	bos	Bosnian	5			ron	Romanian	178	221
	slv	Slovenian	19	17		fra	French	189	233
	mkd	Macedonian	24			spa	Spanish	193	220
	lit	Lithuanian	41			ita	Italian	201	232
	slk	Slovak	60	93	Japonic	jpn	Japanese	201	224
	ces	Czech	101	114	Kartvelian	kat	Georgian	13	
	ukr	Ukrainian	106		Koreanic	kor	Korean	202	
	hrv	Croatian	120		Kra-Dai	tha	Thai	96	83
	srp	Serbian	134		Mongolic	mon	Mongolian	7	
	bul	Bulgarian	172	223	Sino-Tibetan	mya	Burmese	20	
	pol	Polish	173	176		cmn	Chinese	197	232
	rus	Russian	205	178	Turkic	kaz	Kazakh	3	
IE/Germanic	nob	Nor. Bokmal	15			aze	Azerbaijani	5	
	dan	Danish	44			tur	Turkish	179	154
	swe	Swedish	55		Uralic	est	Estonian	10	
	deu	German	165	207		fin	Finnish	23	
	nld	Dutch	181	238		hun	Hungarian	145	240
						eus	Basque	5	

Table 3.1: Languages included in the TED and WIT^3 corpora, along with their respective training sizes (in thousands of sentences). All languages are part of the TED corpus, while only the languages with a size in the WIT^3 column are part of that dataset. The languages are grouped by language family (IE = Indo-European) and sorted in ascending order by their training size in the TED corpus. Basque is an isolated language without an associated language family.

3.5.1.1 Experimental setup

Training Following previous work (Malaviya et al., 2017), we use a Logistic Regression classifier per U_S feature, which is trained with the NMT-learned or SVCCA

(a) Languages and sizes of the WIT^3 dataset



(b) Languages and sizes of the TED dataset

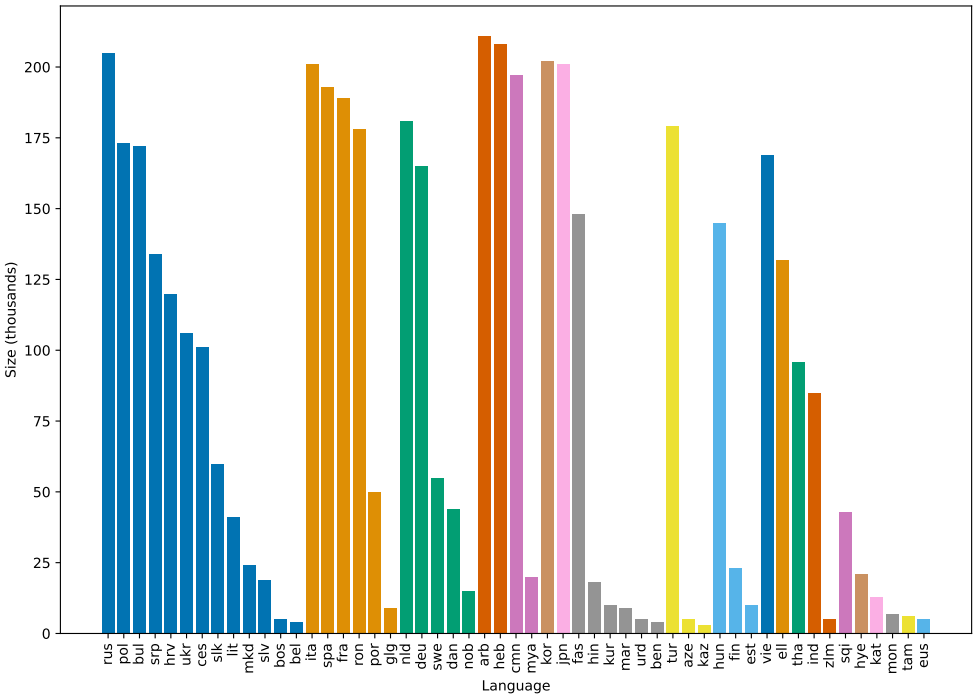


Figure 3.2: Languages and sizes of (a) WIT^3 and (b) TED corpora. The languages are grouped and sorted by family size, and colours are only used to distinguish between consecutive language family groups.

	one-language-out		one-family-out	
	Single	SVCCA	Single	SVCCA
L_B (Bible)	**72.77**	71.68	**72.15**	70.62
L_W (WIT-23)	81.27	**84.83**	79.49	**79.68**
L_T (TED-53)	77.96	**85.37**	76.36	**81.06**

Table 3.2: Avg. % accuracy (↑) of typological feature prediction per NMT-learned and SVCCA(U_S,L_*) setting.

representations in both "one-language-out" and "one-language-family-out" settings, and we average the results.

Given that most of the languages in L_W and L_T are from the Indo-European language family, we decided to split it into subgroups: Albanian, Armenian, Balto-Slavic, Germanic, Hellenic, Indo-Iraninan and Italic/Romance, as shown in Table 3.1.

SVD parameter selection As we noted previously 3.3.2, in the SVD step, we exhaustively search for a threshold of accumulated explained variance of the original dataset that achieves the best performance in the classification task. For a fair comparison, we also transform the data-learned embeddings (L_B, L_W and L_T) with SVD and look for their optimal variance threshold.

Inference For prediction, we use the original NMT-learned embedding or its SVCCA projection as inputs.

3.5.1.2 Results

In Table 3.2, we observe that SVCCA outperformed their NMT-learned counterparts for L_W and L_T, where the performance is significantly better for the one-language-out setting. However, the case for the L_B vectors is different, as we observed that the SVCCA transformation did not improve their performance. One potential explanation for this result is that L_B has been previously trained on a large translated dataset of more than a thousand languages from the Bible, providing it already with a strong sense of language structures for classification tasks (Malaviya et al., 2017). In contrast, L_W and L_T have only been trained on 23 and 53 languages corpora, respectively, making it more challenging for them to capture the complexities of typological features. Nonetheless, another interesting point to note is that the Bible data used to train L_B is

not often used in machine translation research due to its restrictive domain of religious text.

Regarding the SVD parameter selection (see 3.3.2), we use a 0.5 threshold for the NMT-learned vectors of L_B and L_W, and 0.7 for L_T. In case of the SVCCA representation, SVCCA(U_S,L_T) uses [0.75,0.70], whereas SVCCA(U_S,L_B) and SVCCA(U_S,L_W) employ [0.95,0.50] values. The parameter values are the same for both one-language-out and one-family-out settings. We can argue that there is redundancy in the NMT-learned embeddings, as the prediction of typological features with Logistic Regression always prefers a dimensionality-reduced version instead of the original data (threshold at 1.0) in every setting that we are using them.

In conclusion, the results reveal that acquiring specific typological knowledge in language embeddings extracted from multilingual NMT models can be challenging if the training corpora lacks sufficient diversity in terms of the number of languages (L_B versus L_W or L_T). Solely relying on textual corpora to encode linguistic information may prove disadvantageous when dealing with a limited number of languages. However, incorporating typological vectors through the use of SVCCA presents a viable solution for inducing linguistic typology in such scenarios. These results are particularly promising for machine translation research, which frequently employs multilingual datasets encompassing up to a hundred languages, such as OPUS-100 (Zhang et al., 2020).

3.6 Language phylogeny analysis

After analysing the typological knowledge encoded in the SVCCA-based shared space of languages, we turn our attention to another factor for measuring the encoded structural information between languages: phylogenetics. A phylogenetic tree, or phylogeny, is a tree-based structure that represents the evolutionary relationships between nodes, in this case, languages. According to Bjerva et al. (2019b), there is a positive correlation between the language distances in a phylogenetic language tree and a pairwise distance matrix of data-learned representations. Our goal is to investigate whether fusing linguistic typology with SVCCA can preserve or enhance the embedded relationship information. To that end, we examine how well a language phylogeny can be reconstructed from language representations (see §3.6.1), as well as studying the correlation (see §3.6.2).

3.6.1 Inference of a phylogenetic tree

Östling and Tiedemann (2017) were among the first to build a phylogenetic tree using language embeddings extracted from a char-based language model trained on Bible translations, and as a downstream analysis, they conducted a qualitative evaluation of a language tree of the Germanic branch. In contrast, Rabinovich et al. (2017) proposed the task of reconstructing or inferring a language phylogeny from textual corpora by extracting feature vectors from a collection of translated texts. Although they did not use data-learned language embeddings, they systematically compared the resulting trees with a gold standard phylogenetic tree built by Serva and Petroni (2008), which included 17 languages from Germanic, Italic, and Balto-Slavic groups. Building on these previous works, our study aims to reconstruct a phylogeny from various language representations and compare it with an established phylogenetic tree to further investigate the relationship information that is encoded.

3.6.1.1 Experimental design

Following previous work (Rabinovich et al., 2017), we choose the tree of 17 Indo-European languages proposed by Serva and Petroni (2008) as the Gold Standard, which is shown in Figure 3.3a. It is worth noting that we do not generalise the analysis for more languages, as the inferred tree of Serva and Petroni (2008) is only an approximation by lexicostatistic methods (see §3.6.2). Likewise, for the dendrogram inference, we employ agglomerative clustering with variance minimization (Ward Jr, 1963) as the linkage method and cosine similarity as proposed by Bjerva et al. (2019b)[2].

In order to evaluate the effectiveness of different language representations, we have generated a phylogeny from each individual view, including the typological and NMT-learned views, as well as their SVCCA composition. Additionally, we have taken a concatenation ($\oplus$) of the typological and NMT-learned views as a baseline. To ensure the consistency and reliability of our results, we have performed a SVD parameter selection for all vectors, including the concatenation baseline, as described in §3.3.2.

It is worth noting that neither the NMT-learned nor the concatenated vectors contain all 17 language entries of the Gold Standard. This highlights one of the significant advantages of the SVCCA vectors, as they are capable of representing "unknown" or unseen languages using one of the views, as discussed in §3.3.1. For instance, from

[2]More details about the agglomeration clustering algorithm are described in Chapter 4, §4.3.1, where we focus on the language clustering task in multilingual NMT.

the 17 entries in the L_W embeddings of Tan et al. (2019), we lack five language vectors (English, Swedish, Danish, Latvian, Lithuanian), and for our L_T embeddings, there are no entries for English and Latvian.

3.6.1.2 Evaluation metric

Previous studies (Rabinovich et al., 2017) evaluated the task by calculating the distance between the hypothesised tree and the GS tree. This distance was computed as the sum of squared differences between the leaf-pair distances of the two trees. However, interpreting the results solely based on these values, which are between 0 and 1, can be challenging.

In contrast, we propose using a tree edit-distance metric to achieve more easily interpretable results. This type of metric measures the minimum cost of transforming one tree into another by inserting, deleting or modifying (the label of) a node. To this end, we used the All Path Tree Edit Distance algorithm (APTED; Pawlik and Augsten, 2015, 2016), a novel one for the phylogenetic inference task. By using an edit-distance method, we are able to assess the degree of impact of a single change of linkage in the Gold Standard more transparently.

Complementary, as we need to compare inferred pruned trees with different number of nodes, we provide a normalised version given by: $\text{nAPTED} = \text{APTED}\,/(|\,\text{GS}\,| + |\tau|)$, where τ is the inferred tree and the operator $|.|$ indicates the number of nodes. The denominator then is the maximum cost possible of deleting all nodes of τ and inserting each Gold Standard node. By using the normalised metric, we are able to account for the fact that a tree with fewer nodes may require fewer edit operations to transform it into the Gold Standard tree.

3.6.1.3 Results

Table 3.3 shows the results for all settings, and it can be observed that the single-view (U_S, L_B, L_W, and L_T) scores are generally poor. For example, the U_S inferred tree (Fig.3.3c) requires 30 edits to match the Gold Standard. However, the L_T (Fig.3.3d) tree requires half the number of edits to match the Gold Standard.

The best absolute and normalised scores were obtained by fusing the typological vectors (U_S) and the data-learned embeddings (L_T) with SVCCA (Fig. 3.3b). In the resulting tree, English is projected in the Germanic branch, and Latvian is separated from the Balto-Slavic group. Although Bulgarian is misplaced in the SVCCA-based

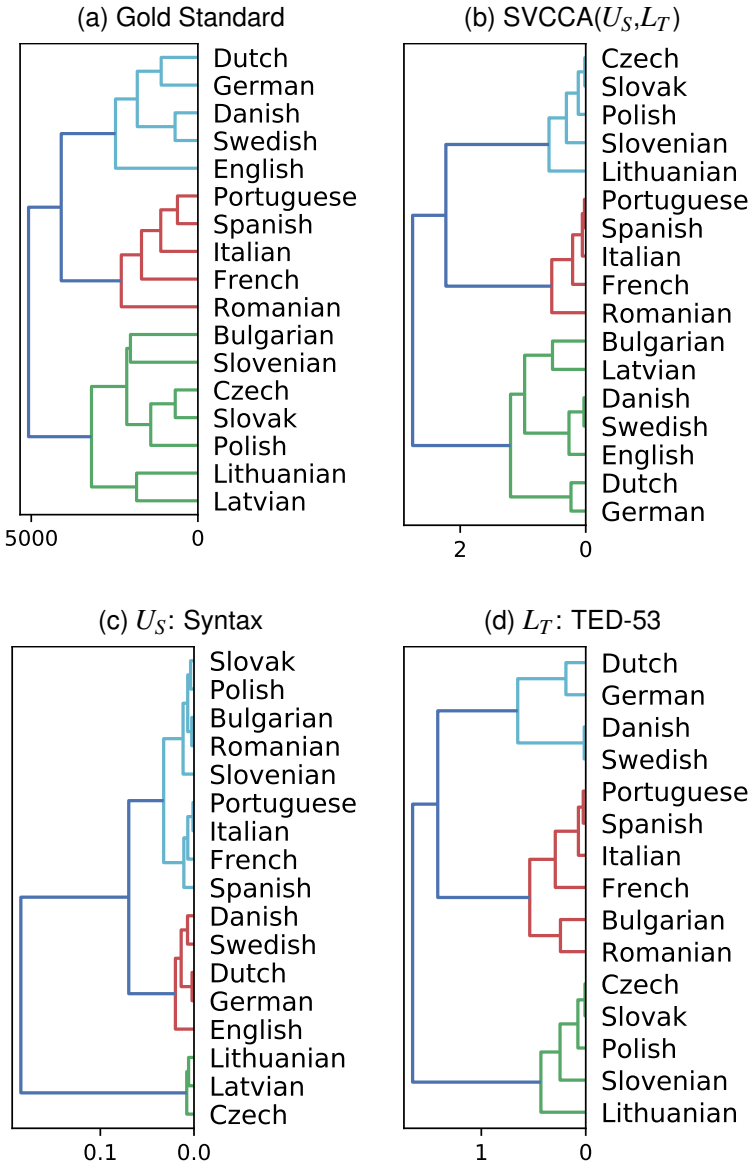


Figure 3.3: Gold Standard phylogeny (a) and reconstructed trees (b-d). L_T is smaller.

	KB-view		
U_S (Syntax)	30 (0.45)		
	NMT-learned	$U_S \oplus L_*$	SVCCA(U_S,L_*)
L_B (Bible)	35 (0.54)	27 (0.42)	23 (0.34)
L_W (WIT-23)	35 (0.62)	23 (0.41)	27 (0.48)
L_T (TED-53)	15 (0.26)	18 (0.29)	**10** (**0.15**)

Table 3.3: APTED (and nAPTED) scores (↓) between the GS and inferred trees from all scenarios. APTED ranges from 0 (no difference) and the size of the tree at most. NMT-learned and concatenation ($\oplus$) can only reconstruct pruned trees of 16 (L_B), 12 (L_W) and 15 (L_T) languages.

tree, it appears to be an issue inherited from the original L_T representation (Fig. 3.3d). Nevertheless, the inferred tree by SVCCA(U_S,L_T) required only ten edits to match the Gold Standard, which is significantly lower than the maximum possible cost of 66. This confirms that our approach is a robust alternative for completing language entries and inferring a language phylogeny.

In conclusion, the use of typological knowledge with SVCCA enhances the language relationships encoded in the NMT-learned embeddings. Moreover, the SVCCA approach generates trees with complete nodes (17), while L_W or L_T (and the concatenation baselines that use them) can only generate 11 or 12 nodes, respectively. In the following section (§3.6.2), we will further discuss the kind of relationships we are representing in the NMT-learned embeddings and SVCCA and study their correlation.

3.6.2 Correlation of SVCCA with genetic similarity

Previous research by Bjerva et al. (2019b) has suggested that raw language embeddings from language modelling can correlate with genetic and structural similarity. They investigated this by correlating a distance matrix with pairwise-leaf-distances of the Gold Standard tree. However, Serva and Petroni (2008) originally inferred the phylogeny by comparing the translated Swadesh list of 200-words with Levenshtein (edit) distances (Dyen et al., 1992). This list is a crafted set of concepts (e.g. "I", "eye", "sleep") for comparative linguistics and is usually processed by lexicostatistics methods to study language relationships through time. Therefore, it is more accurate to argue that corpus-based embeddings could partially encode lexical similarity of languages.

To further investigate the language representations, we correlated the Gold Standard's cophenetic matrix and the pairwise cosine-distance matrices of U_S, L_T, and SVCCA(U_S,L_T) language representations. The cophenetic matrix is a distance matrix that represents the pairwise distances of the hierarchy's leaves (languages) in a tree. Similar to Bjerva et al. (2019b), we used the Spearman correlation coefficient, which is non-parametric, to measure the association between the matrices. Table 3.4 shows the correlation coefficients with p-values < 0.001, where we observe that SVCCA has enhanced the correlation of U_S and L_T.

Our conclusion is that typological knowledge strengthens the representation of lexical similarity within NMT-learned embeddings. However, further investigation is needed to explore the relationship between corpus-based embeddings and structural similarity, which can be done by computing a distance matrix using syntax-dependency-tag counts per language from annotated treebanks (Bjerva et al., 2019b).

Lang. representation	Corr. coeff.
U_S	0.48
L_T	0.68
SVCCA(U_S,L_T)	**0.80**

Table 3.4: Spearman correlation coefficients between the Gold Standard tree's cophenetic matrix and each language representation's pairwise cosine-distance matrix (p-values$<$0.001).

3.7 Conclusion

In this chapter, we have presented a method, based on SVCCA, for constructing a typologically-aware vector space representation of languages by fusing language embeddings extracted from multilingual NMT models with linguistic variables of syntax from typology databases. Our goal was to investigate whether the proposed method for combining typology data and NMT embeddings effectively enhances the vector space representation of languages itself, and whether it successfully preserves both the typology knowledge and data-driven learned information in a single representation.

We tested the effectiveness of our method on a task of typological feature prediction, and found that acquiring specific typological knowledge in language embeddings extracted from multilingual NMT models can be challenging if the training corpora

lacks sufficient diversity in terms of the number of languages. However, incorporating typological vectors through the use of SVCCA presents a viable solution for inducing linguistic typology in such scenarios. These results are particularly promising for machine translation research, which frequently employs multilingual datasets encompassing up to a hundred languages.

Furthermore, we applied our method to the task of phylogenetic tree inference from language representations and found that the use of typological knowledge with SVCCA enhances the language relationships encoded in the NMT-learned embeddings. Additionally, the SVCCA approach generates trees with complete nodes, while other baselines can only generate incomplete trees. Finally, we investigated a further correlation of genetic relationships and found that typological knowledge strengthens the genetic or lexical similarity within NMT-learned embeddings.

Overall, our findings demonstrate that incorporating typological knowledge into multilingual NMT models can effectively enhance the representation of languages and language relationships. However, further research is needed to fully explore the impact of typology on multilingual NMT models (see Chapter 4) and to extend the datasets to a more diverse set of languages (see Chapter 6).

Chapter 4

Language Representations in Multilingual NMT

Our primary objective is to explore the impact of linguistic typology variables and knowledge on machine translation. In the preceding chapter (Chapter 3), we successfully constructed a vector space of languages that integrates typological language representations and NMT-learned embeddings. In this chapter, we investigate the potential of utilizing the language similarity captured in the vector space of the multi-view language representations to improve multilingual NMT applications. To achieve this goal, we focus on a language clustering task in multilingual NMT Tan et al. (2019) and extend the analysis by adapting the ranking approach of Lin et al. (2019) for multilingual models. The content of this chapter is based on Oncevay et al. (2020).

4.1 Introduction

Our question is: **can we leverage a combined language space of typological language vectors and NMT-learned embeddings to improve multilingual machine translation?** Language similarity or relationships between languages have been essential factors for studying transfer learning and multilingual learning approaches in machine translation. For example, some researchers argue for the relevance of language family relationships for selecting parent language-pairs in transfer learning (Zoph et al., 2016), while others suggest that more data or data-based factors are preferable for effective transfer (Kocmi and Bojar, 2018). Likewise, Lin et al. (2019), determined that corpus-based metrics are more relevant than typological similarity or geographical proximity for choosing candidate languages for transfer learning.

Furthermore, in the context of multilingual settings, qualitative analyses of the encoded representations in multilingual NMT models have also shown evidence of capturing traits of language family relationships (Kudugunta et al., 2019). Therefore, identifying pairs of languages that are more related by different factors has been a widely-used approach in these kinds of applications.

To take advantage of our vector space of languages in multilingual NMT, we must consider different approaches that have been taken to improve it. Deepening the model (Zhang et al., 2020), adding language-specific modules (Bapna and Firat, 2019), or dividing the model into shared and language-specific parameters (Lin et al., 2021) have been some of the attempts that have shown overall positive outcomes. However, our focus is on reducing negative interference by creating smaller multilingual models rather than a large one, as proposed by Tan et al. (2019). Specifically, they introduced the language clustering task for multilingual NMT, using data-learned language embeddings to measure similarity and group languages accordingly.

Therefore, we consider that our combined language representation, which integrates NMT-learned language embeddings with typological vectors, and enhances their genetic or lexical relationships as demonstrated in §3.6.2, could have a significant impact on multilingual NMT. In the following section, §4.2, we outline our approach for language clustering, and further argue about the benefits of clustering new languages for a multilingual model using our multi-view language representations in §4.5.4. Afterwards, in §4.6, we discuss that using factored embeddings is the preferred setting for training language representations for clustering tasks. Finally, as a complementary analysis in §4.7, we adapt the ranking approach of Lin et al. (2019) and leverage the distance in our language space to select candidate languages for training multilingual models.

4.2 Language clustering in multilingual NMT

Clustering can be a midway point between managing individual pairwise systems and training large multilingual models, as illustrated in Figure 4.1. Large multilingual NMT models have the potential to improve translations for low-resource languages and enable zero-shot translation, making them a popular approach in the field. Although their training process is simpler, they can be resource-intensive (Johnson et al., 2017). On the other hand, pairwise systems have historically performed better for high-resource language pairs, although the gap has been drastically reduced recently

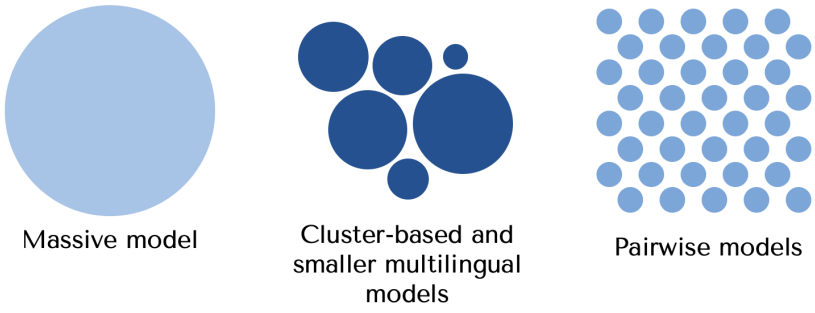


Figure 4.1: Language clustering aims to reduce negative interference by grouping similar languages.

(Bapna et al., 2022). However, in low-resource settings, bilingual systems typically exhibit suboptimal performance due to the limited amount of training data available (Haddow et al., 2022).

In this context, the language clustering task aims to find an optimal group of languages that maximises positive transfer and minimises negative interference in multilingual NMT (Tan et al., 2019). Clustering is a class of unsupervised machine learning algorithms that groups related objects together to discover patterns and structures in data that are difficult to detect manually. In the context of multilingual NMT, clustering algorithms can group languages based on their similarities in a vector space.

Building on the work of Tan et al. (2019), we propose a new approach for clustering related languages in multilingual NMT. While their approach required a new model to be trained for any new language added to the cluster, our multi-view representations can easily handle unseen languages, as explained in §3.3.1. In the following section, §4.3, we elaborate on our methodology for clustering. Then, §4.4 describes the experimental setup used to evaluate our approach, and finally, §4.5 presents and discusses the results of our experiments.

4.3 Methodology for clustering

First, we present the clustering method that we used in our experiments (see §4.3.1) and we detail the criterion that we used to determine the optimal number of clusters (see §4.3.2). Then, we introduce the datasets and languages we worked with (see §4.3.3). Finally, we describe the baselines that we used for comparison and evaluation (see §4.3.4).

4.3.1 Clustering method

Similar to Tan et al. (2019), we use hierarchical agglomeration with average linkage and cosine similarity, which is described in Algorithm 1. Average linkage, also known as UPGMA (Unweighted Pair Group Method with Arithmetic Mean), computes the distance between two clusters as the average distance between all pairs of points in the two clusters.

Algorithm 1: Hierarchical agglomerative clustering with average linkage and cosine similarity

Input: List of language represented as feature vectors

Output: Dendrogram of languages

for *each language representation* **do**
    Initialize a cluster containing only this language;
end
while *number of clusters is greater than 1* **do**
    Compute the cosine similarity between all pairs of clusters;
    Merge the two clusters with the highest similarity using average linkage:

$$d_{\text{avg}}(C_i, C_j) = \frac{1}{n_i n_j} \sum_{x \in C_i} \sum_{y \in C_j} d(x, y),$$

    where d is the cosine similarity and n_i, n_j are the number of elements in clusters C_i, C_j, respectively;
end

The output is a dendrogram, and we must select a level to define the final number of clusters and their respective languages. See Figure 4.2 for a graphic example. However, we differ from Tan et al. (2019) in the criterion for choosing the optimal number of clusters, as we explain in the following part.

4.3.2 Selection of number of clusters

The elbow method (Thorndike, 1953), used by Tan et al. (2019), aims to identify an optimal number of clusters. This is done by plotting the sum of squared distances from the cluster centres of each data point, languages in our case, to the number of clusters. As the number of clusters increases, the sum of squared distances typically decreases; however, this sum will eventually become marginal as the number of clusters increases.

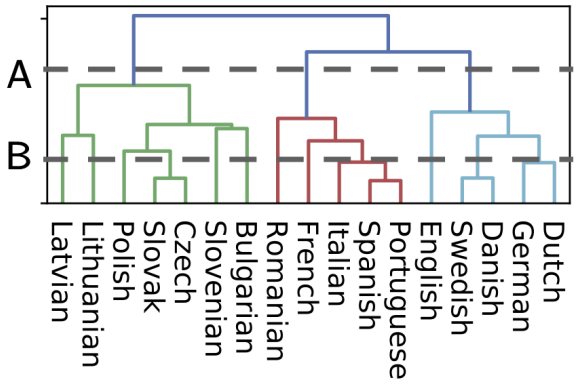


Figure 4.2: Dendrogram of the Gold Standard language phylogeny from Serva and Petroni (2008). To define the number of clusters, threshold A splits the dataset into three clusters (the three Indo-European language branches of Balto-Slavic, Italic and Germanic), whereas threshold B splits more granulated groups (12).

The 'elbow' point on the graph is where the number of clusters is optimal, as this is where we can see a significant decrease in the sum of squared distances from the cluster centres. However, as we can see in Figure 4.3, the heuristic is shallow and might be ambiguous.

Instead, we propose applying the silhouette analysis to determine the optimal number of clusters (Rousseeuw, 1987). It is based on the idea of comparing how similar each element is to its assigned cluster versus how similar it is to the other clusters, and the following formula explains it:

$$silhouette(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{4.1}$$

where i is the index of a language feature vector; $a(i)$ is the average cosine distance between i and all other languages in its cluster; and $b(i)$ is the average cosine distance between i and all languages in the nearest neighbouring cluster.

This method produces a score for each language point in the [-1,1] range that can be used to assess the overall clustering quality. A sample cluster with a silhouette close to 1 indicates that it is cohesive and well-separated. With the average silhouette of all language points, we inspect the different numbers of clusters and look for the peak value above two.

In Figure 4.3, we compare the Elbow method with the silhouette score for clustering the 53 TED languages using our multi-view language representations that combine typological vectors (U_S) and learned embeddings (L_T). The Elbow criterion is ambigu-

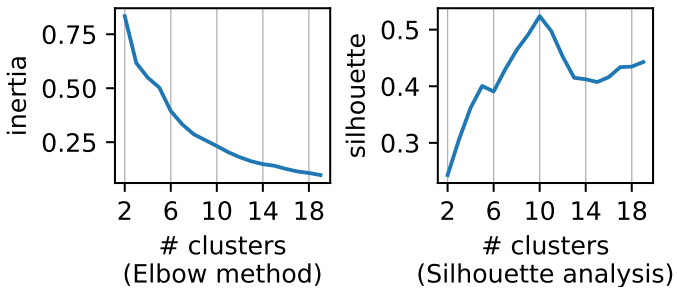


Figure 4.3: Elbow method (left) versus Silhouette analysis (right) for clustering the 53 languages of TED using SVCCA(U_S,L_T).

ous, whereas the silhouette score peaks at 10 clusters, which is the number we select. We later present the dendrogram and the language clusters in the results section (§4.5) and display them in Figure 4.5a.

4.3.3 Dataset and languages

For our dataset and languages, we used the corpus that was processed and tokenised by Qi et al. (2018). This corpus includes 53 languages (TED-53) paired with English and is the same dataset from which we previously learned our L_T embeddings (refer to Chapter 3 for more information). We selected TED-53 instead of the 23 languages used by Tan et al. (2019) to allow for an analysis of how to extend the clusters of languages when we lack data-learned embeddings for some of them.

Regarding processing the dataset, we drop any sentences from the training sets that overlap with any of the test sets. Since we are building many-to-English multilingual systems, this is important, as any such overlap will bias the results.

4.3.4 Clustering baselines and approaches

Continuing with the methodology, we list the baselines and clustering approaches as follows:

1. Individual: A pairwise model per language paired with English, which serves as a significant baseline for high-resource language pairs.

2. Massive: A single model for all languages, which is a relevant baseline for evaluating low-resource language pair performance.

3. Language families or Family: As we mentioned earlier, several studies have focused on language families to analyze transfer learning approaches and multilingual models (Zoph et al., 2016; Lin et al., 2019; Kudugunta et al., 2019). We divide the 33 Indo-European languages from the TED dataset into 7 branches.

4. Syntax or U_S: If we want to assess the impact of our multi-view language representations, we also need to compare the performance of the clusters obtained by each source, in this case, the typological Syntax vectors.

5. Learned or L_T: Similar to Syntax, we compare the performance of the NMT-learned view as a significant baseline. These language embeddings are trained from the same TED-53 dataset. Details about how we extracted them are in §3.4.2.

6. Concatenation or ConCAT or $U_S \oplus L_T$: Based on our previous experiments in language phylogeny (see §3.6), we consider this a strong and straightforward baseline for combining two views.

7. SVCCA-53: Our multi-view representation with SVCCA, which composes both U_S and L_T (53 languages) vectors.

8. SVCCA-23: Similar to the previous setting, but we use the set of 23 language embeddings of L_W instead (Tan et al., 2019) and project the 30 unseen languages using their typological view, according to §3.3.1.

With the last approach, we are interrogating whether SVCCA is a useful method for rapidly increasing the number of languages in a multilingual setting without retraining massive models. In contrast, new languages would require their NMT-learned embeddings for clustering using the L_T and concatenation baselines.

4.4 Experimental setup

To evaluate the performance of the clustering approaches, we focus on a many-to-English multilingual NMT task. This setting is commonly used to simplify the evaluation process, but similar experiments could be conducted in other settings, such as one-to-many multilingual NMT.

Afterwards, we delve into the specifics of our model and training settings in §4.4.1. We then describe the process of fine-tuning the SVD parameter, which is an important step in building our SVCCA representations, in §4.4.2.

4.4.1 Model, training and evaluation

In this part, we list all the configurations used for training the NMT models:

Subword segmentation model We jointly learn 90k shared sub-words with the byte pair encoding (Sennrich et al., 2016b) algorithm built in SentencePiece (Kudo and Richardson, 2018).

NMT architecture Similar to Tan et al. (2019), we train small Transformer models (Vaswani et al., 2017) with 2 encoder and decoder layers.

NMT toolkit We opted to use two different NMT toolkits for our experiments. First, we used Nematus (Sennrich et al., 2017) to train a 53-multilingual NMT model with the TED-53 corpus and extract the factored language embeddings (L_T). Nematus allowed us to configure the factors, which was not possible with other toolkits. For the remaining experiments, we chose the Marian NMT (Junczys-Dowmunt et al., 2018) toolkit as it is efficient in training multiple models due to its optimised code, making it ideal for our large-scale experimental setup[1]. To identify the source language in Marian NMT, we used the initial pseudo-token setting. We did not need to retrieve new language embeddings after training each model, as we only required the NMT-learned embeddings from the large model of 53 language-pairs.

Data sampling and batches Following Tan et al. (2019), we up-sample the training data per language and maintain language-wise proportionality within each batch. Additionally, we let Marian NMT automatically determine the mini-batch size based on the sentence-length and available memory (using the mini-batch-fit parameter).

GPUs and training We train our models with up to four NVIDIA P100 GPUs using Adam optimiser (Kingma and Ba, 2015) with default parameters ($\beta_1 = 0.9, \beta_2 = 0.98, \varepsilon = 10^{-9}$). We chose these default optimiser settings as they have been widely

[1] In the results section, we report all results, including the large model, from training with Marian-NMT.

used in previous works (Tan et al., 2019) and have shown good performance. Additionally, using multiple GPUs for training large models is a common practice in the literature (Johnson et al., 2017; Bapna et al., 2022), as it can significantly reduce the training time without sacrificing performance.

Validation and evaluation metric Finally, we used early stopping at 5 validation steps for the cross-entropy metric, and for evaluating the translations, we used sacreBLEU (Post, 2018) which the following evaluation string:
BLEU+case.mixed+numrefs.1+smooth.exp+tok.13a+version.1.3.7.

4.4.2 SVD explained variance for SVCCA

As we discussed in the previous chapter (Chapter 3, §3.3.2), we fine-tune the threshold of the explained variance ratio for the SVD step to avoid redundancy from the typological vectors or NMT-learned embeddings while computing our multi-view language representation. Particularly for this task, different values of the variance threshold may result in different clustering settings for each representation. However, we cannot perform an exhaustive analysis, as was done in the previous computational typology tasks where we exhaustively searched for the best-performing parameter, since training different cluster-based multilingual NMT models is prohibitively expensive in terms of resources.

To fine-tune the threshold of the explained variance ratio for the SVD step, we take inspiration from the bootstrap clustering methodology (Nerbonne et al., 2008). This approach involves repeatedly sampling a dataset with replacement and calculating a statistic of interest for each sample to estimate its variability. In our case, we use it to assess the stability of our number of clusters given by the peak silhouette score (see §4.3.2 and Eq. 4.1).

For example, we start by obtaining all possible transformations of the 53 NMT-learned language embeddings (U_S) using a range of SVD thresholds from 0.5 to 1.0 with a step of 0.05. Then, we iteratively compute the cluster using 10 to 53 languages, and perform resampling bootstrapping for each iteration (e.g. first we select, 100 times, 10 random languages from the pool of 53). At each step, we calculate the peak silhouette score to determine the optimal number of clusters and plot the number of clusters against the number of languages (see Figure 4.4). Finally, we select the SVD threshold value that shows a steady increase in the number of clusters, or in other words, the

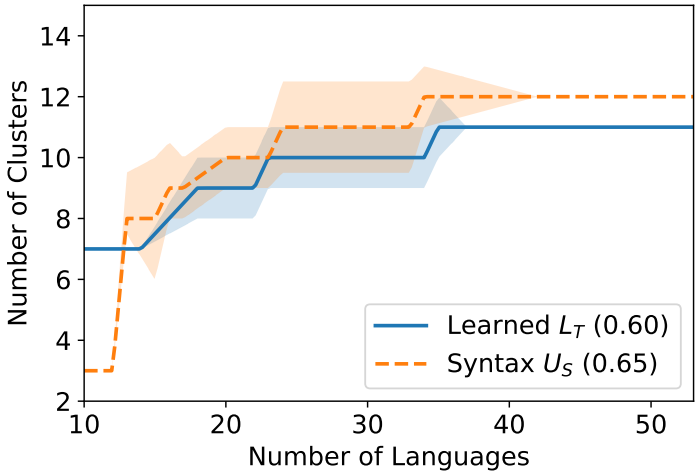


Figure 4.4: Analysis of the number of clusters per total languages using bootstrap clustering and the specific SVD thresholds of explained variance ratio: 0.65 for U_S and 0.60 for L_T. We show the confidence interval computed from the bootstrapping, and we observe that the number of clusters is stable since 36 and 35 languages for U_S and L_T vectors, respectively.

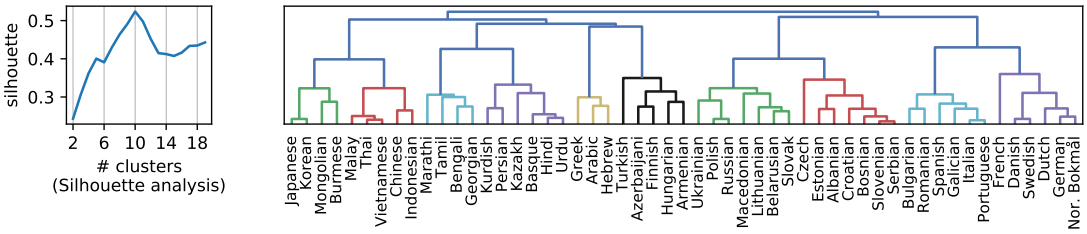
least variability throughout the incremental bootstrapping.

Figure 4.4 shows the selected SVD threshold for U_S and L_T, which are 0.65 and 0.60, respectively. In both cases, we observe that the number of clusters stops increasing after reaching a specific number of languages. Other configurations for the SVD parameter do not exhibit this pattern. These parameter values are applied to the concatenation baseline, single-view vectors, and the SVCCA approaches (see 4.3.4).

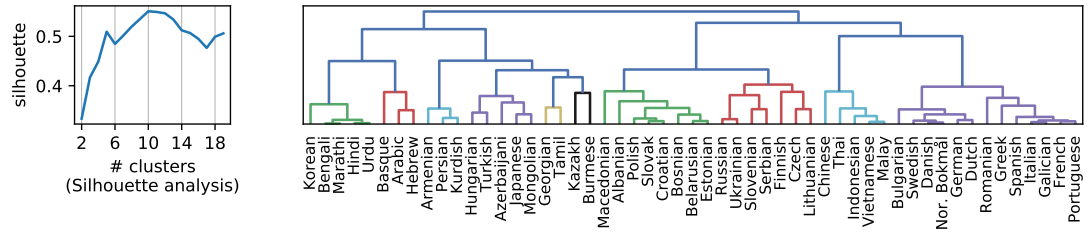
4.5 Language clustering results

In this section, we present the results of our language clustering task for multilingual NMT, following the methodology and experimental setup described in the previous parts. We begin by discussing the composition of the clusters obtained using the SVCCA method in §4.5.1, providing insight into the relationships captured by our approach. Next, in §4.5.2, we analyse the clustering results grouped by training size bins, allowing us to explore the effects of dataset size on the quality and consistency of the clusters. To further investigate the patterns and dynamics of the language relationships in the clustering, we also discuss the results by language families in §4.5.3. Finally, as a reference, we provide the translation results per language and baseline in Table 4.2 at the end of the section. Our goal is to provide a comprehensive evaluation

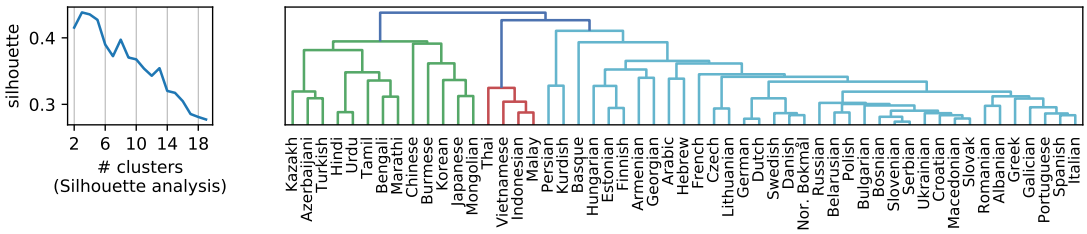
(a) SVCCA-53(U_S, L_T): SVCCA representations of Syntax and TED-53



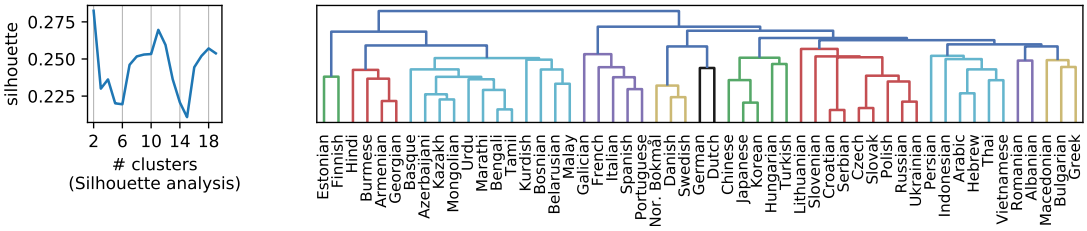
(b) SVCCA-23(U_S, L_W): SVCCA representations of Syntax and WIT-23



(c) U_S: Syntax



(d) L_T: NMT-learned from TED-53 (using factors)



(e) $U_S \oplus L_T$: Concatenation of Syntax and TED-53

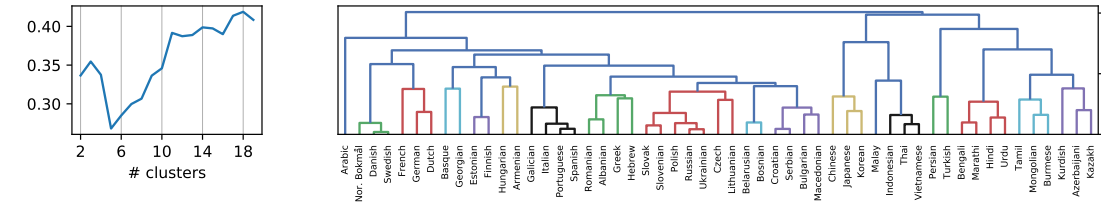


Figure 4.5: Silhouette analysis and dendrograms for (a) SVCCA-53(U_S,L_T); (b) SVCCA-23(U_S,L_W); (c) U_S or Syntax; (d) L_T or NMT-learned embeddings; (e) $U_S \oplus L_T$ or concatenation. Using the silhouette score, we automatically select the highest peak (greater than 2) and represent the language clusters with different consecutive colours. For instance, there are 10 clusters in (a) and (b).

of the effectiveness and potential of our approach, as well as to offer insights into the linguistic diversity and language relationships captured by our methodology.

4.5.1 Cluster composition

In Figure 4.5, we can observe that the silhouette score for SVCCA-53 (Fig. 4.5a) indicates that ten clusters are the optimal setting for achieving more cohesive and better-separated groups, which are shown in the dendrogram by consecutively different colours. It is also notable that the sizes of the clusters do not exhibit high variance, with the smallest one containing only 3 languages (Greek-Arabic-Hebrew) and the largest one having seven entries. In contrast, the concatenation baseline (Fig. 4.5e) is split into 18 clusters by the silhouette analysis, with an average of 2.9 languages per cluster. This is not a practical choice for maintaining NMT systems in production. Arabic, for instance, is left alone in a cluster, whereas neither SVCCA-53 nor SVCCA-23 has a cluster with only one member.

Furthermore, we can observe, in SVCCA-53, that some of the languages are grouped by phylogenetic (e.g., the Spanish-Galician-Italian-Portuguese subgroup) or geographical criteria (e.g., Japanese-Korean or Thai-Vietnamese subgroups). These agglomeration patterns are evident in both the typological view (Fig. 4.5c) and the NMT-learned (Fig. 4.5d) embeddings. This is an encouraging result as it indicates that both sources of information about language relationships are preserved in the combined representation, each with specific language properties and encoded relationships.

Regarding the typological view or U_S (Fig. 4.5c), we observe that it only splits three clusters by its silhouette analysis, with one of them being a large group. This suggests that using only typological information, it is difficult to split highly related languages, such as the large Indo-European group. In contrast, with the NMT-learned embeddings or L_T (Fig. 4.5d), we observe good cohesion in small groups of languages (for sub-families), but high-level connections are missed, with the Indo-European groups segregated and mixed with other language families. However, this is not necessarily a negative outcome, as the NMT-learned embeddings encode a different type of relationship from the corpus, which could be more related to shared vocabulary (or lexical similarity, as we studied in §3.6.2) or even the domain of the texts for each language-pair. However, it is noteworthy that the combined SVCCA-53 representation retains and combines all of these sources of information.

Upon closer inspection, we observe that some entries do not correspond to their

respective family branches. Such observations are not surprising, since they can be attributed to the encoded information from the single-view sources. For instance, the L_T phylogenetic tree (Fig. 3.3d) "misplaced" Bulgarian within Italic languages. However, these unexpected agglomerations may uncover surprising clusters that could help avoid isolating languages without close relatives, such as Basque or even Japanese, which is the only Japonic member in the set of languages. Therefore, the cluster composition of the SVCCA-53 representation can help identify hidden relationships between languages that may not be immediately apparent using traditional typological or NMT-learned embeddings.

Similar patterns emerge in the clustering with the SVCCA-23 language representations, which will be discussed with more detail in a specific section (see §4.5.4).

4.5.2 Training size bins

Next, our aim is to assess the impact of different clustering settings on language pairs of varying sizes, ranging from low to high. To this end, we segregate the language pairs

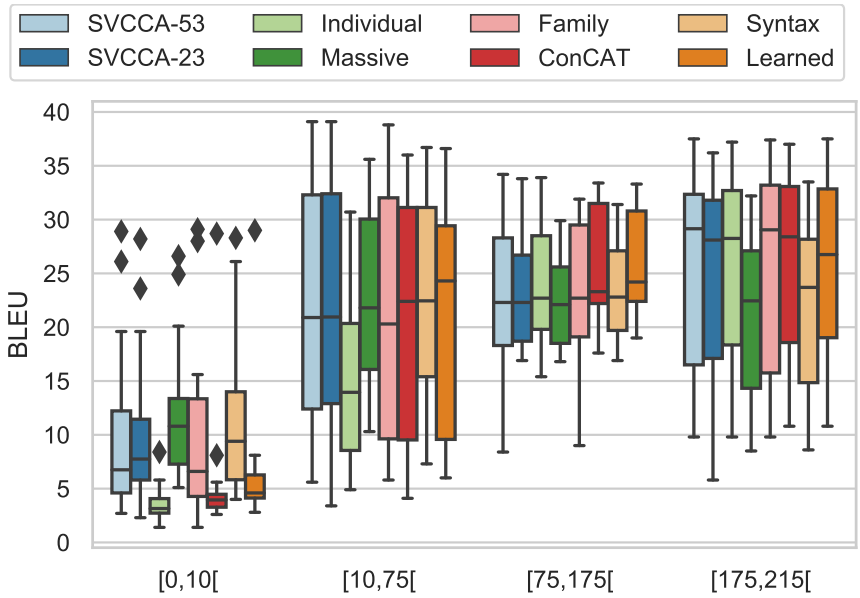


Figure 4.6: Box plots of BLEU scores per training-size bins. Each bin is represented by the range of minimum and maximum training size, and they group 14, 14, 13, and 12 languages, respectively. The box in each plot corresponds to the interquartile range (IQR), covering the middle 50% of the data. The whiskers extend up to 1.5 times the IQR above the third quartile (Q3) and down to 1.5 times the IQR below the first quartile (Q1). Outliers, shown as diamonds, are individual data points that fall beyond this range.

based on the number of training sentences they have, and group them into bins with similar sizes. The upper bounds of the bins are manually defined as [10,75,175,215] thousands of training sentences, resulting in groups of [14,14,13,12] languages, respectively. Figure 4.6 displays the box plots of BLEU scores, allowing us to analyse the mean and variance of each distribution. Complementary, Table 4.2, at the end of the results section, shows the languages per bin and their scores for all the baselines.

In all bins, we observe that both SVCCA-53 and SVCCA-23 consistently achieve competitive performance compared to the best baseline in each group. In other words, their clusters provide stable performance for both low and high-resource languages.

When considering the first two bins, which represents the extremely low-resource and low-resource scenarios, the Massive baseline and Syntax or U_S (with its large cluster of 36 languages) outperformed the rest of the approaches as expected. However, our analysis reveals that the SVCCA-53 aand SVCCA-23 approaches, with smaller clusters of at most 7 or 13 languages, respectively, perform almost as well than U_S in the first bin, and achieve a comparable performance to the two baselines in the second bin. It is worth noting that the other baselines underperformed in the extremely low-resource setting, with exception of Family, which achieved a comparable performance to SVCCA-53. The reason is that most of the languages that belong to the first bin are paired with 2 or more extra languages by their family relationships. The only exceptions are Basque, Tamil and Mongolian, which are unique family members, and Malay which is paired with Indonesian only.

On the other end of the spectrum, in the rightmost bin representing the highest resource languages, Massive and U_S perform as expected by achieving worse results than the SVCCA and other baselines. It is worth noting that the SVCCA approach still demonstrates competitive accuracy compared to the Individual and Family approaches in this bin, despite the expectation that these approaches would perform better for higher-resource languages due to their large number of clusters with only one language, which favours high-resource settings.

Moreover, the family-based clusters maintain steady performance across most bins but the lowest setting, appearing as a robust alternative. However, it is essential to note that the linguistic diversity of multilingual machine translation datasets may not always support this approach. Out of 53 languages, only 20 do not belong to the Indo-European family, and these 20 languages are spread across 12 different language families. This indicates that very few families with more than two languages are represented in the dataset. Additionally, there is one isolated language, Basque, that cannot

be paired with any other language, even if the dataset is extended.

Other approaches, like using the NMT-learned embeddings (L_T) as proposed by Tan et al. (2019) or the concatenation baseline (ConCAT), obtain similar translation results in the last three bins. However, obtaining the NMT-learned embeddings (for L_T and ConCAT) requires training a massive model on all 53 languages first, whereas using SVCCA and a pre-trained smaller set of language embeddings is sufficient for projecting new representations, as we show with our SVCCA-23 approach and will discuss further in §4.5.4. This highlights the practical advantages of SVCCA in terms of efficiency.

4.5.3 Results by language families

Following the guidelines of Anastasopoulos (2019) for evaluating multilingual benchmarks, we have grouped the scores by language families, as shown in Table 4.1. It is worth noting that most of the approaches have obtained clusters with similar overall translation family-weighted accuracy, except for the individual models, which significantly underperform. The poor performance is transferred to the Family baseline, as most of the groups contain only one language, given the low language diversity of the dataset. We note again that having single-language clusters is inefficient for deployment purposes.

The U_S vectors achieve the highest overall accuracy, mainly due to its large cluster of 36 languages covering different language family groups (see Fig. 4.5c). Meanwhile, SVCCA-53 achieves the second-best overall result, by a small margin, with 3 to 7 languages per cluster, which are usually faster to train. Similarly, SVCCA-23 achieves competitive performance. Besides, the Massive model, the L_T embeddings, and the concatenation baseline present competitive scores as well. However, we note again that the first requires more resources to train until convergence, whereas the last two require the 53 pre-trained embeddings from a previously trained massive system.

Upon closer inspection, we can see that for the isolated language Basque, the best-performing cluster for the language, besides the massive model and the large cluster of U_S, is SVCCA-23, followed by SVCCA-53. The former groups Basque with Arabic and Hebrew (see Fig. 4.5b), which are part of the Afroasiatic group and are not genetically related to Basque, although they share some typological similarities, particularly in terms of their rich inflectional morphology. Interestingly, Basque is closer to the Afroasiatic pair in the U_S dendrogram than in the L_T one, for instance (see Fig. 4.5c

Lang. families	# L	Size(k)	Ind.	Mass.	Fam.	U_S	L_T	ConCAT	**SVCCA-53**	**SVCCA-23**
Isolate (Basque)	1	5	2.20	**11.10**	2.20	*10.90*	5.60	3.90	6.40 $_{\Delta\text{-}4.7}$	10.10 $_{\Delta\text{-}1.0}$
Dravidian	1	6	1.40	**5.10**	1.40	*4.00*	2.80	2.60	2.70 $_{\Delta\text{-}2.4}$	2.30 $_{\Delta\text{-}2.8}$
Mongolic	1	7	2.70	**6.90**	2.70	5.70	3.90	3.50	5.20 $_{\Delta\text{-}1.7}$	*6.10* $_{\Delta\text{-}0.8}$
Kartvelian	1	13	5.80	*14.30*	5.80	**14.50**	8.80	4.60	5.60 $_{\Delta\text{-}8.9}$	5.50 $_{\Delta\text{-}9.0}$
IE/Armenian	1	21	9.00	*16.30*	9.00	**16.90**	9.80	13.20	13.30 $_{\Delta\text{-}3.6}$	12.20 $_{\Delta\text{-}4.7}$
IE/Albanian	1	44	20.80	27.80	20.80	*29.10*	28.60	**31.60**	26.30 $_{\Delta\text{-}5.3}$	25.80 $_{\Delta\text{-}5.8}$
Kra-Dai	1	97	15.40	16.80	15.40	16.90	**19.00**	17.60	*17.70* $_{\Delta\text{-}1.3}$	*17.70* $_{\Delta\text{-}1.3}$
IE/Hellenic	1	132	31.90	29.90	31.90	30.90	32.20	*33.40*	**34.20**	32.70 $_{\Delta\text{-}1.5}$
Austroasiatic	1	170	*22.70*	20.30	*22.70*	21.60	**23.60**	22.20	22.30 $_{\Delta\text{-}1.3}$	22.30 $_{\Delta\text{-}1.3}$
Japonic	1	201	9.80	8.50	9.80	8.60	**10.80**	**10.80**	9.80 $_{\Delta\text{-}1.0}$	9.70 $_{\Delta\text{-}1.1}$
Koreanic	1	203	14.40	12.20	14.40	11.90	**15.10**	*15.00*	13.30 $_{\Delta\text{-}1.8}$	5.80 $_{\Delta\text{-}9.3}$
Austronesian	2	91	13.95	22.20	18.50	22.85	17.25	15.55	**23.05**	**23.05**
Sino-Tibetan	2	218	9.90	**11.90**	10.75	9.95	10.90	9.95	*11.20* $_{\Delta\text{-}0.7}$	9.05 $_{\Delta\text{-}2.8}$
Afroasiatic	2	420	29.45	22.45	**30.20**	23.70	27.65	28.40	*29.70* $_{\Delta\text{-}0.5}$	29.10 $_{\Delta\text{-}1.1}$
Uralic	3	180	11.13	15.03	13.00	**15.63**	12.57	11.37	13.93 $_{\Delta\text{-}1.7}$	*15.20* $_{\Delta\text{-}0.4}$
Turkic	3	189	8.27	9.33	8.87	8.87	*9.40*	8.73	9.23 $_{\Delta\text{-}0.2}$	**9.43**
IE/Germanic	5	462	27.70	30.74	34.18	31.90	32.60	32.76	**34.68**	*34.56* $_{\Delta\text{-}0.1}$
IE/Indo-Iranian	6	198	7.20	**12.32**	7.18	*10.07*	8.45	7.70	6.30 $_{\Delta\text{-}6.0}$	8.63 $_{\Delta\text{-}3.7}$
IE/Italic	6	823	28.67	29.15	**34.02**	30.32	33.50	*33.65*	*33.65* $_{\Delta\text{-}0.4}$	32.90 $_{\Delta\text{-}1.1}$
IE/Balto-Slavic	13	1,171	17.74	22.26	**23.88**	*23.24*	22.05	21.30	22.39 $_{\Delta\text{-}1.5}$	21.71 $_{\Delta\text{-}2.2}$
Weighted average →			16.70	19.76	19.79	**20.03**	19.60	19.16	*19.97* $_{\Delta\text{-}0.1}$	19.82 $_{\Delta\text{-}0.2}$
Number of clusters/models →			53	1	20	3	11	18	10	10

Table 4.1: BLEU score average per language family (IE=Indo-European). Every method includes the weighted BLEU average per number of languages (#L) and the number of clusters/models. Bold and italic represent first and second best results per family. Δ for SVCCA indicates the difference with respect to the highest score.

and Fig. 4.5d). In any case, they compose a cluster with 420k sentences for training, which definitely helps the low-resource and isolated language.

In a similar vein, we can examine the case of Mongolian, a unique language belonging to the Mongolic family and a low-resource one. After the Massive setting, the best-performing cluster is obtained by SVCCA-23, which groups Mongolian along with Japanese, Azerbaijani, Turkish and Hungarian (refer to Fig. 4.5b). Although Mongolian and Japanese are grouped together in the typological vectors (as seen in Fig. 4.5c), Mongolian is closer to two Turkic languages, namely Azerbaijani and Kazakh, in the NMT-learned embeddings (Fig. 4.5d). In addition, Hungarian is closer to Turkish in the same setting, which could be a possible reason why they are all grouped to-

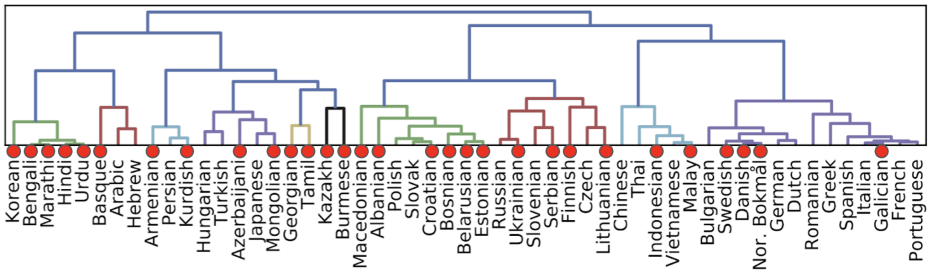


Figure 4.7: Dendrogram computed from SVCCA-23, or SVCCA(U_S,L_W). The red dots indicate the 30 languages that are projected using their typological vector view.

gether by SVCCA-23. This close connection to Turkic languages might be attributed to historical contact and some loanwords or just the similarity in the topics of their respective datasets, which is a confounder in our study.

Furthermore, SVCCA-23 presents itself as a fast and competitive alternative if our aim is to target specific new languages, a topic we delve into in the following section.

4.5.4 Clustering unseen languages

As we have consistently noted in our discussion of the results, the SVCCA representations, particularly the SVCCA-23 setting, offer significant benefits. We should note once again that SVCCA-23, or SVCCA(U_S,L_W), is a combination of 23 languages, combining the typological vectors or U_S with the NMT-learned language embeddings or L_W, which were trained using the WIT^3 dataset of 23 languages paired with English. In this context, if we want to cluster 53 languages using SVCCA-23, we first need to project the 30 languages that were not included in the creation of the shared vector space. We do this by following the steps introduced in §3.3.1, using the typological vectors of the additional 30 languages to project them onto the shared space. The dendrogram obtained by clustering the SVCCA-23 representations with marks in the 30 extra languages projected after the vector space computation is shown in Figure 4.7.

This approach is a fast and efficient way to add a new language to the original set. For example, if we want to add Malay or Indonesian to the 23 language dataset, we only need to retrain the multilingual model of Chinese, Thai, and Vietnamese with either or both of them. Similarly, if we decide to add Galician, we only need to include it in the cluster of Portuguese, French, Italian, Spanish, Greek, and Romanian. Furthermore, if we need to deploy a translation model for Basque, the SVCCA-23 chosen cluster of only three languages (including Arabic and Hebrew) could achieve comparable or

even better accuracy than a massive model.

4.5.5 Overall conclusions

In summary, our study has shown that SVCCA provides reliable performance across various resource settings and outperforms existing baselines. Additionally, the unexpected agglomerations in the resulting clusters may reveal previously unknown linguistic relationships, helping to avoid isolating languages without close relatives. Furthermore, SVCCA is an efficient approach when clustering a new language without the need to extract its NMT-learned embedding from a previously trained large multilingual model.

In the following section, §4.6, we explore how the different approaches for extracting NMT-learned embeddings (factored embeddings or initial pseudo-token, that are previously shown in Figure 3.1) have an impact on the clustering process. Afterwards, we explore the complementary task of language ranking in §4.7. Finally, in §4.8, we briefly describe a tool to facilitate the computation of language representations.

4.6 Factored embeddings for language clustering

In this section, we focus on the configuration for extracting the NMT-learned embeddings, which is a crucial setting for the L_T baseline in the clustering experiments. As discussed in the previous chapter (Chapter 3, §3.2 and §3.4.2, and Figure 3.1), there are two different ways to extract data-learned embeddings: (i) factored embeddings, which were used by Tan et al. (2019) in multilingual NMT and by Östling and Tiedemann (2017) in a char-based neural model, and (ii) adding the language identity vector as an initial pseudo-token, which was used by Malaviya et al. (2017) in NMT for Bible translations. Our question is whether there is a preferred setting for the language embedding configuration in the context of language clustering.

To answer this question, we compute the silhouette score and plot the dendrograms for additional settings in Figure 4.8. Specifically, we examine (a) the L_B embeddings trained with Bible translations and initial pseudo-token, (b) an alternative set of 53 language embeddings (L_T*) from TED-53 but using the initial pseudo-token instead of factored embeddings, and (c) the SVCCA-53* representations using L_T* and U_S. For the first two cases (a-b), we observe that the silhouette score never exceeds 0.2 (1 is best), and the curve keeps degrading as we examine a higher number of clusters,

Table 4.2: List of languages sorted by training size (in thousands), with their BLEU scores per clustering approach. The total average is shown in the last row.

ISO	Size	Bin	Individual	Massive	Family	U_S	L_T	$\oplus$	SVCCA-53	SVCCA-23
kaz	3	1	2.5	5.3	4	4.3	3.3	2.7	3.3	3
bel	4	1	3.1	13	14.3	13.7	4.3	2.8	12.4	10.1
ben	4	1	3.1	10.5	5.9	6.2	4.3	4.6	4.4	5.7
eus	5	1	2.2	11.1	2.2	10.9	5.6	3.9	6.4	10.1
zlm	5	1	4.1	20.1	15.6	19.7	6.5	4.1	19.6	19.6
bos	5	1	4.2	26.6	28	28.3	6.5	4.1	26.1	23.6
urd	5	1	3.9	11.8	7.5	8	5.5	5.6	7.1	6.8
aze	5	1	2.8	8.1	6.4	6.7	4.2	3.2	7.3	7.4
tam	6	1	1.4	5.1	1.4	4	2.8	2.6	2.7	2.3
mon	7	1	2.7	6.9	2.7	5.7	3.9	3.5	5.2	6.1
mar	9	1	3.2	7	5.1	5.2	4.1	4	3.3	4.7
glg	9	1	8.4	24.9	29.1	26.1	29	28.7	28.9	28.2
kur	10	1	4	10.1	6.8	10.8	4.9	3.6	6.3	8.1
est	10	1	5.8	13.5	10.5	14.1	8.1	8.1	11.7	11.9
kat	13	2	5.8	14.3	5.8	14.5	8.8	4.6	5.6	5.5
nob	15	2	19	35.2	38.8	36.4	35	35	39.1	39.1
hin	18	2	8.1	16	8.8	10.5	9.5	6.2	8.3	8.6
slv	19	2	8.7	19.5	19.8	20.2	21.8	19.3	18.1	19.7
mya	20	2	4.9	10.3	7.6	7.3	6	4.1	7.7	3.4
hye	21	2	9	16.3	9	16.9	9.8	13.2	13.3	12.2
fin	23	2	8.5	14.4	11.5	14.9	8.3	8.3	12.1	15
mkd	24	2	15.7	26.8	27.3	27.4	27.2	28	25.1	22.6
lit	41	2	12.2	17.9	19.4	18.4	20	19	17.9	18.6
sqi	43	2	20.8	27.8	20.8	29.1	28.6	31.6	26.3	25.8
dan	44	2	30.7	35.6	38.4	36.7	34.4	34.4	38.9	39
por	50	2	27.2	32.8	36.9	33.7	36.6	36	36.7	36.5
swe	55	2	27	30.8	33.6	31.8	29.7	29.7	34.3	34.6
slk	60	2	18.1	24.1	26	24.7	26.8	25.5	23.7	22.2
ind	85	3	23.8	24.3	21.4	26	28	27	26.5	26.5
tha	96	3	15.4	16.8	15.4	16.9	19	17.6	17.7	17.7
ces	101	3	20.7	22.1	23.9	22.8	24.2	23.3	21.2	22.1

Table 4.2, continued from previous page

ISO	Size	Bin	Individual	Massive	Family	U_S	L_T	$\oplus$	SVCCA-53	SVCCA-23
ukr	106	3	19.8	20.9	22.6	22	23.5	22.5	21.2	21.7
hrv	120	3	28.5	27.5	30.4	28.9	30.8	31.5	28.3	26.7
ell	132	3	31.9	29.9	31.9	30.9	32.2	33.4	34.2	32.7
srp	134	3	26.4	25.6	28.3	27.1	28.8	29.4	26.3	25.4
hun	145	3	19.1	17.2	17	17.9	21.3	17.7	18	18.7
fas	148	3	20.9	18.5	9	19.7	22.4	22.2	8.4	17.9
deu	165	3	30.1	25.5	29.5	26.9	31.4	31.7	29.9	29.6
vie	169	3	22.7	20.3	22.7	21.6	23.6	22.2	22.3	22.3
bul	172	3	33.9	29.9	31.9	31.4	33.3	33.1	34.2	33.8
pol	173	3	18.9	17.4	19.1	18.2	19.3	18.9	18.3	16.9
ron	178	4	30	25.8	30.7	27	28.1	30.8	30.8	29.6
tur	179	4	19.5	14.6	16.2	15.6	20.7	20.3	17.1	17.9
nld	181	4	31.7	26.6	30.6	27.7	32.5	33	31.2	30.5
fra	189	4	35.6	30.6	35.9	32	35.9	36.1	34.3	34.5
spa	193	4	37.2	32.2	37.4	33.5	37.5	37	37.5	36.2
cmn	197	4	14.9	13.5	13.9	12.6	15.8	15.8	14.7	14.7
jpn	201	4	9.8	8.5	9.8	8.6	10.8	10.8	9.8	9.7
ita	201	4	33.6	28.6	34.1	29.6	33.9	33.3	33.7	32.4
kor	202	4	14.4	12.2	14.4	11.9	15.1	15	13.3	5.8
rus	205	4	20.4	18.1	19.4	19	20.1	19.5	18.3	18.8
heb	208	4	32.4	24.4	32.9	25.8	29.9	30.3	31.9	31.6
arb	211	4	26.5	20.5	27.5	21.6	25.4	26.5	27.5	26.6
		Avg.	16.7	19.8	19.8	**20.0**	19.6	19.2	**20.0**	19.8

which contrasts the trend shown in Figure 4.5 and in the silhouette from (c), where the L_T* are fused with typological vectors.

Upon closer inspection of the dendrograms, we observe that the hierarchies in sub-figures (a) and (b) do not define cohesive groups for the languages. They are usually very separated (the right vertical axis shows the distance between the points in the dendrogram), especially in the L_B embeddings, where the distance indicates that the first cluster links are above the 0.5 threshold, unlike SVCCA or other baselines. Furthermore, in sub-figure (c), we note that the SVCCA(U_S,L_T*) is adversely affected by the noisy agglomeration of the original NMT-learned embeddings with initial pseudo-

tokens. This noisy agglomeration results in only a few preserved linguistic relationships from the typological vectors, such as the small cluster of Danish-Swedish-Dutch, although it is unexpectedly associated with Portuguese.

These patterns in the silhouette analysis and the dendrograms suggest that the language points cannot be clustered in a cohesive way. However, this does not imply that the L_B or L_T* language representations are irrelevant. As we have previously demonstrated in §3.5.1, these representations can be used to perform a classification task such as typological feature prediction, which aligns with the findings of Malaviya et al. (2017). We, therefore, argue that a many-to-one multilingual NMT model encodes just enough information in the initial pseudo-token to identify or classify the input sentence by language, but not enough to establish clear distances or relationships between the languages involved.

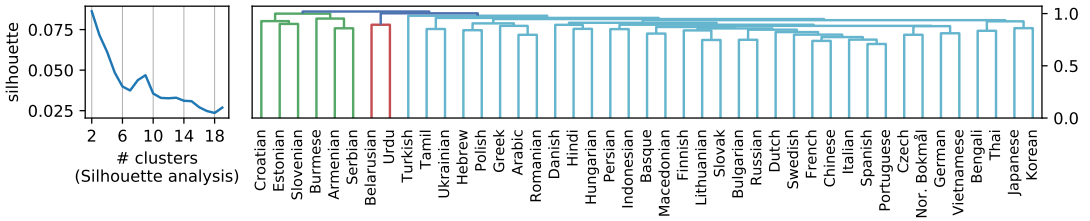
Therefore, we consider it crucial to use factored language embeddings for extracting language relationships and to apply them in language clustering. This finding is consistent with the work of Tan et al. (2019) for multilingual NMT and also for the case of language modelling, as Östling and Tiedemann (2017) could obtain a hierarchy of Germanic languages from their factored embeddings.

4.7 Language ranking for multilingual NMT

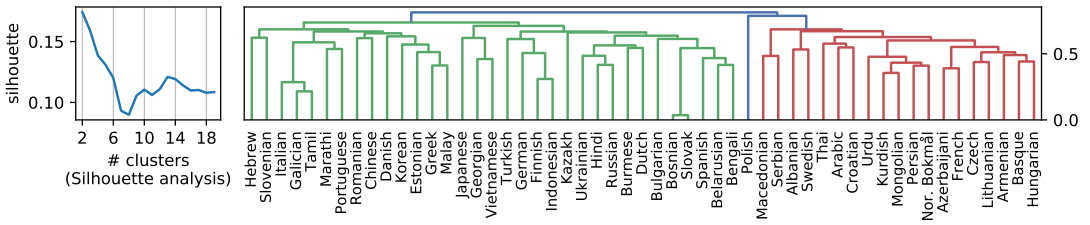
After examining the language clustering task, we now turn our attention to a complementary task proposed by Lin et al. (2019), which involves ranking language candidates for transfer learning. Unlike clustering, we aim to identify the best candidates for building a multilingual NMT model around an initial language seed. This is particularly relevant in cases where we need to work with a new language pair. The idea is to construct a multilingual cluster around the new language that can reduce negative transfer and improve the model's performance. This approach will enable us to continue evaluating the capacity of our language vector space to encode similarity information.

LANGRANK Lin et al. (2019) proposed a method called LANGRANK, which determines the languages that should be transferred to a specific low-resource language for a given NLP task, including NMT. They trained their models by performing exhaustive searches through the potential transfer languages, which were computationally and resource-intensive. They also used a diverse set of crafted features, including

(a) L_B: NMT-learned from Bible



(b) L_T*: NMT-learned from TED-53 but with initial pseudo-tokens



(c) SVCCA-53*(U_S,L_T*): SVCCA representations of Syntax and TED-53 but with initial pseudo-tokens

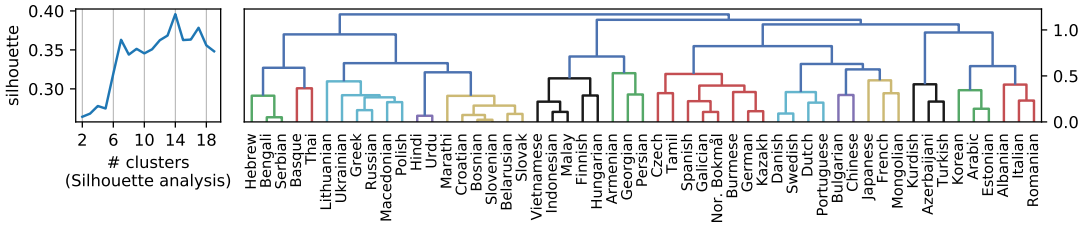


Figure 4.8: Silhouette analysis and dendrograms for clustering the 53 languages of TED-53 using different language representations. In (a) and (b), we note that the silhouette score is below 0.2 (1 is best).

linguistically-informed vectors from lang2vec (Littell et al., 2017) and corpus-based statistics such as word/sub-word overlapping and the ratio of token types or data size between the target child and potential parent candidates. At test time, their model can rapidly predict optimal transfer languages based only on the defined features. One limitation of LANGRANK is that it is only able to rank languages that were included in the training data, which are the languages from the TED corpus (Qi et al., 2018).

Our approach We propose an adapted task of choosing k-related languages for multilingual transfer from a ranked list of candidates. We clarify our proposal for a multilingual approach in Figure 4.9. In our experimental design, k is set to a fixed value. For future work, we encourage the development or testing of automatic selection tech-

niques for this value, such as a silhouette-based heuristic.

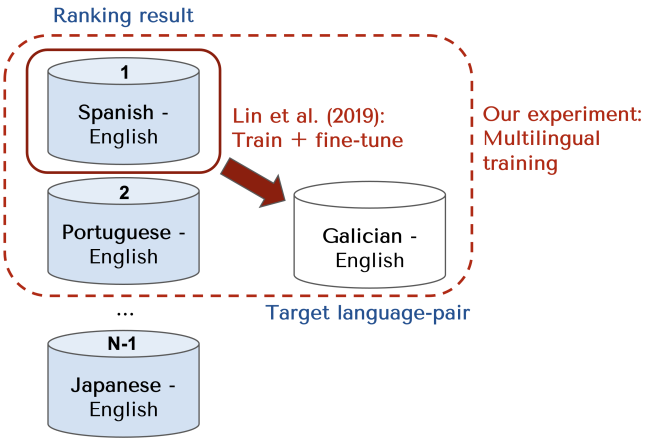


Figure 4.9: From a dataset of N language-pairs, LANGRANK returns a ranked list of $N-1$ language-pairs we should transfer from, given a language-pair as input. Instead of choosing only the top language-pair as in Lin et al. (2019) for training a parent model and fine-tune with the lowest-resource language-pair, we opt to choose k-related languages for multilingual training.

To develop a ranking model for NMT, analogous to LANGRANK, for a dataset of N languages aligned with English, we would need to collect performance scores for $N*(N-1)$ pairwise NMT systems. Additionally, the application of such a model is limited to the original set of N languages. To overcome these limitations, we suggest utilizing our multi-view representations to rank related languages from the vector space using a straightforward similarity metric such as cosine similarity. Our language representations encode typological and lexical relationships, similar to the features considered by Lin et al. (2019). However, we do not need to train a ranking model based on pairwise NMT systems scores, and we can utilize the projection capabilities of SVCCA to generate representations for new languages outside the initial set, as explained in §3.3.1 and further demonstrated in §4.5.4.

4.7.1 Experimental setup

For the ranking experiments, we selected five low-resource languages from the TED-53 dataset: Bosnian (Indo-European/Balto-Slavic), Galician (Indo-European/Italic), Malay (Austronesian), Estonian (Uralic), and Georgian (Kartvelian). These languages have between 5,000 and 13,000 translated sentences with English and were chosen because they showed the most significant improvement from individual to massive

training settings in the previous experiments for clustering (see Table 4.2).

We note again that LANGRANK was trained on TED (Qi et al., 2018), the same dataset we build our multi-view language representations or SVCCA-53. Using their released ranking model, we identified the top-3 related languages (choosing $k = 3$). It should be noted that LANGRANK prefers to select candidates with larger data sizes (Lin et al., 2019). This provides us with a multilingual training set of approximately 500,000 sentences for each case. To ensure a fair comparison regarding data size, an essential factor for transfer learning (Kocmi and Bojar, 2018; Lin et al., 2019), we used SVCCA and cosine similarity to select the n closest languages, which together accounted for a similar amount of parallel sentences.

It is worth noting that our method does not explore the possibility that a single larger dataset in one language may be more beneficial than many smaller datasets. Exploring the possible advantages of combining LANGRANK with our method could be a potential avenue for future work.

4.7.2 Results and discussion

We present the results of our ranking experiments in Table 4.3, which displays the BLEU scores of translations into English for smaller multilingual models that group each language seed with their candidates ranked by LANGRANK and our SVCCA-53 representations. In addition, we include the results of the individual and massive NMT systems from the clustering experiments.

Although the massive system provides a significant improvement over the individual one, we observe that many of the smaller multilingual models outperform the translation accuracy of the massive system. This suggests that the amount of data is not the most crucial factor for supporting multilingual transfer in a low-resource language, which is in line with previous studies (Wang and Neubig, 2019) and with our goal of reducing negative interference in multilingual models.

Regarding the comparison of the two ranking approaches, we observe that SVCCA approximates the performance of LANGRANK in most cases. However, LANGRANK prefers related languages with large datasets, as it only requires three candidates to group around half a million training samples. In contrast, SVCCA suggests including from three to ten languages to achieve a similar amount of parallel sentences. Nevertheless, increasing the number of languages could lead to negative interference in the multilingual model. To mitigate this issue, we could bypass candidate languages that

do not possess a specific amount of training samples, but we leave this analysis for future work.

We argue that our representations provide a robust alternative to determine which languages are suitable for multilingual transfer learning. The significant advantage is that we do not need to pre-train NMT systems from a specific dataset. Moreover, we can extend the coverage of languages easily without re-training the ranking model to consider new language entries (see §3.3.1 and §4.5.4).

Language	Ind.	Mas.	LANGRANK	SVCCA-53
Bosnian	4.2	26.6	28.8 (434)	28.2 [5]
Galician	8.4	24.9	27.7 (443)	28.4 [3]
Malay	4.1	20.1	21.2 (463)	21.0 [4]
Estonian	5.8	13.5	13.5 (533)	12.1 [6]
Georgian	5.8	14.3	13.3 (499)	10.5 [10]

Table 4.3: BLEU scores (L→English) for Individual, Massive and ranking approaches. LANGRANK shows the accumulated training size (in thousands) for the top-3 ($k = 3$) candidates, whereas with SVCCA we approximate the amount of data and include the number of languages (n) between brackets.

4.8 Tool for language representations

After analysing and discussing the potential of multi-view language representations for multilingual NMT, we have also released a command-line interface tool at https://github.com/aoncevay/multiview-langrep, along with our L_T factored embeddings. The tool includes the following functionalities:

- To compute multi-view language representations using the SVCCA method.
- To use other language vectors from lang2vec, such as Phonology or Phonetic Inventory instead of Syntax, as the typological view.
- To upload and use new data-learned language embeddings from any setting or task, such as NMT or language modelling.
- To project language vectors in the shared space given a specific list of languages.

- To generate the dendrogram and perform the silhouette analysis for a given language representation.
- To rank candidate languages for multilingual transfer learning given a language seed.

We believe that this tool could potentially benefit multilingual NLP studies that involve massive datasets of hundreds of languages, such as Named Entity Recognition, Part-of-Speech tagging, and more.

4.9 Conclusion

In conclusion, our study has shown that SVCCA is a valuable tool for multilingual NMT, particularly when aiming to reduce negative interference by constructing clusters of multilingual models. The effectiveness of our SVCCA-based clusters is evident from the stable performance they provide across various resource settings, outperforming or matching some existing and robust baselines. Additionally, the resulting clusters may reveal unique linguistic relationships from typology and NMT-learned embeddings, and help to prevent the isolation of languages without close relatives.

Furthermore, with SVCCA, we can work with unseen languages, making it possible to add new languages to the clusters without the need to retrain the entire model. This feature makes our approach highly flexible, efficient, and adaptable for different use cases.

We have also demonstrated that factored embeddings are more relevant for clustering than using the initial pseudo-token configuration. Finally, we have shown that SVCCA is a strong competitor to LANGRANK when attempting to build a multilingual model given a language seed.

Overall, we consider that our SVCCA-based approach has the potential to improve the multilingual NMT systems by reducing negative transfer, while also uncovering new insights into the relationships between different languages.

Chapter 5

The Impact of Morphological Typology

Continuing the investigation into the impact of linguistic typology knowledge on machine translation, this chapter turns its attention to the role of morphology. We introduce techniques for automatically and semi-automatically computing two key morphological typology indices (fusion and synthesis) and conduct a detailed analysis of these indices at both the word and segment level for translating English-Turkish and English-Spanish. The findings presented in this chapter are based on Oncevay et al. (2022a).

5.1 Introduction

Our next research question is: **are morphological typology variables relevant in the context of machine translation?** Morphology refers to the study of the internal structure of words and how they are formed from smaller units called morphemes. Morphological features are particularly important for machine translation because they can affect the way that words are segmented into subword units, which is an essential step in the translation process (Ataman et al., 2017; Amrhein and Sennrich, 2021). For example, some languages, such as Turkish and Spanish, have complex verb conjugation patterns that can result in a large number of possible word forms, while languages like English and German have fewer inflections. Similarly, some languages like French and Arabic have complex gender systems that require agreement between nouns and adjectives, while others like Chinese and Vietnamese have no gender distinctions at all. These differences in morphological features can pose significant challenges for machine translation systems, which must be able to accurately identify subword units and understand their relationships within the larger context of the sentence.

The complexity of morphology is also reflected in linguistic typology (Comrie, 1989). According to Payne (2017), the morphological typology of a language can be characterised by two phenomena: synthesis and fusion. Synthesis refers to the number of morphemes that make up a single word, while fusion refers to the number of inflections and meanings that can be fused within specific morphemes. These variables can have a significant impact on the subword segmentation process and on machine translation performance in general. Highly synthetic languages, such as Finnish and Turkish, have a relatively large number of morphemes per word, which can make segmentation and decoding longer sequences more challenging (Ataman et al., 2017). In highly fusional languages, such as Russian and Spanish, a large number of inflections can be combined within specific morphemes to convey multiple grammatical meanings. This can be challenging for machine translation because the information is not explicitly provided and needs to be inferred from context (Mager et al., 2018b).

Previous research in NLP that evaluates their studies in typologically-diverse languages has often overlooked the granularity of the indices of fusion and synthesis (see §5.2 for more details). Understanding the dimensions of morphological diversity is crucial in evaluating machine translation models, as it allows us, for instance, to assess whether a model can generate more fusional or synthetic segments for a specific target language. To achieve this, we attempt to quantify a continuous spectrum of morphological diversity at a word and sentence level using the orthogonal variables of synthesis and fusion. By characterising a language along these two dimensions, we could gain a deeper understanding of its morphology, and quantifying the complexity of morphological typology based on synthesis and fusion is the first step towards proposing a solution.

The main contributions of this chapter are as follows:

- To the best of our knowledge, we present the first computational quantification of the morphological typology indices of synthesis and fusion for an NLP task (see §5.3 and §5.4, respectively). For synthesis, we propose baselines to approximate the computation of the morphological index using unsupervised and supervised segmentation methods, and evaluate them in English and German using annotated data in morphological segmentation. For fusion, we perform a semi-automatic annotation of the different Spanish verb forms.

- In §5.5.2 and §5.5.3, we analyze the relationship between these two indices and machine translation quality at the word-level, and observe that a higher degree

of synthesis or fusion usually corresponds to less accurate translations in specific word types (we study nouns and verbs in English-Turkish and verbs in English-Spanish).

- We complement the word-level evaluation with a manual annotation of synthesis and fusion (see §5.5.4).

- In §5.6, we extend our analysis to the segment-level for English-Turkish and English-Spanish in both directions, and identify consistent results.

5.2 Related work

Most studies in NLP that address morphological typology are related to either the development of morphological analysis systems or the evaluation of typologically diverse languages in terms of morphology (e.g. Vania and Lopez (2017); Xu et al. (2020)). However, the typology used to distinguish languages varies across different studies. For instance, Vania and Lopez (2017) considers four phenomena to label languages: fusionality, agglutination, reduplication and root-pattern; whereas Xu et al. (2020) considers more fine-grained elements such as affixation (prefixation, infixation and suffixation) or partial reduplication. Similarly, a fine-grained analysis on non-concatenative morphology for NMT was performed by Amrhein and Sennrich (2021). It is important to note that none of the previous studies has addressed the morphological phenomena of synthesis and fusion as a quantifiable variable. We consider that working with these two indices is a pragmatic approach for quantifying complexity in morphological typology, as there are insights and guidelines grounded in linguistic knowledge (Payne, 2017) to perform this analysis.

Besides, a survey by Ponti et al. (2019), on computational typology for NLP, pointed out that morphological knowledge is potentially helpful for analysing the difficulty in generation tasks such as language modelling and NMT for both unsupervised and supervised settings. More specifically, they suggested that the degree of fusion (related to the index of fusion proposed by Payne (2017)) can impact the rate of less frequent words, which is a relevant parameter for generation tasks.

5.3 Synthesis: automatic computation

In this section, we present our methodology for automatically computing the degree of synthesis in a given language. To achieve our goal, we must perform a reliable morphological segmentation. While a rule-based morphological analyser and disambiguator is a suitable option when available (as we use later for Turkish in §5.5.2), we compare morphologically-supervised and unsupervised segmentation methods in this section. The former relies on texts with morpheme boundary annotations, while the latter is trained on raw text. The methods considered are:

- Byte-Pair-Encoding (BPE) (Sennrich et al., 2016b): an unsupervised method based on iteratively merging the most frequent pairs of consecutive bytes in the input text until a predefined vocabulary size is reached. We use the implementation in SentencePiece (Kudo and Richardson, 2018).

- Unigram Language Model (uniLM) (Kudo, 2018): an unsupervised method that maximises the likelihood of the input text under a unigram language model with respect to subword units. We also use the implementation in SentencePiece (Kudo and Richardson, 2018).

- Morfessor (Creutz and Lagus, 2002): a family of unsupervised segmentation methods that group character sequences together based on their co-occurrence frequencies, while considering their internal structures.

- Pointer Generator Network (PtrNet): a morphologically supervised segmentation method that leverages annotated data to predict morphological boundaries. The method trains a neural network to predict a pointer to the position of the next boundary in a sequence, given the previous context. We use the implementation of Mager et al. (2020), which extends the original model (See et al., 2017) to improve its performance on morphologically-rich languages. The training data is described in §5.3.1.

We analysed several vocabulary sizes (4k, 8k, 16k, 32k, 64k) for BPE and uniLM but reported only the best one, which is 64k for all cases. These segmentation methods were chosen based on their popularity and performance in previous studies on subword segmentation for machine translation, as well as their availability and ease of use. Additionally, we specifically selected Morfessor and Pointer Generator Network for their wide usage in processing morphologically rich languages (Mager et al., 2020).

5.3.1 Datasets and evaluation

To evaluate and compare the performance of the segmentation methods, we have chosen two widely studied languages in machine translation: English and German. These languages have annotated data available, which is essential for training the supervised segmentation methods. In this case, we use the CELEX dataset of segmented words for English and German (Steiner, 2016, 2017). Some examples for English are shown in Table 5.1. To ensure the reliability of the results, we randomly split the dataset into training, development, and test sets in an 80-10-10 ratio. The supervised method was trained on the training set and fine-tuned on the development set to achieve the best performance.

As for the unsupervised methods, we used two corpora: the news-commentary-v15 (Barrault et al., 2019) and EuroParl-v10 (Koehn, 2005). These corpora were chosen because they are commonly used for training and evaluating NMT models. The news-commentary-v15 corpus contains news articles in various languages, while the EuroParl-v10 corpus consists of speeches delivered in the European Parliament.

Furthermore, we propose two metrics to evaluate the performance of the segmentation methods in computing synthesis:

- Morpheme count accuracy: This metric evaluates whether the number of morphemes obtained by the segmentation method matches the number of morphemes in the reference segmentation. This is a strict measurement as synthesis is the number of morphemes per word.

- Exact morpheme segmentation precision: This metric measures whether the morphemes obtained by the segmentation method match those in the reference segmentation. To compute this, we first perform an automatic alignment between the hypothesis and reference segments using the parallel Needleman-Wunsch algorithm for sequences (Naveed et al., 2005), and then calculate the exact match at the morpheme level.

We chose these metrics as they directly relate to the synthesis variable, which is the ratio of morphemes per word. The accuracy count assesses whether the obtained number of morphemes matches the reference, while the exact segmentation precision measures the similarity between split morphemes in the hypothesis and reference segmentations. It is worth noting that the last metric is more suitable for evaluating morphological segmentation rather than computing synthesis, but we included it as a complementary score.

Word	Annotation	PtrNet output
cookie	cookie	cook**m**
polysyllabically	polysyllabic+ally	polysyllabic+ally
polygamy	polygamy	poly**g**gmy
rounders	round+er+s	roun**ders**
sabre-toothed	sabre+tooth+ed	sabre+toothe**d**

Table 5.1: Examples of English entries in CELEX, including the annotation of their segmentation and the output of the best PtrNet model reported in Table 5.2. Morpheme boundaries are marked with "+", and errors generated by the model are in bold.

5.3.2 Results and discussion

Table 5.2 displays the morphological segmentation scores for English and German. Our observation is that both BPE and uniLM underperform when the word is not expected to be segmented (column "1"). This pattern is in line with the findings of Bostrom and Durrett (2020) who noted that unsupervised segmentation methods tend to over-split the roots of words. However, both BPE and uniLM improve their accu-

	English				German			
#morphs.	1	2	3	4	1	2	3	4
#entries	16,914	28,900	1,798	73	13,061	32,007	5,808	360
	Morpheme count accuracy							
uniLM_{64k}	**0.54**	0.52	0.49	**0.59**	**0.35**	0.27	0.21	0.18
BPE_{64k}	0.5	**0.53**	0.5	0.52	0.29	**0.33**	**0.28**	**0.26**
Morfessor	0.22	0.47	**0.55**	0.48	0.17	0.26	**0.28**	0.25
PtrNet	**0.82**	**0.84**	**0.56**	**0.81**	**0.74**	**0.86**	**0.7**	**0.42**
	Exact morpheme segmentation precision							
uniLM_{64k}	**0.54**	0.52	0.6	**0.8**	**0.29**	0.38	0.32	0.22
BPE_{64k}	0.5	0.44	0.56	0.76	0.24	0.33	0.23	0.08
Morfessor	0.21	**0.58**	**0.7**	0.78	0.17	**0.45**	**0.44**	**0.36**
PtrNet	**0.76**	**0.67**	**0.81**	**0.8**	**0.67**	**0.73**	**0.72**	**0.62**

Table 5.2: Accuracy count and segmentation precision for English and German using unsupervised and supervised segmentation methods. Results are grouped by the expected number of morphemes (e.g. "1" means that the word should not be split).

racy and precision when the expected number of morphemes is larger. Interestingly, Morfessor also underperforms in the "1" case for both languages and only outperforms the other unsupervised methods when we measure precision for many morphemes. On the other hand, the PtrNet supervised method consistently outperforms the rest in almost all scenarios, which is an expected result. Nevertheless, it's important to note that methods with high generative power, such as PtrNet, may introduce issues, as illustrated by the examples in Table 5.1, where potential out-of-vocabulary tokens are generated.

Our conclusion is that, to compute synthesis, we should prioritise, besides a rule-based morphological analyser, a supervised segmentation method like PtrNet, particularly when data is available[1]. For the word-level analysis in §5.5, we will focus on Turkish, a language with a rule-based morphological analyser.

5.4 Fusion: Semi-automatic computation

Calculating fusion should be approached in a case-by-case scenario, as there are different considerations provided by Payne (2017), such as the presence of prefixes, infixes, suffixes, ambifixes, non-concatenative processes (reduplication, apophony, and subtractive morphology), and more. Therefore, there is no automatic tool that can obtain the fusion score directly. Even more challenging is the development of a tool that can compute fusion scores in a language-independent manner, making it necessary to develop specific methods for each language.

We then chose to focus on Spanish as a case study for computing fusion, as it is a highly fusional language where verbs and auxiliary verbs are known to contain the highest degree of fusion among all parts-of-speech (POS). For instance, the Spanish verb "había" (meaning "had" in English) is composed of the auxiliary "ha" (third person singular present tense of "tener" meaning "to have") and the imperfect past tense marker "-ía". Additionally, the availability of annotators and machine translation training and evaluation data for Spanish made it a practical choice for our study.

[1] Ideally we would evaluate a morphological analyser for English and German too, but it is worth noting that the most widespread morphological analysers for these languages (e.g. spaCy, which we use later for Spanish) outputs the morphological feature tags but does not perform a morphological segmentation.

5.4.1 Procedure

To begin our study on computing fusion for Spanish verbs, we first constructed a corpus with annotations of fusion. To ensure a balanced representation of verbs from different paradigms, we performed an annotation per paradigm and based on the ending of the verb (-ar, -er, -ir). This approach allowed us to assess the degree of fusion for each verb form regardless of the lemma, with the exception of irregular verbs, which can present limitations and potential noise. To reduce the risk of biases assessments, we conducted a human evaluation in further machine translation experiments, in §5.5.4, for irregular and regular verbs altogether.

Next, given any text collection in Spanish[2], we take the following steps:

1. First, we performed automatic annotation of POS and morphological features using the spaCy model `es_dep_news_trf`[3] which has a high accuracy rate of 0.99 for POS and morphological tagging in the UD Spanish AnCora dataset (Taulé et al., 2008). This model was trained on news texts and has shown to work well in a variety of Spanish language processing tasks.
2. Second, we manually reviewed the automatic annotation to correct any special cases, such as specific verb forms that were mislabeled as adjectives.
3. Finally, we obtained a set of all unique verb paradigms and morphological features in the corpus, taking into account the three different types of verb endings in Spanish as separate elements. Using the Unimorph database (McCarthy et al., 2020) is another alternative for extracting all the possible unique inflections. We aligned and considered both tag sets for the annotation, as shown in Table 5.3.

After obtaining a list of unique verb paradigms and endings, we proceeded to annotate the fusion index. We followed these steps:

1. For each unique verb paradigm and ending, we segmented a verb sample into its morphemes[4]. For example, the verb *habló* ('talked') is split into *habl-ó*, and *habláramos* ('we were to speak') into*habl-ára-mos*.
2. Then, we analysed how many morphological features are fused in each morpheme. We asked ourselves if changing the value of a feature would change

[2]As noted later in §5.5.3, for our experiments, we use the NEWSTEST2013.EN-ES evaluation set from WMT13 (Bojar et al., 2013).

[3]Available at https://spacy.io/models/es#es_dep_news_trf

[4]This annotation is useful for analysing synthesis as well, but synthesis is not the goal for this part of the study.

the surface form or the morpheme. For instance, in *habl-ó*, *-ó* is involved in 5 features (mode (indicative), subject person (third person), subject number (singular), tense (past) and aspect (perfective)). For *habl-ára-mos*, *-ára* includes the past and subjunctive, whereas *-mos* denotes the person and number. If any of aforementioned features changes its value, the surface form will also change.

3. Next, we counted and aggregated the results per morpheme and obtained the fusion for each verb form. For instance, the fusion for *habl-ó* is 4/5 = 0.8, and for *habl-ára-mos* is 2/4 = 0.5.

Finally, with the unique list of verb inflections and endings annotated, we can now extend the degree of fusion to all the verbs in the original Spanish corpus. This enables us to quantify the degree of fusion for each verb form in the corpus. To illustrate the annotation process, an example is shown in Table 5.3.

5.5 Word-level analysis in machine translation

In this section, we delve into the difficulty of machine translation based on the degree of these phenomena. Specifically, we ask the following question: **how challenging is it to translate a word based on its index of synthesis or fusion?**

For the variable of fusion, we focus on Spanish verbs as in the previous section. However, for the case of synthesis, we prefer to work with Turkish, a language that has an available morphological analyser (Sak et al., 2008), and more importantly, presents a high synthesis and agglutination (or low fusion) (Zingler, 2018), which means that words are composed of several morphemes, and the morpheme boundaries are explicit, respectively. Then, we focus on Turkish nouns and verbs as they typically contain more morphemes than other parts-of-speech.

In both cases, for the machine translation analysis, the source language is English, and Turkish and Spanish are the target languages. The reasons to choose English as the paired language are the availability of parallel corpora and the low fusion and low synthesis degree of English. Studying the cases where the source and target language exhibit further combinations of synthesis and fusion is left to future work.

In the upcoming sections, we outline the general experimental design for both synthesis and fusion variables (see §5.5.1). Next, we delve into the specifics of the experimental setup and outcomes for synthesis in the context of machine translation for English-Turkish (see §5.5.2) and fusion for translating English-Spanish (see §5.5.3).

Example (es): **Hablaremos** de la propuesta con la que se **condenó** a la ex primer ministra y fue **apoyada** por 147 diputados en la votación.

Example (en): **We will talk** about the proposal which **condemned** the ex prime minister and was **supported** for 147 congressmen in the vote.

Verbs	Features (spaCy)	Features (UniMorph)	Segmentation	feats. per morph	fusional morph. joints	total joints	Fusion index
hablaremos *(we will talk)*	Mood=Ind, Number=Plur, Person=1, Tense=Fut, VerbForm=Fin	V;IND;FUT;1;PL	habl - are - mos	0 – $2_{(IND;FUT)}$ – $2_{(1;PL)}$	0+(2-1)+(2-1) = 2	2+2 = 4	0.5
condenó *(condemned)*	Mood=Ind, Number=Sing, Person=3, Tense=Past, VerbForm=Fin	V;IND;PST;3;SG;PFV	conden - ó	0 - $5_{(IND;PST;3;SG;PFV)}$	0+(5-1) = 4	4+1 = 5	0.8
apoyada *(supported)*	Gender=Fem, Number=Sing, VerbForm=Part	V.PTCP;PST;FEM;SG	apoy - ada	0 - $3_{(PST;FEM;SG)}$	0+(3-1) = 2	2+1 = 3	0.66

Table 5.3: Annotation example in Spanish. We first identify the verbs (in bold) and obtain their morphological features (using spaCy and the UniMorph schema). Then, we split the verb into its morphemes (segmentation), and identify which features are fused in each morpheme (feats. per morph). Finally, we compute the index of fusion by dividing the fusional morpheme joints by the total joints (which includes the agglutinative or explicit boundaries). On a side note, examples of verbs with zero fusion are in the infinitive (e.g. hablar *(to talk)*) and gerund (e.g. hablando *(talking)*) forms.

Finally, we present supplementary results from a human evaluation in §5.5.4.

5.5.1 Experimental design

To conduct the experiment, we compare a gold standard reference with machine translation system outputs at the word level, following these steps:

1. First, we automatically **tag all the words using a morphological analyser** (the Boun morphological analyser and disambiguator (Sak et al., 2008) for Turkish and a spaCy model trained on the Ancora Universal Dependency parser (Taulé et al., 2008) for Spanish) in both the reference and system output. The part-of-speech annotations are needed to identify the target words: verbs in Spanish and verbs and nouns in Turkish. For synthesis in Turkish, the number of morphemes works as a proxy as we work at the word level, while for fusion in Spanish, we require the inflexion to determine the degree of fusion from the annotated unique list or verb forms (see §5.4).
2. Next, we **align the words** between the reference and system output. For this purpose, we use the word alignment model in the awesome-align tool (Dou and Neubig, 2021), which is based on the multilingual BERT model (Devlin et al., 2019). To fine-tune the model for our purposes, we train it on a parallel corpus consisting of both the reference and the output generated by the machine translation system.
3. Finally, we **calculate the translation accuracy** for the target part of speech. We consider an exact match metric (0 or 1) for a strict evaluation.

After obtaining the translation results at the word level, we categorise the scores based on the various degrees of synthesis (by the different number of morphemes) and fusion (when the fusion is equal to or greater than 0). In addition, we take into account different confounding factors that may affect the translation performance, such as the frequency of the word in the training set and whether the full word is part of the vocabulary input of the model or not. This allows us to provide a more comprehensive analysis of the translation performance and to identify potential sources of error.

5.5.2 Synthesis analysis: English-Turkish

In this part, we introduce further details about the experimental setup for analysing the case of synthesis in English-Turkish. Afterwards, we discuss the results.

	Total	#1	#2	#3	#4	#5+
Verbs	3,834	133	2,265	1,036	308	92
Nouns	10,680	5,899	2,974	1,556	244	7

Table 5.4: Number of nouns and verbs in the Turkish reference set, and their respective number of morphemes.

Data We use the NEWSTEST2018.EN-TR evaluation set from WMT18 (Bojar et al., 2018), with 3,000 samples. On the Turkish side, there are 45,944 tokens and Table 5.4 shows the distribution of the number of morphemes obtained with the morphological analyser of Sak et al. (2008).

Model We use an English-Turkish system trained with the TIL corpus of 39.9M parallel sentences (Mirzakhalov et al., 2021). On the NEWSTEST2018.EN-TR set, the performance is 13.06 and 49.54 in BLEU (Papineni et al., 2002) and chrF (Popović, 2015), respectively.

5.5.2.1 Results and discussion

Figure 5.1 shows the average accuracy (exact translation, 0 or 1) of nouns (top) and verbs (bottom) in NEWSTEST2018.EN-TR, where the number of morphemes in the horizontal axis is a proxy for the index of synthesis (number of morphemes per word).

We observed that the average accuracy usually decreases as the number of morphemes increases from 1 to more, especially for high-frequency words (rightmost subplots). This pattern is more apparent in nouns than in verbs, as the latter have fewer cases to analyse overall. The differences between 2, 3, or more than 4 morphemes are not significant in most cases and are sometimes inconsistent, such as in verbs with the highest frequency. Nevertheless, we can argue that isolating nouns ($synthesis = 1$) are easier to decode than synthetic nouns ($synthesis > 1$) for the English→Turkish language pair.

The observed pattern holds for whether the word is part of the vocabulary of the model or not. Additionally, rare words (leftmost subplots) have generally lower translation accuracy than more frequent words (middle and rightmost subplots), which is an expected outcome. The only exception are Nouns with one morpheme, which have a relatively high translation accuracy. A potential reason is that most of the one-morpheme Nouns are named entities that do not require a translation from the English

side. Overall, the accuracy gap across words with different number of morphemes is reduced in the leftmost subplots. However, the main reason is that the NMT model does not have sufficient examples to learn from, regardless of the synthesis value.

5.5.3 Fusion analysis: English-Spanish

Next, we introduce more details about the experimental setup for analysing the case of fusion in English-Spanish, followed by a discussion of the results.

Data We use the NEWSTEST2013.EN-ES evaluation set from WMT13 (Bojar et al., 2013) with 3,000 samples. On the Spanish side, there are 62,055 tokens, with 6,317 verbs, and where 1,411 of them are more agglutinative ($fusion = 0$) and 4,822 more fusional ($fusion > 0$).

Model For training the model, we use the MarianNMT toolkit (Junczys-Dowmunt et al., 2018), a Transfomer base architecture (Vaswani et al., 2017) with default parameters, and four NVIDIA V100 GPUs. We obtained different English-Spanish models using the newscommentary-v8 (Bojar et al., 2013) and EuroParl (Koehn, 2005) datasets with joint vocabulary sizes of 8k, 16k and 32k, which were tokenized using the unigram-LM algorithm (Kudo, 2018) implemented in SentencePiece (Kudo and Richardson, 2018). After experimenting with different models, we selected the best-performing system: a combination of both datasets (2.2M sentences) with 16k pieces. On the NEWSTEST2013.EN-ES evaluation set, this system achieved a performance of 31.6 BLEU points.

5.5.3.1 Results and discussion

Figure 5.2 shows the average translation accuracy of Spanish verbs in NEWSTEST2013.EN-ES for verbs without and with some degree of fusion.

In the middle and rightmost subplots, where the frequency of the verbs is higher, the average accuracy of non-fusional verbs is consistently higher than the fusional ones. This trend holds regardless of whether the verbs are part of the model's input vocabulary or not.

However, in the leftmost subplot, where the verbs are the least frequent, the model's translation accuracy is lower overall. In this case, the effect of fusion is less clear,

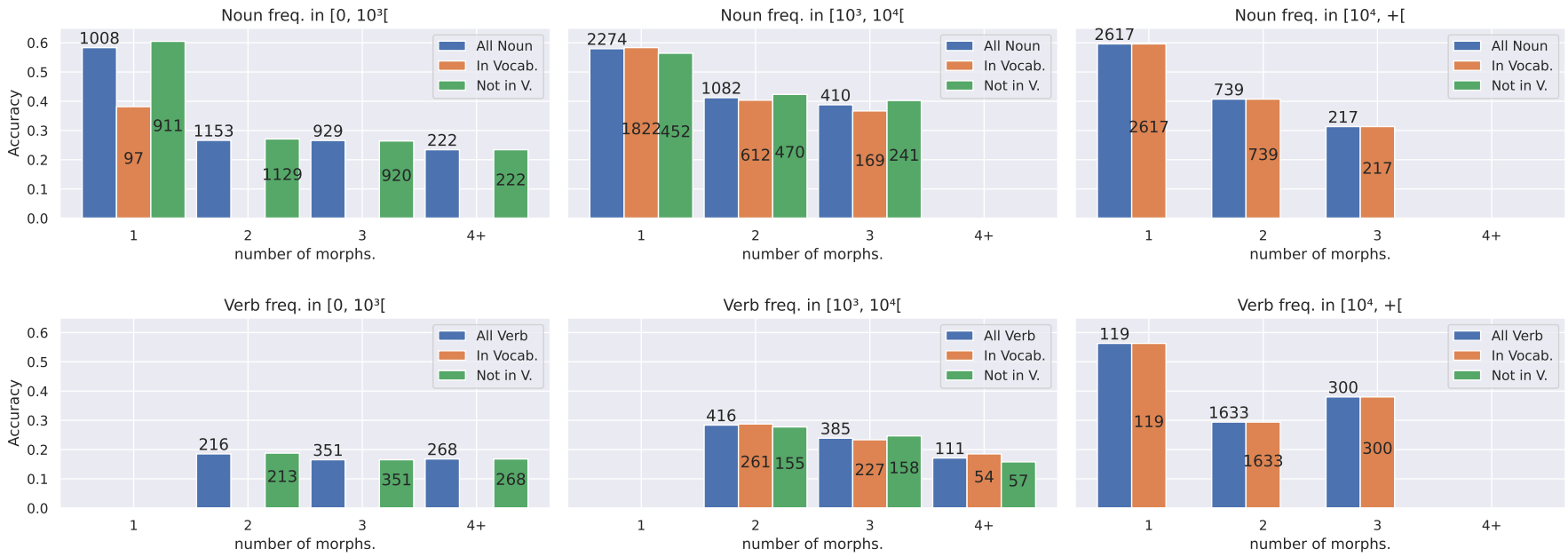


Figure 5.1: Accuracy (exact translation) for Nouns (top) and Verbs (bottom) in the English→Turkish translations. Results are grouped by the training frequency of the words (less to more frequent from left to right), and each subplot presents the scores for all the words, and whether they belong or not to the vocabulary input of the model. The number of samples are stacked in each bar, and we do not show entries with less than 30 samples.

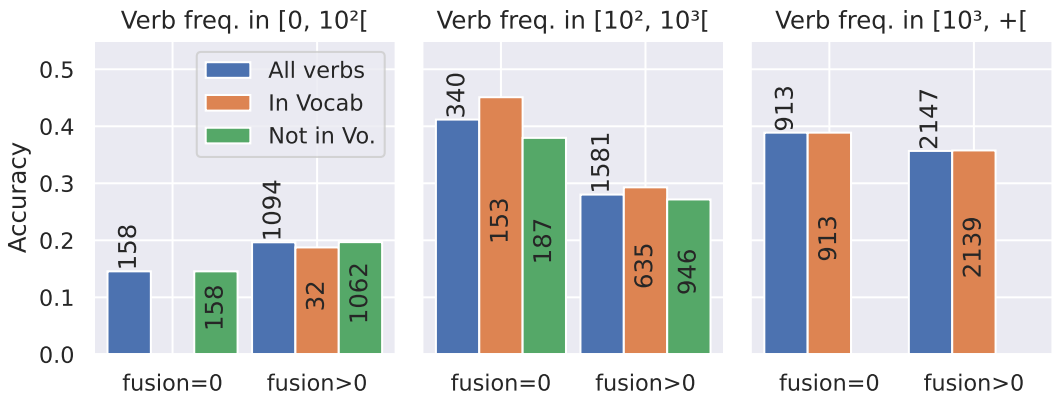


Figure 5.2: Accuracy (exact translation) for Verbs in the English→Spanish translations. Results are grouped by the training frequency and whether the word belongs to the vocabulary of the model (In V) or not (Not in V).

which suggests that the model does not have enough information to learn from, regardless of the degree of fusion.

5.5.4 Human evaluation

Exact translation accuracy has its limitations since there can be acceptable translations depending on the context, such as synonyms. Therefore, to complement the automated evaluation, we conducted a human evaluation of a sample of sentences from 10% of each evaluation set, focusing on two scores:

1. **Semantic score**: evaluates the meaning of the word used in the automatic translation (system output) and how it compares with the gold standard translation. Scale goes from 1 (no relationship at all) to 4 (it is the same lemma).

2. **Grammar score**: evaluates the grammatical form and how it compares with the gold standard translation. Scale goes from 1 (different inflection) to 3 (same inflection).

Afterwards, in §5.5.4.1, we performed a correlation analysis between the automatic translation accuracy (0 or 1) and the Semantic and Grammar scales. The annotation protocol and information about the annotators is shown in Table 5.5.

Table 5.5: Annotation protocol for the Semantic and Grammar scores in the analysis of synthesis and fusion at word-level, and information about the annotators.

Annotation Protocol

This study measures the translation quality of translations generated by a translation system. You are given a list of sentences where one column lists each word in the gold standard (correct) translation and the corresponding column the system-generated translations. The evaluation of the translations will rely on the two scores described below. The scores to use in the evaluation are:

Semantic score: Please assign each word in the output one of the scores you find most appropriate:

1. There is no relationship between the two lemmas
2. The lemmas are different but the translation does not fit well in the context
3. The lemmas are different but it is still an acceptable translation (e.g. synonym)
4. It is the same lemma

Grammar score: Please assign each word in the output one of the scores you find most appropriate:

1. The word is inflected in a different way and it is not necessarily correct
2. The word has different inflection but it is still grammatically correct
3. The words have the same inflection, and it is correct

Please annotate all words in the translations in the file shared with you. In your evaluation try assigning the two scores to each word independently. The inflection of the word measures the morphological feature and should also be evaluated independently from the analyzer output which is automated and may contain errors.

The file contains example annotations for your reference, please ask any questions related to unresolved annotation examples by contacting the project coordinators.

About the annotators

For both Turkish and Spanish, the annotators were contacted directly due to their expertise in morphology (both of them are PhD students in Linguistics and Computational Linguistics, respectively), besides requiring that they are native speakers of the target languages. Also, they were paid more than the minimum wage per hour of annotation of their country of residence, and were told that the annotated data will be released upon acceptance of the study.

5.5.4.1 Correlation between accuracy and human annotations

Given that translation accuracy (0-1) and either Semantic (1-4) or Grammar (1-3) scores are categorical variables with a ranking, we investigate their correlation using Spearman's rank correlation coefficient. This method compares the ranking of the two variables and returns a coefficient ranging from -1 to 1.

More specifically, we perform the correlation while categorising the results by part-of-speech (nouns or verbs) and the number of morphemes for synthesis in Turkish (see §5.5.4.2), as well as the fusion of verbs in Spanish (see §5.5.4.3).

5.5.4.2 Results for synthesis

Table 5.6 shows the Spearman correlation coefficients between translation accuracy and manually annotated semantic or grammar scores for different numbers of morphemes in Turkish nouns and verbs.

The results show that there is a significant and strong positive correlation between translation accuracy and semantic score for Turkish nouns and verbs. Similarly, in terms of grammar scores, there is a significant and moderate to strong correlation between all part-of-speech and number of morphemes. These findings suggest that the human annotations support the automatic analysis.

	Semantic score		**Grammar score**	
#morphs	**Nouns**	**Verbs**	**Nouns**	**Verbs**
1	0.7761	0.7226	0.4997	0.5911
2	0.5875	0.7069	0.4566	0.4148
3	0.5539	0.5666	0.4457	0.4188
4+	0.5882	0.5839	0.5380	0.4060

Table 5.6: Spearman correlation coefficients between translation accuracy and annotated semantic or grammar scores for the index of synthesis in Turkish nouns and verbs. All p-values are lower than 0.05.

Additionally, we calculated the proportion of false positives, or the cases where the annotation for either Semantic or Grammar is the highest (4 or 3, respectively), but the automatic translation accuracy is the lowest (0), for different numbers of morphemes. The results, presented in Table 5.7, suggest that the machine translation system faces increasing difficulty as the number of morphemes increases, in generating a noun or

#morphs	Semantic=4	Grammar=3
1	0.1253	0.2256
2	0.3390	0.4739
3	0.4370	0.5068
4+	0.4375	0.5714

Table 5.7: Proportion of entries with 0 accuracy and maximum Semantic (4) or Grammar (3) score, grouped by number of morphemes.

verb with the same meaning but not the same grammatical structure (as measured by the Semantic score), or with the same grammatical form but a different lemma (as measured by the Grammar score).

5.5.4.3 Results for fusion

Next, Table 5.8 shows the Spearman correlation coefficients between translation accuracy and manually annotated semantic or grammar scores for the index of fusion (0 or more) in Spanish verbs. The results show that there is a significant and strong positive correlation for both semantic and grammar scores, with stronger correlations observed for the former one.

	Semantic score	Grammar score
fusion$=0$	0.7871	0.5833
fusion>0	0.7156	0.6118

Table 5.8: Spearman correlation coefficients between translation accuracy and annotated semantic or grammar scores for the index of fusion in Spanish verbs. All p-values are lower than 0.05.

Finally, in Table 5.9, we analyse the ratio of false positives for the translation accuracy (0) with respect to the highest semantic (4) and grammar scores (3). For the semantic metric, we observe that a higher degree of fusion leads to a higher proportion of false positives, or a higher ratio of translated verbs with the same lemma but not the proper verb. On the other hand, for the grammar metric, we observe that the ratio of false positives is similar across the two levels of fusion. This suggests that the issue of the model generating the correct verb form but could not associate it with the correct lemma is not affected by the degree of fusion.

	Semantic=4	**Grammar=3**
fusion$=0$	0.1714	0.3958
fusion>0	0.2664	0.3907

Table 5.9: Proportion of entries with 0 accuracy and maximum Semantic (4) or Grammar (3) score, grouped by the degree of fusion.

5.5.5 Overall conclusion

In summary, our analysis of synthesis and fusion in word-level machine translation for English-Turkish and English-Spanish revealed a relationship between the degree of synthesis and fusion and translation difficulty, as confirmed by manual evaluation. We observed a strong to moderate correlation between the automatic metric and the manual scores, indicating the need for further investigation to fully understand the impact of synthesis and fusion on translation difficulty. In the following section, we will examine this relationship at the segment level.

5.6 Segment-level analysis in machine translation

We now turn to a segment-level analysis to further explore the relationship between machine translation difficulty and the degree of synthesis or fusion. For this purpose, we process a set of NMT systems for the language pair we want to evaluate. The general steps are as follows:

1. For each system output, we compute **automatic evaluation metrics**, such as chrF (Popović, 2015) and COMET (Rei et al., 2020)), with respect to the reference set on a per-sentence basis. These two metrics complement each other in our analysis: chrF emphasises character-level overlapping, while COMET prioritises semantic correctness and fluency.

2. For each sentence in the evaluation set, we compute a set of **predictor variables** that may affect the performance of the automatic evaluation metrics, such as the degree of synthesis or fusion. We also include the length of the sentence in characters (*char.count*), as longer texts are generally associated with increased translation difficulty.[5]

[5]Initially, we considered other potential predictors, including *word.count*, but these were omitted due to their collinearity with *char.count*. Additionally, there are other potential predictors like the presence

3. We calculate the Variance Inflation Factor (VIF) for each predictor variable. VIF measures how much the variance of an estimated regression coefficient increases if the predictors are correlated.[6] With VIF, we confirm no colinearity between the chosen predictors.

4. We normalise the predictor variables to reduce the impact of their magnitudes.

5. We construct a set of **generalised linear models**[7] for each system output and evaluation metric, where the predictors are used to explain the performance of the automatic metric. The goal is to identify which predictors have a significant effect on the metric's performance. For both evaluation metrics, we assume a Gaussian distribution.

6. After model creation, we extract the significant predictors and their coefficients (with p-value$<$0.05), providing an indication of which variables are most relevant for predicting the metric's performance. The coefficient values indicate the strength and direction of the relationship between each predictor variable and the target variable.

5.6.1 Machine translation models

In the segment-level analysis for English-Turkish and English-Spanish (in both directions), we use various NMT models that were either trained in our previous experiments or provided by other studies. Specifically, we consider models trained with different corpus sizes to analyse different performance levels, and they are as follows:

- EnTr1: A Transformer large model trained on the TIL corpus of 39.9M parallel sentences, which was also used in §5.5.2 for the word-level analysis of synthesis for verbs and nouns in English-Turkish (Mirzakhalov et al., 2021)[8].

- EnTr2: A smaller version of EnTr1 trained with a sample (10%) of the TIL corpus. This model uses a Transformer base architecture (Vaswani et al., 2017)

of infrequent words in the sentence, which could pose translation challenges. However, for the sake of simplicity, we focus on *char.count* and leave other predictors for future analysis.

[6]Generally, a VIF greater than 5-10 indicates a problem with collinearity.

[7]A generalised linear model is a statistical model that extends the ordinary linear regression model to accommodate different types of response variables, or in this case, our predictors. It specifies a linear relationship between the independent variables and a function of the response variable, which in our case are the automatic evaluation metrics for machine translation outputs (Nelder and Wedderburn, 1972).

[8]There is no TrEn1 model in our analysis, as the previous study did not provide a similar model in the opposite translation direction.

with a joint vocabulary size of 8k pieces obtained through uni-LM from SentencePiece (Kudo and Richardson, 2018).

- TrEn2: An NMT system similar to EnTr2, but in the opposite translation direction.
- EnEs1: An NMT system used in §5.5.3 for the word-level analysis of fusion for verbs in English-Spanish. It is a Transformer base (Vaswani et al., 2017) model trained on 2.2M parallel sentences with a joint vocabulary of 16k pieces using uni-LM from SentencePiece (Kudo and Richardson, 2018).
- EsEn1: An NMT system similar to EnEs1, but in the opposite direction.
- EnEs2: Same configuration as EnEs1 (model and vocabulary), but trained on a smaller dataset of only newscommentary-v8 data, consisting of around 300k sentences.
- EsEn2: An NMT system similar to EnEs2, but in the opposite translation direction.

5.6.2 Synthesis on English-Turkish and Turkish-English

We first evaluate the English-Turkish (EnTr1, EnTr2) and Turkish-English (TrEn2) models. Also, as we are studying synthesis in Turkish, both predictors (synthesis and *char.count*) are computed on the Turkish side, regardless of the translation direction. Table 5.10 shows the coefficient estimates for predicting the performance of different English-Turkish translation models in terms of chrF and COMET.

metric	model	**synthesis**	**char.count**
chrF	EnTr1	-1.09 ±0.27	
	EnTr2	-1.1 ±0.27	
	TrEn2	-1.48 ±0.28	0.63 ±0.28
COMET	EnTr1	0.01 ±0.01	-0.20 ±0.01
	EnTr2	-0.03 ±0.01	-0.13 ±0.01
	TrEn2	-0.03 ±0.01	-0.03 ±0.01

Table 5.10: GLM coefficient estimates with confidence intervals for significant predictors in English-Turkish. We only show results with p-value$<$0.05.

We observe that, for chrF, higher synthesis corresponds to lower scores, indicating that increased morphological complexity in terms of morphemes per word tends to reduce character-level F-scores. However, we note that this relationship might not be linear and may involve other unaccounted factors in our analysis. In contrast, COMET scores are affected by synthesis, but the effect size is relatively small. This suggests that COMET, emphasising semantic similarity, is less sensitive to variations in synthesis.

Furthermore, the positive correspondence between chrF and *char.count* in TrEn2 could imply that longer Turkish sentences in the source side provide more context, which helps in character-level matching. However, the negative correspondence between *char.count* and COMET in TrEn2 indicates that, despite achieving good character-level overlap (as reflected by chrF), the translations might not effectively capture the intended meaning or fluency. COMET, being a more holistic metric, accounts for factors beyond character-level similarity and is sensitive to semantic correctness and fluency.

5.6.3 Fusion on English-Spanish and Spanish-English

We also investigated the impact of fusion in English-Spanish (EnEs1, EnEs2) and Spanish-English (EsEn1, EsEn2) models. As our focus was on Spanish, both predictors (fusion and *char.count*) were computed on the Spanish side, irrespective of the translation direction. In Table 5.11, we present the coefficient estimates for the predictors in the context of English-Spanish translation models, analysed through the chrF and COMET metrics.

metric	model	**fusion**	**char.count**
chrF	EnEs1	-1.18 ±0.37	2.45 ±0.37
	EnEs2	-1.49 ±0.35	2.29 ±0.35
	EsEn1	-1.33 ±0.37	2.82 ±0.37
	EsEn2	-1.27 ±0.35	2.89 ±0.35
COMET	EnEs1	-0.02 ±0.01	
	EnEs2	-0.04 ±0.01	-0.05 ±0.01
	EsEn1	-0.03 ±0.01	
	EsEn2	-0.05 ±0.01	

Table 5.11: GLM coefficient estimates with confidence intervals for significant predictors in English-Spanish. We only show results with p-value$<$0.05.

Notably, the results reveal that fusion has a significant impact and a negative correspondence with both the chrF and COMET metrics. However, the effect size on chrF is notably larger than on COMET. This observation highlights the importance of fusion, which measures the amount of grammatical information fused within a single morpheme, in character-level performance (chrF) while also influencing semantic correctness and fluency (COMET).

Regarding the unexpected effect of *char.count* on chrF, the positive correspondence in all models suggests that longer sentences on the Spanish side might provide additional context, aiding character-level matching. Conversely, this effect is not observed in COMET for most of the models, underlining the metric's focus on semantic correctness and fluency rather than character-level measures. It's important to note that while a high chrF score does indicate strong character-level performance, it does not necessarily guarantee good overall translation quality. This underscores the significance of measuring different metrics, as noted in this study, to gain a comprehensive understanding of translation quality and the effect of morphological variables such as fusion.

5.6.4 Overall conclusion

In summary, our analysis emphasised the relationship between morphological typology variables and translation performance measured by chrF and COMET. Specifically, in the English-Turkish MT models, our analysis confirms that synthesis affects both metrics, with a stronger impact on chrF than on COMET, which focuses on semantic similarity. Furthermore, fusion significantly influences both chrF and COMET metrics, with a more substantial impact on the former, in the English-Spanish MT models. Besides, the unexpected effect of *char.count* on chrF in some models suggests a nuanced relationship between sentence length and character-level F-score. These results enhance our understanding of how diverse metrics encompass different aspects of translation quality, and highlight the potential of variables like fusion and synthesis to advance our insights in MT evaluation.

5.7 Limitations

It is important to note the limitations of this study. Overall results do not suggest that translating into more isolating languages, such as Chinese, or more agglutinative ones,

such as Turkish, is necessarily easier than their counterparts. Highly isolating languages can present significant issues related to word coverage and vocabulary size of the model, while highly agglutinative languages are more complex in terms of morphology. Furthermore, it is difficult to isolate the degree of fusion from synthesis completely. For instance, Turkish is a highly synthetic but also highly agglutinative language (the opposite extreme of fusion). Furthermore, some languages exhibit both synthetic and fusional traits, such as Navajo, a Native American language. It is even more challenging for machine translation when dealing with source and target languages that exhibit high fusion and high synthesis, such as between Spanish and other indigenous languages of the Americas, including Asháninka or Guaraní. As we will discuss in the following chapter, Chapter 6, these challenges underscore the importance of ongoing research and development of machine translation systems that can effectively handle the complex morphological typology of diverse languages.

While this study has provided insights into the impact of synthesis and fusion on machine translation performance for a limited number of languages, extending the analysis to other languages may present practical challenges. Synthesis can only be calculated directly if the morphological analyser splits the word into morphemes, and fusion poses several issues, as previously mentioned in §5.5.3. Additionally, Payne (2017) have noted that the discourse can impact the computed variables due to the diversity of the vocabulary. This study focuses on news data only, and it will be relevant to extend it to different domains.

To address these limitations, we believe that our word-level analysis, which targets specific parts of speech, has been essential in enabling the study of the indexes and partially isolating them from each other. We also selected study cases that represent different levels of synthesis/fusion, such as Spanish verbs with few morphemes and Turkish with a more agglutinative morphology. In the future, to rapidly extend the evaluation for new languages and domains, we could follow a less fine-grained analysis of each index. For instance, instead of granulating per number of morphemes, we could compare synthesis$=1$ versus synthesis>1, as we did in this work with fusion$=0$ versus fusion>0.

5.8 Conclusion and future work

In this chapter, we proposed methods for quantifying the indices of synthesis and fusion in an automatic and semi-automatic way, respectively. We applied these methods

to English and German to compute synthesis, and Spanish verbs to calculate fusion. Furthermore, our analysis at the word level for Turkish nouns and verbs in synthesis, and Spanish verbs in fusion, revealed a relationship between these variables and machine translation performance (when translating from English), which was further supported by human evaluation. Additionally, by using generalised linear models, we were able to investigate the impact of these indices on machine translation performance at segment-level, suggesting different degrees of influence on distinct evaluation metrics. However, further analysis is needed for more combinations of languages with isolating-synthetic or agglutinative-fusional traits in either the source or target side of a translation system.

Overall, our study contributes to a better understanding of the impact of synthesis and fusion on machine translation performance, paving the way for future research in developing more effective machine translation systems that take into account the morphological properties of different languages. For instance, as future work, we can ask: are we improving the automatic translation of highly fusional words or segments? Following our methodology, we could stratify evaluation sets to measure how our models perform in different parts of the spectrum.

Chapter 6

Polysynthesis in Endangered Languages: an Extreme Machine Translation Challenge

The previous chapter studied how the variables of synthesis and fusion impact and hinder machine translation performance. Based on those findings, this chapter aims to investigate an extreme case where the challenges of limited data and highly synthetic languages compound, and to determine if we can still improve translation performance for morphologically-complex and extremely low-resource languages. For this purpose, we focus on the development of machine translation corpora for understudied and endangered languages, the implementation of robust machine translation baselines, the analysis of the effect of morphologically supervised versus unsupervised segmentation methods, and the application of a less-data-dependent segmentation approach.

6.1 Motivation

Our question is: **how can we improve machine translation performance for the extreme compounding challenge of highly synthetic and low-resource languages?** A high degree of synthesis, or polysynthesis, is a linguistic phenomenon characterised by complex word structures that incorporate multiple morphemes, or according to Payne (2017), a high ratio of morphemes per word. Languages with polysynthesis can convey multiple meanings within a single word, as we show in the example 5:

(5) *nonkotsitasanomempebentajeibetamanakero*
no-n-kotsi-t-asano-mempe-bent-a-jei-be-t-aman-ak-e-ro
1-IRR-cook-EP-INTENS-pretend-EP-PL-frust-EP-early-PRF-IRR-3F

'We will really pretend to cook for her early in the morning without success'
(Jaime Montoya Samamé, fieldnotes)

where 'kotsi' is the root verb ('to cook') of the 15-morpheme word in Asháninka, a language with polysynthetic traits spoken in Amazonia. Polysynthesis in Asháninka poses a unique challenge for machine translation systems due to its complex morphological structures. This challenge is compounded by the fact that, like many other endangered indigenous languages of the Americas, Asháninka is severely understudied in NLP because of the scarcity of data (Mager et al., 2018a).

To address the challenge of polysynthesis in extremely low-resource scenarios, we first ask: **can small multilingual models be beneficial for translating low-resource polysynthetic languages?** We previously demonstrated that clusters of multilingual models are effective in low-resource machine translation, as they can still leverage the similarities between languages to improve translation performance without needing a massive model with several language pairs. For this purpose, we assemble training data and build evaluation sets for polysynthetic languages spoken in South America, such as Asháninka or Shipibo-Konibo. Additionally, we explore further strategies such as pre-training with high resource language-pairs (Kocmi and Bojar, 2018), back-translation (Sennrich et al., 2016a), and fine-tuning (Neubig and Hu, 2018) to enhance our baseline multilingual systems. However, we found that the extra engineering efforts were insufficient in improving translation performance significantly. Hence, we shift our focus towards segmentation, a relevant step in machine translation.

The importance of segmentation in machine translation is further highlighted in the case of polysynthetic languages. The complex word structures produced in these languages make segmentation crucial in dealing with rare words that are composed of several morphemes. We then ask: **is morphologically-aware segmentation beneficial for machine translation where one language is polysynthetic and extremely low-resourced?** To address this, we compare a wide set of segmentation methods, morphologically-supervised and unsupervised approaches, and apply them to the input of machine translation systems where either the source or the target language is polysynthetic. For this purpose, we also introduce a new morphologically annotated corpus for Shipibo-Konibo.

From the comparison, we found that unsupervised segmentation remains relevant for machine translation in the extreme high synthesis and low-resource language scenario. However, such methods rely heavily on the availability of textual data. We aim to reduce the dependency on the corpus size, so we pose the final question: **can a corpus-independent segmentation method be more effective for machine translation of polysynthetic and low-resource languages?** In response, we propose a linguistically-motivated segmentation approach that leverages syllable units, which are corpus-independent. For this purpose, we apply syllable-based segmentation to machine translation for Shipibo-Konibo, utilising multilingual models to enhance our analysis.

The research presented in this chapter has been previously published in the following academic articles:

- In Oncevay (2021), we describe the development of multilingual machine translation models for polysynthetic and endangered languages of South America. We also contributed to the collection of training and evaluation data for indigenous languages of the Americas as part of a larger initiative known as AmericasNLP (Mager et al., 2021) (see §6.3 and §6.4).

- In Mager et al. (2022)[1], we contributed to the comparison of various unsupervised and morphologically-supervised segmentation methods for machine translation of languages with polysynthetic traits and limited resources (see §6.5).

- In Oncevay et al. (2022b), we propose a linguistically-motivated syllable-based segmentation approach and evaluate its applicability in machine translation for polysynthetic languages (see §6.6).

6.2 Related work

Studying the polysynthesis phenomenon in machine translation is not widespread in the literature (Schwartz et al., 2020). However, there is an exception for Inuktitut, a polysynthetic language spoken in North America, and the English-Inuktitut language pair was included in WMT 2020 news shared task (Barrault et al., 2020). This is due to its available corpus, and it is considered a medium-resource language pair, with

[1] As the second author, we note our contribution explicitly: compilation and processing of a new morphological segmentation dataset for Shipibo-Konibo; training of part of the machine translation experiments with several segmentation methods, including all the statistical significance analysis.

1.3M parallel sentences (Joanis et al., 2020) and 1.4M tokens from monolingual data extracted from CommonCrawl. For the shared task, different submited systems applied pretraining methods, morphological segmentation, or multilingual transfer with related languages, although obtaining mixed results (Bawden et al., 2020; Kocmi, 2020; Knowles et al., 2020; Roest et al., 2020). Besides, Ortega et al. (2020b) used morphological information, such as affixes, to guide the BPE segmentation algorithm (Sennrich et al., 2016b) for Quechua, another highly synthetic language in the Americas. However, their improvement is not statistically significant.

6.3 Languages and Datasets

To develop and evaluate machine translation models for the extremely high synthesis and low-resource language case, we shift our attention toward the Americas continent. Despite the significant number of languages spoken in the Americas, most of them receive little attention from NLP researchers, including those that are considered endangered. Glottolog (Nordhoff and Hammarström, 2012) reports that there are 86 language families and 95 language isolates in the Americas, highlighting the urgent need for the development of language technologies that can support these communities. Through the documentation, promotion, and revitalisation of endangered languages, language technologies have the potential to make a significant impact (Zariquiey et al., 2022).

My contribution is part of a bigger AmericasNLP 2021 Shared Task effort (Mager et al., 2021), where we provided training and evaluation datasets for machine translations in 10 indigenous languages of the Americas (paired with Spanish). The work consisted of two phases (I only note the languages I worked with):

- I collect available **training data** (parallel and monolingual corpora) for the following languages: Shipibo-Konibo[2], Asháninka, Quechua, Aymara (paired with Spanish or English). I then perform basic cleaning steps on the datasets (see §6.4.1).

- The **development** and **test** sets were sampled from XNLI (Conneau et al., 2018). We selected genres that we deemed relatively straightforward to translate into the target languages, including ”face-to-face” conversations, letters, and telephone

[2]Throughout the last part of this chapter, we focus solely on Shipibo-Konibo due to the availability of annotated data and access to experts and native speakers who could support the development and evaluation of our resources and experiments.

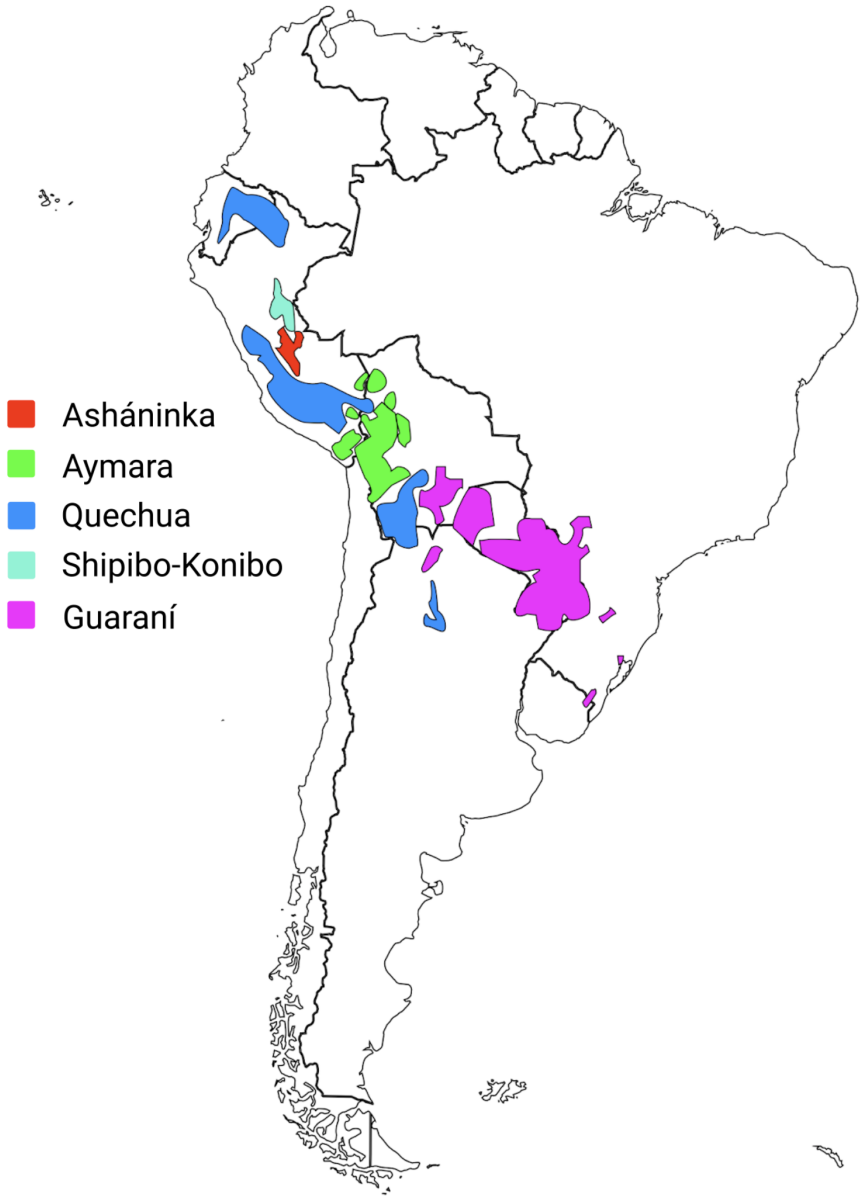


Figure 6.1: South America map with the approximate location of the speaker communities of five indigenous languages included in the AmericasNLP Shared Task. Adapted from Ebrahimi et al. (2022).

Table 6.1: Languages: Details and datasets

Language	**Lang. Family**	**Speakers**	**Variant**	**ISO**	**Resources**
Shipibo-Konibo	Panoan	30,000	Shipibo-Konibo	shp	Language learning (Tatoeba) (Gómez Montoya et al., 2019) Books for bilingual education (Galarreta et al., 2017) Dictionary entries (Loriot et al., 1993) Monolingual educational books (Bustamante et al., 2020)
Asháninka	Arawak	35,000	Asháninka Perené	cni	Traditional stories (Ortega et al., 2020a) Educational texts (Romano and Richer, 2008) Environmental laws (Mihas, 2011) Monolingual educational books (Bustamante et al., 2020)
Quechua	Quechuan	7M	Southern Quechua or Chanka/Ayacucho	quy	JW300 (Agić and Vulić, 2019) Official dictionaries from the Peruvian Government Miscellaneous and dictionary entries (Huarcaya Taquiri, 2020)
Aymara	Aymaran	2M	Central Aymara (La Paz)	aym	News from Global Voices (Prokopidis et al., 2016), published in OPUS (Tiedemann, 2012)

dialogues. I supervise and curate the translations for both Shipibo-Konibo and Asháninka, and provide the translator with specific requests, such as translating only the words and concepts that are well-known in the communities, whereas other terms could be preserved in Spanish. There was constant communication with the translators to keep the punctuation marks, named entities and numeric formats. Moreover, the development and test sets were created following the official writing convention proposed by the Peruvian Government and taught in bilingual schools. Ebrahimi et al. (2022) and Kann et al. (2022) describe more details about further NLI experiments with the translated sets.

By providing these resources, we hope to contribute to the visibility and preservation of these languages and support the efforts of language activists and communities, alongside our research goal. The map in Figure 6.1, adapted from Ebrahimi et al. (2022), displays the approximate location of the speakers of the languages we worked with, as well as Guaraní, another language evaluated in the AmericasNLP Shared Task. It should be noted that the majority of the areas of the four languages we focused on are situated within Peruvian borders. We present the details and training resources for the four languages in Table 6.1, and in Table 6.2, we summarise the dataset sizes.

ISO	Language	Mono.	es	en
aym	Aymara	8,680	5,475	5,045
cni	Ashaninka	13,193	3,753	
quy	Quechua Ayacucho		104,101	14,465
shp	Shipibo-Konibo	23,593	14,437	
quz	Quechua Cusco		97,836	21,760

Table 6.2: Number of sentences in monolingual and parallel corpora aligned with Spanish (es) or English (en). The latter are used for en→es translation and we only noted non-duplicated sentences w.r.t. the *–es corpora. We use Quechua Cusco data as complementary resources for multilingual training.

6.4 Multilingual Machine Translation

First, we address the extremely low-resource scenario with multilingual machine translation. Tan et al. (2019) and our prior research, in Chapter 4, have indicated that a

massive multilingual model consisting of numerous language pairs is not totally necessary to attain the benefits of improving performance for low-resource language pairs. Instead, a smaller group of languages that are similar to some degree can suffice. Although we note that each language in our case study belongs to a different language family, that is not a hindrance for multilingual models. Typically, family-based clusters are not the most effective ones, as we previously observed (see Chapter 4, §4.5).

Concerning the training data available, it is unsurprising that even if all the parallel data are combined, the resulting dataset only contains 225k sentences (see Table 6.2), which is not comparable to the size of language-pairs released in other shared tasks like WMT, where datasets are in the millions. Data is a compounding issue for the languages' complex morphology and the training set's domain disparity. To address these challenges, we also consider further strategies to enhance our multilingual model, such as pretraining with a high-resource language pair, back-translation and fine-tuning to the out-of-domain evaluation set.

6.4.1 Pre-processing

The compiled datasets are potentially noisy and not cleaned. To address this, we reduced the number of sentences using several heuristics, such as identifying sentences where Arabic numbers or punctuation did not match on both sides of the parallel sentence, sentences where there were more symbols or numbers than words, and sentences where the ratio of words from one side was five times larger or shorter than the other.

Table 6.3 shows the original and cleaned data size per language pair along with some statistics. The proportion of cleaned data is smallest for the Spanish-Shipibo-Konibo language pair, which suggests that the original datasets may have already undergone some cleaning before publication. A similar case is observed for the Spanish-Asháninka pair. However, a large amount of data was filtered out from all Quechua datasets, which may be due to sentence alignment issues in the JW300 corpus (Agić and Vulić, 2019) that require specialised tools to address (Kreutzer et al., 2022).

We also calculated the ratio of tokens per sentence (T/S) and the ratio of tokens in the source compared to the target sentence. We found that the T/S value for the source is always greater than the target sentence, which is to be expected given that Indo-European languages, such as Spanish or English, are more isolating and tend to use more tokens per sentence. This is confirmed by the last ratio variable, where a value greater than one indicates that the source language (Indo-European and more isolating)

Lang. pair	S (raw)	S (clean)	% clean	T/S (src)	T/S (tgt)	ratio T src/tgt
es-aym	6,453	5,475	-15.16%	19.27	13.37	1.44
es-cni	3,860	3,753	-2.77%	12.29	6.52	1.89
es-quy	128,583	104,101	-19.04%	14.2	8.17	1.74
es-shp	14,511	14,437	-0.51%	6.05	4.31	1.4
es-quz	130,757	97,836	-25.18%	15.23	8.62	1.77
en-quy	128,330	91,151	-28.97%	15.03	8.68	1.73
en-quz	144,867	100,126	-30.88%	14.84	8.42	1.76
en-aym	8,886	7,689	-13.47%	19.36	13.32	1.45

Table 6.3: Statistics and cleaning for all parallel corpora. We observe that the Shipibo-Konibo and Ashaninka corpora are the least noisy ones. S = number of sentences, T = number of tokens.

has more tokens than the target language (Amerindian and more synthetic). The case of Asháninka is particularly noteworthy, as the factor is almost two, indicating that there are, on average, almost twice as many tokens in a Spanish sentence as in an Asháninka translation.

6.4.2 Evaluation

The training data have been extracted from different domains and sources, which are not necessarily the same as the evaluation sets that we developed for the AmericasNLP shared task. Therefore, we integrated a portion of the development set (995 sentences per language) into the training data. For this purpose, we split the development set into three parts: 25%-25%-50%. The first two parts are our custom dev and devtest sets (which were used prior to testing on the official test set), while the last 50% section is added to the training data to reduce the domain gap. Additionally, we extracted a sample from the training data to double the size of the development set. It was important to avoid any overlapping of the Spanish side with the training set, as we were evaluating multilingual models. We used the same multi-text sentences for evaluation and ensured that the mixed data in the validation set allowed us to evaluate how the model fit with all the domains. The performance of all models was evaluated using BLEU (Papineni et al., 2002) and chrF (Popović, 2015) metrics, and the results on the official test sets (with 1k sentences) are reported as well.

6.4.3 Experimental procedure

For the experiments, we used a Transformer-base model (Vaswani et al., 2017) with the default configuration in Marian NMT (Junczys-Dowmunt et al., 2018). The steps are as follows:

Joint subword segmentation To take advantage of the potential lexical sharing of the languages (e.g. loanwords), we train a unique multilingual segmentation model by sampling all languages with a uniform distribution (including Spanish and English, as Spanish–English is our high-resource language pair for pre-training). We used the unigram language model implementation in SentencePiece (Kudo and Richardson, 2018) with a vocabulary size of 32,000. For simplification purposes, we are overlooking the polysynthetic nature of the Amerindian languages for the segmentation process until the following sections.

Pre-training We chose the Spanish–English language pair and trained MT systems in both directions. English is not related to any of the studied languages but is Spanish's largest paired language corpus. For this purpose, we consider the EuroParl (1.7M sentences) (Koehn, 2005) and the NewsCommentary-v8 (174k sentences) (Bojar et al., 2013) corpora. The en→es and es→en models achieved 34.4 and 32.3 BLEU points, respectively, in the newsdev2013 set (Bojar et al., 2013), which is competitive with state-of-the-art models.

Multilingual fine-tuning Using the pre-trained en→es model, we fine-tuned the first multilingual model many-to-Spanish. Following established practices, we used a uniform sampling for all the language pairs (quz–es included) to avoid under-fitting the low-resource language pairs. Temperature-based sampling (Aharoni et al., 2019) or automatically learned data scorers are more advanced strategies (Wang et al., 2020). However, we left that analysis for further work. Results are in Table 6.4, row (a). Using the es→many direction (row (e)), we replicated this to the es→en model.

Back-translation We attempted to improve the training of models (b) and (f) for polysynthetic languages by back-translating the monolingual data using model (a). However, the resulting models trained on the back-translated data underperformed or failed to converge. This could be due to the noisy nature of the translations produced by model (a).

BLEU	Aymara			Ashaninka			Quechua			Shipibo-Konibo		
→**Spanish**	dev	devtest	test	dev	devtest	test	dev	devtest	test	dev	devtest	test
(a) Multilingual	**11.11**	**9.95**	3.70	**8.40**	**9.37**	**5.21**	12.46	11.03	8.04	**10.34**	**12.72**	**10.07**
(b) Multi+BT	10.76	8.39	2.87	7.30	5.34	3.44	11.48	8.85	7.51	9.13	10.77	7.58
(c) Multi+BT[t]	10.72	8.42	2.86	7.45	5.69	3.15	11.37	10.02	7.12	8.81	10.73	7.18
(d) Pairwise	9.46	7.66	2.04	4.23	3.96	2.38	**15.21**	**14.00**	**8.20**	7.72	9.48	4.44
Spanish→	dev	devtest	test	dev	devtest	test	dev	devtest	test	dev	devtest	test
(e) Multilingual	8.67	6.28	2.19	6.74	11.72	**5.54**	10.04	5.37	**4.51**	10.82	10.44	6.69
(f) Multi+BT	3.31	2.59	0.79	1.29	3.38	2.82	1.36	2.02	1.73	1.63	3.76	2.98
(g) Multi+BT[t]	**10.55**	**6.54**	**2.31**	**7.36**	**13.17**	5.40	**10.77**	5.29	4.23	**11.98**	**11.12**	**7.45**
(h) Pairwise	7.08	4.96	1.65	4.12	8.40	3.82	10.67	**6.11**	3.96	8.76	7.89	6.15

Table 6.4: BLEU scores for the dev and devtest custom partitions and the official test set, including all the multilingual and pairwise MT systems into and from Spanish. BT = Back-translation. BT[t] = Tagged back-translation.

chrF	Aymara			Ashaninka			Quechua			Shipibo-Konibo		
→**Spanish**	dev	devtest	test	dev	devtest	test	dev	devtest	test	dev	devtest	test
(a) Multilingual	**31.73**	**28.82**	**22.01**	**26.78**	**26.82**	**22.27**	32.92	32.99	29.45	**31.41**	**33.49**	**31.26**
(d) Pairwise	28.77	25.03	19.79	20.43	20.40	18.83	**36.01**	**36.06**	**30.90**	27.25	29.91	25.31
Spanish→	dev	devtest	test	dev	devtest	test	dev	devtest	test	dev	devtest	test
(g) Multi+BT[t]	**37.32**	**35.17**	**26.70**	**38.94**	**38.44**	**30.81**	44.60	38.94	**37.80**	**40.67**	**39.47**	**33.43**
(h) Pairwise	28.89	28.23	21.13	32.55	32.29	27.10	**45.77**	**39.68**	36.86	34.97	34.96	27.09

Table 6.5: chrF scores for the dev and devtest custom partitions and the official test sets for the best multilingual setting and the pairwise baseline in each direction.

Tagged back-translation (BT[t]) To address these challenges, we incorporated a special tag for the back-translated data, following the approach proposed by Caswell et al. (2019). With BT[t], we signal to the model the synthetic data, which is potentially less reliable than the original sentences. We report the results of this approach in Table 6.4, rows (c) and (g).

Pairwise baselines We obtained pairwise systems by fine-tuning the same pre-trained models (without any back-translated data). For a straightforward comparison, they used the same multilingual SentencePiece model.

6.4.4 Analysis and discussion

Table 6.4 shows the BLEU scores for all the machine translation systems, and as a reference, we also report the chrF scores in Table 6.5 for the best multilingual setting

and the pairwise baseline.

Additionally, we found that for the translation models into Spanish, the multilingual model without back-translated data consistently outperformed the other models in all languages except Quechua, where the pairwise system achieved the highest translation accuracy. This is likely due to Quechua being the highest-resource language pair in our experiment, as its performance was also negatively impacted in the multilingual setting. This behaviour is commonly observed in multilingual training, and additional approaches such as adapter layers (Bapna and Firat, 2019) or temperature-based sampling (Aharoni et al., 2019) may help mitigate the issue. We observe a similar scenario in the other translation direction from Spanish, where the best multilingual setting with back-translated data could not surpass the performance of the Spanish-Quechua model on the devtest set. Nevertheless, the gains for Aymara, Ashaninka and Shipibo-Konibo by using the multilingual approach are significant in the dev and devtest sets.

Furthermore, we note that the models are not totally overfitted to any of the evaluation sets in both translation directions. Exceptions are Spanish-Aymara and Spanish-Quechua, with a significant performance dropping from dev to devtest in their best settings, meaning that it started to overfit to the training data.

Concerning the results on the official test set, the performance is lower than the results with the custom evaluation sets or devtest. The main potential reason is the different domain of the test set in comparison to the training sets. Another point to highlight, in the official test results, is that the best result in the Spanish–Quechua language-pair is obtained by a multilingual model (the scores between the model (e) and (g) are not significantly different) instead of the pairwise baseline, as in the dev or devtest sets.

Using back-translated data can be useful for providing the model with additional real text in the polysynthetic language. However, this is not the case when translating into Spanish, where the multilingual model without back-translated data performs better in most cases. Nonetheless, decoding a polysynthetic and low-resource language remains a challenging task, and the relatively low BLEU scores do not necessarily indicate proper adequacy or fluency in translation. Similarly, while the best chrF scores we obtained are competitive, they still do not outperform the state-of-the-art reported in the AmericasNLP shared task (Mager et al., 2021). Even in the SOTA systems, the human evaluation results were poor. The results of this study indicate that, despite the use of all available data, machine learning and engineering perspectives alone are insufficient for addressing the challenges of low-resource polysynthetic languages. Thus,

we redirect our focus towards an alternative approach to improve machine translation performance in this context: subword segmentation.

6.5 Unsupervised versus Morphologically-Supervised Segmentation

The previous experiments aimed to address the challenge of low-resource and high synthesis languages in machine translation through engineering efforts. In this section, we focus solely on the segmentation aspect, as the complex morphology of polysynthetic languages may benefit from morphologically-aware segmentation methods. Thus, we compare several segmentation methods, including unsupervised and morphologically-supervised approaches, as follows:

BPEs (Sennrich et al., 2016b) is our baseline segmentation method, and we use the SentencePiece implementation (Kudo and Richardson, 2018).

Morfessor (Smit et al., 2014) is an unsupervised method that uses a statistical model for the discovery of morphemes using minimum description length optimisation. For this analysis, we use Morfessor 2.0.

FlatCat (Grönroos et al., 2014) is a variant of Morfessor. It consists of a category-based hidden Markov model and a flat lexicon structure for segmentation.

LMVR (Ataman et al., 2017) modifies the FlatCat implementation by adding a lexicon size restriction and increasing the tendency of the model to increase the segmentation of commonly seen words.

Seq2seq and s2s+multi are vanilla RNN sequence-to-sequence model with attention. The first variant (seq2seq) employs a supervised neural model, whereas the second method (s2s+multi) uses the most promising extension proposed by Kann et al. (2018) adding random generated strings in an auto-encoding fashion (s2s+multi).

Pointer Generator Networks (PtrNet) (See et al., 2017) is a supervised model that we previously used in Chapter 5, §5.3, for computing the index of synthesis.

Additionally, to isolate the effect of different segmentations, and to facilitate the analysis, we ablate some aspects, such as the out-of-domain evaluation sets. We also choose not to test multilingual models, for two reasons. Firstly, we require data to train morphologically-supervised segmentation models, which is only available for the Shipibo-Konibo language (we describe this new dataset in the following section). Secondly, we use different segmentation approaches for the source and target languages, and therefore need to separate the analysis. In this part of the study, Spanish will be segmented with BPE only.

6.5.1 Datasets

Morphological segmentation To train the supervised methods, we introduce a new morphologically annotated dataset for Shipibo-Konibo. For this purpose, we adapted annotated sentences for lemmatization and part-of-speech tagging (Pereira-Noriega et al., 2017), and from a treebank (Vasquez et al., 2018), which was segmented in morphemes due to a particular phenomenon for clitics in the dependencies annotation. The sources of the annotations are Governmental educational books that follow the official writing standard. Table 6.6 shows some annotated samples of the corpus, while 6.7 contains details and statistics about the new dataset.

In the training set, we noticed that a significant number of words were segmentable, with a proportion of 72% consisting of more than one morpheme. The maximum number of morphemes in a word was found to be 5, and the synthesis variable reached a relatively high value of 2, considering the high proportion of single-morpheme words (28%). If the synthesis value of the whole dataset is not higher, it is because some of the indivisible (one-morpheme) words in the dataset are Spanish loanwords, especially nouns and names.

Input	Output	Notation
kachiokea	kachio-kea	mountain-FROM
jakonmatani	jakon-ma-tani	good.person-NEG-ALMOST
tetebo	tete-bo	hawk-PL

Table 6.6: Annotation samples of the morphological segmentation dataset for Shipibo-Konibo. The notation in English is added for clarity.

	train	dev	test
Words	604	163	329
Segmentable words	437	114	228
Morphemes	1215	321	642
Unique Morphemes	476	181	319
Words+1	0.72	0.69	0.69
Synthesis	2.01	1.97	1.95
Max-Morph	5	5	5
OOV-Morph		93	179

Table 6.7: Statistics of the Shipibo-Konibo dataset for morphological segmentation. Words+1: proportion of words consisting of more than one morpheme; Max-Morph: maximum number of morphemes found in one word; OOV-Morph: morphemes in evaluation not seen in training.

Machine translation We use the collected datasets for Shipibo-Konibo–Spanish for the machine translation experiments. However, we split the training set into development and test sets (587 and 1030 sentences, respectively) to simplify the out-of-domain analysis. The split was stratified according to the different sources of the parallel corpora.

6.5.2 Experimental setup

Metrics For evaluating machine translation, we use the standard BLEU (Papineni et al., 2002) and chrF (Popović, 2015) metrics from the SacreBLEU implementation (Post, 2018), whereas for morphological segmentation, we compare all outputs against the gold annotated test sets using the EMMA F1 metric (Spiegler and Monson, 2010).

Model We use a Transformer model (Vaswani et al., 2017) with the hyperparameters proposed by Guzmán et al. (2019) as a baseline for low-resource languages. After searching for the best vocabulary size using 2k, 4k, 5k, 6k and 8k, we use a 5k vocabulary size for all sides using BPE. Besides, we use the fairseq toolkit (Ott et al., 2019) for all translation experiments. The polysynthetic languages are segmented with the different investigated segmentation methods and Spanish always uses BPE in both translation directions.

6.5.3 Results and discussion

	Morph. Seg.	es-shp		shp-es	
	EMMA F1	BLEU	chrF	BLEU	chrF
BPE	71.41	10.84	36.54	11.85	32.59
Morfessor	59.45	5.00*	33.24*	9.65*	30.69*
FlatCat	67.95	11.68	37.58	**12.29**	**33.38**
LMVR	67.58	**12.84**	**38.99***	11.14	32.60
Seq2seq	82.25	0.77*	25.21*	10.27*	31.10*
s2s+multi	**85.99**	0.13*	22.06*	9.51*	27.79*
PtrNet	78.22	0.06*	22.78*	8.91*	27.97*

Table 6.8: Morphological segmentation results for Shipibo-Konibo (left) and machine translation performance for Spanish–Shipibo-Konibo (right) with different segmentation methods. Maximum scores are in bold. For MT, we run a paired approximation test with 10000 trials using the BPE-based system output as the baseline, and "*" indicates a p-value < 0.05.

Table 6.8 presents contrasting results. While BPE is the best-performing unsupervised approach for morphological segmentation, it is outperformed by the more morphologically supervised segmentation models. However, the supervised segmentation models perform poorly in the machine translation task compared to the unsupervised ones in both directions. We argue that the supervised segmentation methods often innovate new subwords in their output, which only adds noise to the input of the machine translation model.[3] For instance, in Table 5.1 from Chapter 5, we included some PtrNet outputs of English words, such as *cookm* (from *cookie*) and *polyggmy* (from *polygamy*).

Regarding the unsupervised segmentation methods for machine translation, LMVR is the best method for translating into the polysynthetic Shipibo-Konibo, while FlatCat is the best for the opposite direction. In both cases, most outcomes are comparable to the BPE results, with only one case (es-shp with chrF) being statistically superior.

Apart from the numerical results, we observe that most unsupervised segmentation

[3]In Mager et al. (2022), we analysed the occurrence of out-of-vocabulary (UNK) tokens within the machine translation test set for four photosynthetic languages, including Shipibo-Konibo. Our findings revealed that supervised models, such as s2s+multi, exhibited the highest count of UNK tokens, whereas unsupervised methods like LMVR demonstrated a slightly lower count. Notably, the BPE approach demonstrated the most comprehensive token coverage among the studied methods.

methods are still relevant for handling languages with high synthesis and limited data. However, these methods rely on the size of textual data to extract patterns and take advantage of frequent subwords. In an extremely low-resource scenario, it is desirable to reduce the dependence of the segmentation method on the available textual data. For this reason, in the following section, we propose a less data-dependent segmentation approach.

6.6 Syllable-based segmentation

In machine translation, we rely on subword segmentation as a widespread approach to generate rare subword units (Sennrich et al., 2016b). However, lacking large textual collections can constrain the learning of unsupervised segmentation methods. Alternatively, we could use character-level modelling since they also have access to subword information (Kim et al., 2016), but further issues can arise, such as long-term dependencies, longer training time to converge, as well as the need to deepen the model (Gao et al., 2020).

For this reason, syllables are introduced as an alternative since they are speech units ("A syl-la-ble con-tains a sin-gle vow-el u-nit") and behave as a mapping function to reduce the length of the sequence with a larger "alphabet" or syllabary. Syllables are fundamental phonological units that participate in important word prosodic patterns, such as stress assignment, and are more linguistically relevant units than characters. Their extraction can be rule-based and corpus-independent, but data-driven methods or hyphenation using dictionaries can approximate them as well.

Therefore, this section evaluates whether syllables are useful for encoding or decoding a highly synthetic and extremely low-resource language, and we choose Shipibo-Konibo (paired with Spanish) as our case study because the language exhibits high synthesis and low-resource traits, but also presents a shallow (or transparent) orthography, a property that could be beneficial to syllabification.

Orthographic depth, which is the degree of grapheme-phoneme correspondence (Borgwaldt et al., 2005), can increase complexity to syllabification (Marjou, 2021). For example, English has a deep orthography (weak correspondence), whereas Finnish is more transparent (Ziegler et al., 2010). In the case of Shipibo-Konibo, as with any other language that is going through a process of revitalisation, the alphabet has been recently standardised, and the grapheme-phoneme ambiguity has been reduced (Alva and Oncevay, 2017). For example, in Spanish, there is ambiguity in the sounds for 'ca'

and ‘ka’, whereas for Shipibo-Konibo, it has been simplified into ‘ka’ only.

Since Spanish does not exhibit a transparent orthography as Shipibo-Konibo, syllables are considered alongside state-of-the-art segmentation methods such as BPE to segment Spanish and other languages, such as English. Additionally, multilingual NMT systems are used to provide robust baselines, as they outperformed pairwise systems for the chosen language pair, as noted in §6.4.

6.6.1 Related work

For syllable-based machine translation, there are mostly studies for related paired languages, such as Indic languages (in statistical MT without subword-based baselines: Kunchukuttan and Bhattacharyya (2016)), Tibetan–Chinese (Lai et al., 2018), and Myanmar–Rakhine (Myint Oo et al., 2019). Instead, Spanish–Shipibo-Konibo is a non-related language-pair. The only distant pair was English–Myanmar (ShweSin et al., 2019), but they did not compare it with unsupervised subword segmentation approaches. Furthermore, neither of these studies analysed multilingual scenarios.

6.6.2 Syllabification in Shipibo-Konibo

For Shipibo-Konibo, we adapt the rule-based syllabification tool developed by Alva and Oncevay (2017), which follows the standardised orthographic patterns of the language. The original method employs syllabification to validate whether a word is composed of orthographically consistent syllables, primarily for spell-checking purposes. In simpler terms, if a word cannot be segmented according to these orthographic patterns, it may potentially be a misspelling, a loanword, or a named entity from another language. To illustrate this concept, we provide the following examples:

1. The Shipibo-Konibo word *atipana*, meaning “can”, is syllabified as *a-ti-pa-na*.
2. The Spanish word *pasaporte*, meaning “passport”, cannot be properly syllabified.
3. The name *John*, originating from English, also defies syllabification.

In the context of machine translation, we apply the syllabification tool as a segmentation method for Shipibo-Konibo texts. However, there could be a presence of loanwords and named entities from diverse languages that lack direct counterparts in the Shipibo-Konibo language. In cases where a word cannot be segmented, we resort

to either splitting it into individual characters (e.g. *pasaporte* into *p-a-s-a-p-o-r-t-e*) or employing an alternative segmentation method. For example, we can employ a joint BPE segmentation model trained with the whole parallel corpus (e.g. *pasaporte* into *pasa-porte* or *pas-a-porte*). It's important to note that the parallel texts from the paired languages are segmented not by syllables, but rather by standard segmentation methods like BPE.[4]

For our experiments in the following sections, we complement syllables with BPE, creating an approach referred to as SYL+BPE. Additionally, we compare this approach against using BPE alone, which serves as a baseline referred to as BPE-ALL. We provide more details about the baselines and segmentation approaches in the following section and Table 6.9.

6.6.3 Machine translation systems with syllables

In contrast to prior work (see §6.6.1), we (i) study syllable-based NMT for a distant and low-resource language-pair, Spanish–Shipibo-Konibo, and where one language (Shipibo-Konibo) exhibits polysynthetic traits; (ii) compare syllables against the most widespread unsupervised segmentation method (BPE, Sennrich et al., 2016b) with automatic metrics and human evaluation; and (iii) analyse the applicability of syllables on multilingual NMT systems. The last element is significant, as multilingual models are a significant approach for leveraging low-resource language-pairs performance (Siddhant et al., 2022). For the multilingual setting, we include Spanish-English as it is the highest-resource language pair available that includes Spanish.

For these reasons, we focus on MT settings and segmentation approaches that compare syllables with BPE, using either SYL+BPE or BPE-ALL. We enumerate the settings as follows and show details of their segmentation in Table 6.9.

1. MONO-SYSTEM: a pairwise NMT model where each source and target is segmented with a different method.

2. JOINT-SYSTEM: another pairwise NMT model where the BPE-ALL baseline is jointly trained with the source and target data.

3. O2M-SYSTEM: a multilingual one-to-many NMT model where the BPE-ALL baseline is jointly trained with all the languages. We add Spanish–English for

[4]We attempted to use syllables in other languages as well, such as Spanish and English, but with negative results. With large data, unsupervised segmentation methods like BPE can obtain more significant and overlapping subwords from source and target.

system	segmentation	Shipibo-Konibo	Spanish	English
MONO-SYSTEM	SYL+BPE	Syllables (+characters)	BPE_{es}	
	BPE-ALL	BPE_{shp}	BPE_{es}	
JOINT-SYSTEM	SYL+BPE	Syllables (+$BPE_{shp,es}$)	$BPE_{shp,es}$	
	BPE-ALL	$BPE_{shp,es}$	$BPE_{shp,es}$	
O2M-SYSTEM	SYL+BPE	Syllables (+$BPE_{shp,es,en}$)	$BPE_{shp,es,en}$	$BPE_{shp,es,en}$
	BPE-ALL	$BPE_{shp,es,en}$	$BPE_{shp,es,en}$	$BPE_{shp,es,en}$

Table 6.9: Description of the segmentation method used for each language in all MT systems and baselines. $BPE_{language(s)}$ is a (joint) BPE segmentation model that is trained with the data of the specified language(s).

multilingual training to leverage the overall performance of the low-resource language pair.[5]

To better illustrate the segmentation achieved with SYL+BPE in the three MT settings, we show some examples in Table 6.10. Specifically, we can observe how the term *pasaporte* ("passport" in Spanish) is split in distinct ways on the Shipibo-Konibo side. In the MONO-SYSTEM, the word is segmented into individual characters, while in the other two systems, it is processed through a joint BPE model. Additionally, all other words on the Shipibo-Konibo side maintain consistent syllable-based segmentation across all settings.

system	Spanish	Shipibo-Konibo
MONO-SYSTEM	_¿ puedo _ver _su _pasa por te ?	_¿ min _p a s a p o r t e _en _o in ti _a ti pa na ?
JOINT-SYSTEM	_¿ puedo _ver _su _pas a por te ?	_¿ min _pas a por te _en _o in ti _a ti pa na ?
O2M-SYSTEM	_¿ p uedo _ver _su _pasa porte ?	_¿ min _pasa porte _en _o in ti _a ti pa na ?

Table 6.10: Segmentation examples with the SYL+BPE approach for the three MT settings. The English translation is: "Can I see your passport?"

6.6.4 Experimental setup

Data For Spanish–Shipibo-Konibo (es–shp), we use the dataset compiled in the previous section (see §6.3), and perform the same split as in the morphological segmen-

[5] We do not consider the many-to-one direction due to resource constraints and because we observed that the improvements by syllables are noted when translating into Shipibo-Konibo only.

tation experiments (see §6.5) for the development and test subsets, to make the results comparable. For the multilingual models, we use the Spanish–English (es–en) train set from EuroParl (Koehn, 2005) and newscommentary-v8 (2.2M parallel sentences in total), and the NEWSTEST2013.ES-EN (Bojar et al., 2013) evaluation sets.

BPE settings We use the implementation of SentencePiece (Kudo and Richardson, 2018). Following our experiments in §6.5, we fix the best vocabulary size at 5000 pieces for the MONO-SYSTEM, after trying different values from 1k to 10k. The segmentation for JOINT-SYSTEM and O2M-SYSTEM use 5000 and 16000 pieces, respectively.

Model and training We reproduce the settings of our experiments in the previous section (see §6.5) by using the fairseq toolkit (Ott et al., 2019), and a Transfomer model (Vaswani et al., 2017) with smaller dimensions. For the pairwise systems, we train up to 100 epochs with an early stopping policy of 5 (validating every 5 epochs), whereas for the multilingual systems we train up to 30 epochs. For all the experiments, we use 4 NVIDIA GeForce GTX 1080 Ti GPUs. For the multilingual O2M-SYSTEM, we use a sampling approach with temperature of 5 (Aharoni et al., 2019).

Automatic evaluation We use chrF (Popović, 2015) from SACREBLEU (Post, 2018).[6]

In summary, we will compare three different system approaches: MONO-SYSTEM, JOINT-SYSTEM and O2M-SYSTEM. And for each of them, we will apply a different segmentation methodology: BPE-ALL and SYL+BPE. Our goal is to understand whether a syllable-based segmentation approach is effective and compatible with the widespread BPE method.

6.6.5 Results and discussion

Table 6.11 shows the translation performance in all settings with two different segmentation methods. We observe that SYL+BPE are statistically better than the BPE-ALL baseline when translating from Spanish into Shipibo-Konibo, but not in the other direction. This fact indicates that syllables support the decoding more than the encoding step of a language with a transparent orthography. One possible reason is that, like characters, syllables do not carry semantic or grammatical information. We argue that

[6] chrF2+numchars.6+space.false+v.1.5.0.

Model	Segmentation method			
	BPE-ALL	SYL+BPE	BPE-ALL	SYL+BPE
	es→shp		**es→en**	
MONO-SYSTEM	37.62±1.87	**41.27***±0.54		
JOINT-SYSTEM	40.41±0.82	**41.74***±0.95		
O2M-SYSTEM	48.30	**51.25***	53.99	53.85
	shp→es			
MONO-SYSTEM	**33.37**±0.79	32.85*±1.22		
JOINT-SYSTEM	**34.55**±0.56	33.13*±0.75		

Table 6.11: chrF scores in the test subsets. For the first two settings, we run three experiments and present the mean and standard deviation. The latter only has one run due to resource constraints, and we report es–en scores as a reference. Syllabification (in SYL+BPE) is only applied on the Shipibo-Konibo side. (*) indicates a p-value ≤ 0.05 against the BPE baseline.

adding more encoder layers, as is done in character-models (Gao et al., 2020), may address this limitation, but we leave further investigation for future work.

Concerning the NMT systems trained with joint vocabulary models, we observe that using the JOINT-SYSTEM setting reduces the gap between BPE-ALL and SYL+BPE, probably due to the shared roots between the two languages (i.e., loanwords from Spanish into Shipibo-Konibo). Furthermore, we note that the impact of syllables is not minimised in a multilingual system (O2M-SYSTEM), where the performance for es→shp has drastically improved, and the other language-pair (es→en) retains a comparable result between BPE-ALL and SYL+BPE. These findings suggest that syllable-based subwords can be effectively combined with BPE pieces in larger multilingual NMT models

Regarding the experiments with MONO-SYSTEM, we note that they are comparable with the study of our previous section (see §6.5), where we tested several unsupervised and supervised morphological segmentation methods against BPE for machine translation. Our result with syllables in es→shp outperforms all other approaches, such as LMVR (Ataman et al., 2017), with a 38.99 chrF score. This indicates that syllables are a robust alternative to morphologically-aware methods when we are dealing with limited data and translating into a polysynthetic language.

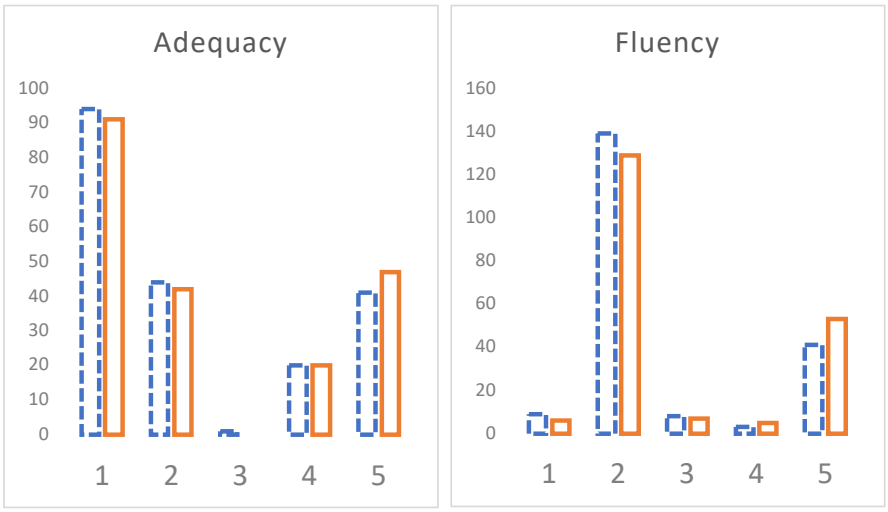


Figure 6.2: Adequacy and fluency scores (1-5) for 200 outputs of two approaches: BPE-ALL (dashed blue) and SYL+BPE (solid orange), from the best es→shp given by O2M-SYSTEM.

6.6.6 Human evaluation

We also conducted a small human evaluation of system outputs to compare the segmentation approaches (BPE-ALL versus SYL+BPE) for the best performing model (O2M-SYSTEM) using a 5-points scale for the adequacy and fluency of the Spanish→Shipibo-Konibo translation, which is the translation direction that benefited from the syllable segmentation. The annotation protocol and annotator's information are provided in Table 6.12.

6.6.6.1 Results

Figure 6.2 shows the scores annotated for adequacy and fluency, where we compare BPE-ALL and SYL+BPE for O2M-SYSTEM, which obtained the best performance for both segmentation approaches in es→shp. We observe that the adequacy is very poor for both systems (1-2), but there is an advantage for SYL+BPE in the smaller batch of highest adequacy (5), with 3% more of the total samples. Regarding fluency, both systems mostly obtain a low score (2), but there is a consistent advantage for SYL+BPE over BPE-ALL in the highest value (5), with 6.5% more of the total samples. The differences are very small to determine whether a segmentation works better than the other from human judgement, but they are consistent with the automatic evaluation provided previously. A larger sample, an extra annotator, or more robust systems could aid in clarifying other potential benefits.

Table 6.12: Annotation protocol and details about the annotator.

Annotation protocol

The expert received the source sentence in Spanish, the reference in Shipibo-Konibo, and an anonymized system output, which includes the baseline (BPE-ALL) and our syllable-based system (SYL+BPE). The expert received only 200 samples (per system, same entries) that were randomly selected and shuffled. They were asked to annotate **Adequacy** (Does the output sentence express the meaning of the reference?) from 1 to 5 (extremely bad to excellent), and **Fluency** (Is the output sentence easily readable and looks like a human-produced text?) from 1 to 5 as well. The following were the descriptions of the ratings as provided to the expert annotator in Spanish (translated into English here for convenience):

Adequacy The output sentence expresses the meaning of the reference.

1. Extremely bad: The original meaning is not contained at all.
2. Bad: Some words or phrases allow to guess the content.
3. Neutral.
4. Sufficiently good: The original meaning is understandable, but some parts are unclear or incorrect.
5. Excellent: The meaning of the output is the same as that of the reference.

Fluency The output sentence is easily readable and looks like a human-produced text.

1. Extremely bad: The output text does not belong to the target language.
2. Bad: The output sentence is hardly readable.
3. Neutral.
4. Sufficiently good: The output seems like a human-produced text in the target language, but contains weird mistakes.
5. Excellent: The output seems like a human-produced text in the target language, and is readable without issues.

About the annotator The annotator is a native speaker of Shipibo-Konibo, a certified and professional translator, and a bilingual teacher in Peru. The annotator has experience translating corpus for MT research and performing human evaluation for Spanish–Shipibo-Konibo. This expertise is almost unique for Shipibo-Konibo, and we could not identify a second annotator with the same expertise to obtain inter-annotation agreement.

6.6.7 Open-vocabulary language modelling with syllables

We also performed experiments with syllable-based segmentation on 21 languages for an open-vocabulary language modelling task. We compared the segmentation against characters, BPE subwords, Morfessor-based pieces, and annotated morphemes from Universal Dependency treebanks, and we found that syllables consistently outperform other baselines (measuring with a comparable character-level perplexity). As language modelling is not a relevant task in this thesis, we refer the reader to Oncevay et al. (2022b) for more details.

6.6.8 Limitations and opportunities

Syllables only cannot offer a universal solution to the subword segmentation problem for all languages, as the syllabification tools are language-dependent. Besides, the analysis should be extended to more scripts and morphological types. Furthermore, we do not encode any semantics in the syllable-vector space, with a few exceptions like in Korean (Choi et al., 2017).

Nevertheless, building a syllable splitter might require less effort than annotating morphemes to train a robust supervised tool. For instance, we can build a syllabification tool for English following five general rules from: https://www.howmanysyllables.com/divideintosyllables. Their implementation should take less effort than annotating morphological segmentation datasets or building a Finite-State-Transducer for morphological analysis. Although it is worth noting that the benefits of using syllables were only observed in languages with transparent orthography, which is not the case of English.

Specifically for machine translation, syllables could be useful when: (i) we are dealing with extremely low-resource data, which affects unsupervised word segmentation, (ii) we are translating into a language with a high synthesis, which has been observed as a factor that impacts on NMT performance, and (iii) we are working with a language with a transparent orthography. This is the scenario for several languages from the Americas, where their writing systems have been recently standardised for documentation and revitalisation purposes (Mager et al., 2018a), and some resources for machine translation have been compiled (Mager et al., 2021).

6.7 Conclusion

In this chapter, we have explored the extreme challenge of machine translation for languages with high synthesis and low resources. We have focused on endangered languages of the Americas, which often exhibit these traits, and aimed to develop machine translation resources for them. Through our experimentation, we have explored the efficacy of multilingual approaches, which have outperformed pairwise baselines, but are still constrained by the availability of parallel or monolingual data. Furthermore, while focusing on the subword segmentation methods, we identified that unsupervised methods like BPE are effective in these cases, in contrast to morphologically-supervised methods.

Our analysis has led us to propose a new syllable-based segmentation approach that is less data-dependent, which is an essential feature for these extremely low-resource languages, given that other unsupervised methods rely heavily on the availability of textual data. This new approach has shown positive outcomes in our case study, specifically for the translation direction into the polysynthetic and low-resource language, and including multilingual settings.

Chapter 7

Conclusions

In conclusion, this thesis aimed to investigate the impact of linguistic typology on neural machine translation performance. The research questions addressed in this study were: (1) whether a combined language representation can incorporate complementary sources of information from typological databases and NMT-learned representations, (2) whether the combined language space can improve multilingual machine translation tasks, (3) whether morphological typology variables of synthesis and fusion are relevant for machine translation performance, and (4) how machine translation performance can be improved for high-synthesis and low-resource languages.

To answer these questions, we proposed a method to compute language representations that encode typological features, specifically syntax variables, and showed that the combination of linguistic typology databases and pre-trained language embeddings leads to better results in computational typology tasks such as typological feature prediction and phylogenetic tree inference. The proposed language representations were also evaluated in multilingual machine translation tasks and demonstrated to perform as well as strong baselines across low-resource or high-resource languages, identifying relevant related languages to train a multilingual model and reduce negative transfer. Furthermore, we observed that they are an efficient alternative when we need to add new languages to a multilingual setting.

We then investigated the impact of morphological typology variables of fusion and synthesis on machine translation performance. We proposed methods to automatically and semi-automatically quantify these variables, and then analyse that there is a relationship between words with higher synthesis or fusion and translation performance in our case studies of Turkish and Spanish, respectively. This was consistent with a further analysis at the segment level. After identifying that a high value of synthesis

can hinder translation performance, we examined the extreme case of high-synthesis and low-resource languages for machine translation, and proposed a new segmentation approach using syllable units that improved translation performance against strong baselines and offered promising results in our case study.

Finally, in addition to our technical contributions, our work aimed to broaden the scope of attention given to understudied and endangered languages of the Americas, contributing towards the diversity of machine translation research. By developing new datasets and conducting experiments on them, we hope to advance the field of machine translation from different angles and expand the understanding of machine translation in such languages.

Overall, the thesis has demonstrated the importance of linguistic typology in machine translation and proposed various methods and analyses to incorporate typological knowledge into machine translation research.

7.1 Future work

While this thesis has made significant contributions to studying the impact of linguistic typology on machine translation, there are still many avenues for further research that could broaden and deepen our understanding of this field. Therefore, we describe some possible directions for future work:

- Our methodology for analysing the impact of morphological typology on machine translation, specifically the variables of synthesis and fusion, could be valuable to explore other important linguistic typological features. For example, it would be interesting to quantify word order features on a continuous scale, as some approaches have previously attempted (Futrell et al., 2015; Guzmán Naranjo and Becker, 2018), and analyse how much the word order divergence between the source and target language impacts NMT performance.

- The indices of synthesis and fusion could also be used as parameters for sampling in multilingual NMT models or curriculum learning (Platanios et al., 2019), where different measurements of complexity sort training samples. These variables could also be used for evaluating NMT performance in automatic metrics or quality estimation.

- We have shown that a less data-dependent segmentation approach based on syllable units is relevant in our low-resource and high-synthesis language case study.

It would be interesting to investigate adapting widespread unsupervised segmentation methods, such as BPE (Sennrich et al., 2016b) or unigram language modelling (Kudo, 2018), to include specific subwords, such as syllables or crafted affixes from lexicons.

- We also suggest focusing on highly fusional languages, where more morphological information is encoded in fewer morphemes. Making this encoded information explicit, such as in pseudo-tokens for the input sentence, as Goldwater and McClosky (2005) have done for statistical MT, or trying factored-based architectures (Sennrich and Haddow, 2016; Armengol-Estapé et al., 2021), could be a way to deal with this challenge.

Finally, we encourage the research community to continue developing machine translation resources for understudied and endangered languages. Diversity is crucial to extending our research in machine translation and ensuring that these languages are not left behind.

Bibliography

Agić, Ž. and Vulić, I. (2019). JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210, Florence, Italy. Association for Computational Linguistics.

Aharoni, R., Johnson, M., and Firat, O. (2019). Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

Alva, C. and Oncevay, A. (2017). Spell-checking based on syllabification and character-level graphs for a Peruvian agglutinative language. In *Proceedings of the First Workshop on Subword and Character Level Models in NLP*, pages 109–116, Copenhagen, Denmark. Association for Computational Linguistics.

Amrhein, C. and Sennrich, R. (2021). How suitable are subword segmentation strategies for translating non-concatenative morphology? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 689–705, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Anastasopoulos, A. (2019). A note on evaluating multilingual benchmarks. Available in: http://www.cs.cmu.edu/~aanastas/evaluating_multilingual.html.

Armengol-Estapé, J., Costa-jussà, M. R., and Escolano, C. (2021). Enriching the transformer with linguistic factors for low-resource machine translation. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 73–78, Held Online. INCOMA Ltd.

Ataman, D., Negri, M., Turchi, M., and Federico, M. (2017). Linguistically motivated

vocabulary reduction for neural machine translation from turkish to english. *The Prague Bulletin of Mathematical Linguistics*, 108:331 – 342.

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In Bengio, Y. and LeCun, Y., editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Baker, M. C. (1996). *The polysynthesis parameter*. Oxford University Press.

Bapna, A., Caswell, I., Kreutzer, J., Firat, O., van Esch, D., Siddhant, A., Niu, M., Baljekar, P., Garcia, X., Macherey, W., Breiner, T., Axelrod, V., Riesa, J., Cao, Y., Chen, M. X., Macherey, K., Krikun, M., Wang, P., Gutkin, A., Shah, A., Huang, Y., Chen, Z., Wu, Y., and Hughes, M. (2022). Building machine translation systems for the next thousand languages.

Bapna, A. and Firat, O. (2019). Simple, scalable adaptation for neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1538–1548, Hong Kong, China. Association for Computational Linguistics.

Barrault, L., Biesialska, M., Bojar, O., Costa-jussà, M. R., Federmann, C., Graham, Y., Grundkiewicz, R., Haddow, B., Huck, M., Joanis, E., Kocmi, T., Koehn, P., Lo, C.-k., Ljubešić, N., Monz, C., Morishita, M., Nagata, M., Nakazawa, T., Pal, S., Post, M., and Zampieri, M. (2020). Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

Barrault, L., Bojar, O., Costa-jussà, M. R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Koehn, P., Malmasi, S., Monz, C., Müller, M., Pal, S., Post, M., and Zampieri, M. (2019). Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Bawden, R., Birch, A., Dobreva, R., Oncevay, A., Miceli Barone, A. V., and Williams, P. (2020). The University of Edinburgh's English-Tamil and English-Inuktitut submissions to the WMT20 news translation task. In *Proceedings of the Fifth Confer-*

ence on Machine Translation, pages 92–99, Online. Association for Computational Linguistics.

Bjerva, J. and Augenstein, I. (2018a). From phonology to syntax: Unsupervised linguistic typology at different levels with language embeddings. 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 907–916, New Orleans, Louisiana. Association for Computational Linguistics.

Bjerva, J. and Augenstein, I. (2018b). Tracking typological traits of uralic languages in distributed language representations. In *Proceedings of the Fourth International Workshop on Computational Linguistics of Uralic Languages*, pages 76–86, Helsinki, Finland. Association for Computational Linguistics.

Bjerva, J., Kementchedjhieva, Y., Cotterell, R., and Augenstein, I. (2019a). A probabilistic generative model of linguistic typology. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1529–1540, Minneapolis, Minnesota. Association for Computational Linguistics.

Bjerva, J., Östling, R., Veiga, M. H., Tiedemann, J., and Augenstein, I. (2019b). What do language representations really represent? *Computational Linguistics*, 45(2):381–389.

Bojar, O., Buck, C., Callison-Burch, C., Federmann, C., Haddow, B., Koehn, P., Monz, C., Post, M., Soricut, R., and Specia, L. (2013). Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 1–44, Sofia, Bulgaria. Association for Computational Linguistics.

Bojar, O., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Koehn, P., and Monz, C. (2018). Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels. Association for Computational Linguistics.

Borgwaldt, S. R., Hellwig, F. M., and De Groot, A. M. (2005). Onset entropy matters–letter-to-phoneme mappings in seven languages. *Reading and Writing*, 18(3):211–229.

Bostrom, K. and Durrett, G. (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.

Bustamante, G., Oncevay, A., and Zariquiey, R. (2020). No data to crawl? monolingual corpus creation from PDF files of truly low-resource languages in Peru. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2914–2923, Marseille, France. European Language Resources Association.

Caswell, I., Chelba, C., and Grangier, D. (2019). Tagged back-translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 53–63, Florence, Italy. Association for Computational Linguistics.

Cettolo, M., Girardi, C., and Federico, M. (2012). WIT^3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, pages 261–268, Trento, Italy.

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Choi, S., Kim, T., Seol, J., and Lee, S.-g. (2017). A syllable-based technique for word embeddings of Korean words. In *Proceedings of the First Workshop on Subword and Character Level Models in NLP*, pages 36–40, Copenhagen, Denmark. Association for Computational Linguistics.

Collins, C. and Kayne, R. (2011). *Syntactic Structures of the World's Languages*. New York University.

Comrie, B. (1989). *Language universals and linguistic typology: Syntax and morphology*. University of Chicago press.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting*

of the Association for Computational Linguistics, pages 8440–8451, Online. Association for Computational Linguistics.

Conneau, A., Rinott, R., Lample, G., Williams, A., Bowman, S., Schwenk, H., and Stoyanov, V. (2018). XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Creutz, M. and Lagus, K. (2002). Unsupervised discovery of morphemes. In *Proceedings of the ACL-02 Workshop on Morphological and Phonological Learning*, pages 21–30. Association for Computational Linguistics.

Daiber, J., Stanojević, M., and Sima'an, K. (2016). Universal reordering via linguistic typology. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3167–3176, Osaka, Japan. The COLING 2016 Organizing Committee.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Dhillon, P. S., Foster, D. P., and Ungar, L. H. (2015). Eigenwords: Spectral word embeddings. *The Journal of Machine Learning Research*, 16(1):3035–3078.

Dou, Z.-Y. and Neubig, G. (2021). Word alignment by fine-tuning embeddings on parallel corpora. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2112–2128, Online. Association for Computational Linguistics.

Dryer, M. S. and Haspelmath, M., editors (2013). *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Dumais, S. T. (2004). Latent semantic analysis. *Annual review of information science and technology*, 38(1):188–230.

Dyen, I., Kruskal, J. B., and Black, P. (1992). An indoeuropean classification: A lexicostatistical experiment. *Transactions of the American Philosophical society*, 82(5):iii–132.

Ebrahimi, A., Mager, M., Oncevay, A., Chaudhary, V., Chiruzzo, L., Fan, A., Ortega, J., Ramos, R., Rios, A., Meza Ruiz, I. V., Giménez-Lugo, G., Mager, E., Neubig, G., Palmer, A., Coto-Solano, R., Vu, T., and Kann, K. (2022). AmericasNLI: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6279–6299, Dublin, Ireland. Association for Computational Linguistics.

Faruqui, M. and Dyer, C. (2014). Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 462–471, Gothenburg, Sweden. Association for Computational Linguistics.

Futrell, R., Mahowald, K., and Gibson, E. (2015). Quantifying word order freedom in dependency corpora. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 91–100, Uppsala, Sweden. Uppsala University, Uppsala, Sweden.

Galarreta, A.-P., Melgar, A., and Oncevay, A. (2017). Corpus creation and initial SMT experiments between Spanish and Shipibo-konibo. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 238–244, Varna, Bulgaria. INCOMA Ltd.

Gao, Y., Nikolov, N. I., Hu, Y., and Hahnloser, R. H. (2020). Character-level translation with self-attention. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1591–1604, Online. Association for Computational Linguistics.

Goldwater, S. and McClosky, D. (2005). Improving statistical MT through morphological analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 676–683, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Gómez Montoya, H. E., Rojas, K. D. R., and Oncevay, A. (2019). A continuous improvement framework of machine translation for Shipibo-konibo. In *Proceedings*

of the 2nd Workshop on Technologies for MT of Low Resource Languages, pages 17–23, Dublin, Ireland. European Association for Machine Translation.

Greenberg, J. H. (1960). A quantitative approach to the morphological typology of language. *International journal of American linguistics*, 26(3):178–194.

Grönroos, S.-A., Virpioja, S., Smit, P., and Kurimo, M. (2014). Morfessor flatcat: An hmm-based method for unsupervised and semi-supervised learning of morphology. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1177–1185.

Guzmán, F., Chen, P.-J., Ott, M., Pino, J., Lample, G., Koehn, P., Chaudhary, V., and Ranzato, M. (2019). The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6098–6111, Hong Kong, China. Association for Computational Linguistics.

Guzmán Naranjo, M. and Becker, L. (2018). Quantitative word order typology with UD. In *Proceedings of the 17th International Workshop on Treebanks and Linguistic Theories (TLT 2018)*, pages 91–104.

Haddow, B., Bawden, R., Miceli Barone, A. V., Helcl, J., and Birch, A. (2022). Survey of low-resource machine translation. *Computational Linguistics*, 48(3):673–732.

Haghighi, A., Liang, P., Berg-Kirkpatrick, T., and Klein, D. (2008). Learning bilingual lexicons from monolingual corpora. In *Proceedings of ACL-08: HLT*, pages 771–779, Columbus, Ohio. Association for Computational Linguistics.

Hardoon, D. R., Szedmak, S., and Shawe-Taylor, J. (2004). Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16(12):2639–2664.

Harris, Z. S. (1951). *Methods in structural linguistics.* University of Chicago Press.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Huarcaya Taquiri, D. (2020). Traducción automática neuronal para lengua nativa peruana. Bachelor's thesis, Universidad Peruana Unión.

Joanis, E., Knowles, R., Kuhn, R., Larkin, S., Littell, P., Lo, C.-k., Stewart, D., and Micher, J. (2020). The Nunavut Hansard Inuktitut–English parallel corpus 3.0 with preliminary machine translation results. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2562–2572, Marseille, France. European Language Resources Association.

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., Hughes, M., and Dean, J. (2017). Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Joshi, P., Santy, S., Budhiraja, A., Bali, K., and Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Junczys-Dowmunt, M., Grundkiewicz, R., Dwojak, T., Hoang, H., Heafield, K., Neckermann, T., Seide, F., Germann, U., Aji, A. F., Bogoychev, N., Martins, A. F. T., and Birch, A. (2018). Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Jurafsky, D. and Martin, J. H. (2019). *Speech and Language Processing*. Pearson Education, Inc., 3rd edition.

Kann, K., Ebrahimi, A., Mager, M., Oncevay, A., Ortega, J. E., Rios, A., Fan, A., Gutierrez-Vasques, X., Chiruzzo, L., Giménez-Lugo, G. A., Ramos, R., Meza Ruiz, I. V., Mager, E., Chaudhary, V., Neubig, G., Palmer, A., Coto-Solano, R., and Vu, N. T. (2022). AmericasNLI: Machine translation and natural language inference systems for indigenous languages of the americas. *Frontiers in Artificial Intelligence*, 5.

Kann, K., Mager Hois, J. M., Meza-Ruiz, I. V., and Schütze, H. (2018). Fortification of neural morphological segmentation models for polysynthetic minimal-resource languages. 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 47–57, New Orleans, Louisiana. Association for Computational Linguistics.

Kim, Y., Jernite, Y., Sontag, D., and Rush, A. M. (2016). Character-aware neural language models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, pages 2741–2749. AAAI Press.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In Bengio, Y. and LeCun, Y., editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Knowles, R., Stewart, D., Larkin, S., and Littell, P. (2020). NRC systems for the 2020 Inuktitut-English news translation task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 156–170, Online. Association for Computational Linguistics.

Kocmi, T. (2020). CUNI submission for the Inuktitut language in WMT news 2020. In *Proceedings of the Fifth Conference on Machine Translation*, pages 171–174, Online. Association for Computational Linguistics.

Kocmi, T. and Bojar, O. (2018). Trivial transfer learning for low-resource neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 244–252, Belgium, Brussels. Association for Computational Linguistics.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Koehn, P. and Knowles, R. (2017). Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Kreutzer, J., Caswell, I., Wang, L., Wahab, A., van Esch, D., Ulzii-Orshikh, N., Tapo, A., Subramani, N., Sokolov, A., Sikasote, C., Setyawan, M., Sarin, S., Samb, S., Sagot, B., Rivera, C., Rios, A., Papadimitriou, I., Osei, S., Suarez, P. O., Orife, I., Ogueji, K., Rubungo, A. N., Nguyen, T. Q., Müller, M., Müller, A., Muhammad, S. H., Muhammad, N., Mnyakeni, A., Mirzakhalov, J., Matangira, T., Leong, C., Lawson, N., Kudugunta, S., Jernite, Y., Jenny, M., Firat, O., Dossou, B. F. P., Dlamini, S., de Silva, N., Çabuk Ballı, S., Biderman, S., Battisti, A., Baruwa, A., Bapna, A., Baljekar, P., Azime, I. A., Awokoya, A., Ataman, D., Ahia, O., Ahia,

O., Agrawal, S., and Adeyemi, M. (2022). Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Kudo, T. (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.

Kudo, T. and Richardson, J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Kudugunta, S., Bapna, A., Caswell, I., and Firat, O. (2019). Investigating multilingual NMT representations at scale. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1565–1575, Hong Kong, China. Association for Computational Linguistics.

Kunchukuttan, A. and Bhattacharyya, P. (2016). Orthographic syllable as basic unit for SMT between related languages. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1912–1917, Austin, Texas. Association for Computational Linguistics.

Lai, W., Zhao, X., and Bao, W. (2018). Tibetan-Chinese neural machine translation based on syllable segmentation. In *Proceedings of the AMTA 2018 Workshop on Technologies for MT of Low Resource Languages (LoResMT 2018)*, pages 21–29, Boston, MA. Association for Machine Translation in the Americas.

Libovický, J., Schmid, H., and Fraser, A. (2022). Why don't people use character-level machine translation? In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2470–2485, Dublin, Ireland. Association for Computational Linguistics.

Lin, Y.-H., Chen, C.-Y., Lee, J., Li, Z., Zhang, Y., Xia, M., Rijhwani, S., He, J., Zhang, Z., Ma, X., Anastasopoulos, A., Littell, P., and Neubig, G. (2019). Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of*

the Association for Computational Linguistics, pages 3125–3135, Florence, Italy. Association for Computational Linguistics.

Lin, Z., Wu, L., Wang, M., and Li, L. (2021). Learning language specific sub-network for multilingual machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 293–305, Online. Association for Computational Linguistics.

Littell, P., Mortensen, D. R., Lin, K., Kairis, K., Turner, C., and Levin, L. (2017). URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14, Valencia, Spain. Association for Computational Linguistics.

Loriot, J., Lauriault, E., and Day, D. (1993). *Diccionario Shipibo-Castellano*. Instituto Lingüístico de Verano.

Mager, M., Çetinoğlu, Ö., and Kann, K. (2020). Tackling the low-resource challenge for canonical segmentation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5237–5250, Online. Association for Computational Linguistics.

Mager, M., Gutierrez-Vasques, X., Sierra, G., and Meza-Ruiz, I. (2018a). Challenges of language technologies for the indigenous languages of the Americas. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 55–69, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Mager, M., Mager, E., Medina-Urrea, A., Meza Ruiz, I. V., and Kann, K. (2018b). Lost in translation: Analysis of information loss during machine translation between polysynthetic and fusional languages. In *Proceedings of the Workshop on Computational Modeling of Polysynthetic Languages*, pages 73–83, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Mager, M., Oncevay, A., Ebrahimi, A., Ortega, J., Rios, A., Fan, A., Gutierrez-Vasques, X., Chiruzzo, L., Giménez-Lugo, G., Ramos, R., Meza Ruiz, I. V., Coto-Solano, R., Palmer, A., Mager-Hois, E., Chaudhary, V., Neubig, G., Vu, N. T., and Kann, K. (2021). Findings of the AmericasNLP 2021 shared task on open machine translation for indigenous languages of the Americas. In *Proceedings of the First*

Workshop on Natural Language Processing for Indigenous Languages of the Americas, pages 202–217, Online. Association for Computational Linguistics.

Mager, M., Oncevay, A., Mager, E., Kann, K., and Vu, T. (2022). BPE vs. morphological segmentation: A case study on machine translation of four polysynthetic languages. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 961–971, Dublin, Ireland. Association for Computational Linguistics.

Malaviya, C., Neubig, G., and Littell, P. (2017). Learning language representations for typology prediction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2529–2535, Copenhagen, Denmark. Association for Computational Linguistics.

Marjou, X. (2021). OTEANN: Estimating the transparency of orthographies with an artificial neural network. In *Proceedings of the Third Workshop on Computational Typology and Multilingual NLP*, pages 1–9, Online. Association for Computational Linguistics.

McCarthy, A. D., Kirov, C., Grella, M., Nidhi, A., Xia, P., Gorman, K., Vylomova, E., Mielke, S. J., Nicolai, G., Silfverberg, M., Arkhangelskiy, T., Krizhanovsky, N., Krizhanovsky, A., Klyachko, E., Sorokin, A., Mansfield, J., Ernštreits, V., Pinter, Y., Jacobs, C. L., Cotterell, R., Hulden, M., and Yarowsky, D. (2020). UniMorph 3.0: Universal Morphology. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3922–3931, Marseille, France. European Language Resources Association.

Mihas, E. (2011). *Añaani katonkosatzi parenini, El idioma del alto Perené.* WI:Clarks Graphics.

Mirzakhalov, J., Babu, A., Ataman, D., Kariev, S., Tyers, F., Abduraufov, O., Hajili, M., Ivanova, S., Khaytbaev, A., Laverghetta Jr., A., Moydinboyev, B., Onal, E., Pulatova, S., Wahab, A., Firat, O., and Chellappan, S. (2021). A large-scale study of machine translation in Turkic languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5876–5890, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mithun, M. (1986). On the nature of noun incorporation. *Language*, 62(1):32–37.

Murawaki, Y. (2015). Continuous space representations of linguistic typology and their application to phylogenetic inference. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 324–334, Denver, Colorado. Association for Computational Linguistics.

Murawaki, Y. (2017). Diachrony-aware induction of binary latent representations from typological features. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 451–461, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Murawaki, Y. (2018). Analyzing correlated evolution of multiple features using latent representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4371–4382, Brussels, Belgium. Association for Computational Linguistics.

Murthy, R., Kunchukuttan, A., and Bhattacharyya, P. (2019). Addressing word-order divergence in multilingual neural machine translation for extremely low resource languages. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3868–3873, Minneapolis, Minnesota. Association for Computational Linguistics.

Myint Oo, T., Kyaw Thu, Y., and Mar Soe, K. (2019). Neural machine translation between Myanmar (Burmese) and Rakhine (Arakanese). In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 80–88, Ann Arbor, Michigan. Association for Computational Linguistics.

Naveed, T., Siddiqui, I. S., and Ahmed, S. (2005). Parallel needleman-wunsch algorithm for grid. In *Proceedings of the PAK-US International Symposium on High Capacity Optical Networks and Enabling Technologies (HONET 2005), Islamabad, Pakistan*.

Nekoto, W., Marivate, V., Matsila, T., Fasubaa, T., Fagbohungbe, T., Akinola, S. O., Muhammad, S., Kabongo Kabenamualu, S., Osei, S., Sackey, F., Niyongabo, R. A., Macharm, R., Ogayo, P., Ahia, O., Berhe, M. M., Adeyemi, M., Mokgesi-Selinga, M., Okegbemi, L., Martinus, L., Tajudeen, K., Degila, K., Ogueji, K., Siminyu, K., Kreutzer, J., Webster, J., Ali, J. T., Abbott, J., Orife, I., Ezeani, I., Dangana,

I. A., Kamper, H., Elsahar, H., Duru, G., Kioko, G., Espoir, M., van Biljon, E., Whitenack, D., Onyefuluchi, C., Emezue, C. C., Dossou, B. F. P., Sibanda, B., Bassey, B., Olabiyi, A., Ramkilowan, A., Öktem, A., Akinfaderin, A., and Bashir, A. (2020). Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, Online. Association for Computational Linguistics.

Nelder, J. A. and Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384.

Nerbonne, J., Kleiweg, P., Heeringa, W., and Manni, F. (2008). Projecting dialect distances to geography: Bootstrap clustering vs. noisy clustering. In Preisach, C., Burkhardt, H., Schmidt-Thieme, L., and Decker, R., editors, *Data Analysis, Machine Learning and Applications*, pages 647–654, Berlin, Heidelberg. Springer Berlin Heidelberg.

Neubig, G. and Hu, J. (2018). Rapid adaptation of neural machine translation to new languages. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 875–880, Brussels, Belgium. Association for Computational Linguistics.

Nordhoff, S. and Hammarström, H. (2012). Glottolog/Langdoc:Increasing the visibility of grey literature for low-density languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3289–3294, Istanbul, Turkey. European Language Resources Association (ELRA).

O'Horan, H., Berzak, Y., Vulić, I., Reichart, R., and Korhonen, A. (2016). Survey on the use of typological information in natural language processing. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1297–1308, Osaka, Japan. The COLING 2016 Organizing Committee.

Oncevay, A. (2021). Peru is multilingual, its machine translation should be too? In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 194–201, Online. Association for Computational Linguistics.

Oncevay, A., Ataman, D., Van Berkel, N., Haddow, B., Birch, A., and Bjerva, J. (2022a). Quantifying synthesis and fusion and their impact on machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1308–1321, Seattle, United States. Association for Computational Linguistics.

Oncevay, A., Haddow, B., and Birch, A. (2020). Bridging linguistic typology and multilingual machine translation with multi-view language representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2391–2406, Online. Association for Computational Linguistics.

Oncevay, A., Rivas Rojas, K. D., Chavez Sanchez, L. K., and Zariquiey, R. (2022b). Revisiting syllables in language modelling and their application on low-resource machine translation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4258–4267, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Ortega, J., Castro-Mamani, R. A., and Montoya Samame, J. R. (2020a). Overcoming resistance: The normalization of an Amazonian tribal language. In *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*, pages 1–13, Suzhou, China. Association for Computational Linguistics.

Ortega, J. E., Mamani, R. C., and Cho, K. (2020b). Neural machine translation with a polysynthetic low resource language. *Machine Translation*, 34(4):325–346.

Osborne, D., Narayan, S., and Cohen, S. B. (2016). Encoding prior knowledge with eigenword embeddings. *Transactions of the Association for Computational Linguistics*, 4:417–430.

Östling, R. and Tiedemann, J. (2017). Continuous multilinguality with language vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 644–649, Valencia, Spain. Association for Computational Linguistics.

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 the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Pawlik, M. and Augsten, N. (2015). Efficient computation of the tree edit distance. *ACM Transactions on Database Systems (TODS)*, pages 3:1–3:40.

Pawlik, M. and Augsten, N. (2016). Tree edit distance: Robust and memory-efficient. *Information Systems*, 56:157–173.

Payne, T. E. (2017). Morphological typology. In *The Cambridge Handbook of Linguistic Typology*, pages 78–94. Cambridge University Press.

Pereira-Noriega, J., Mercado-Gonzales, R., Melgar, A., Sobrevilla-Cabezudo, M., and Oncevay-Marcos, A. (2017). Ship-lemmatagger: Building an nlp toolkit for a peruvian native language. In Ekštein, K. and Matoušek, V., editors, *Text, Speech, and Dialogue*, pages 473–481, Cham. Springer International Publishing.

Pires, T., Schlinger, E., and Garrette, D. (2019). How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Platanios, E. A., Stretcu, O., Neubig, G., Poczos, B., and Mitchell, T. (2019). Competence-based curriculum learning for neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1162–1172, Minneapolis, Minnesota. Association for Computational Linguistics.

Ponti, E. M., O'Horan, H., Berzak, Y., Vulić, I., Reichart, R., Poibeau, T., Shutova, E., and Korhonen, A. (2019). Modeling language variation and universals: A survey on typological linguistics for natural language processing. *Computational Linguistics*, 45(3):559–601.

Ponti, E. M., Reichart, R., Korhonen, A., and Vulić, I. (2018). Isomorphic transfer of syntactic structures in cross-lingual NLP. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*,

pages 1531–1542, Melbourne, Australia. Association for Computational Linguistics.

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.

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.

Prokopidis, P., Papavassiliou, V., and Piperidis, S. (2016). Parallel Global Voices: a Collection of Multilingual Corpora with Citizen Media Stories. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 900–905, Portorož, Slovenia. European Language Resources Association (ELRA).

Qi, Y., Sachan, D., Felix, M., Padmanabhan, S., and Neubig, G. (2018). When and why are pre-trained word embeddings useful 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 2 (Short Papers)*, pages 529–535, New Orleans, Louisiana. Association for Computational Linguistics.

Rabinovich, E., Ordan, N., and Wintner, S. (2017). Found in translation: Reconstructing phylogenetic language trees from translations. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 530–540, Vancouver, Canada. Association for Computational Linguistics.

Raghu, M., Gilmer, J., Yosinski, J., and Sohl-Dickstein, J. (2017). SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems 30*, pages 6076–6085. Curran Associates, Inc.

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.

Roest, C., Edman, L., Minnema, G., Kelly, K., Spenader, J., and Toral, A. (2020). Machine translation for English–Inuktitut with segmentation, data acquisition and pre-training. In *Proceedings of the Fifth Conference on Machine Translation*, pages 274–281, Online. Association for Computational Linguistics.

Romano, R. and Richer, S. (2008). Ñaantsipeta asháninkaki birakochaki. Available in: www.lengamer.org/publicaciones/diccionarios/.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

Sak, H., Güngör, T., and Saraçlar, M. (2008). Turkish language resources: Morphological parser, morphological disambiguator and web corpus. In *International Conference on Natural Language Processing*, pages 417–427. Springer.

Sapir, E. (1921). *Types of linguistic structure.* Harcourt Brace & Company.

Schwartz, L., Tyers, F. M., Levin, L. S., Kirov, C., Littell, P., Lo, C., Prud'hommeaux, E., Park, H. H., Steimel, K., Knowles, R., Micher, J., Strunk, L., Liu, H., Haley, C., Zhang, K. J., Jimerson, R., Andriyanets, V., Muis, A. O., Otani, N., Park, J. H., and Zhang, Z. (2020). Neural polysynthetic language modelling. *CoRR*, abs/2005.05477.

See, A., Liu, P. J., and Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Sennrich, R., Firat, O., Cho, K., Birch, A., Haddow, B., Hitschler, J., Junczys-Dowmunt, M., Läubli, S., Miceli Barone, A. V., Mokry, J., and Nadejde, M. (2017). Nematus: a toolkit for neural machine translation. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 65–68, Valencia, Spain. Association for Computational Linguistics.

Sennrich, R. and Haddow, B. (2016). Linguistic input features improve neural machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 1, Research Papers*, pages 83–91, Berlin, Germany. Association for Computational Linguistics.

Sennrich, R., Haddow, B., and Birch, A. (2016a). Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Sennrich, R., Haddow, B., and Birch, A. (2016b). 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.

Serva, M. and Petroni, F. (2008). Indo-European languages tree by Levenshtein distance. *EPL (Europhysics Letters)*, 81(6):68005.

ShweSin, Y., Pa, W. P., and Soe, K. (2019). UCSYNLP-lab machine translation systems for WAT 2019. In *Proceedings of the 6th Workshop on Asian Translation*, pages 195–199, Hong Kong, China. Association for Computational Linguistics.

Siddhant, A., Bapna, A., Firat, O., Cao, Y., Chen, M. X., Caswell, I., and Garcia, X. (2022). Towards the next 1000 languages in multilingual machine translation: Exploring the synergy between supervised and self-supervised learning.

Simons, G. F. and Fenning, C. D., editors (2019). *Ethnologue: Languages of the World. Twenty-second edition*. Dallas Texas: SIL international. Online version: http://www.ethnologue.com.

Smit, P., Virpioja, S., Grönroos, S.-A., and Kurimo, M. (2014). Morfessor 2.0: Toolkit for statistical morphological segmentation. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 21–24, Gothenburg, Sweden. Association for Computational Linguistics.

Spiegler, S. and Monson, C. (2010). EMMA: A novel evaluation metric for morphological analysis. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1029–1037, Beijing, China. Coling 2010 Organizing Committee.

Steiner, P. (2016). Refurbishing a morphological database for German. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*

(LREC'16), pages 1103–1108, Portorož, Slovenia. European Language Resources Association (ELRA).

Steiner, P. (2017). Merging the trees - building a morphological treebank for German from two resources. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories*, pages 146–160, Prague, Czech Republic.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 3104–3112, Cambridge, MA, USA. MIT Press.

Tan, X., Chen, J., He, D., Xia, Y., QIN, T., and Liu, T.-Y. (2019). Multilingual neural machine translation with language clustering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 963–973, Hong Kong, China. Association for Computational Linguistics.

Taulé, M., Martí, M. A., and Recasens, M. (2008). AnCora: Multilevel annotated corpora for Catalan and Spanish. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Thorndike, R. L. (1953). Who belongs in the family? *Psychometrika*, 18(4):267–276.

Tiedemann, J. (2012). Parallel data, tools and interfaces in opus. In Chair), N. C. C., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Tsvetkov, Y., Sitaram, S., Faruqui, M., Lample, G., Littell, P., Mortensen, D., Black, A. W., Levin, L., and Dyer, C. (2016). Polyglot neural language models: A case study in cross-lingual phonetic representation learning. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1357–1366, San Diego, California. Association for Computational Linguistics.

Vania, C. and Lopez, A. (2017). From characters to words to in between: Do we capture morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2016–2027, Vancouver, Canada. Association for Computational Linguistics.

Vasquez, A., Ego Aguirre, R., Angulo, C., Miller, J., Villanueva, C., Agić, Ž., Zariquiey, R., and Oncevay, A. (2018). Toward Universal Dependencies for Shipibo-konibo. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 151–161, Brussels, Belgium. Association for Computational Linguistics.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Wang, X. and Neubig, G. (2019). Target conditioned sampling: Optimizing data selection for multilingual neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5823–5828, Florence, Italy. Association for Computational Linguistics.

Wang, X., Tsvetkov, Y., and Neubig, G. (2020). Balancing training for multilingual neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8526–8537, Online. Association for Computational Linguistics.

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244.

Xu, H., Kodner, J., Marcus, M., and Yang, C. (2020). Modeling morphological typology for unsupervised learning of language morphology. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6672–6681, Online. Association for Computational Linguistics.

Zariquiey, R., Oncevay, A., and Vera, J. (2022). CLD^2 language documentation meets natural language processing for revitalising endangered languages. In *Proceedings*

of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages, pages 20–30, Dublin, Ireland. Association for Computational Linguistics.

Zhang, B., Williams, P., Titov, I., and Sennrich, R. (2020). Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics.

Zhou, C., Ma, X., Hu, J., and Neubig, G. (2019). Handling syntactic divergence in low-resource machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1388–1394, Hong Kong, China. Association for Computational Linguistics.

Ziegler, J. C., Bertrand, D., Tóth, D., Csépe, V., Reis, A., Faísca, L., Saine, N., Lyytinen, H., Vaessen, A., and Blomert, L. (2010). Orthographic depth and its impact on universal predictors of reading: A cross-language investigation. *Psychological science*, 21(4):551–559.

Zingler, T. (2018). Reduction without fusion: Grammaticalization and wordhood in Turkish. *Folia Linguistica*, 52(2):415–447.

Zoph, B., Yuret, D., May, J., and Knight, K. (2016). Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.