Low Resource NLP for Polysynthetic Languages: Morphological Segmentation and Machine Translation

Von der Fakultät Informatik, Elektrotechnik und Informationstechnik der Universität Stuttgart zur Erlangung der Würde eines Doktors der Naturwissenschaften (Dr. rer. nat.) genehmigte Abhandlung.

Vorgelegt von

Jesus Manuel Mager Hois

aus Mexico.

Hauptberichter	Prof. Dr. Ngoc Thang Vu
Mitberichter	Prof. Dr. Katharina von der Wense
Mitberichter	Prof. Dr. Alexander Fraser

Tag der mündlichen Prüfung: 16. Februar 2024

Institut für Maschinelle Sprachverarbeitung
der Universität Stuttgart

2024

Erklärung (Statement of Authorship)
Hiermit erkläre ich, dass ich die vorliegende Arbeit selbständig verfasst habe und dabei keine andere als die angegebene Literatur verwendet habe. Alle Zitate und sinngemäßen Entlehnungen sind als solche unter genauer Angabe der Quelle gekennzeichnet.

Zusammenfassung

Diese Dissertation untersucht die maschinelle Sprachverarbeitung (NLP) für polysynthetische indigene Sprachen Amerikas in extrem ressourcenarmen Szenarien, mit Schwerpunkt auf morphologischer Segmentierung und maschineller Übersetzung. Polysynthetische Sprachen stellen aufgrund ihrer komplexen Morphologie besondere Herausforderungen dar, da ein einzelnes Wort Informationen kodieren kann, die in anderen Sprachen mehrere Wörter erfordern würden. Die Arbeit untersucht drei Hauptforschungsfragen: (1) die Machbarkeit der Modellierung polysynthetischer Morphologie mittels neuronaler Netze in ressourcenarmen Umgebungen, (2) den Einfluss morphologischer Segmentierung auf die Qualität maschineller Übersetzung und (3) Ansätze zur Handhabung und Nutzung von Code-Switching-Phänomenen in diesen Sprachen.

Die Forschung führt neue Datensätze für morphologische Segmentierung und maschinelle Übersetzung für verschiedene indigene Sprachen ein, darunter Rarámuri, Tepehua und Shipibo-Konibo. Durch umfangreiche Experimente zeigt die Studie, dass neuronale Sequence-to-Sequence-Modelle auch mit begrenzten Trainingsdaten effektiv morphologische Segmentierung durchführen können. Für die maschinelle Übersetzung zeigt die Forschung, dass unüberwachte morphologische Segmentierungsmethoden bei polysynthetischen Sprachen bessere Ergebnisse erzielen als Standardansätze wie Byte-Pair-Encoding (BPE). Die Arbeit untersucht auch Code-Switching, führt eine neuartige Sprachidentifikationsaufgabe auf Subwortebene ein und zeigt, dass synthetische Code-Switching-Daten die Leistung der maschinellen Übersetzung verbessern können.

Darüber hinaus behandelt die Dissertation ethische Überlegungen bei der Entwicklung von Sprachtechnologien für indigene Gemeinschaften, basierend auf Interviews mit Sprachaktivisten und Gemeindeführern. Die Ergebnisse unterstreichen die Bedeutung der Einbindung der Gemeinschaft und der Datensouveränität bei der Entwicklung indigener Sprachtechnologie. Insgesamt erweitert diese Arbeit unser Verständnis computergestützter Ansätze zur Verarbeitung polysynthetischer Sprachen und schafft gleichzeitig Grundlagen für ethisch verantwortungsvolle Forschung in diesem Bereich.

Abstract

This dissertation investigates Natural Language Processing (NLP) for polysynthetic indigenous languages of the Americas in extremely low-resource scenarios, with a focus on morphological segmentation and machine translation. Polysynthetic languages present unique challenges due to their complex morphology, where a single word can encode information that would require multiple words in other languages. The work examines three main research questions: (1) the feasibility of modeling polysynthetic morphology using neural networks in low-resource settings, (2) the impact of morphological segmentation on machine translation performance, and (3) approaches for handling and leveraging code-switching phenomena in these languages.

The research introduces new datasets for morphological segmentation and machine translation for several indigenous languages including Rarámuri, Tepehua, and Shipibo-Konibo. Through extensive experiments, the study demonstrates that neural sequence-to-sequence models can effectively perform morphological segmentation even with limited training data. For machine translation, the research shows that unsupervised morphological segmentation methods outperform standard approaches like byte-pair encoding (BPE) for polysynthetic languages. The work also examines code-switching, introducing a novel sub-word level language identification task and demonstrating that synthetic code-switched data can improve machine translation performance.

Additionally, the dissertation addresses ethical considerations in developing language technologies for indigenous communities, based on interviews with language activists and community leaders. The findings emphasize the importance of community involvement and data sovereignty in indigenous language technology development. Overall, this work advances our understanding of computational approaches for processing polysynthetic languages while establishing foundations for ethically responsible research in this domain.

To all the people who dare to create a better world every day, because a better world is possible. To all the people of my community: Wixarika, and to my village Zoquipan. To my family and especially to my Mother who with her example and strength, has taught me the way. To my friends who have supported me in each step and to all my coauthors and advisors who taught me how to research. Pampariyuzi!

Contents

1. Introduction

In recent years, the usage of Natural Language Processing (NLP) techniques in real-world applications has impacted the world in all fields. Technologies like machine translation (MT) have achieved levels that make them able to be used daily to connect speakers from multiple languages around the world. In theory, such a system will allow us to break through the language barriers and make a mutually intelligible multilingual world possible. However, these advances are not yet available for most world languages, even if some current neural models claim to be language-independent (Feng et al., 2016; Al-Badrashiny & Diab, 2016b).

In this work, we explore the complexities of the processing of morphologically rich languages (more precisely, polysynthetic languages) of the . First, we investigate the modeling capabilities of neural networks to perform segmentation of polysynthetic languages. Then, we apply this knowledge to study the MT task. In all cases, we work with the real scenario of scarce data, creating or using existing real, low-resourced, and noisy datasets. We also consider that indigenous languages are not the majority and are exposed to a dominant language. It is natural for them to have a strong code-switching phenomenon. Therefore, we also explore the code-switching scenario common in multilingual and noisy scenarios like ours (Skiba, 1997; Crystal, 1987).

1.1. Motivation

The American continent is linguistically diverse; it comprises many indigenous languages that are nowadays spoken from North to South America. A wide range of linguistic families exhibit linguistic phenomena that differ from the most common languages usually studied in NLP. There are approximately 28 million[1] people who self-identify as members of an indigenous group (Wagner, 2016) and they speak around 900[2] native or indigenous languages (Mager, Gutierrez-Vasques, et al., 2018b). This represents an important

[1] Each country has its own methodology and criteria to estimate the number of speakers. This is the sum of all estimations.

[2] This number varies depending on the classification criteria used on different studies.

cultural and linguistic richness. This richness was captured by the following quote from McQuown (1955) "in one small portion of the area, in Mexico just north of the Isthmus of Tehuantepec; one finds a diversity of linguistic types hard to match on an entire continent in the Old World". Despite this, few language technologies have been developed for these languages. Moreover, many indigenous languages spoken in the Americas face a risk of language extinction (Hale, 1992).

Since indigenous languages are digitally scarce, developing technologies can have a positive social impact on the communities that depend on these languages (Mager, Bhatnagar, et al., 2023). However, the great diversity of these languages poses interesting scientific challenges, e.g., adapting well-established approaches, creating new methods, and collecting new data. Handling these challenges contributes to building more general computational language models and gaining a deeper insight into human language understanding. Moreover, many statistical NLP methods seek to achieve language independence; however, they often lack linguistic knowledge or do not cover the broad diversity of languages (Bender, 2011). This tendency is slowly changing, and new efforts are starting to include more languages of the world. Examples of these efforts are Masakhane (Nekoto et al., 2020; Adelani et al., 2022), AmericasNLP (Mager et al., 2021), and huge dataset collections like FLORES (Team, 2022; Costa-jussà et al., 2022).

1.2. Challenges

NLP research has centered most of its efforts around Indo-European languages and English (Ruder, 2020). However, these languages have less morphological richness than agglutinative and polysynthetic languages, with characteristics that share many indigenous languages of the Americas (ILA). As a result, most studies have a limited ability to handle these morphological features (Schwartz et al., 2020). Progress has been achieved with sub-word units such as BPEs(Sennrich et al., 2016b) managed to introduce methods that target the sparsity arising from the morphological complexities.

The second issue is the lack of large annotated or even unannotated datasets. State-of-the-art models take advantage of vast amounts of monolingual raw text, with self-supervised training like BERT (Devlin et al., 2019), GPT-x (Radford et al., 2019; Brown et al., 2020), BART (M. Lewis et al., 2019), and T5 (Raffel et al., 2020). This is also true for instruction-based large language models (LLMs) such as GPT4 (OpenAI, 2023) and Llama2 (Touvron et al., 2023). This is impossible for our targeted languages, as the existing monolingual data are extremely limited or non-existent. Regarding supervised

machine translation, few experiments have been performed in real low-resource scenarios (§4.4). Most of the experiments rely on simulated scenarios (Sennrich & Zhang, 2019). However, true low-resource languages have additional challenges: noisy text and lack of general NLP resources (i.e., normalizers, tokenizers, language identification tools, POS taggers, etc.). Overall, the low-resource scenario is challenging and a hard problem to tackle.

Finally, we find additional problems for processing our languages that are not seen as often in other languages: the lack of orthographic normalization and the existence of a wide dialectical variability inside each language (Mager et al., 2021). We should find robust models or ways to reduce this data sparsity for these problems.

1.3. Scope of this work

In this work, we are focusing on NLP for indigenous languages of the that have a polysynthetic typology. Performing an extensive study on NLP for these languages will help to improve our understanding of these languages in the context of NLP and set a foundation for future studies. As most of these languages are low-resource for most of the NLP tasks, we also focus on the low-resource scenario. We delimit the study in this thesis to the tasks of machine translation (MT), language identification (LID), and morphological segmentation (MS), and additionally explore these challenges in a code-switching (CS). Machine translation is important to handle, so it is possible to transfer knowledge from and to these languages. However, as the languages have a high morphological richness, we also study the tasks of morphological segmentation. Finally, as these languages are minority languages and most of the speakers have a certain degree of bilingualism, we also study the impact of code-switching. The a subset of languages where data is possible to collect. Due to the challenges of data collection, the sample small. Languages included in this thesis vary depending on the task and are the following: Mexicanero, Nahuatl, Popoluca, Tepehua, Shipibo Konibo, RarÃ¡muri, Yorem Noki, and Wixarika (§2.1.3).

1.4. Research questions and hypothesis

The research questions of this work are as follows:

1. **Is it possible to model morphological segmentation of polysynthetic languages with neural networks in low-resource scenarios?** Neural networks

have shown a strong performance in high-resource scenarios. We start from the intuition that with different techniques (like data augmentation, multi-task training, and transfer-learning), it is possible to obtain better results with neural networks than with other techniques, for the MS tasks, even in extremely low resource scenarios.

2. **To what extent does morphological segmentation influence the machine translation performance?** We hypothesize that morphological-based segmentation can outperform frequency-based segmentation when applied to the MT task in the extreme scenario of polysynthetic languages.

3. **How can we handle Code-Switching in these languages and use this phenomenon to improve the performance of MT models?** Code-switching is a common phenomenon in minority languages, such as the indigenous languages of the Americas. We hypothesize that code-switching is strongly present not only at the word level but also at the sub-word level, that it is possible to segment, and that this type of text can help improve the MT task.

With this thesis, we aim to study the application of neural NLP methods to the indigenous languages of the that have a polysynthetic typology. In doing so, we the interest in the NLP community and start to model the particular linguistic phenomenon of these languages. To do so, we datasets to perform our investigations for the tasks of LID, MT, and MS. This allows us to study these tasks and continue this research for other scientific community members. We also a the performance of the current models, explore the best-fitting models for these tasks, and analyze their performance. With all this said, our goal with this work is to encourage a broad community of researchers, language activists, native speakers, and activists to work on our languages. In doing so, we also understand that this thesis is just part of a bigger picture and that the work, rather than being individual, is a communal effort.

1.5. Chapter organization

To present this research, we the current thesis into six chapters. First, we have this introduction, followed by **chapter 2**, where we introduce the basic concepts for this work: endangered and low-resourced languages (§2.1.1); some important linguistically features of the ILA (§2.1.2); we present a short description of the languages we are

working with (§2.1.3); we introduce the basics of the models used in this work (§2.2); and we also present a brief survey of the current previous work on machine translation for low-resourced languages (§2.4), as well as on morphological segmentation (§3).

In **chapter 3** we present our experiments and findings on the modeling of polysynthetic languages in low-resource scenarios. In section 3 we introduce the task of surface segmentation (MSS)[3], we suggest a set of improvements to the vanilla neural models and discuss our experimental results. Some languages have a strong fusion phenomenon between morphemes. For these cases, canonical segmentation (MCS)[4] is more suited. In section 3.5, we show the experiments for the MCS task and discuss our findings.

The experimental results for machine translation (MT) are shown in **chapter 4**. First, we introduce the challenges for machine translation when it comes to the ILA (§4.1); then, we explore the low-resource methods for MT (§4.3). In section 4.4 we describe the advances in MT for our languages and the existing parallel datasets, which includes the AmericasNLI set of 11 indigenous languages of the Americas (ILA) and the findings of the collective effort[5] to improve MT for these languages (§5.3.4). Then, we explore the impact of morphological segmentation in machine translation (§4.5). To finalize the machine translation chapter, we conduct a study with indigenous leaders, teachers, and activists on the ethical implications of creating MT models for their languages (§4.6).

In **chapter 5** we model code-switching in a rich morphological context and explore ways to use code-switched text to improve machine translation systems. We divide this chapter into two sections: first, we model the intra-word code-switching together with the general language identification (LID) task (§5.2); and then, we compare the effectiveness of real and synthetic code-switching data for machine translation (§5.3).

Finally, in **chapter 6** we show this work's conclusions and give an overview of this work's contributions and findings.

1.6. Contributions and related publications

The work presented in the current thesis has been peer-reviewed and accepted to top tier conferences of the Natural Language Processing field. The papers published in top-tier conferences with the author of this thesis as first author are five: Mager, Çetinoğlu,

[3]The surface segmentation task aims to find the boundaries between morphemes on the surface form of a word.

[4]The canonical segmentation task reconstructs the underlying morpheme forms given an inflected surface form.

[5]We co-organized the AmericasNLP shared task on Machine Translation.

& Kann (2019) at NAACL-HLT, Mager et al. (2020) at EMNLP, Mager, Mager, et al. (2023) at ACL, and Mager et al. (2022) at Findings of ACL; and as second author A. Ebrahimi et al. (2022) at ACL and Gaser et al. (2023) at EACL. In addition, two first author journal articles (Mager, Rosales, et al., 2019; Mager & Meza, 2021a); and five peer-reviewed workshops (Mager et al., 2021; Mager & Kann, 2020; Mager, Bhatnagar, et al., 2023; Denisov et al., 2021; A. Ebrahimi et al., 2023) are part of this work.

Here we enumerate the contributions of this work, their related publications, and the sections where they are discussed. First, we start with datasets that were curated, collected, or created: we create a morphological canonical segmentation dataset for Popoluca and Tepehua (Mager et al., 2020) in chapter 3; a morphological surface segmentation dataset for RarÃ¡muri and Shipibo-Konibo languages (Mager et al., 2022); parallel datasets for the RarÃ¡muri–Spanish (Mager et al., 2022) in chapter 4; Wixarika–Spanish CS languages pairs in chapter 5; a sub-word language identification code-switching dataset for the Wixarika–Spanish language pair (Mager, Çetinoğlu, & Kann, 2019) in chapter 5. Additionally, we curated a dataset of 10 ILA with Spanish for the AmericasNLP Shared task 2021 (Mager et al., 2021; A. Ebrahimi et al., 2022) and we added Chatino–Spanish for the AmericasNLP 2023 Shared Task (A. Ebrahimi et al., 2023)[chapter 4].

We also contribute to improving the performance of NLP systems on the following tasks: we propose a new sub-word level language identification (SLID) task for code-switching and adapt the segmental RNN (SegRNN) model to the task (Mager, Çetinoğlu, & Kann, 2019) [chapter 5]; we propose new-to-the-task sequence-to-sequence imitation learning (IL) and pointer-generator network (PtrGenNet) models for canonical morphological segmentation (Mager et al., 2020) in chapter 3; and we create a new semi-supervised approach for the sequence-to-sequence (seq-to-seq) morphological surface segmentation task in chapter 3.

We perform studies on existing tasks and methods to understand the challenges to further improve our systems. First, we investigate the performance of the diverse morphological segmentation methods in a simulated low-resource setting. We then apply this method to real low-resourced languages and found that neural networks are a great option for the task (Mager et al., 2020) [chapter 3]. With these results, we also study the impact of sub-word segmentation on machine translation in extreme low-resource scenarios (Mager et al., 2020) in chapter 4. We compare frequency-based sub-word segmentation with a broad set of morphology-inspired segmentation methods. We also study challenges and current methods for machine translation in low-resourced scenarios

extensively, and in particular for the case of the indigenous languages of the Americas (Mager & Meza, 2021a; Mager, Bhatnagar, et al., 2023) in chapter 4. Finally, we also study the difference between synthetic and real code-switching data to improve machine translation [chapter 5]. Finally, we explore the ethical implications of applying NLP and machine translation modeling to the indigenous languages of the Americas (Mager, Mager, et al., 2023). This is the first study that performs a systematic survey and interviews of native speakers language activists, and community leaders from the Americas, on the application of MT models for their languages.

2. Background

In the current chapter, we define the key concepts for this work and introduce the basic architectures and frameworks used in the following chapters.

2.1. Indigenous Languages of the Americas

2.1.1. Endangered and low-resource languages

Terms frequently used in NLP are *low-resource language*, *resource-poor language*, and *low-resource setting*. Those terms do not highlight that many low-resource languages are also endangered (Z. Liu et al., 2022). Instead, they emphasize the critical machine learning problem of getting a data-driven approach to perform well with a smaller-than-ideal amount of available data (or just fewer data than what has been used for other languages). In this case, algorithmic or technological innovations are needed to close the performance gap between high-resource and resource-poor languages. This further implies that being low-resourced is not a property of a language but a term that only makes sense in the context of a particular task or task. However, it is also true that if a language has an overall lack of resources (monolingual and annotated data), we can refer to it as a low-resource language (LRL). We can additionally refer to extreme cases, where almost no data is available, as extreme low-resource language (eLRL)

In contrast, the term endangered language (EL) refers to a language that has a certain degree of danger for its existence.[1] According to the UNESCO classification, (Moseley, 2010) languages can be sorted into the following different categories:

- *safe*: spoken by all generations;
- *vulnerable*: restricted just to a certain domain (i.e. inside the family);
- *definitely endangered*: it has no kids that speake the language;

[1] In this work, we will discuss only non-artificially created languages.

- *severely endangered*: only elder people speak it;
- *critical endangered*: there are only speakers left with partial knowledge and they use it infrequently;
- *extinct*, when there are no persons able to speak the language anymore.

Languages can become endangered due to social, cultural, and political reasons; most commonly conquests and wars, economic pressures, language policies from political powers, assimilation of the dominant culture, discrimination, and language standardization (Austin & Sallabank, 2013).

It is important to distinguish both terms, as a LRL can be a widely spoken language, while a EL might have more data for a specific task that one that is not endangered. For example, Inuktitut is not low resourced in MT, but is considered endangered. However, in this work, all the languages that we study have certain degree endangered, and are also extremely low-resourced.

On the machine learning side, an additional challenge arises when a language: data for endangered languages is not easily available (or, in fact, available at all), as these languages have limited media production, for example TV shows, literature, internet blogs, etc. (Hämäläinen, 2021). One possible source of data for these languages are already existing documents in form of books, records and archives Bustamante et al. (2020).

2.1.2. Indigenous languages of the Americas

To have a broader understanding of this term, the United Nations (UN, 2022) created an understanding of the term based on the following: An indigenous language is a language spoken or used to be spoken as a mother tongue by indigenous people. However, who is considered an indigenous person? The answer to this question is not easy, as many factors come into play. For example, in the USA, a racial heritage of blood percentage (R. W. Schmidt, 2011) defines if someone belongs to a particular tribe. In Mexico, an indigenous person is one who self-recognizes belonging to a group that has roots before the Spanish colonization (Gobierno, 2022). However, these definitions are specific to each country's regional and historical reality.

In an attempt to have a broader understanding of this term, the United Nations (UN, 2022) created an understanding of the term based on the following:

Considering the diversity of indigenous peoples an official definition of "indigenous" has not been adopted by any UN-system body. Instead, the system has developed a modern understanding of this term based on the following:

- Self-identification as indigenous peoples at the individual level and accepted by the community as their member.
- Historical continuity with pre-colonial and/or pre-settler societies
- Strong link to territories and surrounding natural resources
- Distinct social, economic, or political systems
- Distinct language, culture, and beliefs
- Form non-dominant groups of society
- Resolve to maintain and reproduce their ancestral environments and systems as distinctive peoples and communities. A question of identity
- According to the UN, the most correct approach is identifying, rather than defining, indigenous peoples. This is based on the fundamental criterion of self-identification, as underlined in a number of human rights documents.

This work will focus on the indigenous languages of the Americas (ILA). The languages defined under this criteria can differ. However, our primary goal is to work on the languages directly related to the ones spoken on the American continent before the European conquest and still spoken nowadays (or its derived languages). The language variety in these groups is wide. Therefore, we will work with some samples, mostly from languages spoken in Mexico. These languages have a diverse number of linguistic phenomena. In this work, we will introduce some of the primary and most essential features found among languages in the region.

Tonal languages use the tone (pitch) to distinguish lexical meanings or grammatical functions (Moira, 2002). The same morpheme or word can have different meanings depending on its tone. Some languages in the region (i.e., Chatino) have even 13 tonnes.

In our first example, we see how tones can determine the meaning of a specific word. All examples for tonal languages are taken from Cruz (2011) for the San Juan Quiahije Chatino. The superscripts $^{\mathrm{L}}$, $^{\mathrm{HL+0}}$, $^{\mathrm{M}}$, $^{\mathrm{MH}}$ and $^{\mathrm{M0}}$ are tone marks.

Chatino	*English*
kta^{M}	flour
kta^{MH}	chepil (edible herb)
kta^{L}	tobacco

As mentioned previously, the tones can also mark certain inflectional and derivational categories

Chatino	*English*
kwa^{MH}	she/he swept
kwa^{H}	you swept
$\text{kwa}^{\text{HL+0}}$	she/he will sweep
kwa^{M0}	you will sweep

Fusion and Agglutinating languages Agglutination refers to the morphological process of concatenating morphemes, conserving the canonical forms of the morphemes (Bickel & Nichols, 2005). These languages are polymorphemic, where each morpheme corresponds to a single lexical meaning (Johan et al., 2001).

In the following example, we show the concatenating property of Huallaga Quechua (Weber, 1989):

Yapya|y|ta	usha|na|n|ta|shi	huyarayka|n
plow|INF|OBJ	finish|sub|3P|OBJ|IND	be:waiting|3

He is waiting for him to finish plowing

On the other side of the spectrum, we have fusional languages. These languages modify the final surface form of the morphemes when a morphological process starts. It is also important to note that a single morpheme expresses several different meanings or grammatical functions (Johan et al., 2001).

To illustrate this phenomenon, we show an example in Texitepec Popoluca (Wichmann, 2007).

bic?u	myoŋta?
bic?u	ny|boŋ|ta?
2 pro.aux.:ipfv.	A1|sleep|1,2pl.

You are sleeping

We can see that the morphemes *ny*, *boŋ*, and *ta?* have an underlying canonical form; however, when they are combined in the word *myoŋta?*, the surface form changes due to the interaction of all morphemes.

Polysynthetic languages The following linguistic features define a polysynthetic language: These languages have head marking (Nichols, 1986) with a head-final (SOV) or an initial head structure (VSO or SVO) or a free word order (Greenberg, 1963). The verb in a polysynthetic language must have an agreement with the subject, objects, and indirect objects (M. C. Baker, 1996); nouns can be incorporated into the complex verb morphology (Mithun, 1986); therefore, polysynthetic languages have agreement morphemes, pronominal affixes, and incorporated roots in the verb (M. C. Baker, 1996), and also encode their relations and characterizations into that verb. These languages' most common word orders are SOV, VSO, SVO, and free order. Polysynthetic languages will have fusional or agglutinating processes in their morphological system, as seen in sections 3.5 and 3.

Namely, we analyze the phrase "She always asks us for tortillas." in an example taken from Gómez & López (1999), of the Wixarika language[2]:

m+k+	pa:pa	ya	p+|ta|ti|u|ti|wawi|ri|wa
Ella	tortilla	enf	asi|1pl:o|its|vis|pl:a|pedir|apl|hab

She always asks us for tortillas

From the syntax point of view, Wixarika employs a head-final structure (SOV), as seen in the example. Therefore, we have in the third place the emphatic factor "ya", which realizes the agreement of the initial subject and the direct object; also, we need an asserter for the indirect object "ta", which in this case is the morpheme "p+". The verb exists of different prefixes collocated before the verb stem "wawi": the morpheme "ti" is an intensifier of the visibility of the ambit of the speaker, expressed with the morpheme "u"; the prefix "ti" on the first place of the verb refers to the plurality of the action and the plural of the direct object. Therefore, we can speak of incorporating the object into the verb.

2.1.3. Languages

In this section, we will briefly describe the languages we work on. Most languages are spoken mainly in Mexico but also in Guatemala (Náhuatl) and Perú (Shipibo-Konibo). This does not include the 10 languages of the AmericasNLP shared task (Mager et al., 2021), but most languages overlap with this dataset (except Popoluca and Tepehua). A detailed description of the datasets used for this work will be included with their

[2]The analysis is taken from our previous to this thesis work Mager, Mager, et al. (2018)

respective experiments. Figure 2.1 is a map that shows the regions and countries where each studied language is spoken[3].

Mexicanero is a Western Peripheral Nahuatl variant, spoken in the Mexican state of Durango by approximately one thousand people (Pottier, 2002). This dialect is isolated from the rest of the other branches and has a strong process of Spanish stem incorporation while also having borrowed some suffixes from that language (Vanhove et al., 2012). It is common to see Spanish words mixed with Nahuatl agglutinations. In the following example, we can see an intrasentential mixing of Spanish (*in uppercases*) and Mexicanero:

u|ni|ye MALO – *I was sick*

Nahuatl is a large subgroup of the Yuto-Aztecan language family and, including all of its variants, the most spoken native language in Mexico. Its almost two million native speakers live mainly in Puebla, Guerrero, Hidalgo, Veracruz, and San Luis Potosi, but also in Oaxaca, Durango, Modelos, Mexico City, Tlaxcala, Michoacan, Nayarit, and the State of Mexico. There are 30 recognized variants of Nahuatl spoken by over 1.5 million speakers across Mexico, where Nahuatl is recognized as an official language SEGOB (2020). Three main dialectical groups are known: Central Nahuatl, Occidental Nahuatl, and Oriental Nahuatl. The data collected for the morphological segmentation task belongs to the Oriental branch spoken by 70 thousand people in Northern Puebla[4].

Like all languages of the Yuto-Aztecan family, Nahuatl is agglutinative and polysynthetic. Usually, the verb functions as the stem and gets extended by morphemes specifying, e.g., subject, patient, object, or indirect object. The most common syntax sequence for Nahuatl is SOV. An example word is:

o|ne|mo|kokowa|ya
I was sick

This Nahuatl variant has five vowels $\{i, u, e, o, a\}$ and does not distinguish if they are short or large. As Nahuatl does not have an extended accepted orthography, our alphabet was adopted from Lastra de Suárez (1980) with simplifications.

[3]The map is designed and created by Rebeca Guerrero.
[4]This data was collected in our paper Kann et al. (2018), a work previous to this thesis.

Wixarika is a language spoken in the states of Jalisco, Nayarit, Durango, and Zacatecas in Central West Mexico by approximately fifty thousand people. It belongs to the Coracholan group of languages within the Yuto-Aztecan family. Wixarika has five vowels {a,e,i,+[5],u} with long and short variants. An example of a word in the language is:

ne|p+|ti|kuye|kai – *I was sick*

Like Nahuatl, it has an SOV syntax, with heavy agglutination on the verb. Wixarika is morphologically more complex than other languages from the same family because it incorporates more information into the verb (Leza & López, 2006).

Yorem Nokki is part of Taracachita subgroup of the Yuto-Aztecan language family and is also known as Yaqui. Its Southern dialect is spoken by close to forty thousand people in the Mexican states of Sinaloa and Sonora, while its Northern dialect has about twenty thousand speakers (INEGI, 2020). In this work, we consider the Southern dialect. The nominal morphology of Yorem Nokki is rather simple, but, like in the other Yuto-Aztecan languages, the verb is highly complex.

ko'kore|ye|ne
I was sick

Shipibo-Konibo Shipibo-Konibo is a polysynthetic Panoan language spoken by around 35,000 native speakers in the Amazon region of Peru. It is a language with agglutinative processes, most of which are suffixes. However, clitics are also used and are a widespread element in Panoan literature (Valenzuela, 2003). Shipibo-Konibo uses an SOV word order (Faust, 1973) and postpositions (Vasquez et al., 2018). We use the standard writing supported by the Ministry of Education in Peru. The following example shows a phrase in Shipibo-Konibo[6]. To have a better intuition about the language, we show the following example:

ne|ri|a bai|koma iki.
Over here the way is not good.

[5]While linguists often use a dashed i (i̇) to denote this vowel, in practice, almost all native speakers use a plus symbol (+). In this work, we use the latter.
[6]Examples taken from (Valenzuela, 2003)

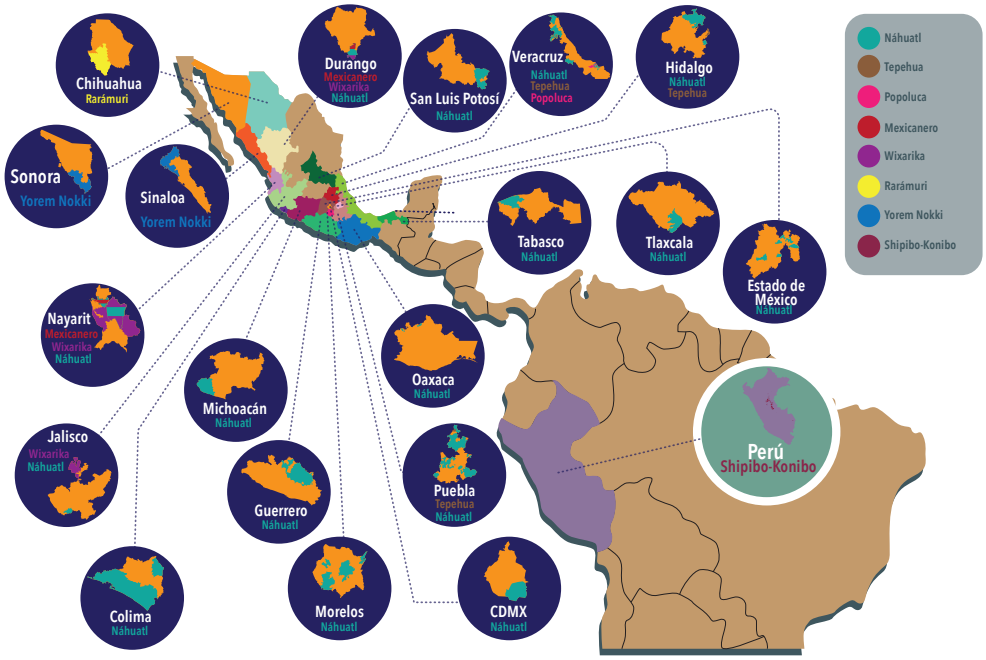

Figure 2.1.: Localization of the studied languages in the context of the Latin-American region. Blue-circled maps show a detailed view of Mexican states. The Mexican territory marked for each language is based on the municipalities listed in INALI (2022) and the information on MC (2022) for Peru.

Rarámuri The Rarámuri language, also known as *Tarahumara*, which means *light foot* (INALI, 2017), which comes from the rarámuri word "rare" = foot in Spanish and from "muri" in rarámuri which means to run in Spanish , sometimes it is also written as "ralámuli" or as "Rarómariraicha", "ralámuliraicha" and "rarámariraicha" (Pintado-Cortina, 2004). This language belongs to the Taracahitan subgroup of the Uto-Aztecan language family and to the subgroup Taracahitan (Goddard, 1996). The variants of this language are: western, northern, highland, central, and southern (INEGI, 2008a). Rarámuri is an official language of Mexico, spoken mainly in the Sierra Madre Occidental region in the state of Chihuahua (in 17 municipalities), Mexico, in 1,552 localities by a total of 89,503 speakers (INEGI, 2020). Rarámuri is a polysynthetic language characterized by a head-marking structure, according to the terms of Nichols (1986). The variant we are going to use is the Highlands variant (INEGI, 2008a). Translation orthography and word boundaries are similar to (Caballero, 2008).

oshí|ki|ma
She will write him

Popoluca. Popoluca of Texistepec (language code: poq[7]) is part of the Mixe-Zoquean family. Its morphology is classified as polysynthetic, and it mostly follows a verb, subject, object (VSO) word order (Dryer & Haspelmath, 2013). This language is almost extinct, with only one native speaker alive reported in 2005 (Gordon Jr, 2005). However, attempts for language revival have been reported (INEGI, 2008b). Efforts made for language revitalization can benefit from advances in NLP. Thus, creating and developing accurate models for those languages are highly important.

Here we show an example of canonical segmentation in Popoluca and its English gloss. The plus symbol is part of the alphabet of the language. We use a | as a morpheme delimiter.

kki:?mba: → ky|k+:?m|ba:
You are small

Tepehua. Tepehua (language code: tpp) belongs to the Totonacan language family. It is spoken in three Mexican regions: in the northeastern part of the state of Hidalgo (around 3000 speakers), in the villages of Pisaflores (around 4000 speakers), and in Tlachichilco in the state of Veracruz (around 3000 speakers) (Gordon Jr, 2005). It is also polysynthetic. Tepehua permits free word order but has a preference for a subject, verb, object (SVO) configuration (Dryer & Haspelmath, 2013).

An example for canonical segmentation is

iklakadíkdi → ik|laka|tikti
I am small

The language variant used in our dataset is spoken in Pisaflores, Veracruz.

2.2. Modeling

This section will introduce the base frameworks and models and the used notation that we apply in our core experiments.

2.2.1. The Encoder-Decoder paradigm

This paradigm is an upper set for many architectures. Lets X and Y be variable vectors of size $N > 0$ and $M > 0$. The goal is to learn a function that maps $X \rightarrow Y$. As X and

[7] We use the languages codes defined in the ISO 639-3 standard.

Y we need a two-stage network. As shown in figure 2.2, the first of these stages is the encoder, represented by an encoding function $z = \text{Enc}(X)$ that transforms the input into vector X into a latent representation Z. The size of Z is a given as a hyper-parameter. The second stage is the decoder, which predicts $\hat{Y}$ from the latent space given a learned function $\hat{Y} = \text{Dec}(Z)$.

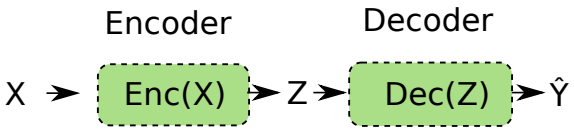

Figure 2.2.: Encoder-Decoder paradigm.

The Encoder-Decoder paradigm is especially useful when X and Y are sequences of symbols with variable lengths. The transformation of a sequence to another one, where $M \neq N$, with an Encoder-Decoder approach is known as the sequence-to-sequence (seq-to-seq) framework.

2.2.2. Neural Recurrent Networks

A Recurrent Neural Network (RNN)(Jordan, 1986; Elman, 1990) is a neural network that process a variable length input $X = x_1, x_2, \ldots, x_T$, and maintains a hidden vector $\mathbf{h}$ over time. At each time step t the hidden state $\mathbf{h_t}$ is updated by a non-linear activation function f:

$$\mathbf{h}_t = f(\mathbf{h}_{t-1}, x_t). \tag{2.1}$$

The RNN can learn a probability distribution when being trained or predict the next symbol of a sequence.

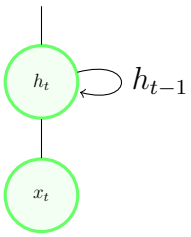

Figure 2.3.: basic RNN concept

To illustrate this point, we show a simple RNN that was proposed by Elman (1990). The input is $x_t \in X$ for each time step t. This input is multiplied with a matrix $\mathbf{V}$ and

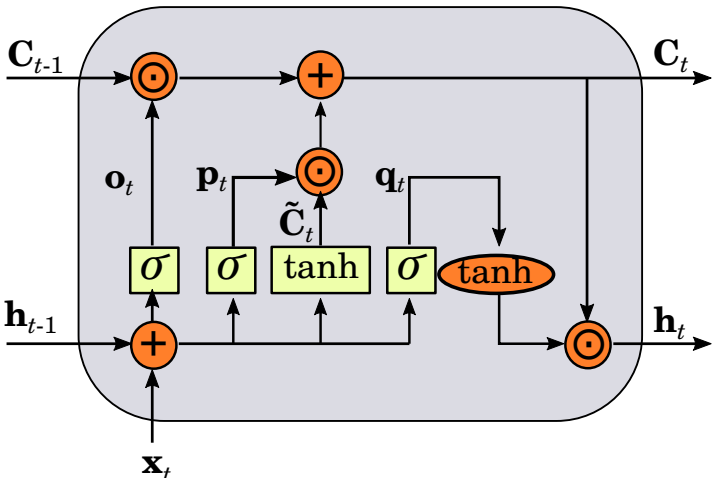

Figure 2.4.: Diagram showing the LSTM cell.

summed to the previous hidden state h_{t-1} that is also multiplied with a weight matrix $\mathbf{W}$. This input is the input for the non-linear function tanh to obtain the new hidden state.

$$\mathbf{h}_t = \tanh(\mathbf{W}\mathbf{h}_{t-1} + \mathbf{U}x_t + \mathbf{b}) \tag{2.2}$$

$$\hat{\mathbf{y}}_t = \text{softmax}(\mathbf{c} + \mathbf{V}\mathbf{h}_t) \tag{2.3}$$

The problem with this simple method is that they have two ways to go wrong: the gradient can vanish, or it can exploit Hochreiter (1991). It requires finding the optimal $\mathbf{h}$ dimensions to avoid, if even possible to avoid this problem. In practice, it is hard to train longer sequences.

LSTM To handle these problems Hochreiter & Schmidhuber (1997) introduced the long short term memory (LSTM) framework. The main idea is to have two hidden vectors: the traditional short memory vector $\mathbf{h}_t$ that will be passed directly to the next time step and transformed completely, and an additional context vector $\mathbf{c}_t$. This context vector will contain information in the form of long-term memory and be modified just slightly during each time step. The context vector is modified with the help of gates, which will learn how much should be added or forgotten. Figure 2.4 shows how this RNN cell is organized.

Forget gate layer (eq. 2.4) was introduced by Gers et al. (2000). This gate's main function is to learn to forget elements of $\mathbf{C}_{t-1}$. The input gate (eq. 2.7) determines how

much information is added to the context vector. Finally to calculate the new hidden layer $\mathbf{h}_t$ we use information from $\mathbf{h}_{t-1}$, $\mathbf{x}_t$ and the context $\mathbf{c}_t$ (eq. 2.9). The forward pass of this network is

$$\mathbf{o}_t = \sigma(\mathbf{W}_o\mathbf{h}_{t-1} + \mathbf{U}_o\mathbf{x}_t + \mathbf{b}_o) \tag{2.4}$$
$$\mathbf{p}_t = \sigma(\mathbf{W}_p\mathbf{h}_{t-1} + \mathbf{U}_p\mathbf{x}_t + \mathbf{b}_p) \tag{2.5}$$
$$\tilde{\mathbf{c}}_t = \tanh(\mathbf{W}_{\tilde{c}}\mathbf{h}_{t-1} + \mathbf{U}_{\tilde{c}}\mathbf{x}_t + b_{\tilde{c}}) \tag{2.6}$$
$$\mathbf{c}_t = \mathbf{o}_t \odot \mathbf{C}_{t-1} + \mathbf{p}_t \odot \tilde{\mathbf{c}}_t \tag{2.7}$$
$$\mathbf{q}_t = \sigma(\mathbf{W}_q\mathbf{h}_{t-1} + \mathbf{b}_q) \tag{2.8}$$
$$\mathbf{h}_t = \mathbf{q}_t \odot \tanh(\mathbf{c}_t) \tag{2.9}$$

where $\mathbf{W} \in \mathbb{R}^{h\times d}, \mathbf{U} \in \mathbb{R}^{h\times h}$ are weighted matrices and $b \in \mathbb{R}^h$ is a bias vector. All these vectors are the parameters to learn during training. The input vector $\mathbf{x}_t \in \mathbb{R}^d$ has d size and the context $\mathbf{c}_t \in \mathbb{R}^h$ and hidden state $\mathbf{c}_t \in (-1, 1)^h$ have a size of h. The $\odot$ operation is an element-wise product.

A popular variation of LSTM is the gated recurrent unit (GRU) Cho et al. (2014) model. It uses one single gate for the forget and input gates and also uses a single vector for the state and the hidden vector. In practice, LSTM's and GRU's have almost the same performance, with fewer parameters usage by the latter.

The RNN sequence-to-sequence model. A RNN seq-to-seq model is based on the Encoder-Decoder framework[8]. It is capable of learning the conditional distribution over a variable-length sequence $X = x_1, \ldots, x_T$ given another variable-length sequence $Y = y_1, \ldots, y_{T'}$ (Cho et al., 2014).

Figure 2.5 show the architecture of the RNN seq-to-seq model. The encoder is an RNN (or a Bidirectional RNN (Graves et al., 2005)) that reads sequentially X and ends with a fixed-length representation $\mathbf{z}$ of it. The decoder is another RNN that generates a sequence Y, such as

$$P(y_t|y_{t-1}, y_{t_2}, \ldots, \mathbf{z}) = g(\mathbf{h}_{t-1}, y_{t-1}, \mathbf{z})). \tag{2.10}$$

[8] We base the notation of this section on (Cho et al., 2014) and adapt it to fit our overall notation schema.

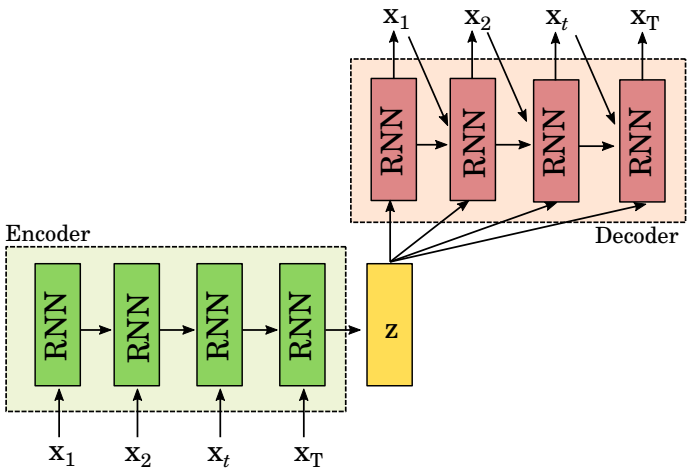

Figure 2.5.: The RNN encoder-decoder architecture.

The attention mechanism. RNN seq-to-seq models have a strong potential for many tasks. However, in practice, it has been shown that at a certain decoding step y_t it was not possible to look back at the sequence X was a problem for tasks like alignment. Therefore, Bahdanau et al. (2015a) introduced a method to search through X at decoding time[9]. For an overview of the change in the architecture, we refer the reader to figure 2.6 First, the attention mechanism changes the decoder equation 2.10 to the following form:

$$p(y_i|y_1, \ldots, y_{i-1}, \mathbf{x}) = g(y_{i-1}, \mathbf{h}_i^{(y)}, \mathbf{z}_i). \tag{2.11}$$

$h_i^{(y)}$ is the hidden state of the RNN decoder, which is computed for each time step i as follows:

$$\mathbf{h}_i^{(y)} = f(\mathbf{h}_{i-1}^{(y)}, y_{i-1}, \mathbf{z}_i). \tag{2.12}$$

where $\mathbf{z}_t$ is the context vector. This context vector is different from $\mathbf{z}$ from the vanilla seq-to-seq, as for each y_t a z_t is computed. This context vector is computed with the hidden vectors from the encoder $\mathbf{h}^{(x)} = (\mathbf{h}_{1^d}^{(x)}, \mathbf{h}_t^{(x)}, \ldots, \mathbf{h}_{T_x}^{(x)})$ is defined as

[9] We base our notation on the one used by Bahdanau et al. (2015a)

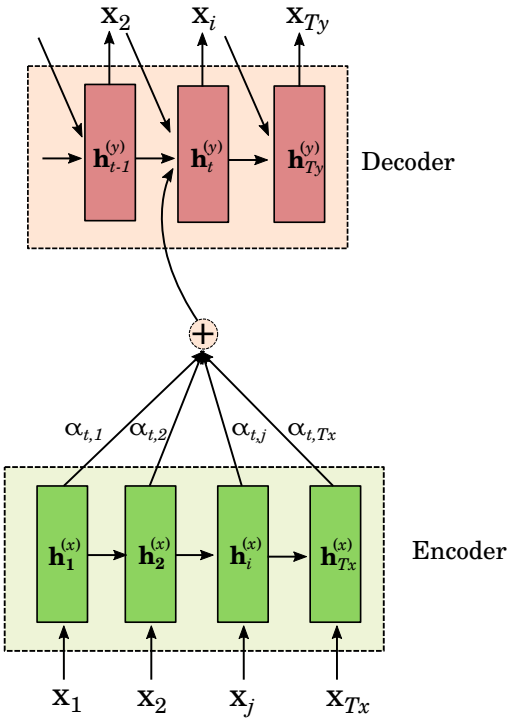

Figure 2.6.: Sequence-to-sequence model with an attention mechanism.

$$\mathbf{z}_i = \sum_{j=1}^{T_x} \alpha_{ij} \mathbf{h}_j^{(x)}. \tag{2.13}$$

The α_{ij} values come from a matrix that is calculated as:

$$\alpha_{ij} = \frac{\exp(\mathbf{e}_{ij})}{\sum_{k=1}^{T_x} \exp(\mathbf{e}_{ik})} \tag{2.14}$$

where e_{ij} comes from an attention model that predicts how well an element in x with position, j fits an element in y with position i.

$$\mathbf{e}_{ij} = a(\mathbf{h}_{i-1}^{(y)}, \mathbf{h}_j^{(x)}) \tag{2.15}$$

To parameterize the attention model $a()$ it can be done as a feed froward network

$$\mathbf{e}_{ij} = a(\mathbf{h}_{i-1}^{(y)}, \mathbf{h}_{j}^{(x)}) \tag{2.16}$$

$$= \mathbf{V}\tanh(\mathbf{W}_1\mathbf{h}_{i-1}^{(y)}, \mathbf{W}_2\mathbf{h}_{j}^{(x)}). \tag{2.17}$$

2.2.3. The transformer architecture

The transformer architecture, first introduced by Vaswani et al. (2017) is the current *de facto* architecture for many NLP tasks, including MT. We use the encoder-decoder paradigm for MT and MS. The first advantage of the transformer architecture over the previously discussed RNN is that they are easily parallelizable. A sequence can be processed simultaneously, and therefore, it is unnecessary to iterate over time steps as with the RNN architectures. This is especially helpful when we train with huge datasets of for tasks where a relatively big dataset is needed (i.e., MT). This brings us to the following advantages: the fact that the entire sequence is processed simultaneously allows the model to overcome the previous limitation of gradient vanishing over time and, therefore, allows it to work with longer sequences.

Furthermore, as the size of the transformer hidden states change with the size of the input sequence, the problem of having one non-variable vector that encodes the sentence is not a problem anymore. Moreover, altogether also makes it possible to scale the training to handle huge datasets (H. Wang et al., 2022). In this subsection, we will introduce the basics of the model[10].

Transformer input . All the input elements x_i in $\mathbf{x}$ are converted into a tensor $\widetilde{\mathbf{x}}$ with the help of a finite lookup table, called the vocabulary (see 2.7). This vocabulary contains unique words and relates them to an embedding (usually initialized randomly). If an element is not found in the vocabulary, it gets assigned with an Out Of Vocabulary (OOV) symbol. The size of each vector is E.

However, as the model will not be able to handle the order of the input sequence, the transformer use *positional encodings* (**PE**) to inject the information about the position of the element in the input sequence. There are many strategies of performing this task Gehring et al. (2017). To keep it simple, we show the positional encodings introduced

[10]The notation used for the transformer section is based on Ramon Astudillo's work, with minor adaptations.

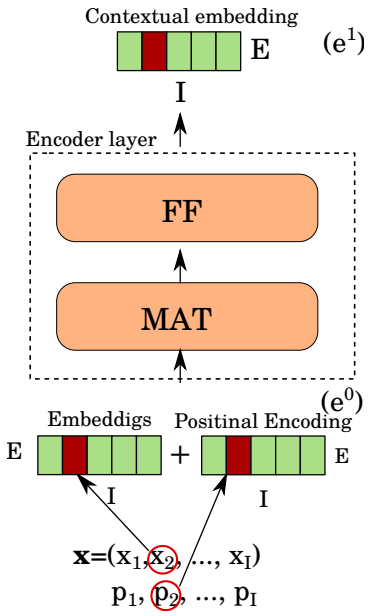

Figure 2.7.: The interaction of the positional encoder, the embedding layer, an encoder layer, and the resulting contextual representation.

by the original paper. The position embedding p_i is calculated

$$\mathbf{PE}_{(,2j)} = \sin(i/1000^{2i/E}) \tag{2.18}$$

$$\mathbf{PE}_{(i,2j)} = \cos(i/1000^{2i/E}) \tag{2.19}$$

where i is the position and j the dimension. As $\mathbf{PE}, \tilde{\mathbf{x}} \in \mathbb{R}^{E \times I}$, it is easy to obtain our final input embedding $\mathbf{e}^0$:

$$\mathbf{e}^0 = \tilde{\mathbf{x}} + \mathbf{PE}. \tag{2.20}$$

Attention The input to a attention layer AT needs three tensors: a query q, a key k, and a value v. This "output is computed as a weighted sum of the values, where the weight assigned to each value is computed by a compatibility function of the query with the corresponding key" Vaswani et al. (2017). Attention layer The AT is defined as follows:

$$AT_h(\mathbf{q}, \mathbf{k}, \mathbf{v}) = \mathbf{W}_h^V \cdot \mathbf{k} \cdot \text{softmax}\left(\frac{1}{A}(\mathbf{W}_h^K \mathbf{v})^T \mathbf{W}_h^Q \mathbf{q}\right). \tag{2.21}$$

In the case of self-attention, k, q, and v are the same contextual vector e^n, or e^0 if it is the first layer. This substituted in from 2.21 we obtain

$$\text{AT}_h(\mathbf{e}^n, \mathbf{e}^n, \mathbf{e}^n) = \mathbf{W}_h^V \cdot \mathbf{e}^n \cdot \text{softmax}\left(\frac{1}{A}(\mathbf{W}_h^K \mathbf{e}^n)^T \mathbf{W}_h^Q \mathbf{e}^n\right). \tag{2.22}$$

The Multi-head Attention Layer With AT defined in equation 2.21, we concatenate a H number of attention heads. The resulting operation is called the "multi-head attention" layer(MAT):

$$\text{MAT}(\mathbf{q}, \mathbf{k}, \mathbf{v}) = \mathbf{W}^o \begin{bmatrix} \text{AT}_1(q, k, v) \\ \text{AT}_2(q, k, v) \\ \dots \\ \text{AT}_H(q, k, v) \end{bmatrix} \tag{2.23}$$

with $\mathbf{W}^o \in \mathbb{R}^{E \times H \cdot A}$ where A is the attention dimension of AT. It is important to note that when we use AT in a encoder-decoder framework, we use for the encoder $AT(\mathbf{e}^n, \mathbf{e}^n, \mathbf{e}^n)$, and $AT(\mathbf{d}^n, \mathbf{e}^n, \mathbf{e}^n)$ for the decoder.

The feed-forward network The next component shown in figure 2.7 is the feed forward network FF is defined as

$$\text{FF}(z^n) = \mathbf{W}^2 \cdot \mathbb{M}^{\text{drop}} \odot \text{relu}(\mathbf{W}^1 \cdot \mathbf{x}^n) \tag{2.24}$$

with expanding $\mathbf{W}^1 \in \mathbb{R}^{2 \cdot E \times E}$ and $\mathbf{W}^2 \in \mathbb{R}^{E \times 2 \cdot E}$. These layers can be seen as feature detectors. $\mathbf{M}^{\text{drop}}$ is a dropout mask applied element-wise ($\odot$). This leaves us with the final value of the next context vector:

$$\mathbf{e}^{n+1} = \text{FF}(\text{MAT}(\mathbf{e}^n, \mathbf{e}^n, \mathbf{e}^n)) \tag{2.25}$$

The transformer layer includes residual connections to avoid gradient vanishing, dropout layers to avoid overfitting, and normalization. This is included in the MAT and FF networks as follows:

$$\tilde{\text{MAT}}(\mathbf{q},\mathbf{k},\mathbf{v}) = \text{layernorm}(\mathbf{M}^{\text{drop}} \odot \text{MAT}(\mathbf{q},\mathbf{k},\mathbf{v}) + \mathbf{q}) \tag{2.26}$$

$$\tilde{\text{FF}}(\mathbf{z}) = \text{layernorm}(\mathbf{M}^{\text{drop}} \odot \text{FF}(\mathbf{z}) + \mathbf{z}). \tag{2.27}$$

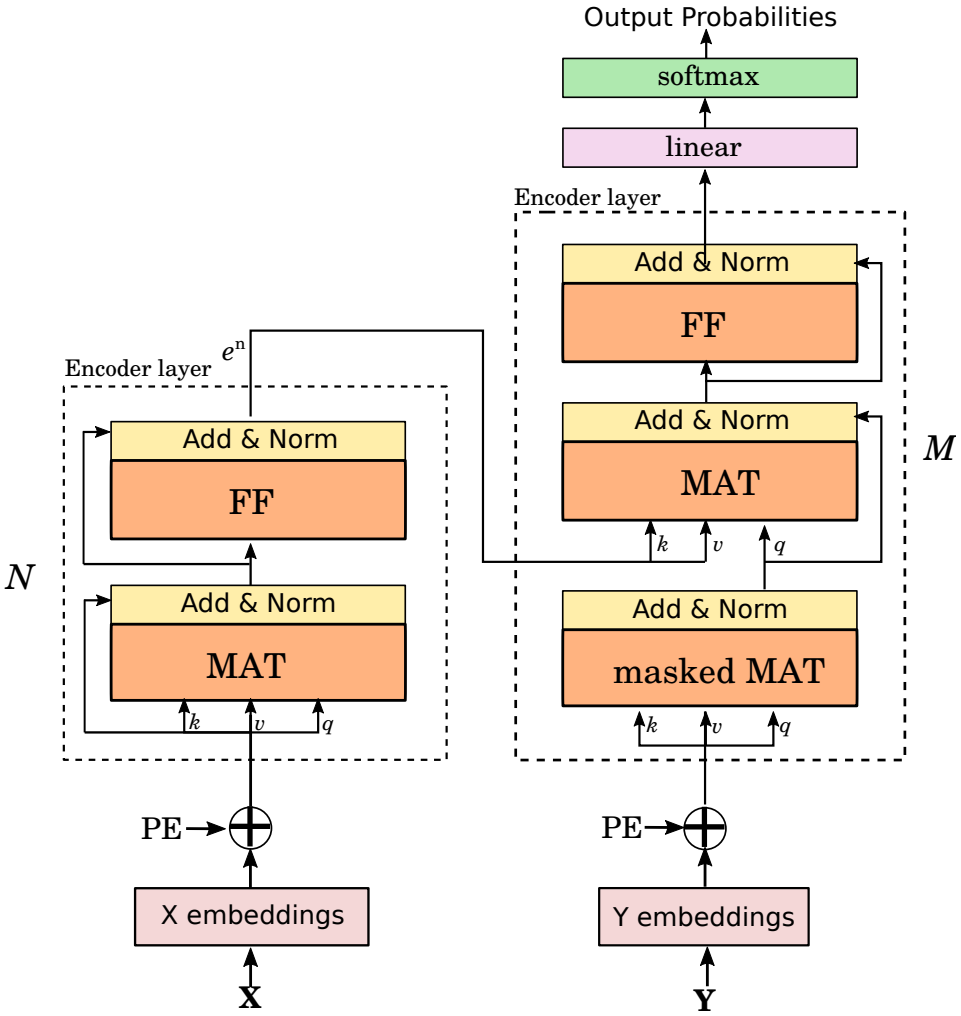

Figure 2.8.: The transformer architecture in an encoder-decoder setup.

The decoder layer . As mentioned before, at the decoding time, we will use the contextual embedding of the last layer of the encoder ($\mathbf{e}^N$) for k and v, but the query is the self-attention on the previous attention layer. This is also called cross-attention.

$$d^{m+1} = \text{FF}(\text{MAT}(\text{MAT}(d^m, d^m, d^m), e^N, e^N) \tag{2.28}$$

The first decoder layer . As we want to generate the next token from a target sequence $\mathbf{y}$, we want to hide the rest of the sequence at training time to the right of our time step. Therefore, the input vector of the decoder d^0 needs to be cut from $\mathbf{y}$.

$$\mathbf{d}^0 = [\mathbf{s}, \mathbf{y}_{1:J-1}] + \text{PE} \tag{2.29}$$

where $\mathbf{s}$ is a sentence start symbol.

As $MAT(d^n, d^n, d^n)$ provides infinite look-back, the future not seen elements of the input sequence need to be masked with the masking tensor $\mathbf{M}$.

$$\mathbf{W}_h^V \cdot \mathbf{d}^m \cdot \underset{i\rightarrow}{\text{softmax}} \left(\frac{1}{A} (\mathbf{W}_h^K \mathbf{d}^m)^T \mathbf{W}_h^Q \mathbf{d}^m \cdot \mathbf{M} \right) \tag{2.30}$$

Finally, figure 2.8 shows the entire diagram of an encoder-decoder version of the transformer.

2.3. Tasks

In this section we describe the existing tasks that we address in this thesis, together with the metrics that we use to evaluate the performance of our metrics. We additionally introduce a new task: intra-word code-switching that we formally describe in section 5.1.2.

2.3.1. The morphological segmentation task

Morphological segmentation denotes the task of dividing words into their constituting morphemes, i.e., their smallest meaning-bearing units (Wiemerslage et al., 2022), and has been studied extensively in natural language processing (Ruokolainen et al., 2016). The most common form of segmentation consists of separating morphemes at the surface level and is called morphological segmentation (MS). This kind of segmentation is better suited for languages with agglutinative properties that concatenate morphemes into

words without changing their forms. This challenge got relevant with the Morpho-challenge 2005 [11] , English, Turkish, and Finnish (Kurimo et al., 2006, 2010). In the following example, we can see a simple surface segmentation of an English word.

surface form	orchestration
segmentation	orchestra|tion

However, surface segmentation (MSS) is not always well suited for all languages: in fusional languages, morphemes are merged during word formation and, thereby, change their surface forms. If we want to retrieve valuable information about the underlying morphemes that form this word, we must discover them and reconstruct their original forms. This task is known as canonical segmentation (MCS). This task can be defined as jointly segmenting and restoring the original forms of the underlying morphemes. It is important to note that this joint task can also be performed end-to-end. In the following example, we can see the same word as in the previous example again, followed by a restoration step and segmenting of the resulting string afterward. The restored character is bolded.

surface form	orchestration
original form restoration	orchestrat**e**ion
segmentation	orchestrate|ion

Evaluation

To evaluate morphology, we compare all outputs against the gold annotated test sets calculating accuracy and the EMMA F1 metric (Spiegler & Monson, 2010). This metric handles the alignment problem between the predicted segmentation and the ones of the gold annotated data. For this proposal, this metric uses the optimal maximum matching algorithm in a bipartite graph. This matching algorithm is performed with integer linear programming, such as

$$\text{argmax}_B \sum_{i,j} (C \cdot B)_{ij} \tag{2.31}$$

$$s.t. \sum_i b_{ij} \leq 1, \sum_i b_{ij} \leq 1, \sum_j b_{ij} \leq 1, b_{ij} \geq 0, \tag{2.32}$$

[11] http://morpho.aalto.fi/events/morphochallenge2005/

where $C = \{c_{ij}\}$ is a matrix that represents the weight of each potential assignment edge, and the matric $B = \{b_{ij}\}$ is a binary matrix. The value of c_{ij} is an integer representing the number of words sharing the morpheme a_i in the key p_j. If an edge exists from a_i to p_j, then $b_{ij} = 1$.

Finally, with the assignment problem solved, the EMMA score uses a classical f-measure harmonic mean:

$$f_measure = \frac{1 \cdot precision \cdot recall}{precision + recall} \tag{2.33}$$

We refer the reader to the original paper (Spiegler & Monson, 2010) for a more detailed description. We chose this metric over a simple accuracy metric, as it is robust and allows us to consider true positives and negatives, as well as false positives and negatives, all on a morpheme level. Accuracy, on the other hand, would only consider the cases where a prediction achieved an exact match with the gold annotation.

Additional metrics We also use two complementary metrics. The first one is **accuracy**, i.e., the proportion of entirely correctly segmented words, to get a better understanding of partially right segmentation. To get more information about subword-level errors, we also employ **edit distance** on the character level. This is particularly useful to penalize big mistakes in a single word.

2.4. Modeling Machine Translation

Formally, the task of MT consists of converting text X in a source language L_x into text Y in a target language L_y that conveys the same meaning.[12] Translating text $X \in L_x$ into $Y \in L_y$ can be described as a function (Neubig, 2017):

$$Y = \texttt{MT}(X). \tag{2.34}$$

X and Y can be of variable lengths, such as phrases, sentences, or even documents.

If other languages are used during translation, e.g., as pivots, we denote them as $L_1, \ldots, L_n$. We refer to a corpus of monolingual sentences in language L_i as $M^{L_i} = S_1, ..., S_n$.

[12] This is an approximation since it is, in general, not possible to map the meaning of text exactly into another language (Nida, 1945; Sechrest et al., 1972; M. Baker, 2018).

Probabilistic Modeling and Data When using probabilistic MT models, the goal is to find $Y \in L_y$ with the highest conditional probability, given $X \in L_x$. Under the supervised machine learning paradigm, a parallel corpus $C_{parallel} = (X_1, Y_1), ..., (X_n, Y_n)$ is used to learn a set of parameters θ, which define a probability distribution over possible translations. Given $C_{parallel}$, the training objective of an NMT model is generally to maximize the log-likelihood $\mathcal{L}$ with respect to θ:

$$\mathcal{L}_\theta = \sum_{(X_i,Y_i)\in C_{parallel}} \log p(Y_i|X_i;\theta). \tag{2.35}$$

Within this overall framework, there are a number of design decisions one has to make, such as which model architecture to use, how to generate translations, and how to evaluate. In the specific case of LRL, we are constantly faced with a lack of sufficient parallel training data and also have to deal with languages of vastly different typologies and scripts, and we will highlight some of the issues caused by these issues.

Decoding Decoding refers to generating output $\hat{Y}$, given the parameters θ and an input X. Often, decoding is done by approximately solving the following maximization problem:

$$\text{argmax}_{\hat{Y}} p(\hat{Y}|X;\theta) \tag{2.36}$$

Most NMT systems factorize the probability of $\hat{Y} = \hat{y_1}, ..., \hat{y_T}$ in a left-to-right fashion:

$$p(\hat{Y}) = \prod_{t=1}^{T} p(\hat{y}_t|\hat{y}_{<t}, X, \theta) \tag{2.37}$$

Thus, the probability of token $\hat{y}_t$ at time step t is computed using the previously generated tokens $\hat{y}_{<t}$, the source sentence X and the model parameters θ. Common algorithms for finding a high-probability translation are greedy decoding, i.e., picking the token with the highest probability at each time step, and beam search (Lowerre, 1976). Furthermore, A* (Hart et al., 1968) and multiple variants of beam search have been explored (R. Zhou & Hansen, 2005; Furcy & Koenig, 2005). Finally, non-autoregressive decoding has been investigated recently (Gu et al., 2018; Y. Wang et al., 2019; Guo et al., 2020).

2.4.1. Input Representations

The texts X and Y are input into an NMT system as sequences of continuous vectors. However, defining which units should be represented as such vectors is non-trivial. The traditional way is to represent each *word* within X and Y as a vector (or embedding). However, in a low-resource setting, not all vocabulary items often appear in the training data (Jean et al., 2015; Luong et al., 2015). This issue affects languages with a rich inflectional morphology (Sennrich et al., 2016d): as many word forms can represent the same lemma, the vocabulary coverage decreases drastically. Furthermore, for many LRLs, boundaries between words or morphemes are only easy to obtain or well-defined in languages with a standard orthography. Alternative input units have been explored, such as characters (Ling et al., 2015), byte pair encoding (BPE; Sennrich et al., 2016b), morphological representations (Vania & Lopez, 2017b; Ataman & Federico, 2018a), syllables (Z. Zhang et al., 2019), or, recently, a visual representation of rendered text (Salesky et al., 2021). No clear advantage has been discovered for using morphological segmentations over BPEs when testing them on LRLs (Saleva & Lignos, 2021; Gaser, 2022; Gaser et al., 2023). Even if these methods aim to achieve an open vocabulary, they are still sensitive to noise and unknown symbols (Belinkov & Bisk, 2018).

Input representations can be pretrained. The two most common options are i) word embeddings, where each type is represented by a vector, e.g., Word2Vec (Mikolov et al., 2013), Glove (Pennington et al., 2014), or Fasttext (Bojanowski et al., 2017)) embeddings, and ii) contextualized word representations, where entire sentences are being encoded at a time, e.g., ELMo (Peters et al., 2018), GPT-2 (Radford et al., 2019), or BERT (Devlin et al., 2019). However, training of these methods requires large monolingual training corpora, which may not be readily available for LRLs. As most indigenous languages of the Americas have rich morphology, this topic has gathered special interest. The discussion about the usage of segmented morphological input for NMT models is recurrent. (Mager et al., 2022) show that the unsupervised morphological inspired models outperform BPE pre-processing (experimented on 4 languages pares). Similar experiments were done on Quechua–Spanish and Inuktitut–English (Schwartz et al., 2020), comparing BPEs against Morfessor (Smit, Virpioja, Grönroos, Kurimo, et al., 2014). Also, (Ortega et al., 2020) improves the SOTA (state-of-the-art) for Quechua–Spanish MT using a morphological guided BPE algorithm.

2.4.2. Architectures

NMT models typically are sequence-to-sequence (seq-to-seq) models (cf. §2.2.1). They encode a variable-length sequence into a vector or matrix representation, which they then decode back into a variable-length sequence (Cho et al., 2014). The two most frequent architectures are: i) Recurrent Neural Network (RNN), such as LSTM (Hochreiter & Schmidhuber, 1997) or GRU Cho et al. (2014), both with attention (Bahdanau et al., 2015b), and ii) transformers (Vaswani et al., 2017), which define the current state of the art in the high-resource setting (cf. §2.2.3).

As for most neural network models, training an NMT system with one of the mentioned architectures on a limited number of instances in $C_{parallel}$ is challenging (Fernández-Delgado et al., 2014). There are common problems that arise from limited data in the training set. One major advantage of neural models is their ability to learn representations from raw data, in contrast to manually engineered features (Barron, 1993). However, problems arise when not enough data is provided to enable effective learning of features. Another strength of neural networks is their generalization capacity (Kawaguchi et al., 2017). Training a neural network on a small dataset leads to overfitting (Rolnick et al., 2017). Recent studies, however, show empirically that this does not necessarily happen if the network is tuned correctly (Olson et al., 2018).

2.4.3. Evaluation

Accurately judging translation quality is difficult and, thus, often still needs to be done manually. The gold standard evaluation for MT is done by humans. This ideal case would be that bilingual speakers judge the quality of a translation given certain assign scores according to provided criteria such as fluency and adequacy (*Does the output have the same meaning as the input?*). However, manual evaluation is expensive and slow. Moreover, in the case of endangered languages, bilingual speakers can be hard or impossible to find.

An alternative to human evaluation is to use automatic metrics that measure the quality of the output. Automatic metrics provide an alternative.[13] These metrics assign a score to system output, given one or more ground truth reference translations. However, there are a number of challenges in this approach as well (Baisa, 2009; Post, 2018). The most widely used metric is BLEU (Papineni et al., 2002), which relies on token-level

[13]For a detailed overview of automatic metrics for MT we refer the interested reader to specialized reviews (Han, 2016; Celikyilmaz et al., 2020; Chatzikoumi, 2020).

n-gram matches between the translation to be rated and one or more gold-standard translations. For morphologically rich languages, character-level metrics, such as chrF (Popović, 2015) or chrF(Popović, 2015), are often more suitable, as they allow for more flexibility. In the AmericasNLP ST (Mager et al., 2021), this metric was used over BLEU, as it fits better to the rich morphology of many indigenous languages of the Americas.

To have a concrete example, let's have the following Wixarika phrase with an English translation:

yu-huta-me ne-p+-we-'iwa
an-two-ns 1sg:s-asi-2pl:o-brother
I have two brothers

As discussed in Mager, Mager, et al. (2018) it is difficult to translate back from Spanish (or another fusional language) the morpheme *p+* as it has no equivalent in these languages. So if we ignored this morpheme at all, BLEU would penalize the entire word *nep+we'iwa*. In contrast, chrF would credit the translation, even if the *p+* is missing.

Another advantage of looking at individual characters is that this avoids tokenization differences in languages without a clear standard. Another popular metric is METEOR (Denkowski & Lavie, 2011) which handles the problem of synonyms using flexible word matching and takes recall into account, in addition to precision. One shortcoming of these evaluation metrics is that the evaluation depends on the surface forms and not on the ultimate goal of semantic similarity and fluency.

Recent work uses pre-trained models to evaluate the semantic similarity between translations and the gold standard (T. Zhang et al., 2020), but these methods are limited to languages for which such models are available. This is not possible for the indigenous languages of the Americas, as the amount of monolingual data is not enough to train a reliable pre-trained language model[14].

chrF In our experiments, we use chrF as our automatic metric for MT evaluation, given its discussed advantages. The chrF score is n-gram F-score, not requiring additional data or tools. The score is calculated with the harmonic mean:

[14] One exception to this is Quechua, which has a large enough monolingual dataset to train a BERT-like model (Zevallos et al., 2022)

$$\text{chrF}\beta = (1 + \beta^2)\frac{chrP \cdot chrR}{\beta^2 \cdot chrF + chrfR} \tag{2.38}$$

where chrP is the proportion of character n-grams in the hypothesis that also appears in the reference (Precision), and chrR is the proportion of character n-grams in the reference that also appears in the hypothesis (Recall). β is a hyper-parameter that weighs the importance of the precision towards the recall. We refer to the original paper Popović (2015) for an in-depth description of this metric.

3. Morphological segmentation

In this chapter, we discuss the sequence-to-sequence supervised approach to perform morphological segmentation for surface segmentation and canonical segmentation. We test a set of models on the indigenous languages of the and study their performance. We also introduce improvements to these models, analyze the results, and introduce new datasets.

3.1. Previous work

The work on segmentation has an extensive history. The task was first introduced to computational linguistics by Harris (1951). We will explore the main approaches for morphological segmentation. We will start with surface segmentation (MSS) has been the most studied approach. In this recount, we will focus on the low-resource setting and on applications to the ILA (Mager, Gutierrez-Vasques, et al., 2018b).

Rule based approaches Rule-based approaches have been frequently used for low-resource morphological segmentation becausethey do not requirelarge amounts of data. They have been developed, e.g., with finite state transducer (FST) tools like FOMA (Hulden, 2009) or HFST (Lindén et al., 2011). However, this kind of system requires both time and linguistic knowledge. We explore data-driven approaches for the low-resource setting to overcome this limitation. On the other hand, the advantage of these approaches is that no annotated datasets are needed, and therefore, they are highly useful if a grammar for that language exists. Some examples of rule-based methods applied to the ILA are the Finite State approaches to model the morphology of a language: plains Cree (Arppe et al., 2017; Harrigan et al., 2017; Wolfart & Pardo, 1973; Snoek et al., 2014), East Cree (Arppe et al., 2017), for the East Odawa dialect of Ojibwe (Bowers et al., 2017), for Mohawk (Iroquoian language family) (Assini, 2014), for the Bribri (Chibchense language family) (Solórzano, 2017) using the FOMA tool, Quechua (Vilca et al., 2012; Monson et al., 2006), Mapudungun (Monson et al., 2006), and the Argentinian

branch of Quechua and Toba (Porta, 2010). More recently, a new hybrid approach of FST with statistical inference is part of the *Basic Language Technology Toolkit for Quechua* (Rios, 2016). For the Uto-Aztecan languages, a computational tool called "*chachalaca*" performs morphological analysis (Thouvenot, 2011) of Nahuatl. This is a rule-based software focused on Classical Nahuatl. It can generate more than one morphological analysis candidate per word. It is based on grammar that describes most of the 16th-century word constructions. Additionally, Mager, Carrillo, & Meza (2018a) proposes a morphological segmentation tool for the Wixarika language with a supervised approach, using previously given morphological tables and a probabilistic model to infer the inherent morphological rules.

Unsupervised morphological segmentation The task of morphological segmentation was introduced by Harris (1951). However, most work has considered the surface segmentation task, for which unsupervised methods like LINGUISTICA (Goldsmith, 2001) and MORFESSOR (Creutz & Lagus, 2002, 2007; Poon et al., 2009) played an important role. Regarding unsupervised methods, neural methods have been used to tackle the rich morphology of the continent's languages. For the Uto-Aztecan language Tarahumara and the Mayan language Chuj, some studies have attempted to automatically discover affixes through unsupervised approaches (Medina-Urrea, 2007, 2008; Medina Urrea & García, 2006).

Adaptor grammars were first introduced by Johnson et al. (2006). These are nonparametric Bayesian models that modify the generalized probabilistic context-free grammars PCFGs with adaptor parameters that can induce dependencies. These grammars can generally be used to identify grammar structures in an unsupervised fashion (Johnson, 2008). The model has been used extensively for word segmentation. Botha & Blunsom (2013) use them for non-concatenative morphology. Eskander et al. (2016) explores the transfer of morphological knowledge to unseen languages. Sirts & Goldwater (2013) introduce an election method with minimal supervision. Finally, Eskander et al. (2020) introduced MorphAGram, a python framework that allows an easy-to-use system to train adaptor grammars.

Supervised morphological segmentation . In this category, we can find two main approaches. The first is a **tagging based segmentation**. Ruokolainen et al. (2013) cast the task as a sequence-labeling problem using conditional random fields (CRFs; Lafferty et al., 2001). A similar approach was suggested by L. Wang et al. (2016), who

employed a LSTM for tagging. For low resource segmentation, Moeng et al. (2022) use an LSTM-CRF tagger for the Nguni languages. Micher (2017) propose a segmental RNN (SegRNN) for segmenting and tagging Inuktitut.

The second method is to model the problem as a **sequence-to-sequence based segmentation**. This approach was first introduced by us (Kann et al., 2018). This was done in low-resource settings for the ILA Nahuatl, Mexicanero, Wixarika, and Yorem Nokki languages because these supervised methods were shown to perform acceptably even in the low-resource setting (Grönroos et al., 2019). Recent work also included a context to improve morphological disambiguation (Can & Manandhar, 2018; Sakakini et al., 2017). Y. Yang et al. (2019) proposed a pointer network to find surface segmentation boundaries.

Semi-supervised morphological segmentation Ruokolainen et al. (2014) extend the CRF labeling method in a semi-supervised fashion. The same was done with the morphs or systems, where Kohonen et al. (2010) expanded the Morphessor base, and Grönroos et al. (2014) introduced semi-supervised methods with the Flatcat segmenter. Semi-supervised segmentation approaches have been applied to Nahuatl (Gutierrez-Vasques, 2017) using the semi-supervised version of Morfessor. For the lo resource setting Grönroos et al. (2019) used a semi-supervised labeling method to segment Noth Sami words.

Canonical Segmentation For fusional languages, surface segmentation is not very effective. Therefore, restoring morphemes to their canonical form was previously discussed in linguistics Kay (1977) and in the NLP literature. Previous approaches include unsupervised (Naradowsky & Goldwater, 2009), as well as joint models for segmentation and transduction (Cotterell, Vieira, & Schütze, 2016) and neural encoder-decoder models (Kann et al., 2016a; Ruzsics & Samardzic, 2017). However, up to now, supervised models have only been explored in the high-resource setting. No previous work on the canonical segmentation for ILA exists. With our work (Mager et al., 2020), which we will describe in this chapter, we can handle the extreme low-resource scenario with a challenging morphology. On the low-resource side Moeng et al. (2022) use an RNN seq-to-seq approach for the Nguni languages.

Morphological Generation In recent years, morphological generation has experienced substantial progress, with various methods that can be used for the canonical segmentation task. Kann et al. (2016a) used a sequence-to-sequence model to inflect a word

given a set of morphological tags. Sharma et al. (2018a) proposed a pointer-generator model more suitable for the low-resource setting. Aharoni & Goldberg (2017) proposed a neural transducer with hard monotonic attention. Makarov et al. (2017) extended this approach and added a copy operation, and Makarov & Clematide (2018b) proposed imitation learning (Daumé et al., 2009) for training it. Here, we explore the applicability of the models by Sharma et al. (2018a), and Makarov & Clematide (2018b) to low-resource canonical segmentation. The CoNLL-SIGMORPHON Shared Task (Cotterell, Kirov, et al., 2016; Cotterell et al., 2017) released a dataset for the reinfection of 52 languages, including 3 Native American languages.

3.2. Neural Models for Morphological Surface Segmentation

In this section, we describe the basic approach of neural surface segmentation sequence-to-sequence models and then explore different methods to improve them. This includes the previously proposed data augmentation and multi-task training.

3.2.1. Modeling morphological segmentation

We base our study on the encoder-decoder sequence-to-sequence model (§2.2.1) to model the morphological segmentation following work on canonical segmentation by Kann et al. (2016b) and surface segmentation by Kann et al. (2018). The idea is to adapt the seq-to-seq model to a character based RNN encode-decoder model with attention (Bahdanau et al., 2015a) where each token in the input is a character c belonging to a word W. The model generates a sequence of characters with additional morpheme boundary symbols for the output. If the task is canonical segmentation, the model tries to reconstruct the underlying morphemes, predict the morpheme boundaries, and copy the unmodified characters; if the task is surface segmentation, the model will only try to infer the morpheme boundaries and copy the characters of the input string. The following example shows the task of surface segmentation in Raramuri (1) and canonical segmentation in Shipibo-Konibo (2):

3.3. Neural Models for Morphological Surface Segmentation

In this section, we will describe the basic approach of neural surface segmentation sequence-to-sequence models and then explore different methods to improve them. This includes previous proposed data augmentation and multi-task training.

3.3.1. Modeling morphological segmentation

We base our study on the encoder-decoder sequence-to-sequence model (§2.2.1) to model the morphological segmentation following work on canonical segmentation by Kann et al. (2016b) and surface segmentation by Kann et al. (2018). The idea is to adapt the seq-to-seq model to a character based RNN encode-decoder model with attention (Bahdanau et al., 2015a) where each token in the input is a character c belonging to a word W. The model generates a sequence of characters with additional morpheme boundary symbols for the output. If the task is canonical segmentation, the model tries to reconstruct the underlying morphemes, predict the morpheme boundaries, and copy the unmodified characters; if the task is surface segmentation, the model will only try to infer morpheme boundaries and copy the characters of the input string. The following example shows the task of surface segmentation in Rarámuri (1) and canonical segmentation in Shipibo-Konibo (2):

1) k o ' n á r s i a → k o ' - n á r - s i - a
2) o i n x o n a → o i n t i - x o n

Encoder. The first part of our model is a bidirectional Recurrent Neural Network (RNN) which encodes the input sequence, i.e., the sequence of characters of a given the word $X^w = x_1^w, x_2^w, \ldots, x_{T_v}^w$, represented by the corresponding embedding vectors $v_{w_1}^w, ..., v_{w_{T_v}}^w$.

Encoding with this bidirectional RNN yields the forward hidden state $\overrightarrow{h}_i = f\left(\overrightarrow{h}_{i-1}, v_i\right)$ and the backward hidden state $\overleftarrow{h}_i = f\left(\overleftarrow{h}_{i+1}, v_i\right)$, for a non-linear activation function f. Their concatenation $h_i = \left[\overrightarrow{h_i}; \overleftarrow{h_i}\right]$ is passed to the decoder. If encoded with a transformer architecture, we would convert the input $\mathbf{e}^n = \text{FF}(\text{MAT}(\mathbf{X}^w, \mathbf{X}^w, \mathbf{X}^w))$. The encoded tensor $\mathbf{e}^n$ will be used later in the decoding step.

Decoder. The second part of our network, the decoder, defines a probability distribution over strings in $(\Sigma \cup S)^*$, for an Σ and a separation symbol S:

$$p(c \mid w) = \prod_{t=1}^{T_c} p(c_t \mid c_1, \ldots, c_{t-1}, w), \tag{3.1}$$

where $p(c_t \mid c_1, \ldots, c_{t-1}, w)$ is computed using an attention (see equation 2.21) and an output softmax layer over $\Sigma \cup S$. A more detailed description of the model is provided in section 2.2.2.

This approach can also be modified to use a transformer architecture instead (see §2.2.3), using the contextual embedding from the decoder layer $\mathbf{e}^n$, together with the inputs $\mathbf{d}^m$, we can calculate the self-attention $\mathbf{d}^{m+1}$ shown in equation 2.28. For the output layer, we also use a softmax function over $\Sigma \cup S$.

Given the seq-to-seq model explained in the past section 3.3.1, we will now improve this approach in a semi-supervised manner. We first explain how we can use multi-task training for our task. Next, we explain a data augmentation approach. Both methods correspond to one of our previous works (Kann et al., 2018).

Multi-Task Training We want to use additional data to train the neural network so that the network can improve its performance, even with an extremely low-resource scenario. This obeys the following reasoning: multi-task training should act as a regularizer; the segmentation task consists in large parts of learning to copy the input character sequence to the output; in the case of unlabeled data, we expect the character language model in the decoder to improve since it is trained on additional data. First, we have to choose what external data can be used for training. There are two main options: use monolingual data for each language and the usage of randomly generated strings x^r, such as $x^r \in \Sigma^*$.

To leverage these data during training, we first define an autoencoding auxiliary task, which consists of encoding the input and decoding an output that is identical to the original string. Then, our multi-task training objective is to maximize the joint log-likelihood of this auxiliary task and our main segmentation task:

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{(w,c)\in\mathcal{T}} \log \grave{} \left(c \mid e(w)\right) + \sum_{a\in\mathcal{A}} \log \grave{} \left(a \mid e(a)\right)$$

$\mathcal{T}$ denotes the segmentation training data with examples of a word w and its seg-

mentation c. $\mathcal{A}$ denotes either a set of words in the language of the system or a set of random strings. The function e describes the encoder and depends on the model parameters $\boldsymbol{\theta}$, which are shared across the two tasks. As shown in table 3.1, we insert a special character for each sub-task at the beginning of each training instance.

Input	Output
[SEG] n e k i	n e | k i
[COPY] k w a n i s k	k w a n i s k

Table 3.1.: [SEG] and [COPY] are special symbols that we use to mark each sub-task for the multi-task training. [SEG] is the original main segmentation task, while [COPY] is used to copy either the random string or the unlabeled word.

Data Augmentation The other option to use additional data for our task is to append the randomly generated words, or unlabeled data, to the segmentation data. In this case, the system takes this input as a single task. The input in this approach is shown in table 3.2.

Input	Output
n e k i	n e | k i
k w a n i s k	k w a n i s k

Table 3.2.: Segmentation labeled data and unlabeled data are appended and used for training.

The size of the additional training examples is set before training as a hyper-parameter. A primary disadvantage of this method is that it can also teach the model, not to segment words that should be segmented, as no difference is marked.

3.4. Improving morphological surface segmentation with pre-training

Supervised morphological segmentation relies on clean gold annotated data done by linguists. This can be a drawback when these segmentations are applied to real-life data (i.e., parallel data), or when they diverge in orthography or even dialectical variations. Therefore, we extend in a semi-supervised fashion the sequence-to-sequence model introduced by Kann et al. (2018) for surface segmentation and can also be applied for canonical segmentation (Kann et al., 2016b; Mager et al., 2020).

To include non-annotated words, we propose two extensions to the previously described approach (§3.3). Our first approach is to extend the multi-task variant of Kann et al. (2018). In addition to the auxiliary task of auto-encoding a random string, we use words from a noisy text as a second auto-encoding task. Our extended multi-task training objective is to maximize the joint log-likelihood of the two auxiliary tasks and our main segmentation task:

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{(w,c)\in\mathcal{T}} \log \grave{} (c \mid e(w)) + \sum_{a\in\mathcal{A}} \log \grave{} (a \mid e(a)) + \sum_{a\in\mathcal{R}} \log \grave{} (a \mid e(a))$$

where $\mathcal{T}$ denotes the gold annotated training data, w the unsegmented word and c the segmented output; $\mathcal{A}$ the randomly generated strings for the first autoencoding task; and $\mathcal{R}$ is the set of words taken without annotation. The function e describes the encoder. The sizes of $\mathcal{A}$ and $\mathcal{R}$ are hyper-parameters.

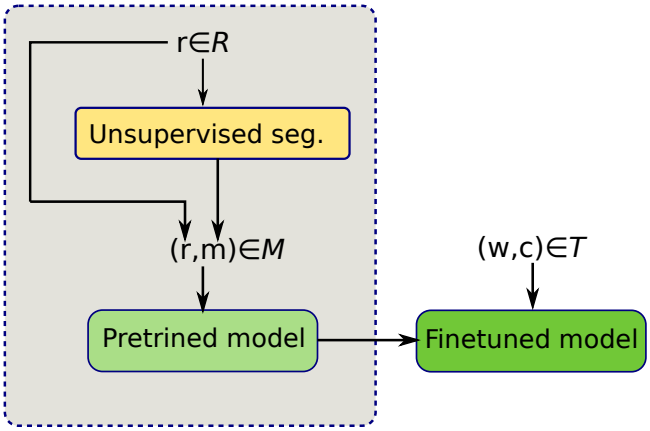

Figure 3.1.: Training algorithm, using noisy labels generated by an unsupervised segmentation model from a raw text R, and fine-tuned with golden labeled data T.

3.4.1. Pretraining-finetuning

Our second approach involves pretraining a sequence-to-sequence segmenter (§3.3.1) and finetuning it with annotated data. This is done with a not annotated raw text $r \in \mathcal{R}$, where $\mathcal{R}$ is a list of words and r is a single word. R is used as the input for an unsupervised morphological segmentation model (i.e., Morfessor Smit, Virpioja, Grönroos, & Kurimo (2014)). Then, a sequence-to-sequence model is trained with the

	Mexicanero	Nahuatl	Wixarika	Rarámuri	Yorem N.
train	427	540	665	604	511
dev	106	134	176	134	127
test	355	449	553	449	425
total	888	1123	1394	1187	1063

Table 3.3.: Number of examples in the final data splits for all languages.

ordered pair of $(r, m) \in \mathcal{M}$, where m is the segmented output of the unsupervised model. To avoid possible orthographic variations among datasets, we define the vocabulary as $\Sigma = \Sigma_R \cup \Sigma_\mathcal{T}$, where $\Sigma_\mathcal{R}$ is the alphabet of the raw text $\mathcal{R}$ and $\Sigma_\mathcal{T}$ the one of gold annotated T. Afterward, we fine-tune the original model with the data of $\mathcal{T}$. The training algorithm is shown in figure 3.1.

3.4.2. Experimental setup

To gain a better insight into how neural supervised methods perform when trained in such a low-resource setup for polysynthetic languages, we experiment on five indigenous languages: Mexicanero, Nahuatl, Wixarika, Rarámuri, and Yorem Nokki.

Datasets

For Raramuri, we curate a new machine-readable dataset. We manually extracted the segmented morphemes from a specialized linguistics paper (Caballero, 2010) and thesis (Caballero, 2008) that contain segmented and non-segmented words. Both sources annotate the Raramuri variant of the village of Choguita.

For the other languages, we use the dataset from our previous work (Kann et al., 2018). It is based on books from the collection *Archive of Indigenous Languages* in Mexicanero (Canger, 2001), Nahuatl (Lastra de Suárez, 1980), Wixarika (Gómez & López, 1999), and Yorem Nokki (Freeze, 1989). We include segmentable and non-segmentable words in our datasets to ensure that our methods can correctly decide against splitting up single morphemes. The phrases in all languages are mostly parallel, such that the corpora are roughly equivalent.

Table 3.4 shows the most common morphemes from all languages that work. The morphemes tend to be short in polysynthetic languages, with a few examples. In table 3.5, we see that the most morphologically rich language from the group is Wixarika (with 3.25 morphemes per word). This is also in line with the 81.6% of words that are

segmentable for this language. This higher morphological complexity naturally produces data sparsity at the token level. On the other hand, we have the example of Mexicanero, which only has 60% segmentable words. In general, the morphological richness of all these languages is higher that of the other languages considered morphological rich, for example, Arabic with 1.22 morphemes per word and 18.6% of segmentable words.

We use the monolingual side corresponding to each language from the parallel data for unlabeled monolingual data. These data are better described in §4.5.2. Unfortunately, we could only collect parallel data for Wixarika, Nahuatl, Rarámuri, and Shipibo-Konibo. For Mexicanero and Yorem Nokki, we cannot work with unsupervised or semi-supervised methods that would require unlabeled data.

Final splits. To make follow-up work on minimal-resource settings for morphological segmentation easily comparable, we provide predefined splits of our datasets[1]. 40% of the data constitute the test sets. Of the remaining data, we use 20% for development and the rest for training. The final numbers of words per dataset and language are shown in Table 3.3.

Baselines

BPEs were applied for the first time for sub-word level segmentation by (Sennrich et al., 2016d). This method is an unsupervised way to perform segmentations, but these segments are not intended to be morphemes. It was introduced to reduce the sparsity in machine translation. We use these methods to use it further as a comparison point for the machine translation task.

Morfessor (Smit, Virpioja, Grönroos, & Kurimo, 2014) As an unsupervised method, we use Morfessor 2.0, a statistical model for discovering morphemes using minimum description length optimization.

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) modify the `FC` implementation by adding a lexicon size restriction and increasing the tendency of the model to increase the segmentation of commonly seen words.

[1] Our datasets can be found together with the code of our models at `http://turing.iimas.unam.mx/wix/MexSeg` .

Mexicanero		Nahuatl		Wixarika		Rarámuri		Yorem N.	
frq.	m.	frq.	m.	frq.	m.	frq.	m.	frq.	m.
136	ni	155	o	327	p+	r	98	102	k
128	ki	99	ni	230	ne	ma	97	88	m
114	ti	84	ti	173	p	ri	94	87	ne
105	u	81	k	169	ti	a	94	83	ka
70	s	61	tl	167	ka	n	93	79	ta
44	mo	59	mo	98	u	ti	91	54	po
42	ka	55	s	97	ta	m	65	50	e’
39	a	52	ki	95	a	ki	48	36	ye
31	nich	48	i	92	pe	si	44	36	su
31	$i	43	tla	91	e	ra	44	36	ri
24	ta	39	’ke	80	r	i	42	34	a
24	l	34	nech	74	wa	k	39	31	me
22	tahtanili	31	no	69	me	ni	36	30	wa
21	no	27	ya	68	ni	t	36	30	re
17	ya	27	tli	68	ke	b	35	27	na
17	t	24	x	66	eu	p	34	24	wi
17	ke	23	tlanilia	58	ye	s	33	24	ßa
17	ita	23	e	57	ri	po	31	23	te
16	piya	21	tika	52	tsi	w	28	20	si
15	an	21	n	52	te	o	27	16	’wi

Table 3.4.: The most frequent morphs (*m.*) with their frequencies (*frq.*) in our datasets.

CRFs As our first supervised model, we use the conditional random fields (Lafferty et al., 2001) segmentation model of Ruokolainen et al. (2014).

semiCRF We also investigate the capabilities of `semiCRFs` Sarawagi & Cohen (2005) for this particular task. For this, we use the Chipmunk implementation Cotterell et al. (2015b).

Pointer Generator Network (See et al., 2017) is an extension of seq-to-seq, that introduces a learned probability of copying and decides when to copy or generate (see §3.5.2).

Seq2seq We also use a vanilla RNN Sequence-to-Sequence (S2S) model with attention. The first variant (`s2s`) employs a supervised neural model. Additionally, we use the most promising extension proposed by Kann et al. (2018), adding randomly generated strings

	Mexcanero.	Nahuatl	Wixarika	Rarámuri	Yorem N.
Words	888	1123	1385	916	1063
SegWords	539	746	1131	591	774
Morphs	1889	2467	4502	1870	2266
UniMorphs	602	810	653	787	662
Seg/W	0.606	0.664	0.816	0.645	0.728
Morphs/W	2.127	2.196	3.250	2.041	2.131
MaxMorphs	7	6	10	5	10

Table 3.5.: Number of words, segmentable words (SegWords), total morphs (Morphs), and unique morphs (UniMorphs) in our datasets. Seg/W: proportion of words consisting or more than one morpheme; Morphs/W: morphemes per word; MaxMorphs: maximum number of morphemes found in one word.

in an auto-encoding fashion (`s2s+multi`).

3.4.3. Results

Table 3.6 shows the results for the surface segmentation task for all languages. Not all languages. As mentioned before, Mexicanero and Yorem Nokki have only available annotated data; therefore, only the results for the supervised methods can be reported. The proposed fine-tuning method performed better for two out of three languages: for `hch` seq2seq-finetuned achieved 83.20 EMMA F1, compared to the next best system crf, with 82.43 EMMA F1. For Raráumri, we could observe the same, with seq2seq-finetuned performing better than the next best pointer generator network (pointernet) with a 90.13 EMMA F1 score. The exception for this was `nah` with CRFs outperforming all neural networks, with 87.83 EMMA score. Regarding the systems without non-labeled monolingual data, we found that seq2aeq+multilingual (93.18 for `azd`) and the pointer generator network for 92.12 for `mfy` achieved the best results. This can be explained because both methods attempt to solve the copy string problem.

Overall, we saw a clear advantage for the seq2seq+finetune method when external un-labeled monolingual data is available. However, if that is not the case, relaying on a neural network with a multitask + random strings is the best alternative. Surprisingly, the combination of multitasking and pre-training had no advantage over the other methods. The same is true for the Pointer-Generator-Network, which could not achieve any improvements from pretraining.

system	azd	hch	mfy	nah	tar
bpe.	-	53.17	-	53.38	62.54
morfessor	-	61.51	-	60.48	59.05
flatcat.	-	62.28	-	58.94	64.65
lmvr.	-	61.27	-	60.55	65.46
semiCRF.	88.95	68.10	82.84	81.92	81.22
crfs.	92.81	82.43	91.10	**87.83**	89.79.
seq2seq.	93.09	82.42	89.71	84.62	88.47.
seq2seq-multitask.	*93.18*	81.75	90.50	84.90	88.37.
pointernet	92.23	65.60	*92.12*	83.85	90.13.
seq2seq-multitask+raw	-	80.88	-	84.68	88.91
seq2seq-finetuned.	-	**83.29**	-	86.58	**91.74**
pointernet-f	-	81.45	-	84.58	88.20

Table 3.6.: Surface segmentation results on the test set for hah, nah, and tar. Canonical segmentation results for shp. F1 score is calculated using EMMA. Bold letter numbers are the best systems when labeled, and unlabeled data are available. Numbers in italic font refer to the best scores when only supervised data is available.

3.4.4. Discussion

From the results discussed in the previous section, we find that:

- The sequence-to-sequence models are especially good for modeling the morphology of polysynthetic languages, even in low-resource languages.
- The low-resource performance can further be further improved with additional unlabeled data. The semi-supervised methods for morphological segmentation improve their supervised seq-to-seq counterparts in all cases. However, for nah, the CRFs methods outperform the neural methods.
- The lack of monolingual data is a problem for modeling languages that are not the main focus of NLP. In our case, we experienced this limitation for Yorem Nokki and Shipibo-Konibio.
- It is helpful for seq-to-seq models to use random-string in a multi-task training fashion. However, it should also be stated that it depends on the characteristics of each language.

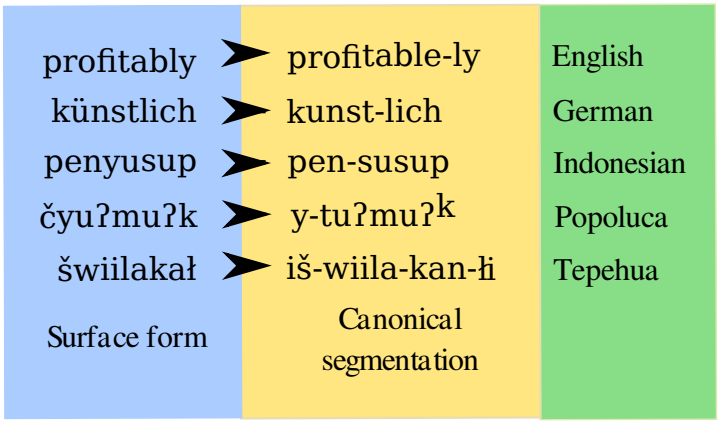

Figure 3.2.: Canonical segmentation examples for all languages in our experiments.

- Finally, we also want to note that the task of morphological segmentation for polysynthetic languages requires a broad collaboration between native speakers, linguists, and the NLP community.

3.5. Canonical Morphological Segmentation for low resource settings

Neural models have neen shown to perform well on the canonical segmentation (MCS) task when large amounts of training data are available (Kann et al., 2016a; Ruzsics & Samardzic, 2017). Nevertheless, datasets with morphological annotations are difficult to obtain becuase they require expert annotators. Furthermore, many languages with complex morphology are spoken by a limited number of people or are listed as endangered languages (Mager, Gutierrez-Vasques, et al., 2018a), which further reduces the possible annotator pool even more. However, morphological segmentation is important for downstream tasks like machine translation (Conforti et al., 2018; Vania & Lopez, 2017a), dependency parsing (Seeker & Çetinoğlu, 2015; Vania et al., 2018), or semantic role labeling (Sahin & Steedman, 2018). Moreover, high performance on these tasks can yield more language-independent NLP models (Gerz et al., 2018). Additionally, morphological segmentation is also important when supporting the work of field linguists (Ide, 2017).

Here, we focus on low-resource canonical segmentation. We propose two new models for the task, which are successfully applied to a related morphological generation task called *morphological inflection*. The approaches we investigate are (i) an LSTM pointer-

generator model (Sharma et al., 2018a), and (ii) a neural transducer trained with imitation learning (IL; Makarov & Clematide, 2018b). Since canonical segmentation and morphological inflection are character-level string transduction tasks, we hypothesize that models that can learn one from limited data will also be able to do so for the other.

We experiment on three benchmark datasets in German, English, and Indonesian but simulate a low-resource scenario by reducing the number of training examples. We further evaluate our models on datasets for two *truly* low-resource polysynthetic languages: Popoluca and Tepehua. We find that our new models indeed outperform previous approaches on all languages in the low-resource scenario. For additional insight, we also evaluate the performance of all models for varying amounts of training data from the high-resource languages and find that the neural transducer with imitation learning outperforms all other models in all but one set with up to 600 training examples. Using the entire training set for English, German, and Indonesian, the state-of-the-art LSTM sequence-to-sequence model performs best. However, the difference to our proposed models is below 3.3% accuracy for all languages and models. Figure 3.2 provides examples for all five languages we experiment on.

3.5.1. Datasets for Popoluca and Tepehua

We release two new datasets for low-resource canonical segmentation in Popoluca (POQ) and Tepehua (TPP)[2]. We use these two languages to shed light on polysynthetic languages that also exhibit fusional phenomena. The high-resource datasets introduced by Cotterell, Schütze, & Eisner (2016) cover fusional (German), analytic (English), and agglutinative (Indonesian) languages.

We collect words for our datasets from two books belonging to the Archive of Indigenous Languages (ALI-Colmex) of the College of Mexico (*Colegio de México*). For Popoluca, we used the book by Wichmann (2007) and for Tepehua, that by MacKay & Trechsel (2010). We include segmentable and non-segmentable words to avoid oversegmentation by our systems. A set of Spanish sentences are used to elicit the data for both languages. This set of sentences is the same across the entire ALI-Colmex collection. Then, for each language, the authors of the books asked native speakers to translate the sentences into the respective languages (elicited data). Afterward, they performed a glossing of the translated text. For more details, we refer the reader to the original books.

[2]The dataset is available at http://turing.iimas.unam.mx/wix/canseg

ENG		DEU		IND		POQ		TTP	
ly	7.53	er	15.66	men	8.65	y	6.08	ya	8.58
ness	3.41	in	10.38	nya	8.29	∅	6.08	łi	6.27
er	2.99	ung	8.14	an	7.18	n	3.98	ka	4.46
ion	1.87	lich	4.37	kan	6.61	ny	3.56	ta	4.29
y	1.50	keit	3.96	di	5.31	k	3.35	ti	3.80
ity	1.24	ig	3.78	pen	4.14	p	2.94	ik	2.81
ation	0.99	los	1.23	ber	2.81	t+k	2.52	ni	2.64
un	0.88	chen	1.16	i	2.45	ky	2.31	ča	2.31
ic	0.85	bar	1.13	ter	1.91	wat	2.10	la	1.82
al	0.81	ver	0.81	per	1.25	aʔ	2.10	maa	1.82
ist	0.76	un	0.77	se	0.72	taʔ	1.89	kin	1.82
able	0.74	e	0.49	ke	0.71	ʔeš	1.26	waa	1.82

Table 3.7.: Relative frequencies of the 12 most common morphemes for each language; ENG = English; DEU = German; IND = Indonesian; POQ = Popoluca; TTP = Tepehua.

	>3Morph.	Surf.	Canon.	NoSeg.	M./W.	Ch./W.
ENG	00.01	36.40	22.83	41.37	01.60	08.18
DEU	01.86	46.07	53.86	00.00	02.20	12.48
IND	05.57	46.21	23.66	30.14	02.07	08.65
POQ	12.12	23.74	56.57	19.70	02.41	06.78
TTP	32.00	21.50	63.00	15.50	03.03	08.62

Table 3.8.: Statistics for all five canonical segmentation datasets. Percentages of words with more than 3 morphemes (>3 Morph.), surface segmentation (Surf.), canonical segmentation (Canon.), and without segmentation (NoSeg.), as well as the average number of morphemes per word (M./W.) and characters per word (Ch./W.).

Table 3.8 shows statistics for all five datasets used in this paper. Importantly, the German dataset only contains multi-morpheme words. Additionally, we observe that most of the Indonesian words only require surface segmentation, while English is the language with the highest ratio of words that do not require any segmentation. On the other hand, Popoluca and Tepehua have the highest proportion of words that require both the splitting and restoration of the canonical forms. Moreover, both languages have a high number of words that contain more than 3 morphemes per word and the highest morphemes-per-word rate. Adding to these facts, the small amount of data available for these languages makes morphological segmentation even harder. To better understand the underlying morphemes seen in each language, we extract the 15 most

common morphemes for each dataset. These morphemes, together with their relative frequencies in our datasets, are shown in Table 3.7.

3.5.2. Models

Inspired by the successes of the two models for low-resource morphological inflection, we propose to apply these architectures to canonical segmentation with limited training data. In this section, we introduce the models.

Pointer-Generator Network

Motivation. The first model we apply to low-resource canonical segmentation is a pointer-generator network (PtrGenNet) (See et al., 2017), i.e., a Sequence-to-Sequence (S2S) model with a mechanism to copy input elements over to the output. We believe that this should make the learning problem easier and help in settings with limited training data. The pointer-generator network can be considered a hybrid between an attention-based sequence-to-sequence model (Bahdanau et al., 2015b) and a pointer network (Vinyals et al., 2015).

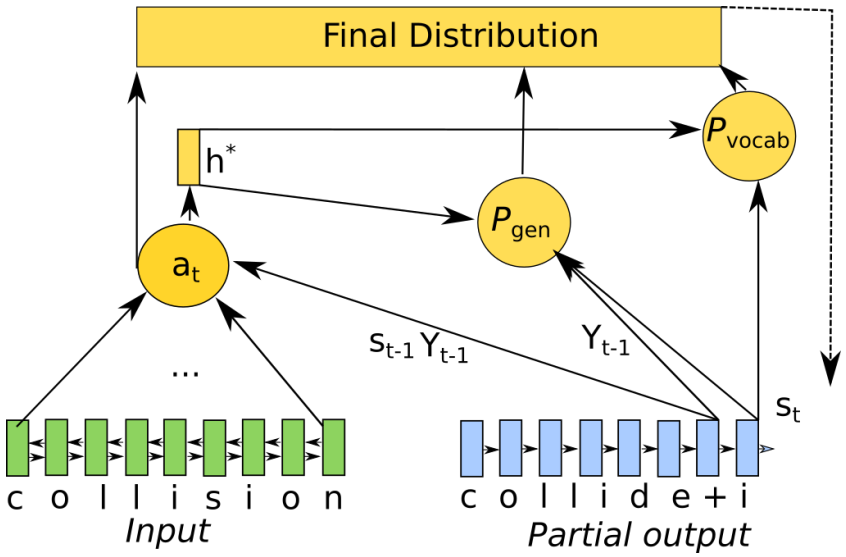

Figure 3.3.: Pointer Generator Network architecture.

Model description. Our pointer-generator network consists of a bidirectional LSTM (Hochreiter & Schmidhuber, 1997) encoder and a unidirectional LSTM decoder with an attention mechanism. We cast the task of canonical segmentation as a character-based

sequence-to-sequence problem, with the characters of the original word as the input and the characters of the restored morphemes in combination with segment boundary markers as the output. Both our encoder and decoder operate at the character level.

The pointer-generator network differs from the standard sequence-to-sequence architecture in that the decoder calculates a probability of copying an element from the input to the output instead of generating. Here, we follow Sharma et al. (2018b) and use two separate encoders: one for the lemma and one for the morphological tags. The decoder then computes the probability distribution of the output at each time step as a weighted sum of the probability distribution over the output vocabulary and the attention distribution over the input characters. The weights can be seen as the probability to generate or copy, respectively, and are computed by a feed-forward network, given the last decoder's hidden state.

The overall architecture is shown in Figure 3.3. The probability of generating a character is computed as

$$p_{gen} = \sigma(w_h h_t^* + w_s s_t + w_y y_{t-1} + b) \tag{3.2}$$

for a context vector h_t^*, a decoder hidden state s_t, the last predicted character y_{t-1}, weights w_h, w_s, w_y, and a bias vector b.

The probability distribution over the characters is calculated as

$$P_{vocab} = \text{softmax}(V[s_t; h_t^*]) \tag{3.3}$$

where V is a learnable function. Then, the final probability of predicting a certain output character c is computed as the sum of p_{gen} multiplied by the probability of c given by the decoder and the total attention distribution $\sum a^t$ multiplied with the probability of copying $1 - p_{gen}$ character c:

$$P(c) = g_{gen} P_{vocab}(c) + (1 - p_{gen}) \sum a_i^t \tag{3.4}$$

This copying mechanism is especially beneficial in low-resource settings, as we can see in the results §3.9. For a more detailed description, we refer the interested reader to Sharma et al. (2018b). We will refer to this model as `PGNet`.

Additionally, there is the smaller difference that, instead of encoding tags and characters as one sequence, employing a single encoder, this model disposes of 2 encoders: One processes (i) the language tag and (ii) the morphological tags of the target form.

The second one encodes the characters of the input word. Two attention mechanisms are used, and a concatenation of the resulting context vectors forms the input to the decoder.

Hyperparameters. All encoder and decoder hidden states are 100-dimensional, and our embeddings are of size 100. For training, we use Adam Kingma & Ba (2014) with a learning rate of 0.001 and a mini-batch size of 32. To avoid overfitting, we use dropout Srivastava et al. (2014) with a coefficient of 0.3 for the high-resource setting and 0.5 for the low-resource setting. We train our model for 100 and 300 epochs and use early stopping with a patience of 10 and 100 for the high-resource and the low-resource setting, respectively.

Neural Transducer with Imitation Learning

Motivation. Hard monotonic attention networks Aharoni & Goldberg (2017) have performed well on morphological generation in the low-resource setting. These systems use a nearly-monotonic alignment between the source and output characters. For our second model, we employ the variant proposed by Makarov & Clematide (2018c), which uses imitation learning for end-to-end training and, thus avoids, error propagation.

Model description. This model is a sequence-to-sequence model with hard monotonic attention Aharoni & Goldberg (2017), which transduces an input sequence of characters into an output sequence by performing edit operations. Following Makarov & Clematide (2018a), it can perform three operations: insertion, deletion, and copy. However, instead of using maximum likelihood estimation (MLE), training is done with imitation learning. The idea is to train a model to imitate an expert policy that maps the training configurations to a set of optimal actions. We aim to minimize the sequence-level loss and action-level loss.

The training is composed of two steps: a roll-in and a roll-out stage. In the roll-in stage, the model gathers actions by sampling from the expert policy. This process returns a set of decoder outputs called configurations. A sequence-level loss is computed for each valid action per configuration for the roll-out stage. For that, the action is executed and is compared to the optimal action sequence of the expert. This loss is defined in terms of Levenshtein distance Levenshtein (1966) between the prediction and the target and the cost of the actions. The cost function uses the information from a character aligner. After calculating the sentence-level loss, it is fed into an action-

level loss. This loss expresses how much certain action suffers relative to the optimal action under the current policy. This is achieved by minimizing the negative marginal log-likelihood of all optimal actions Makarov & Clematide (2018a).

Hyperparameters. For the encoder and the decoder of this model, we use one layer with a 200-dimensional size, with a dropout of 0.5. For optimization, we use ADADELTA Zeiler (2012) with a learning rate of 0.1. As the RNN unit, we use an LSTM. We train the model for 30 epochs, with a patience of 10 epochs. For IL training, we use an inverse sigmoid and a decay rate of 12. For decoding, we employ beam search with a beam of width 4.

3.5.3. Experiments

We now describe the experiments we conduct to explore the performance of our models both in a high-resource and low-resource setting.

	English			German			Indonesian		
	Acc.	ED	F1	Acc.	ED	F1	Acc.	ED	F1
`SemiCRF`	64.7	64.3	76.6	41.9	108.3	74.1	70.4	46.3	84.3
`joint`	72.0	98.0	76.0	59.0	101.0	76.0	90.0	15.0	80.0
`s2s`	♦78.0	**41.2**	88.4	♦77.1	**47.8**	**89.3**	♦94.3	**7.6**	**97.9**
`PGNet`	77.5	42.4	**88.5**	74.8	52.1	88.2	92.9	10.0	97.5
`IL`	76.7	42.9	87.2	73.8	52.3	87.2	93.4	8.4	97.6

Table 3.9.: Results for `semiCRF`, `joint`, `s2s`, `PGNet`, and `IL` for the high-resource setting of English, German, and Indonesian. Lower scores in the ED columns are better. For accuracy, ♦ indicates statistical significance at $p < .01$.

Data

The canonical segmentation datasets for English (`eng`), German (`deu`), and Indonesian (`ind`) by Cotterell, Vieira, & Schütze (2016) each consist of 8000 training, 1000 development, and 1000 test examples. We consider the complete training set to be high-resource. The datasets feature a splitting into 10-folds for cross-validation. For our low-resource experiments, we randomly take a subset of words from each training fold but keep the development and test sets unchanged.

The high-resource datasets cover three languages: English, German, and Indonesian. English is an analytic language from the Indo-European family (Konig & Van der Auw-

era, 2013), German exhibits fusional typology (Hawkins, 2015), while Indonesian is an agglutinative language whose morphology involves the use of affixation, reduplication and cliticization (Hiroki Nomoto & Bond, 2018).

We also experiment with two polysynthetic low-resource languages: Tepehua and Popoluca (cf. Section 3.8). As those datasets are small (900 words for each language), we divide the datasets into 9-folds, each containing 100 training, 100 development, and 700 test examples.

Baselines

We compare the neural transducer with imitation-learning (`IL`) and the pointer-generator network (`PGNet`) to three strong baselines, including the current state of the art for the canonical segmentation task.

Encoder-Decoder (`s2s`). Our first baseline is a character-based encoder-decoder Recurrent Neural Network (RNN) architecture with attention as proposed by Kann et al. (2016a). It defines (in combination with a reranker which we omit here since it is orthogonal to our work) the state of the art on high-resource datasets. To perform experiments in the low-resource setting, we re-implement this model using OpenNMT (Klein et al., 2017). The hyperparameters suggested by Kann et al. (2016a) are as follows: the RNNs of the encoder and decoder have 100 hidden units each; the embedding size is 300. For optimization, we use ADADELTA (Zeiler, 2012) with a minibatch size of 20.

Semi-Markov CRF (`semiCRF`). Our first non-neural baseline is the ChipMunk Cotterell et al. (2015a) implementation of a semi-Markov CRF Sarawagi & Cohen (2005). Although the system can use additional complementary information like morphological tags or dictionaries, we decide not to include those to make our results comparable across all languages and systems.

Joint log-linear model (`joint`) As a second non-neural system, we use a log-linear model that jointly segments and generates underlying representations of the input words Cotterell, Vieira, & Schütze (2016). For segmentation, it uses the `semiCRF`, and for the transduction of the underlying forms, it uses a probabilistic final state transducer Cotterell et al. (2014).

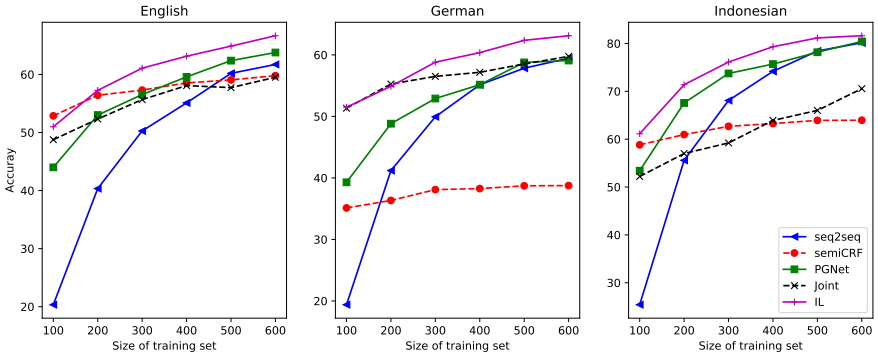

Figure 3.4.: Accuracy for different simulated low-resource settings for our high-resource languages.

Training Details

We choose the hyperparameters for all models following the mentioned previous work. All neural models and the `semiCRF` were trained on a server with 2 Intel(R) Xeon(R) CPU v4@ 2.20GHz, with 4 Nvidia GTX 1080ti graphic cards. A MacBook Pro 2009 laptop was used to train the joint log-linear model. Links to the repositories we use are listed in the complementary material.

	Tepehua			Popoluca		
Model	Acc.	ED	F1	Acc.	ED	F1
`SemiCRF`	21.9	285.3	35.9	26.0	215.0	41.4
`joint`	11.2	335.4	29.5	14.6	393.6	36.8
`s2s`	4.1	532.4	7.7	13.2	309.4	23.3
`PGNet`	17.2	321.7	29.3	27.0	211.0	42.5
`IL`	♦28.4	**242.6**	**44.0**	♦37.4	**158.8**	**54.7**

Table 3.10.: Results for the low-resource languages Popoluca and Tepehua. For accuracy, ♦ indicates statistical significance at $p < .01$.

Results

Low-resource *simulation*. Figure 3.4 shows the accuracy of all systems for different low-resource training set sizes (100, 200, 300, 400, 500, and 600 examples) for English, German, and Indonesian. To ensure statistical significance, we use McNemar's test McNemar (1947) for all accuracy results (Tables 3.9 and 3.10, Figure 3.4) comparing the best and the second best systems. All results are significant at $p < 0.01$. The

scores of all systems vary across languages. However, IL consistently is among the two best systems in terms of accuracy in all settings. For 100 training examples IL is the second-best performing system with 50.99% for English, just behind the semiCRF with 52.87%. For German 51.49% IL slightly outperforms Joint (51.33%) and obtains the best score for Indonesian with 61.14%, where the second best system is semiCRF (58.82%). Moreover, from 300 examples up to 600, IL strongly outperforms all other systems, including non-neural ones.

If we compare the performance of our two proposed systems with s2s, PGNet strongly outperforms s2s with improvements of 22.27%, 17.84%, and 25.66% absolute accuracy for English, German, and Indonesian, respectively, in the setting with 100 training examples; while IL have even bigger gains with improvements of 30.65%, 32.1% and 35.73% of accuracy, respectively.

Looking at the learning curves for each model for increasing training set sizes, we can see that both proposed systems show monotonically increasing performance: they take advantage of more data well, but still achieve decent performance in the low-resource setting, even outperforming all non-neural systems in some settings. On the contrary, the non-neural models joint and semiCRF have, in many cases, a good start but only benefit to a limited extent from additional data. A table listing all individual results for this experiment is included in the supplementary material.

Low-resource *languages*. Results for Popoluca and Tepehua are shown in Table 3.10 and confirm most of the tendencies seen in our low-resource simulation experiment. s2s barely predicts any correct segmentation for Tepehua, and only obtains 4.14% absolute accuracy and 13.19 F_1 score. Similarly, for Popoluca, s2s reaches only 13.19% accuracy. The performance of IL is consistently better on all metrics, with substantial gains for Tepehua of 6.5% accuracy over the closest system (semiCRF) and 10.4% accuracy over PGNet.

The performance of PGNet is consistently better than that of s2s, with gains of 13.03% and 13.77% accuracy for Tepehua and Popoluca, respectively. joint surprisingly shows a low performance for our two low-resource languages, obtaining a 17.2% lower accuracy than the best model for Tepehua (IL), and a 22.8% lower accuracy than the best system for Popoluca (IL).

Overall, all systems perform notably worse for Tepehua and Popoluca than for the high-resource languages. This could be due to their high morphological complexity, as shown in Table 3.8.

High-resource setting. Table 3.9 shows results for `IL`, `PGNet`, `s2s`, `joint`, and `semiCRF` for the high-resource experiment. The `s2s` model gets the best results in this setting with 78.02%, 77.06%, and 94.30% accuracy for English, German, and Indonesian, respectively. However, it only obtains slightly higher accuracy than `PGNet` and the differences in F_1 scores are similarly small. Overall, the pointer-generator network achieves results that are comparable with state-of-the-art in the high-resource setting. In contrast to the good performance for low-resource settings, `IL` under-performs on all metrics compared to `s2s` and `PGNet`. The `joint` model is the best non-neural system and performs clearly worse than both neural systems. Compared to `PGNet`, its accuracy is 5.54% lower for English, 15.80% lower for German, and 2.90% lower for Indonesian. `semiCRF` performs even worse.

3.5.4. Error Analysis

To obtain a better understanding of the results obtained using our neural models, we perform an error analysis on the output for the development sets of all folds. By manual inspection, we identify five not mutually exclusive types of errors: **Oversegmentation (Overseg.)** arises when the number of morpheme boundaries in the prediction is higher than that in the gold standard annotation. **Undersegmentation (Underseg.)** occurs when the number of morpheme boundaries is lower than that in the gold standard. **Restoration error (Res.)** occurs when the prediction does not match the gold annotation and the predicted word without boundaries does not match the input. These are errors that occur in words that undergo orthographic changes during word formation. **Overrestoration (Overres.)** refers to outputs with errors where the correct output needs only segmentation and a copy of the input to the output. **Wrong segmentation (Wrong seg.)** arises when the morpheme boundaries in the prediction are not the same as in gold. From each segmented word, we extract the indices within the word where the segmentation is performed. If the segmentation indices from the gold standard and the prediction are not equal, it counts as this error.

Table 3.12 shows the percentage of errors in all languages for both experimental settings (100 examples in the low-resource setting). For the high-resource experiments, the results for oversegmentation and undersegmentation errors are mixed: for English, `s2s` avoids generating too many segmentation boundaries, but this also has the drawback of not segmenting sufficiently when it is needed. The opposite happens for German, where `IL` performs better as well, with respect to oversegmentation but fails regarding undersegmentation. `PGNet` shows no strong wins or problems regarding these errors, except

	Oversegmentation
Input	internationalisierung
Gold	internationale isier ung
Error	internationale *is i er* ung
Description	The morpheme `isier` is segmented wrongly into three morphemes.
	Undersegmentation
Input	internationalisierung
Gold	internationale isier ung
Error	internationale *isieung*
Description	The morphemes `isier` and `ung` are lacking of a segmentation boundary.
	Restoration Error
Input	internationalisierung
Gold	internationale isier ung
Error	*international* isier ung
Description	The system did not perform the needed restoration for the stem `internationale`.
	Overrestoration
Input	internationalisierung
Gold	internationale isier ung
Error	internationaler *isierer* ung
Description	The systems perfomed a restoration on a morpheme that is not supposed to be restored.
	Wrong segmentation
Input	internationalisierung
Gold	internationale isier ung
Error	internationale *isi erung*
Description	The segmentation was done with the exact number of morphemes as in gold, however, the segmentation points are wrongly placed. In this error count all instances that do not match the exact segmentation boundaries.

Table 3.11.: Examples of error types. Wrong parts are marked in italics.

for English, where it performs better for undersegmentation. `s2s` performs better for restoration errors except for English, where again `PGNet` improves. Concerning oversegmentation errors, `IL` wins on all languages compared to the other neural systems. As Indonesian has a relatively regular morphology, all error types are much less frequent for this language. If we only consider the exact segmentation point prediction, `s2s` performs better for all languages. However, the differences between the observed error rates are relatively small between `s2s` and `PGNet` models. Overall, wrong segmentation errors are

the most common error type for all languages in the high-resource setting.

		Overseg.			Underseg.			Res.		
		IL	PGNet	s2s	IL	PGNet	s2s	IL	PGNet	s2s
High	eng	5.54	05.60	**05.28**	08.13	**06.58**	07.37	08.04	**05.86**	06.28
	deu	**4.17**	04.42	04.88	09.30	08.11	**07.02**	09.24	07.42	**06.94**
	ind	2.26	02.52	**01.91**	01.76	01.67	**01.50**	00.46	00.52	**00.45**
Low	eng	**5.84**	07.52	11.97	26.06	**18.82**	21.75	18.94	10.39	**04.96**
	deu	**1.40**	04.11	07.79	17.56	**14.83**	15.70	32.01	16.26	**07.81**
	ind	**10.94**	11.03	15.47	15.24	**10.61**	14.00	4.91	03.19	**01.46**
	tpp	**15.86**	27.75	34.45	42.43	**07.56**	23.42	32.16	**07.58**	14.10
	poq	**15.86**	21.88	26.10	28.43	**10.22**	25.18	22.86	**10.29**	17.68

		Overres.			Wrong seg.		
		IL	PGNet	s2s	IL	PGNet	s2s
High	eng	**02.34**	05.30	04.00	21.67	19.68	**17.01**
	deu	**06.08**	07.55	06.40	25.46	23.65	**20.49**
	ind	**00.58**	01.22	00.79	05.16	05.29	**03.26**
Low	eng	**02.56**	20.48	49.43	**46.92**	48.35	70.19
	deu	**03.94**	21.88	33.93	**41.66**	51.52	71.78
	ind	**02.96**	19.90	50.00	34.64	**34.25**	36.16
	tpp	**03.80**	25.04	44.20	**69.52**	73.39	86.34
	poq	**07.86**	22.17	49.42	**55.71**	57.54	76.81

Table 3.12.: Error types found in the development set. The high resource configuration includes three languages, while the low-resourced setting refers to the model performance using 100 training examples. This error analysis was performed for all five languages.

In the low-resources experiments, IL excels over all other models for oversegmentation and overrestoration, and for all languages except Indonesian for wrong segmentation errors. This low error rate explains the important gains that this model shows for low-resource languages. PGNet shows, however, better performance avoiding undersegmentation errors in all languages. It also performs better for Popoluca and Tepehua for restoration errors, while s2s has the lowest restoration errors for English, German, and Indonesian.

Finally, we also perform an error analysis of joint (cf. supplementary material). In our low-resource simulation experiments, we notice a surprisingly good performance of joint for German. The data for this language is special since all words contained in

the set are segmentable. We find that `joint` has no undersegmentation errors at all. Also, it makes very few copy errors (9.5%, compared to 21.7% of `PGNet`). For our new datasets, this model obtains a high rate of wrong segmentation (88.87% for Popoluca and 91.57% for Tepehua). It further seems to not easily be able to decide which words should or should not be segmented. This is shown by the high undersegmentation rate (50.14% for Popoluca and 62.43% for Tepehua). Thus, the low performance of `joint` on those languages can be explained by this error type and the high morphemes-per-word rate of those languages as shown in Table 3.8.

3.6. Findings

In this chapter, we explored research **question 1**. For this purpose, we focused our efforts in two morphological segmentation tasks: canonical segmentation and surface segmentation. We first curated two new publicly available datasets for both tasks for Rarámuri, Tepehua, and Popoluca. With these datasets, we aim to draw more attention from the NLP community to the indigenous languages of the Americas. Afterward, we experimented with a set of neural morphological methods. The findings are as follows:

1. Neural sequence-to-sequence models achieve state-of-the-art results for the MCS and MSS for supervised and semi-supervised training, even in extreme low-resource scenarios.

2. We show that the simulated morphological canonical segmentation is not the best approach when studying endangered and low-resource (for the task) languages. This was shown in the performance gap between Popoluca and Tepehua compared to the high-resource setting. The best experimental scenario is to use datasets with truly low-resourced languages.

3. It is useful to use random strings and non-labeled data to improve the neural segmentation methods. However, the real limitation of these models relies on the general scarcity of raw text. Most of the analyzed languages do not have Wikipedia entries or a strong written tradition. In the case of using non-labeled data, we saw that pre-training and fine-tuning is the best option to achieve good results.

4. We also show that frameworks that model string copying (ptrseg), as well as neural transducers (IL) are more effective in extremely low resource settings and are still

capable of achieving good or comparable research with vanilla seq-to-seq models when more resources are available.

5. Even with our promising results, we also recognize that the performance of all systems, when applied to polysynthetic languages in low and extreme low scenarios, from good and still has room for improvement.

Overall, we found that neural models perform well for morphological segmentation tasks, but more than these promising results are needed to have production-ready systems. Also, the central problem of the lack of annotated and non-annotated data is not easy to solve. We strongly believe that this problem can only be solved with a strong community that gathers NLP researchers, linguists, and/or native speakers.

4. Machine Translation

In this chapter, we study machine translation in extremely low scenarios and explore if it is possible to use morphological knowledge to improve the final results. First, we introduce the basic machine translation concepts and key methods for low-resource scenarios. Then, we apply this knowledge to a set of four indigenous languages.

4.1. Challenges of MT for Indigenous Languages

In an overview of the datasets and recent studies of MT for the indigenous languages of the Americas, we found the following main issues to be handled (see Appendix A.1).

Extreme low-resource parallel datasets Even with the recent advances, the resources available to train MT systems are extremely scarce, with training sets between 4k and 20k sentences (see §4.2), with notable exceptions to Inuktitut, Guarani, and Quechua (Joanis et al., 2020; Ortega et al., 2020).

Lack of monolingual data Most of these languages are mostly used in spoken form. In recent years, with the advancement and democratization of mobile technologies, indigenous languages have seen a slight increase in massaging systems and private spheres Rosales et al. (2019). However, the usage of these languages on the internet is rather limited. Even Wikipedia has a limited number of these languages (Mager, Gutierrez-Vasques, et al., 2018b).

Low domain diversity . As most parallel datasets are scarce, they are restricted to a few domains, making it challenging to adapt or aim for general translation models. This has been recognized as a major problem during the AmericasNLP ST (Mager et al., 2021).

Rich morphology An important number of these languages are morphologically rich. We often find polysynthetic, with or highly agglutinative languages (Kann et al., 2018) or even fusional phenomenon (Mager et al., 2020).

Distant paired language The most common languages that we find that indigenous languages of the Americas is translated into are Spanish, English, and Portuguese. However, these languages are distantly related to the indigenous languages of the Americas, and have entirely different linguistic phenomenons (Campbell, 2000; Romero et al., 2016).

Noisy text environments Monolingual texts, if they exist, are found in social media that often use a non-canonical witting (Rosales et al., 2019).

Code-Switching This phenomenon is strongly present in indigenous languages of the Americas, as all these languages are minority languages in their own countries. As a result, the bilingualism among their communities is strong (and CS is a common phenomenon in this setup (Çetinoğlu, 2017)). The final result of this phenomenon is the inclusion of code-switching on a common base (Mager, Çetinoğlu, & Kann, 2019) in their language.

Lack of orthographic normalization The usage of indigenous languages of the Americas faces the problem of having a unified orthographic standard. This is only sometimes possible, as the suggestions of linguists and official entities sometimes match the day-to-day writing of the speakers. Moreover, in some cases, special symbols in the orthographic standards are not accessible in English or Spanish keyboard and need to be replaced with other symbols. The winner of the AmericasNLP ST made significant improvements using orthographic normalizers explicitly developed for each American language (Vázquez et al., 2021).

Dialectal variety The indigenous languages have a strong dialectal variety, making it hard for native speakers to understand even speakers from neighboring villages. The linguistic richness of entire regions is so diverse that even a single state like the Mexican Oaxaca could correspond to the diversity in the whole of Europe (McQuown, 1955).

Dataset	**Paired-languages**	Authors
AmericasNLI	Aymara, Asháninka, Bribri, Guaraní, Nahuatl, Otomí, Quechua, Rarámuri, Shipibo-Konibo, Wixarika	A. Ebrahimi et al. (2022)
CPML	Ch'ol, Maya, Mazatec, Mixtec, Nahuatl and Otomi	Sierra Martínez et al. (2020)
OPUS	*	Tiedemann (2016)
New testament Bible	*	McCarthy et al. (2020b)

Table 4.1.: Parallel dataset collections that contain one or more indigenous languages of the Americas

4.2. Available MT datasets for the ILA

The parallel datasets available for MT have increased during the last few years. At this moment, we can show in two folds the development of these resources: as shown in table 4.2 work on specific language has emerged, but also broader datasets have started to cover the indigenous languages of the Americas (see table 4.1).

Language-specific corpus collection work has been performed for many languages, with a parallel corpus as the main component. In recent times, we have seen Cherokee–English (OPUS) (S. Zhang et al., 2020), Wixarika–Spanish (Mager, Carrillo, & Meza, 2018b), Shipio–Konibo (Feldman & Coto-Solano, 2020), and others (see table 4.2). The Inuktitut– English parallel data is the most prominent of these datasets. The last version of this dataset corpora (Joanis et al., 2020) is medium size with 1,450,094 sentences. Previous versions of this corpus are (Martin et al., 2003). This dataset was used for the WMT 2020 Shared Task on Unsupervised and Low Resourced MT (Barrault et al., 2020).

For wide-spoken languages like Guarani, it is possible to collect a web-crawled dataset, including news articles and social media parallel aligned data (Chiruzzo et al., 2020; Góngora et al., 2021). This dataset also includes the monolingual data. This is possible as Guaraní is one of the most spoken indigenous languages of the continent.

In contrast to the language-specific datasets, we find broader approaches (see table 4.1). The broadest multilingual dataset, which contains the Bible's New Testament, includes about 1600 languages (Mayer & Cysouw, 2014; McCarthy et al., 2020b) of the 2,508 that have been collected by the Summer Institute of Linguistics (SIL) Anderson & Anderson (2012). Another remarkable effort to obtain broad language coverage is the PanLex project (Kamholz et al., 2014), which has gathered lexical translation

Language	Paried-language	ISO	Family	Sentences	Domain
Asháninka	Spanish	cni	Arawak	3883	
Bribri	Spanish	bzd	Chibchan	5923	
Guarani	Spanish	gn	Tupi-Guarani		News, Blogs
Guarani	Spanish	gn	Tupi-Guarani	14,531	News, Blogs
Guarani	Spanish	gn	Tupi-Guarani	14,792	News, Social Media
Nahuatl	Spanish	nah	Uto-Aztecan	16145	Diverse Books
Otomí	Spanish	oto	Oto-Manguean	4889	Diverse Books
Rarámuri	Spanish	tar	Uto-Aztecan	14721	Dictionary Examples
Shipibo-Konibo	Spanish	shp	Panoan	14592	Educational, Religious
Wixarika	Spanish	hch	Uto-Aztecan	8966	Literature
Cherokee	English	chr	Uto-Aztecan		OPUS
Inuktitut	English	iku	Eskimo–Aleut	1,450,094	Legislative

Table 4.2.: Parallel datasets that have been released focusing on one indigenous language

dictionaries for over 5,700 languages. However, for most languages, PanLex contains only a few dozen words. Duan et al. (2020) show that such dictionaries can be used to create an NMT system, making bilingual dictionaries relevant for further studies. Recently, community-driven research groups have started the creation of their own parallel datasets, such as Masakhane (Orife et al., 2020; Nekoto et al., 2020) for African languages, and AmericasNLP for indigenous languages of the Americas (A. Ebrahimi et al., 2022; Mager et al., 2021). The AmericasNLI dataset is an important effort to have a common evaluation benchmark for the 10 indigenous languages of the Americas for the MT and NLI tasks.

Given the constitutional rights of indigenous languages in many countries of the Americas, it is possible to access this data. Vázquez et al. (2021) made available this resource during their shared task system development.

Finally, it is important to mention that many of the languages spoken in the Americas have Wikipedia's set of articles available[1].

Collection of New Data A common way to create parallel data with the help of bilingual speakers is via elicitation (translating the foreign text into another language). However, it has the disadvantage of biasing the created text to forms and topics, culture, and even grammatical forms toward the source language (Lörscher, 2005). A method that avoids this problem is language documentation, which consists of storing and an-

[1]The available languages in Wikipedia can be consulted at: https://es.wikipedia.org/wiki/Portal:Lenguas_indígenas_de_América. Until the publication of this article, there were only entries in Nahuatl, Navajo, Guarani, Aymara, Klaalisut, Esquimal, Inukitut, Cherokee, and Cree.

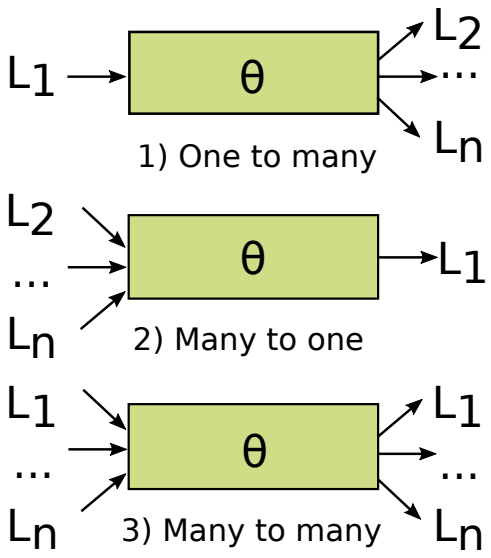

Figure 4.1.: An overview of different multilingual setups.

notating commonly used speech or text (Himmelmann, 2008). However, it is costly and requires specialists. In this process, involving the community members who are bilingual speakers is important (Bird, 2020).

4.3. Low-resource MT paradigms

Most languages of the Americas do not have a large amount of data for MT. Therefore, we introduce the essential paradigms to improve low-resource machine translation. Figure 4.2 shows a general overview of the methods and options to improve LRL MT. For a more detailed understanding of these techniques, we refer the reader to specialized low-resource MT surveys (Haddow et al., 2022; R. Wang et al., 2021; Ranathunga et al., 2021).

4.3.1. Multilingual Supervised Training

With a multilingual set of parallel data $D_{parallel}$ between different language pairs $\{(L_1, L_2), \ldots, (L_m, L_n)\}$, we can train a model that can map a sentence from any source language L_x into any target language L_y contained in $D_{parallel}$. These multilingual NMT models have grown in popularity and efficiency in recent years. We now cover the different training algorithms for these models: These models can be trained in three setups (cf. Figure 4.1) many source languages and one target language (*many-to-one*), 2) one source and many target languages (*one-to-many*), and 3) many source languages and many target

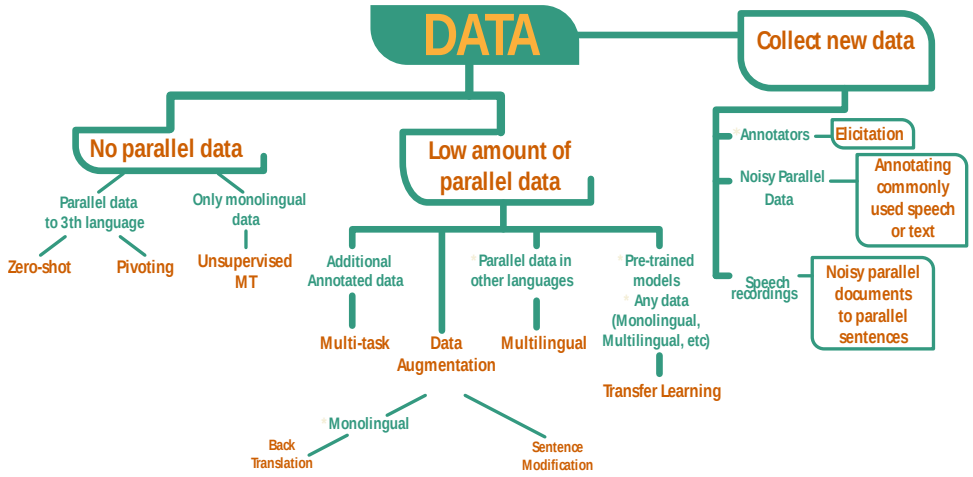

Figure 4.2.: What to do when we have little or no data to train our machine translation models? This diagram shows the basic scenarios, solutions, and common requirements for each method, with the section describing the method.

languages (*many-to-many*). For a general overview of multilingual MT, we refer the reader to surveys dedicated to this topic (Tan et al., 2019; Dabre et al., 2019). Johnson et al. (2017) is the first to introduce a multilingual NMT model, trained on translating from many languages to English and in the opposite direction. The authors show that these models improve over single-language pair models for LRLs.

4.3.2. Multi-task Training

Multi-task training (Caruana, 1997) aims to improve the performance of the main task – MT in our case – by adding one or more additional tasks to the training. The main question is how to best leverage the auxiliary task. The easiest way is to share all network parameters using the ideas already explored in multilingual NMT (§4.3.1). This can be done with a special flag in the input that specifies the current task. It is also possible to share only the encoder and have two separate decoders for each task.

Multilingual Modeling To handle multilingualism, it is also possible to adapt and modify the NMT models. The main proposals to do so has been: sharing all parameters except the attention mechanism of an RNN NMT model (Blackwood et al., 2018); parameter sharing in the transformer architecture (Sachan & Neubig, 2018).

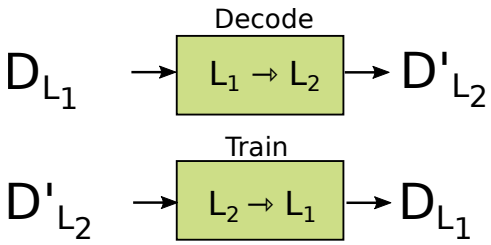

Figure 4.3.: Backtranslation

4.3.3. Data Augmentation

The methods in this subsection require small amounts of parallel data between two languages, along with monolingual data. The main ideas are 1) back-translation, 2) sentence modification, and 3) pivoting.

Back-Translation A straightforward way to leverage monolingual data for low-resource MT is to generate a meaningful signal with the help of an already initialized MT model. This method is called back-translation (BT; Sennrich et al., 2016a) and consists of translating monolingual data and using the resulting pseudo-parallel data as supervision to further improve the translation in the opposite direction by generating a meaningful signal. In Figure 4.3, we can see that With monolingual data M^{L_x} in source language L_x and a trained model that can translate from L_x into the target language L_y, we can generate a translation M'^{L_y}. This pseudo-parallel data (M^{L_x}, M'^{L_y}) is then used to train a new model in the opposite direction. This process can be applied iteratively to improve the translation (Hoang et al., 2018).

Sentence Modification Other methods to generate more parallel sentences are based on lexical substitution. (Fadaee et al., 2017) explores replacing frequent words with low-frequency ones in both source and target to improve the translation of rare words. This is done using language models (LMs) and automatic alignment.

Pivoting If no parallel corpus between languages L_x and L_y is available, but both have parallel corpora with a third language L_p, pivoting is an option. The basic idea is to train two MT systems: one that translates $L_x \rightarrow L_p$ and another for $L_p \rightarrow L_y$. Pivoting was first introduced for SMT (H. Wu & Wang, 2007; Cohn & Lapata, 2007; Utiyama & Isahara, 2007). This allows not only to perform a translation between these two sources and target languages but also to improve language pairs with a low-resource setting.

4.3.4. Semi-supervised and Unsupervised MT

Significant progress has been made by training NMT models in a semi-supervised fashion (Gibadullin et al., 2019) by taking advantage of monolingual text. In addition, the monolingual data can also be leveraged for UMT. In this section, we will explore these two strategies.

Transfer Learning via Pretraining Transfer learning refers to using knowledge learned from one task to improve performance on a related task Weiss et al. (2016). Recently, thisapproach has gained popularity with large multilingual models such as Conneau & Lample (2019) that propose training the encoder and the decoder separately to obtain cross-language representations (XLM). This idea has been further extended by K. Song, Tan, et al. (2019, MASS) to masking a *sequence* of tokens from the input (multilingual MASS (Siddhant et al., 2020)). Another approach is to train the entire transformer model as a denoising autoencoder (BART; M. Lewis et al., 2019) (BART (mBART; Y. Liu et al., 2020)). It is also possible to pre-train a transformer in a multi-task, text-to-text fashion, where one of the tasks is MT (T5; Raffel et al., 2020) (multilingual version (Xue et al., 2021)).

Unsupervised MT UMT covers approaches that do *not* require any parallel text, relying only on monolingual data. This differs from zero-shot translation, which uses parallel data for other language pairs. In recent years, unsupervised MT (UMT) has gained relevance with increasingly good results. Early approaches tackled the problem with an auto-encoder with adversarial training (Lample et al., 2017) or with auto-encoders with a shared encoding space and separate decoders for each target language (Artetxe et al., 2018). The main problem with these approaches is the need for a large monolingual dataset that is unavailable for most indigenous languages of the Americas. The idea is to use a bilingual word embedding space (Artetxe et al., 2017; Zou et al., 2013) to generate a word-to-word translation and then improve the output with an LM. The iterative steps of BT follow this. The latter was also adapted to SMT (Artetxe et al., 2018) with important gains on top of NMT. After this, more methods have been proposed: a combination of SMT and NMT cold starts (Marie et al., 2019) or SMT for initialization, followed by NMT with multiple iterations of BT (Artetxe et al., 2019). Finally, an SMT can be added to those steps (Ren et al., 2019) for regularization.

4.4. Advances in MT for the indigenous languages of the Americas

In recent years, the interest in MT for indigenous languages of the Americas has increased. The task takes work. The first usage of NMT systems has not been successful (Mager & Meza, 2021b). However, we have witnessed great improvements with the use of the LRL MT methods.

The Cherokee–English (S. Zhang et al., 2020) language pair has been explored using a pre-trained BERT (Devlin et al., 2019) for the English side. A system demonstration of this approach is also accessible (S. Zhang et al., 2021). The back-translation strategy for Bribri–Spanish NMT transformers has been explored (Feldman & Coto-Solano, 2020) and by Oncevay (2021) (for four Peruvian languages to Spanish) with good results. The scarce indigenous language monolingual text can be replaced with some extent with Spanish text or extracted from PDFs and other sources (Bustamante et al., 2020).

One of the main challenges for the complex morphological languages in the area has been the prepossessing step. Schwartz et al. (2020) show that even if morphological segmentation has less perplexity a the language modeling time, it is still underperforming or equivalent against BPEs for MT (for Inuktitut-–English, Yupik—English Data, Guaraní—Spanish Data). A more comprehensive (on the segmentation modeling side) was done by Mager et al. (2022) exploring a wide array of segmentation models for Wixarika, Nahuatl, Shipibo-Konibio, and Rarámuri, to Spanish (see §4.5). The latter study showed that supervised morphological segmentation under-performs unsupervised. However, unsupervised morphological segmentation like LMVR (Ataman et al., 2017) and FlatCat (Grönroos et al., 2014) perform better than BPEs. (Ngoc Le & Sadat, 2020) studied how better to perform word segmentation for the Inuktitut–English pair. We found that for this language pair, a morphological segmentation, or a combination of BPEs and morphological segmentation, works better than just applying vanilla BPEs. In addition, training word embeddings for Guarani–Spanish translation is an excellent opportunity to increase the MT performance of these languages (Góngora et al., 2022).

The use of transfer learning from multilingual systems has been attempted, with limited results (Nagoudi et al., 2021) (training an own T5 model for indigenous languages) and (F. Zheng et al., 2021). However, pertaining a Spanish–English model together with indigenous languages of the Americas, and then fine-tuning it (together with a careful prepossessing and filtering step) has been the most successful strategy (Vázquez et al., 2021).

The quality of MT systems for indigenous languages of the Americas has been a constant debate. However, in A. Ebrahimi et al. (2022) we show that the quality of MT for these languages is sufficient to improve other tasks like natural language inference (NLI).

Inuktitut–English ST The WMT 2020 news translation task included Inuktitut–English translation Barrault et al. (2020). In addition, the participating systems explored the difficulties of working with a polysynthetic language in a medium resource scenario. The participating teams in this competition were: Kocmi (2020); Hernandez & Nguyen (2020); Scherrer et al. (2020); Roest et al. (2020); Lo (2020); Knowles et al. (2020); Y. Zhang et al. (2020); Krubiński et al. (2020).

AmericasNLP 2021 ST In 2021, the AmericasNLP[2] community organized a workshop on Machine Translation for 10 indigenous languages of the Americas (Mager et al., 2021). The AmericasNLP shared task winner was (Vázquez et al., 2021). Other participants in this shared task are (Nagoudi et al., 2021; Bollmann et al., 2021; F. Zheng et al., 2021; Knowles et al., 2021; Parida et al., 2021; Nagoudi et al., 2021). It is important to point out the importance of clean data. For Quechua, Moreno (2021) got the best results, generating an additional amount of clean data.

4.5. Importance of Morphological segmentation

Polysynthetic languages are known because of their rich morphology, which encodes most parts of the semantics into verbs, leading to a high morpheme-per-word rate. The resulting combinations of morphemes and roots (as polysynthesis allows multiple roots (M. C. Baker, 1996)) result in extreme type sparsity. Thus, polysynthetic languages represent a challenging environment for NLP methods (Klavans, 2018). To handle this issue, subword segmentation has been a common method to reduce sparsity (Vania & Lopez, 2017a). Moreover, as these languages are mostly extreme low-resource language (eLRL), the challenge is even harder. Some reasons behind this is that most of them are endangered and spoken by minority groups (Mager, Gutierrez-Vasques, et al., 2018b; Littell et al., 2018).

[2] In 2022, a new edition of the AmericasNLP competition is running and features three tasks: Speech-to-text translation, Automatic Speech Recognition (ASR); and automatic speech synthetics. The results are not discussed in this thesis, as the competition has not finished at this time at this moment.

However, what impact does morphological segmentation have on downstream tasks like machine translation (MT), when translating from or into fusional languages? Linguistically inspired segmentation was considered to be the best option to handle rich morphology (Koehn et al., 2005; Virpioja et al., 2007) until the appearance of Byte-Pair Encodings (BPEs; Sennrich et al., 2016c) and has been adopted as the default segmentation technique. BPEs earned this status for good results, unsupervised training, and language independence. Saleva & Lignos (2021) show no significant gain when using an unsupervised morphological segmentation for the input over BPEs when evaluating those methods in moderate LR scenarios for Nepali–English and Kazakh–English, contradicting the initial findings of Ataman & Federico (2018b). However, how would BPEs perform for polysynthetic languages in ELR scenarios? Schwartz et al. (2020) compare BPE, with Morfessor (Smit, Virpioja, Grönroos, & Kurimo, 2014) and Rule-Based morphological analyzers for medium resourced Inuktitut–English and, for the ELR Yupik–English and Guarani–Spanish. Their results show that BPEs outperform Morfessor and the morphological analyzer in all MT cases (but with better Language Modeling capabilities of morphological models over BPEs). These results are later confirmed by C. Liu et al. (2020), for Yupik–English. However, most of these studies only rely on a limited set of segmentation methods and do not consider the quality of the used morphological segmentation methods.

Ortega et al. (2020) use a morphological guided BPE version for the agglutinative Quechua language, hinting that morphological segmentation might still be worth exploring.

As most polysynthetic languages are endangered and have minimal parallel and monolingual data, we want to investigate this real-world case scenario.

This section aims to answer the following specific research questions: i) is morphological segmentation beneficial for MT where one language is polysynthetic and ELR?; and ii) is higher morphological segmentation quality correlated with higher MT scores?

To answer these questions, we perform segmentation experiments on four polysynthetic languages:[3] Nahuatl (`nah`), Raramuri (`tar`), Shipibo-Konibo (`shp`) and Wixarika (`hch`) and apply those segmentations to the MT paired with Spanish (`spa`). First, we revisit a comprehensive set of supervised and unsupervised methods and apply them to the input of MT transformer models. This study is the first to show that unsupervised solid morphological approaches outperform BPEs consistently on ELR polysynthetic languages,

[3] We choose the languages for this study based on the availability of a morphological segmentation dataset.

except for nah. These results are related to Ortega et al. (2020), which found that a morphologically guided BPE can improve the MT performance for Guarani–Spanish. On the other hand, even when supervised morphological segmentation methods achieve better results for the segmentation task regarding MT systems, they underperform all other approaches. We hypothesize that this might be due to over-fitting the clean and out-of-domain morphological training set. To make all these experiments possible, we also introduce two new morphologically annotated datasets for tar and shp; and one parallel dataset for spa–tar[4].

4.5.1. Description of the Raramuri–Spanish Parallel Dataset

For the dataset, we manually extracted phrases that had a translation into Spanish from the Brambila (1976) dictionary. In addition, given that the orthography in this book is out of use, we normalized it to a modern version used in Caballero (2008). The book does not specify the dialect of the sentences. However, from the input of native speakers, we identified the variation as the central dialect. We also identified the examples as part of the conversational domain, with topics mostly related to rural life. Table 4.3 shows the characteristics of the dataset and the dataset splits. The dataset contains a total of 14,719 phrases, with considerably fewer tokens per sentence than in Spanish (i.e., in training, a 1.7 N_{es}/N_{tar} ratio). It is also important to note that the Out Of Vocabulary (OOV) issue is higher for Raramuri than, for Spanish (see OOV in table 4.3). We attribute this phenomenon to the polysynthetic typology of the Raramuri language.

4.5.2. Experimental Setup

Now that we have introduced the main ideas of this chapter, we begin explaining our experimental setup.

Resources

For the machine translation experiment, we use the following parallel datasets: the hch–spa translation of the fairy tales of Hans Christian Andersen (Mager et al., 2017); the Shipibo-Konibo–Spanish translations from a bilingual dictionary and educational material Galarreta et al. (2017); and for nah–spa, the Axolotl dataset (Gutierrez-Vasques et al., 2016). This dataset contains several variants of Nahuatl. In addition, we also use

[4]The datasets are available under http://turing.iimas.unam.mx/wix/mexseg

	train		dev		test	
	tar	es	tar	es	tar	es
S	13,102		587		1,030	
N_{es}/N_{tar}	1.692		1.794		1.689	
N	73,022	93,410	3,183	4,133	5,847	7,547
V	19,044	16,220	1,713	1,771	2,793	2,803
V1	12,894	10,021	1,402	1,365	2,221	2,120
V/N	0.261	0.174	0.538	0.429	0.478	0.371
V1/N	0.177	0.107	0.440	0.330	0.380	0.281
OOV			573	434	1,037	779
%OOV			0.334	0.245	0.371	0.277

Table 4.3.: Parallel corpus' description: S = number of sentences; N_{es}/N_{tar} = ratio of tokens between Spanish and Rarámuri; N = number of tokens; V = vocabulary size; V1 = number of tokens occurring once (hapax); V/N = vocabulary growth rate; V1/N = hapax growth rate; OOV = out-of-vocabulary words w.r.t. train set.

our collected `tar–spa` Parallel corpora (§4.2). The details of the data splitting are described in Table 4.4 in the appendix. For morphological segmentation, we use the `nah`, and `hch` annotated datasets from Kann et al. (2018), and additionally, we use the `shp` and `tar` datasets introduced in section 3.3.1. We use the same splits as reported by the original sources.

	Train	Dev.	Test
`hch–spa`	7442	447	1075
`nah–spa`	14208	644	1291
`tar–spa`	12987	582	1021
`shp–spa`	13102	587	1030

Table 4.4.: Data splitting (in number of phrases) used for our Machine Translation experiments, from and to Spanish.

Metrics, baselines

As previously discussed, we use the chrF Popović (2015) metric, as it is a better fit for polysynthetic languages and, for languages that are morphologically rich. We train all models using a transformer model, with the hyperparameters of the FLORES dataset Guzmán et al. (2019), that is optimized for low-resources settings.

4.5.3. Machine translation results

Figure 4.4 shows the chrF score difference against the BPEs baseline in all directions, but with a subset of systems. Table 4.5 shows the complete translation results using chrF[5]. We first observe that the supervised segmentation approaches underperform in contrast with the unsupervised ones in the settings in all language pairs.

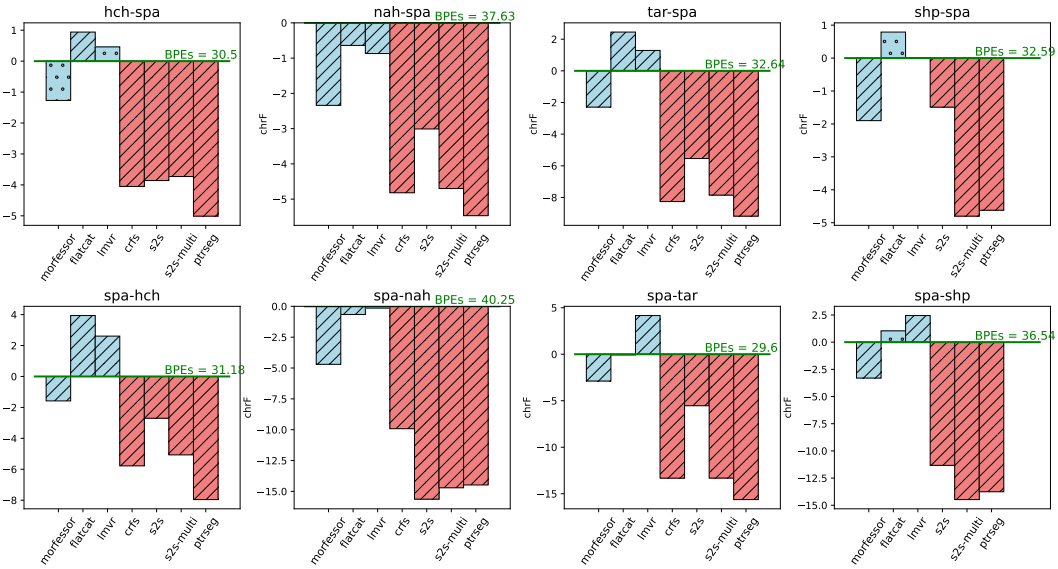

Figure 4.4.: chrF score difference for all morphological segmentation compared to BPEs on the test sets for both translation directions. We run a paired approximation test with 10000 trials using the BPEs system output as the baseline. Diagonals indicate a p-value ≤ 0.05, while stars indicate a p-value > 0.05. The blue systems are unsupervised, while the red ones are supervised.

With the **polysynthetic languages in the source side**, we see that the unsupervised morphological segmentation methods outperform all other systems. In concrete, FC has a significantly (with $p < 0.05$) higher score for hch-spa (31.44 chrF) and tar-spa (35.09 chrF), and a statistical tie for shp-spa with FC(33.38 chrF) and BPEs(32.59 chrf). On the other hand, we can also see that LMVR and BPEsscore difference is not statistically significant for hch-spa (30.96 chrF) and shp-spa.

Despite the good results of s2s, s2s+multi or PtrSeg in morphological segmentation (cf. 3.4.2), for MT they have the worst performance compared to other paradigms with an average of 3.47 chrF lower than BPEswhen comparing to s2s. The same is also true for our non-neuronal supervised method CRFs, which under-performs in average BPEsby

[5]SacreBLEU: chrF2 + numchars.6 + space.false + v.1.5.0

5.46 chrF. We argue that this method creates new subwords in their output, which can aid in morphological segmentation, but MT only adds noise to the input for the model.

system	hch-spa	nah-spa	tar-spa	shp-spa
bpe	30.50	**37.63**	32.64	**32.59**
morfessor	29.23	35.29	30.35	30.69
flatcat	♦31.44	36.99	♦35.09	**33.38**
lmvr	30.96	36.76	33.93	32.60
crfs	26.45	32.81	24.38	-
seq2seq	26.64	34.62	27.10	31.10
seq2seq-rand-mt	26.77	32.93	24.79	27.79
seq2seq-raw-mt	25.04	33.00	26.25	27.61
seq2seq-finetuned	28.71	31.25	30.24	29.88
pointernet	25.49	32.16	23.46	27.97
pointernet.finetune	29.06	35.15	29.75	30.34
system	spa-hch	spa-nah	spa-tar	spa-shp
bpe	31.18	**40.25**	29.60	36.54
morfessor	29.60	35.55	26.72	33.24
flatcat	♦35.12	39.59	29.52	37.58
lmvr	33.79	**40.11**	♦33.76	**38.99**
crfs	25.40	30.33	16.28	-
seq2seq	28.48	24.62	24.06	25.21
seq2seq-rand-mt	26.11	25.54	16.29	22.06
seq2seq-raw-mt	30.58	33.71	29.36	36.88
seq2seq-finetuned	30.43	36.15	31.86	**38.87**
pointernet	23.22	25.77	13.97	22.78
pointernet.finetune	32.16	35.59	29.10	36.01

Table 4.5.: Translation results on the test for both directions. Maximum scores are in bold. We ran a paired approximation test with 10000 trials using the `BPEs` system output of the best systems and compared them to the second-best system. The ♦ symbol indicates a p-value < 0.05.

Our newly introduced semi-supervised methods show strong improvements when using the pre-training algorithm along with `PtrSeg`. The gap between `PtrSeg` +finetune and `BPEs`is reduced to an average of 2.26 chrF. Moreover, finetuning strongly outperforms the vanilla `PtrSeg` version for all language pairs; on average, the difference between both versions is 3.8 chrF.

In the inverse direction, with the **polysynthetic languages on the target side**, `LMVR` is the method that significantly surpasses the baseline for more language pairs:

spa-tar (33.76 chrF) and spa-shp (38.99 chrF); whereas FC obtains the maximum score in spa-hch (35.12 chrF). For spa-nah LMVR (40.11 chrF) and BPEs(40.25 chrF), the two top-performing systems had no significant score difference. Interestingly, for spa-shp, we found that the unsupervised LMVR (38.99 chrF) is tied with the semi-supervised s2s +finetuning system (38.87 chrF).

We could see the same phenomenon of the poor performance of supervised systems. We see this tendency when comparing BPEs(34.39 c chrF) to the (in average) best-supervised method s2s(25.59 chrF). The same is true also for CRFs(24.0). However, when boosted with unlabeled data in a pre-train + finetune fashion, the s2s model achieves close results to BPEs, with an average of 34.32. This means that the semi-supervised version outperforms its supervised version by 8.73.

Overall, we notice that segmentation methods matter for polysynthetic languages in contrast to other languages (Saleva & Lignos, 2021; Gaser, 2022). Poorly suited methods can strongly decrease the performance of downstream tasks like MT. The question of which segmentation method is better for MT is still open. The results show that a specific method helps improve translation in different settings. However, with the current results, we see a tendency to have unsupervised methods outperforming supervised ones. Also, using non-labeled data helps improve in general, such as supervised methods. The impact of this last point is stronger if we translate it into the polysynthetic language.

4.5.4. Analysis

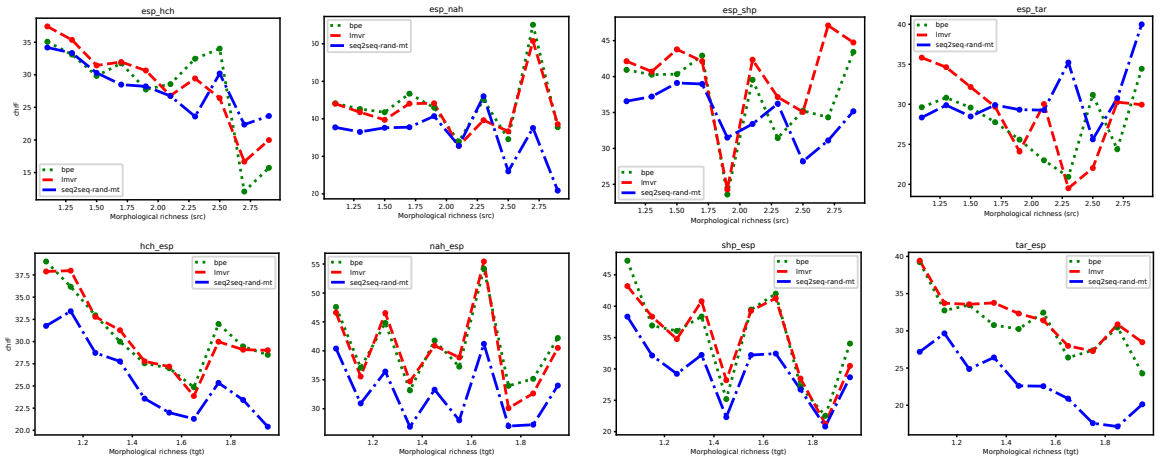

Figure 4.5.: Relation between morphological richness of each polysynthetic language with relation to its chrF score in each translation direction. The scores are analyzed for BPEs, LMVR , and s2s+multi.

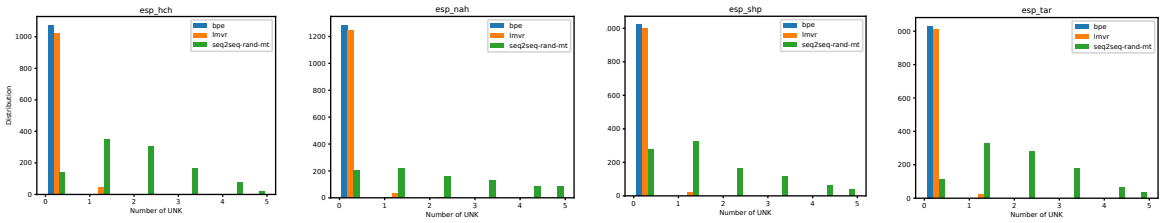

Figure 4.6.: Number of out-of-vocabulary tokens (UNK) found for each polysynthetic language classified by the system. The scores are analyzed for `BPEs`, `LMVR` , and `s2s+multi`.

To better understand the current results, we explore the outputs of different systems. For simplicity, we choose the best-performing segmentation system for each segmentation paradigm. For unsupervised morphological-inspired segmentation, we use `LMVR`, `s2s+multi` for supervised morphological segmentation, and `BPEs` for frequency-based segmentation.

First, we explore the impact of morphological richness on each system. We use `Morfessor` to infer the segmentation for each polysynthetic language data point and divide the number of found morphemes by the total number of tokens. Figure 4.5 shows that there is no clear correlation between morphological richness and the systems' performance for `nah` and `shp`. However, for `hch` we observe that a richer morphology implies a loss in the translation quality. The exact correlation can be seen for the `tar-esp` direction. This correlation is stronger when the polysynthetic language is in the source and weaker when it is in the target. Similar behavior can be observed between `LMVR` and `BPEs`.

Second, we explore the impact of the out-of-vocabulary (UNK) tokens that each segmentation model introduces because having many UNK tokens can negatively influence the MT results. In figure 4.6, we show the number of UNK tokens each segmentation has when used with the dictionary of an MT system. The supervised `s2s+multi` has the highest number of UNK symbols. We suggest that this phenomenon could be the strong generative power of such systems and the well-known artifacts that such models introduce (i.e., string repetitions). However, `LMVR` has a slightly higher number of UNK tokens, leaving `BPEs` the best vocabulary coverage. This can explain the surprisingly low performance of the supervised models.

4.6. Ethics of Machine Translation for Indigenous languages

Research on machine translation and natural language processing (NLP) more generally is moving toward low-resource setups and multilingual models. Thus, the NLP community needs to open the discussion of repercussions and best practices for research on indigenous languages (that in most cases are also low-resourced) since non-artificial languages cannot exist without a community of people that use (or have traditionally used) them to communicate.

Indigenous languages further differ from the more widely used ones in a crucial way: they are commonly spoken by small communities, and many communities use their language (besides other features) as a delimiter to define their own identity Palacios (2008); Enríquez (2019), and have in many cases also a certain degree of endangerment. Furthermore, in some cases, highly sensitive information – such as secret aspects of their religion – has been encoded with the help of their language Barron-Romero et al. (2016). This is why, in recent years, discussions on ethical approaches to studying endangered languages have been started Smith (2021); Z. Liu et al. (2022). When we consider the past (and present) of some of the communities that speak these languages, we will find a colonial history, where research is not the exception Bird (2020). Therefore, it is possible to trespass on ethical limits when using typical NLP and data collection methodologies Dwyer (2006).

In this section, we explore the basic concepts of ethics related to the MT of endangered languages with a special focus on Indigenous communities, surveying previous work on the topic. To better understand the expectations and concerns related to the development of MT systems for Indigenous communities, we conducted an interview study with 22 language activists, language teachers, and community leaders who are members of Indigenous communities from the Americas. Additionally, we also performed 1:1 dialogs with two study participants to deepen our understanding of the matter. The goal is to answer the following research questions: *How do community members want to be involved in the MT process, and why? Are there sensible topics that are not ethical to translate, model, or collect data without the community's explicit permission? How can we collect data in an ethical way?*

Surprisingly, most survey participants positively view MT for their languages. However, they believe that research on their languages should be done in close collaboration with community members. Open access to research discoveries and resources is also

highly valued, and the high quality of the resulting translations. The personal interviews also confirmed this. Thus, our most important finding is that it is crucial to work closely with the communities to understand delicate ethical topics when developing MT systems for endangered languages.

4.6.1. Ethics and Data

The study of endangered languages in indigenous communities has a long history, with the most prominent questions being focused mainly on the challenge of data collection (Smith, 2021).

One of the common forms of this is to use normative ethics (deontology). Examples of relevant guidelines include those from The Australian Institute of Aboriginal and Torres Strait Islander Studies;[6] the Ethical statement of the Linguistic Society of America;[7] and the DOBES code of conduct.[8] These lists are the results of broad discussions that have taken place over decades. In this debate also, indigenous voices inside academia raised (Smith, 2021).

But why do we have so many attempts to set up an ethical code for linguistic fieldwork? Regarding working with human societies, there are no easy solutions for the ethical dilemmas that arise from (Dwyer, 2006). Every situation requires a unique treatment and compromise. This is why, in addition to the creation of a framework that is as general as possible, the concrete application of such principles involves continued discussion. Dwyer (2006) suggests documenting the ethical issues and concerns that arise during a research project and the way these issues are addressed, such that other researchers can learn from the experience. While a code of conduct or principles is good, it runs the risk of introducing either overly complicated – or even inadequate – regulations, relegating this needed discussion.

Overall, we can summarize the principles that appear in all suggested lists under three major themes:

- *Consultation, Negotiation and Mutual Understanding.* The right to consultation of Indigenous people is stipulated in convention 167 of the International Labor Organization (Ilo, 1989) and states that they "have the right to preserve and develop their own institutions, languages, and cultures". Therefore, informing the commu-

[6] https://www.jstor.org/stable/pdf/26479543.pdf
[7] https://www.linguisticsociety.org/content/lsa-revised-ethics-statement-approved-july-2019
[8] https://dobes.mpi.nl/ethical_legal_aspects/DOBES-coc-v2.pdf

nity about the planned research, negotiating a possible outcome, and reaching a mutual agreement on the directions and details of the project should happen in all cases.

- *Respect of the local culture and involvement.* As each community has its own culture and view of the world, researchers – as well as any governing organizations interested in the project – should be familiar with the history and traditions of the community. In addition, it should be recommended that local researchers, speakers, or internal governments should be involved in the project.

- *Sharing and distribution of data and research.* The product of the research should be available for use by the community, so they can take advantage of the generated materials, like papers, books, or data.

Some of these commonly agreed-on principles need to be adapted to concrete situations, which might not be easy to do via a general approach. For instance, the documentation process will create data, and the ownership of this data is a major source of discussion (cf. Sections 4.6.4, **??**). Here, the traditional views of the communities might contradict the juridical system of a country (Daes, 1993). This problem does not have a simple solution and needs to be carefully considered when collecting data.

An additional call from these sources is to decolonize research and stop viewing Indigenous communities as sources of data, but rather as people with their own history Smith (2021). The current divorce between researchers and the cultural units of the communities can lead to reinforcing colonial legacy (Leonard, 2020).

As a final remark, we want to discuss the common assumption that any Ethical discussion must end with a normative setup for a field. It reduces indigenous institutions' collective to norms that allow an individual approach to the matter (Meza Salcedo, 2017). This would also not allow understanding the ethical questions with their own Indigenous communal cosmovision (Salcedo, 2016). Therefore, in this text, we aim to open the MT ethical debate to the NLP researchers and the Indigenous communities based on inclusion and dialog.

4.6.2. Ethics and *Human* Translation

For a successful translation, the inclusion of all participants is important, requiring their equal, informal, and understanding-oriented participation (Nissing & Müller, 2009). For Rachels & Rachels (1986), the minimum conception of morality is that when we give

"equal weight to the interests of each individual affected by one's decision." The question is how authors' intentions relate to the source culture's otherness, with their culturally-specific values (Chesterman, 2001). According to Doherty (2016), "the translation process studies emerged to focus on the translator and the process of translation rather than on the end product," incorporating mixed-method designs to obtain objective observations. A well-documented example of the non-ethical misuse of translation is the application of translation as an instrument for colonial domination. The main aim of this colonialist vision was to "civilize the savages" (Ludescher, 2001). For example, the summer institute of linguistics (SIL International)[9] was used for this goal during the 20th century in countries with Indigenous cultures, translating the Bible and trying to provoke a cultural change[10] in these communities (DelValls, 1978; Errington, 2001; Carey, 2010). Of course, these practices are not new and can be found throughout history (Gilmour, 2007). It is essential to note that non-ethical research can still deliver useful material and knowledge, e.g., for language revitalization (Premsrirat & Malone, 2003), but might inflict harm on the targeted community.

4.6.3. Ethics and *Machine* Translation

In the context of NLP research, the speakers are not directly involved when a model is trained (Pavlick et al., 2014). In contrast, the data collection processes (Fort et al., 2011) and human evaluation (Couillault et al., 2014) directly interact with the speakers and, therefore, have central importance regarding ethics. This is also true for the final translation service, which will interact with the broad public.

Data collection is the first and most evident issue regarding translation. Modern neural MT systems require a large amount of parallel data to be trained optimally (Junczys-Dowmunt, 2019). One way to obtain data is from crowd-sourcing (Fort et al., 2011).

[9] SIL International describes itself as ".. a global, faith-based nonprofit that works with local communities around the world to develop language solutions that expand possibilities for a better life. SIL's core contribution areas are Bible translation, literacy, education, development, linguistic research, and language tools.". https://www.sil.org/

[10] The role of SIL is controversial and, can not be summarized with one single statement. In our approach, we only refer to the role played in relation to cultural change. In many cases, the communities that got religious texts translated were already Christians, given previous colonization actions. However, there are also cases where, non-christian communities had Bibles and other religious texts translated into their language, with missionary aims. This triggered community divisions. For example, the translation of the religious texts to Wixarika (Fernández, 2022). This also happened in the Community of Zoquipan (in the Mexican state of Nayarit), where Christians, using the SIL-translated Bible, triggered an internal conflict in the community (the first author is part of this community). For the interested reader, we also recommend Dobrin (2009) introductory article.

However, this kind of job can be ill-paid and might constitute a problem for the living conditions of the workers (F. A. Schmidt, 2013). Also, data privacy is not trivial to handle. Systems must be able to filter sensitive information.

The problem of encoding biases[11], like gender bias (Stanovsky et al., 2019), is also an ethical concern for MT. It is also necessary to disclose the limitations and issues with certain systems (Leidner & Plachouras, 2017).

NLP research can also be used as a political instrument of power, where we can observe the mutual relationships between language, society, and the individual that "are also the source for the societal impact factors of NLP" (Horváth et al., 2017). In this way, NLP translation can be applied as an instrument to change the culture of minorities as in traditional translation (cf. Section 4.6.2). Thus, colonizers used translation as a means of imperial control and expropriation (Cheyfitz, 1997; Niranjana, 1992). The asymmetry of power causes domination, where subaltern cultures being flooded with "foreign materials and foreign language impositions" is a real danger for minority cultures (Tymoczko, 2006). Schwartz (2022) discuss the need to decolonize the scientific approach of the NLP community as a whole, expressing the need for researchers to be cognizant of the history and the cultural aspects of the communities that use the languages they are working with. Additionally, he proposes that our research should have an obligation to provide some benefit from our studies to the communities, an obligation of accountability (and therefore be in direct contact with their governing organizations), and an obligation of non-maleficence. The fact that many translation systems nowadays are multilingual[12] also result in more multi-cultural challenges (Hershcovich et al., 2022).

Finally, we also want to highlight the importance of discussing MT systems in a text-to-text setup. The usage of text is constrained to certain topics and varies from community to community. For instance, Wixarika and Quechua, languages that are spoken across all generations, are used in a written fashion mostly in private messaging apps (like WhatsApp) but also have a prolific Meme and Facebook publication generation[13]. Even if a certain community does not widely adopt the written tradition, there are, at minimum, legal obligations of the States toward indigenous languages. For example, some constitutions recognize indigenous languages as national languages (e.g., Mexico and Bolivia), binding the state to the responsibility of translating all official

[11] It is also important to note the typological features that might make this challenging. One example is polysynthetic languages and languages without gender coding (Klavans, 2018).

[12] Multilingual systems refer in NLP to systems capable of translating a set of languages from and to English. In some cases, they they also translate between languages where English is not involved.

[13] For example, Wixarika memes: https://www.facebook.com/memeswixarika2019, Quechua speaking group: https://www.facebook.com/groups/711230846397383/

pages, documents, laws, etc., to indigenous languages. This has not been implemented, and this case is a highly valuable application case for machine translation to assist human translation. However, our findings also apply to speech-to-text translation and speech-to-speech tasks that would cover all languages, even with no written tradition.

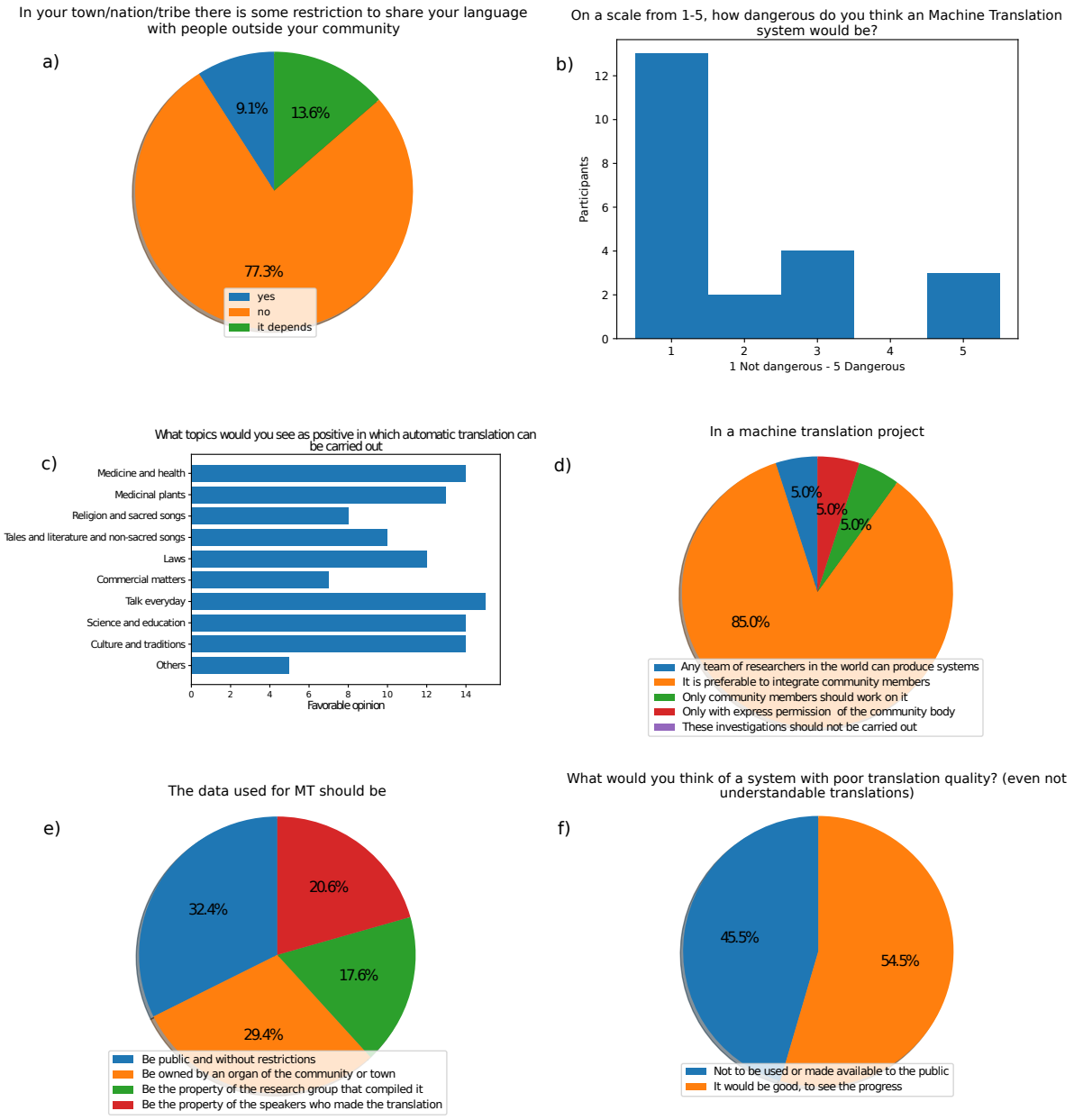

Figure 4.7.: Study performed on 22 participants who are members of Indigenous communities from the Americas.

4.6.4. The Speakers' Opinions

It is important to include the opinion and vision of speakers of endangered languages in NLP research, especially for topics such as MT. Therefore, we conduct a survey study with 22 language activists, teachers, and community leaders from the Americas. Importantly, our primary goal is not only to gather quantitative input on the ethical questions regarding MT for their languages but also to collect qualitative input by asking them to expand on their answers. Additionally, we also perform an interview with a subset of two participants of the initial interview study.

4.6.5. Study Design

We focus our study on the Americas,[14] selecting the following communities: Aymara, Chatino, Maya, Mazatec, Mixe, Nahua, Otomí, Quechua, Tenek, Tepehuano, Kichwa of Otavalo, and Zapotec. We want to note that our study does not aim to represent the general opinion of all Indigenous tribes, nor is it a final general statement on the issue. It is a case study that reveal the opinions of specific groups of speakers of Indigenous languages. Furthermore, the views of the interviewed individuals are their own and do not necessarily represent the views of their tribes, nations, or communities.

Quantitative and Qualitative aspects For the quantitative study, we used a survey. Surveys are a well-established technique to be used with Indigenous communities with an extensive history and are used and documented by classics like Edward Tylor, Anthony Wallace, Lewis Henry Morgan. This is also true for well-recognized Mexican (Indigenous engaged) social anthropologists (Jiménez & Ramos, 1985; Alfredo & Alberto, 1978).

For the qualitative part, we revisit the existing positional papers and articles of Indigenous researchers and activists. In addition, we use open questions in the survey, allowing extending the pure quantitative view to a qualitative one. Finally, we performed two 1-to-1 interviews with an activist (Mixe) and a linguist (Chatino).

Participant Recruitment We contact potential participants online in three ways. Our first approach is to establish communication through the potential participants' official project websites or public online accounts. This includes e-mail, Twitter, Facebook, and Instagram pages. Our second approach is to directly contact people in our target group

[14] Different parts of the world have very different levels of wariness, not just from colonial history but precisely due to interactions with field workers.

with whom at least one of the co-authors has already established a relationship. Finally, we also published a call for participation on social media and check if the volunteers belong to our target group. The goals of our research, and the reach and data handling, are explained directly to each participant and are also included in the final form. We do not gather any personal information about the participants, such as name, gender, age, etc. All study participants are volunteers.

Questionnaire Our study consists of 12 questions. The first three questions are rather general: they ask for the tribe, nation, or Indigenous people the participant belongs to if they self-identify as an activist, community leader, or teacher, and for their fluency in their language. The remaining questions target data policies, inclusion policies, benefits and dangers of MT systems, and best research practices. The full questionnaire is available in the appendix. The questions are available in English and Spanish, but only one form has been filled in English, while the rest has been completed in Spanish. Therefore, the authors have automatically translated all the comments.

4.6.6. Results

The results of the study can be seen in Figure 4.7. Additionally, we also discuss the open answers to each question to provide more insight.

Inclusion of Native Speakers and Permissions to Study the Language Figure 4.7(a) shows that 77.3% of the participants report that their community has no restrictions regarding the sharing of their language with outside people. The comments for this question show that many participants are proud of their language and heritage: "We are supportive and share our roots. Proud of who visits us" We even find stronger statements against the prohibition to share: "No one has the right to restrict the spread of the language". However, there also do exist communities with restrictions. Thus, we conclude that researchers cannot assume by default that all Indigenous groups would agree to share information about their language or would be happy about conducting research on it.

Benefits and Dangers of MT Systems Figure 4.7(b) shows that a strong majority of our participants think that an MT system for their language would be beneficial. However, there is also an important number of people who see at least some degree of danger. In this case, we should look at the participants' comments to understand their

worries. First, we find that a main concern for the participants is the translation quality. The fear of inadequate translations of cultural terms is also important. In Table 4.7, we can see a set of comments that illustrate these fears. One interesting comment refers to the fear of standardization of the participant's language, which could lead to a loss of diversity. In the same table, we can also see the benefits the participants expect, mostly in education and in elevating the status and usefulness of their languages.

Table 4.6 shows some answers to the open question on possible topics that might cause damage to the community. Most answers could not identify any possible topic that could be dangerous. However, the second most frequent answer was related to religion. Some answers are concerned that ancient ceremonial secrets could be revealed. Others also show concerns about the influence of Western religions. This brings us to the question if the Bible (Christodouloupoulos & Steedman, 2015; McCarthy et al., 2020a; Agić & Vulić, 2019) is suited to use as our default corpora for MT, when an indigenous language is involved. Finally, also few answers expressed that the usage of indigenous languages in the internal organization of the community could be in danger with MT systems. In contrast, figure 4.7(c) shows the topics that the most positive evaluation registered: everyday talks (15), science and education (14), culture and traditions (14), and medicine and health (14).

What would you see as damaging topics that should not be machine translated?
Anything ceremonial
Laws, medicine and health, science, mercantile matters, religion and sacred songs.
Issues that threaten organic life.
Western religion
Political situations and religions unless it is in the interest of the person.
Sacred songs, like those of a healer.

Table 4.6.: Some answers to the open question on possible dangers of MT for indigenous languages.

Participation of Members of Indigenous Communities in Research Figure 4.7(d) shows that our study participants think it is important to include people from the targeted communities in the research projects. This confirms the experience in linguistics, where they found a similar pattern (Smith, 2021) (see §4.6.1). It is important to note that only one answer stated that official permission is needed to perform the study. In the comments, the right of consulting was mentioned, together with the advantages of involving community members in research: "It is preferable [to integrate people from the

Can you think of any dangers to the language and culture, if so, which?
There are cultural linguistic concepts that are only understood in our native language.
The existence of so many variants would make the project little or not profitable and would lead the "experts" to an attempt to standardize language, which would be a tremendous mistake.
There are cultural elements that must be taken into account.
They could undoubtedly distort the proper use of the language.
What advantages would you see with an automatic translation system?
The use of automatic translators in spaces such as hospitals, government offices, etc.
Perhaps a contribution of modernity to the community, preservation of the native language.
It would contribute to the status of indigenous languages
It would contribute to the social use of our language
It would facilitate teaching because you would have many support tools.

Table 4.7.: Open answers of speakers to questions on dangers and benefits of MT systems for their communities.

community] to obtain a good system, and not just to have approximations, because only the members of the culture know how the language is used."; “So that the vocabulary is enriched and some words that do not exist are not imposed.”; “Carry out activities where the community can be involved, win-win.”.

Data Usage and Translation Quality Regarding data ownership and accessibility, we find diverse sets of responses. First, Figure 4.7(e) shows different opinions. Overall, we can say that a strong feeling exists that the data should be publicly available. However, regarding the property of the data, opinions are more diverse. Surprisingly, an important number of participants (17%) think that the external research group should own the data. Nevertheless, a higher number of participants think that the data should be owned by the community (29.4%), and 20.6% thinks it should be owned by the speakers who participate in the research. This is a difficult topic, as traditional norms and modern law system interact (cf. Section 4.6.1). In the comments, we find sad examples of mistrust in academic institutions. For example, one comment talks about previous problems of their tribe, as recordings and other material taken by linguists are not accessible to them: “Wary of academic institutions since we currently have issues accessing recordings that belong to academics and libraries and are not publicly accessible.” However, in general, we see a wide range of opinions: “The work of the few who take linguistic identity seriously should be valued”, “It could be public but always with the endorsement and consent of the community.” This diversity demonstrates that there is

a need for researchers to have a close relationship with the communities to understand the background and the aims of each particular case.

As discussed above, the quality of the final system is an important concern for many participants. In Figure 4.7(f) we can see that publishing an experimental MT system is also controversial. The possibility of using an experimental system is liked by 54.8% of our participants, which is slightly higher than the number of participants who are against this (45.5%). Some opinions against it are in line with earlier worries about incorrect translations of cultural content: "Something that is devoid of structure and cultural objectivity cannot be made available to the public" and "...damage would be caused to the language and its representatives since the learners would learn in the wrong way." Most people with a positive opinion agree that an initially poor system could be improved over time: "If it could be improved and corrected, that would be excellent."

4.6.7. Discussion

We survey the ongoing debate on ethics in documentation, translation, and MT before, presenting an interview study in Section 4.6.4. Now we discuss some of the most important issues we have identified in the last section in more depth.

Need for Consultations with Communities Previous experiences (Bird, 2020; Z. Liu et al., 2022) as well our study highlight the need for consultation with Indigenous communities when performing research involving their languages[15]. In some cases, the minimal expressed requirement is to inform speakers about new technological advances. Feedback and quality checks are also crucial for MT systems and important to the members of the communities. This consultation should include intercultural dialog as it has been a central instrument in the decision-making of indigenous communities (Beauclair, 2010). We recommend doing this by integrating community members into the loop while, of course, giving them the credit they deserve.

Legal systems vs. Traditional Views of Communal Knowledge Ownership Legal systems and, with that, copyright laws vary by country. However, legal rules sometimes conflict with the traditional views of Indigenous people (Dwyer, 2006). Thus, when working with Indigenous communities, we recommend discussing and agreeing upon ownership rights with annotators or study participants prior to starting the work to find an arrangement with which everyone is happy with. We would also like to point out

[15] An example of a community-engaged fieldwork is Czaykowska-Higgins (2009)

that, according to our case study, a general feeling is that data and research results need to be accessible to the community speaking the language. This contradicts the practice of some documentation efforts that close the collected data to the public and even to the speakers of the community (Avelino, 2021). Some participants in our study even suggest the usage of Creative Commons (CC)[16] for data. However, the use of CC might not be the best licensing option, as it not designed specifically for the needs of Indigenous. Finally, whenever the collected data are used for commercial usage, special agreements involving financial aspects are crucial.

Permissions Some communities require that a permit from their governing entity be obtained when someone, not a member, wants to study their language. This might be difficult as sometimes there is no central authority. Figuring out from whom to get permission can be challenging in such scenarios. However, as we see in this study, many communities do not require this permission. A promising project that aims to simplify this topic is the KP labels[17]. It is a set of labels that communities can use to express their permissions and willingness to cooperate with researchers and external projects.

Personal Data From the free-text answers, we further learn that, for many speakers, using their own language in their daily environment helps them protect their privacy: Their conversations can only be understood by their family or close environment. This concern of data handling is, however, also valid for other languages.

Concerns about Private Information of the Community The previous point can further be extended to assemblies and other organizational meetings, where the language barrier is used to keep their decisions or strategies private. This is one worry that the communities have with MT and the possible topics that might be harmful for them. Some communities also have general concerns about sharing their language with people who do not belong to them (e.g., the Hopi Dictionary controversy (Hill, 2002)). For this case, it is important not to approach this issue from a Western legal point of view and go toward traditional internal governance practices and norms and consultation with the communities.

Religion and the Bible Regarding problematic domains for MT, multiple survey participants mentioned religion. This is quite relevant for the NLP community, as the

[16] https://creativecommons.org/licenses/
[17] https://localcontexts.org/labels/traditional-knowledge-labels/

broadest resource currently available for minority languages is the Bible. As seen in Section 4.6.2, the colonial usage of the translation of religious texts (Niranjana, 1990) is precisely the origin of these detests. Thus, we recommend that NLP and MT researchers use the Bible carefully, through a consultation process, and consider its impacts. Nevertheless, without a close relationship with each community (e.g., in a massive multilingual MT experiment), the recommendation is to void using the Bible.

Technology and data Sovereignty Having technology for their own languages is well seen by most study participants. However, we also find a strong desire to participate directly in the development of MT systems. This requires more inclusion of Indigenous researchers in NLP. Therefore, training Indigenous researchers and engineers is an important task that we recommend should be valued more highly by the NLP and MT communities. We are aware that existing inequalities cannot be removed immediately or in isolation, but everyone can be supportive.[18] The creation of a collaborative process is a proposal emerging from the communities themselves: "Technology as Tequio; technological creation and innovation as a common good" (Aguilar-Gil, 2020). However, it is not possible to build contemporary data-driven NLP technologies without data. Furthermore, this opens the discussion regarding Data Sovereignty. First, it is important to mention that the communities have the right to self-determination, and this includes the data that they create. Applying this sovereignty to data refers to having control over the data, knowledge[19] and cultural expressions created by these communities. As discussed in this section, it is important to reach agreements with the communities through consultation and direct collaboration. This includes the licensing and ownership of the final data products.

Our Findings and Previous Work Finally, we want to relate our findings to similar discussions in prior work. Most previous concerns and suggestions related to including and consulting people from the communities (Bird, 2020; Z. Liu et al., 2022) are aligned with the wishes and desires of the participants in our study. The inclusion of community members as co-authors (Z. Liu et al., 2022) should not be an artificial mechanic but rather a broad inclusion process, including data and technology sovereignty. This is also aligned with the community building aimed at by S. Zhang et al. (2022). Additionally,

[18] Tech sovereignty is a central topic for the Natives in Tech conference in 2022: https://nativesintech.org/conference/2022

[19] See https://indigenousinnovate.org/downloads/indigenous-knowledges-and-data-governance-protocol_may-2021.pdf

we should consider that there might exist problematic topics and not underestimate the importance of high-quality translations.

4.7. Findings for MT

This section discussed and studied our research question 3: "Is morphological segmentation useful for the machine translation of polysynthetic languages?". To answer this question, we first introduced in detail the basics of MT and discussed the essential techniques and issues when applied to low-resource scenarios. Afterward, we also analyzed the challenges and current situation of MT for the indigenous languages of the Americas (ILA). Having this context, we empirically study the impact of morphological segmentation on the machine translation task when one of the languages is a polysynthetic language. To make that possible, we curated a new Raramuri–Spanish dataset and a morphological segmentation dataset for Shipibo–Konibo, and re-used some of the morphological datasets described in chapter 3. The experimental results from this chapter, together with our in-depth literature review, show the following findings:

- Machine translation for the indigenous languages of the Americas is a challenging task. The languages have strong dialectic varieties, also the lack of normalization[20], low-resource scenarios and the endangered situation of some languages complicate the recollection of data and the creation of computational models. Nevertheless, community projects, with the active involvement of native speakers, can help overcome some of these issues. This model helped us promote the AmericasNLP community (among others) we co-founded and organized.

- Unsupervised morphological segmentation algorithms are an excellent alternative to byte pair encoding (BPE). Flatcat and LMVR are robust alternatives for translating from and to a polysynthetic language. In our experimental results, they even consistently outperformed BPEs. This is relevant, because it shows that morphology can be helpful in the extreme case of polysynthetic languages and extremely low-resource scenarios.

- In contrast, supervised morphological methods are sub-optimal when applied to the downstream machine translation task, even if these methods achieve the best

[20] We consider this as a challenge, but not we are not promoting the normalization into the languages. In the opinion of language activists and linguists language normalization can lead to the destruction of the natural diversity of languages. Rather than changing the language, we believe that the computational models should be sufficiently robust to handle these scenarios.

experimental results on morphological benchmark datasets. With the help of in-depth analysis, we noticed a significant increase of OOV symbols when applying these segmentation models. We hypothesize that this can be due to the domain mismatch between the curated morphology and the noisier parallel dataset.

- Neural supervised methods can be strongly improved with pre-training on unlabeled data, as shown in chapter 3. This same performance improvement is also observed when segmenting the parallel data for MT. Such methods can close the gap with purely supervised methods.

- Machine Translation for Indigenous languages should always consider the needs and opinions of the native speakers and communities. This inclusion should be at all levels: developing community researchers, annotators, and consultation consultations with speakers or governing entities.

In answering the research question, we found evidence that morphological segmentation can improve MT for polysynthetic languages when using unsupervised morphological segmentation. However, it can be subject to other factors, even if we try to reduce the external variables. As discussed, the noise in the dataset, the lack of orthographic normalization, and the dialectical variation (even between villages) can influence the final result. This improvement could also disappear if the size of the training data increases (as shown by (Gaser, 2022)). The lack of large and diverse datasets adds to the challenge.

5. Handling Code-switching

This chapter introduces the phenomenon of code-switching (CS), alternating multiple languages during a text or conversation. Minority languages are mostly spoken by communities, nations, or in general by human groups that share a common language and live in a country where the language is not the most spoken one (Poplack, 2008). This can refer to migrants, indigenous groups, and minority languages, among others. These strong multilingual environments trigger the code-switching phenomena (Skiba, 1997; Crystal, 1987). CS appears daily in written and oral forms Gardner-Chloros et al. (2009) and presents a challenge when processing it with computational models (Çetinoğlu et al., 2016).

Therefore, this is an essential aspect of the polysynthetic indigenous languages of the Americas, as their speakers have a solid bilingual tradition due to their close relations with the dominant national languages (i.e., Spanish, Portuguese, and English). First, we study the implications of sub-word level code-switching (swCS), which is particularly relevant for polysynthetic languages. In the second part of this chapter, we will explore the labeling and segmentation task to handle swCS. Finally, we will also study the impact of CS in MT.

5.1. Code-Switching

5.1.1. Definitions

CS alternates between two or more languages during a conversation of a piece of text (Çetinoğlu et al., 2016). Therefore, code-switching can occur when all participants (or at least one) of the communicative process are bilingual or multilingual. However, this is not the only way than other languages can be incorporated into another one. A close phenomenon is lexical borrowing. The exact boundaries between code-switching and lexical borrowing remain an open question in linguistics (Syawkany, 2022). The classic definition is provided by Poplack & Meechan (1995): lexical borrowing occurs if some

	Example
Inter-sentencial CS	Me *ne*helmoso. Pensando en ty.
Word-level CS	*hik+rix+a* ya pude soñar ami *tuxeri*
Sub-word CS	Sabe que esta aula del SABER SER es un excelente espacio educativo!!! *Pam+pariyutsi Waut+a Kuruxi Manuwe mat+a*!!!

Table 5.1.: Examples of different types of CS switching. Words in Wixarika are marked in italics. All examples do not have orthographic corrections or any other normalization process.

items from another language are morphosyntactically integrated language. The extreme case is the loan words that are fully integrated into the native language and are widely used by native monolingual speakers. If that is not the case, we can refer to it as code-switching (Syawkany, 2022). However, linguists like Myers-Scotton (1993) and Bentahila & Davies (1983) dispute that there is a difference between code-switching and lexical borrowing as in both cases must exist a morphosyntactical integration. In this work, we will focus on the existence of language mixing at any degree that occurs in a single text.

The causes for code-switching are multiple. We can include globalization, different kinds of migrations, colonialism, and language revival, among many other cases (Milroy et al., 1995). This phenomenon is not new, as contacts between languages are present throughout humankind's history (Campbell, 2000). We can define bilingualism as the knowledge of two languages at a certain degree. This leads the bilingual speaker to a new configuration, not just the sum of two monolingual understandings (Mackey et al., 1995).

We can identify three main levels of code-switching in a text:

- **Inter-sentential CS**. This form was the first to be described. The idea is to have large chunks of texts written or spoken in different languages. The delimiter for this kind of code-switching is the sentence.

- **Word-level CS**. This type of code-switching is the most common one explored by the NLP community. It occurs when parts of a single sentence contain words of more than one language. In this case, the delimitation between languages is set on the word boundaries.

- **Sub-word CS**. Finally, we can also see that the morphemes of two, or even more languages can, be combined, forming a mixed word. This last phenomenon can

mostly be found if at least a language is morphologically rich. An example of intra-word CS is between the Romance language Spanish and the Yuto-Aztecan language Wixarika.

From a linguistic point of view, a diverse array of theories attempt to model the code-switching phenomenon. Poplack's work is the most prominent of the pioneers in CS research based on her analysis of Spanish–English. In this work, she proposes her linear order constraints model, which is composed of two constraints: "code-switches will tend to occur at points in discourse where the juxtaposition of L1 and L2 elements does not violate a syntactic rule of either language and" (Poplack, 1980) and "codes may be switched after any constituent in discourse provided that constituent is not a bound morpheme" (Poplack, 1980). Future research proved that these constraints were often violated in other language pairs. The first model is the formulation of the existence of a matrix model. This means that in the CS phenomenon, languages have an asymmetry and that one of the languages is dominant. Therefore, these languages embed the second language in their grammar. This was proposed by Myers-Scotton's Matrix Language Frame model (Myers-Scotton, 1993). Nortier (1995) explains that code-switching does not always follow the matrix language's constraints and that extra grammatical factors can influence this phenomenon. An alternative to the matrix language model is the structural approach. It tries to find a universally applicable and predictive grammar that constrains CS. Until now, finding such a grammar(Gardner-Chloros & Edwards, 2004) has yet to be successful.

5.1.2. NLP perspective on code-switching

The research on CS has been studied primarily task-based. The first task to be studied was LID, as it allowed us to better understand the behavior of code-switching and model possible switching points (Solorio & Liu, 2008). Recently, a set of new tasks attracted the attention of researchers, like machine translation, named entity recognition (NER), sentiment analysis, and dependency parsing. As a result, the number of tasks and datasets available for code-switching has grown, but we now have a more diverse set of analyzed language pairs. As bilingual communities and speakers exist worldwide, code-switching is not an isolated phenomenon and occurs when endangered languages interact with the dominant languages of that region. Therefore in chapter 5, we introduce datasets with code-switching for the Wixarika–Spanish language pair. We explore and study tasks are machine translation and sub-word level language identification.

5.1.3. Code-switching challenges

Çetinoğlu et al. (2016) discussed the general challenges that NLP faces when working with CS data. In this section, we relate these challenges with the ones proper to the indigenous languages of the Americas that are polysynthetic.

Low resources. Code-switching is a low-resource task (up to now) for all languages, even for the most studied English-Spanish pair. The main reason is the difficulty of gathering texts or recordings that contain true, everyday code-switching data, as these data is used in spontaneous conversations. The ethical, legal, and technical difficulties are not trivial. For the ILA we found that most of the data contain a certain degree of code-switching (CS)[1], but it is mainly used with private messaging systems (i.e., WhatsApp). Nearly no official books, newspapers, blogs, encyclopedias, or other forms of written monolingual text are available, with some remarkable exceptions (cf. 5.2.2). Therefore, in the case of ILA, the CS data collection issue has the same implications as the collection of monolingual data.

Linguistic Richness As mentioned before, most languages require an official language and orthographic normalization, and the linguistic richness of these languages (with huge dialectical variations) makes it even harder for computational models to process them. At the same time, we have also borrowing words that are phonologically integrated into the ILA. All this makes it harder to process.

Noisy text . As the CS texts usually appear in the user-generated text, the non-standard text also appears in the non-indigenous language (in our case, in Spanish). We also add to the data scarcity in the low-resource scenarios.

In general, independently from the studied task, each model that works with CS text that involves ILA must deal with the challenges listed here.

5.2. Handling sub-word level Code-Switching

In settings where multilingual speakers share more than one language, mixing two or more languages within a single text, for example, a tweet, is becoming increasingly

[1]During our data recollection, the results show that 43.2% of the sentences that we collected for Wixarika-Spanish, with more than three words, contain at least one morpheme in the form of code-switching.

(a)

ne'iwa	*pecansadox-i*
`hch`	`mix`
my.brother	you|are.tired.PPFV

(b)

ne'iwa	*pe*	*cansado*	*x-i*
`hch`	`hch`	ES	`hch`
my.brother	you|are	tired	PPFV

'My brother, you are tired.'

Figure 5.1.: Intra-word CS between Spanish and Wixarika, **(a)** standard LID for CS, **(b)** our task. PPFV stands for past perfective.

common Grosjean (2010). This constitutes a challenge for natural language processing (NLP) systems because they are commonly designed to handle one language at a time.

CS can be found in multiple non-exclusive variants. For instance, sentences in different languages can be mixed within one text, or words from different languages can be combined into sentences (cf. 5.1). CS can also occur at the subword level when speakers combine morphemes from different languages (swCS). This last phenomenon can mostly be found if at least a language is morphologically rich. An example of intra-word CS between the Romance language Spanish and the Yuto-Aztecan language Wixarika is shown in Figure 5.1.

The language identification (LID) task (i.e., predicting the language of each token in a text) has attracted much attention in recent years (cf. Solorio et al. (2014); Molina et al. (2016)). However, swCS is mostly not handled explicitly: words with morphemes from more than one language are simply tagged with a `mix`label.

While this works reasonably well for previously studied language pairs, overlooking intra-word CS leads to a relevant loss of information for highly polysynthetic languages.

A mixed word is unknown for NLP systems, yet a single word contains much more information than in less agglutinative languages, cf. Figure 5.1(a). The information it contains could be central for a downstream task, e.g., in Figure 5.1(b). The word *cansado* (*tired*) is valuable for sentiment analysis. Furthermore, we find intra-word CS to be much more frequent for Spanish–Wixarika than for previously studied language pairs, so handling it is crucial.

For example, in Table 5.1, we see mixed words composed of a Spanish root and Wixarika affixes. In this case, the word *taprimori* is equivalent to two words in English

(*our cousins*) while *hakiewa* is the Wixarika word for *where* and *is* [2]. In addition, we will show that for at least one language pair involving a polysynthetic language (i.e., Spanish–Wixarika), intra-word CS is much more frequent than for previously studied pairs.

For the German-Turkish (GR-TR) pair (O. Çetinoğlu, 2016) we can find a relatively higher number of unique mixed tokens (197 1.16% of all tokens). However, this number should be considered due to the amount of information compressed into those tokens.

Due to the importance of these words for the overall meaning of a sentence, they should not be handled as Out Of Vocabulary (OOV) tokens. In languages with rich morphology, intra-word code switching occurs more often, and in some extreme cases, can encode most of a sentence in a single mixed word (as shown in Table 5.1).

We introduce a new CS dataset for Spanish–Wixarika (`esp–hch`) and modify an existing German–Turkish (DE–TR) CS corpus (O. Çetinoğlu, 2016) to be processed for sub-word level language identification (SLID). We then apply a segmental RNN (SegRNN) model for the task, which we compare against several strong baselines. Our experiments show clear advantages of SegRNNs over all baselines for intra-word CS.

5.2.1. Related Work

For the LID task, we have seen a continuous interest (Al-Badrashiny & Diab, 2016a; Rijhwani et al., 2017; Y. Zhang et al., 2018), including two shared tasks on the topic (Solorio et al., 2014; Molina et al., 2016) which covered a wide range of language pairs (Spanish–English, Nepali–English, Mandarin–English, Modern Standard Arabic–Arabic dialects). This is reflected in the increased number of available datasets for language detection in CS (Jose et al., 2020).

Most research has been done on LID on the word and token level. Examples of such work and datasets are: Barman et al. (2014) for English, Bengali and Hindi; Maharjan et al. (2015) for Nepali–English and Spanish–English; Kusampudi et al. (2021) for Telugu–English; Mandal et al. (2018) Bengali–English.

The most similar work to ours is D. Nguyen & Cornips (2016), which focused on detecting intra-word CS for Dutch–Limburgish (D. Nguyen et al., 2015). The authors utilized Morfessor (Creutz & Lagus, 2002) to segment all words into morphemes and Wikipedia to assign LID probabilities to each morpheme. However, their task definition and evaluation are on the word level. Furthermore, as this method relies on large

[2] as this word is only used in questions, it also include the question mark that is not explicitly written in this language

monolingual resources, it only applies to low-resource languages like Wixarika, which has its own Wikipedia edition. Additionally, CS between Hindi and English has attracted attention (Rudra et al., 2016; Patro et al., 2017; Pratapa et al., 2018). Posterior works to ours are Sabty et al. (2021) for Egyptian Arabic–English; and Taguchi et al. (2022) that use sub-word level language identification (SLID) for creating a universal dependency treebank. On the application side of sub-word level language identification (SLID), Taguchi et al. (2021) improve Cyrillic-Latin transliteration of the Tartar language on the sub-word level.

Finally, we refer the interested reader to specialized surveys of the current trends, methods, tasks, and datasets in code switching (Winata et al., 2022; Sitaram et al., 2019; Doğruöz et al., 2021). For a language identification CS focused review we suggest Hidayatullah et al. (2022).

5.2.2. Task and Data Description

We introduce the task and two datasets: the German–Trukish (`deu-tur`) and Spanish–Wixarika (`esp-hch`) language pairs.

Task Description

Formally, the task of sub-word level language identification (SLID) consists of producing two sequences, given an input sequence of tokens $X = \langle x_1, \ldots, x_i, \ldots, x_{|X|} \rangle$. The first sequence contains all words and splits $X^s = \langle x_1^s, \ldots, x_i^s, \ldots, x_{|X|}^s \rangle$, where each x_i^s is an m-tuple of variable length $0 < m \leq |x_i|$, where $|x_i|$ is the number of characters in x_i. The second sequence is such that $T^s = \langle t_1^s, \ldots, t_i^s, \ldots, t_{|X|}^s \rangle$, where $|T^s| = |X^s| = |X|$ and each $t_i^s \in T^s$ is an n-tuple of tags from a given set of LID tags. An input-output example for a DE–TR mixed phrase is shown in Figure 5.2.

Input	⟨ 'Yerim',	'seni', ',',	'danke', 'Schatzym'⟩
Output	⟨ (Yerim),	(seni), (,),	(danke), (Schatzy, m)⟩
	⟨ (`tur`),	(`tur`), (`others`),	(`tur`), (`deu`, `tur`)⟩

Figure 5.2.: Subword-level LID in German–Turkish.

Datasets

German–Turkish The German–Turkish Twitter Corpus (Ö. Çetinoğlu & Çöltekin, 2016) consists of 1029 tweets with 17K tokens. They are manually normalized, tokenized,

and annotated with language IDs. The language ID tag set consists of `tur` (Turkish), `deu` (German), `lang3` (another language), `mix`(intra-word CS), `ambig` (ambiguous language ID in context), and `other` (punctuation, numbers, emoticons, symbols, etc.). Named entities are tagged with a combination of `ne` and their language ID: `ne.tur`, `ne.deu`, `ne.lang3`. In the original corpus, some Turkish and mixed words undergo a morphosyntactic split,[3] with splitting points not usually corresponding to language boundaries. For subword-level LID, these morphosyntactic splits are merged back into single words. We manually segment `mix` words at language boundaries and replace their labels with more fine-grained language ID tags. However, the percentage of sentences with mixed words is 15.66%. The complete dataset statistics can be found in Table 5.2.

Tokens	All	%	Unique	Unique %
deu	3992	20.37	1360	20.43
tur	9913	50.59	4071	61.16
lang3	112	0.57	83	1.25
ambig	32	0.16	23	0.18
other	4345	22.17	294	4.42
ne.tur	417	2.13	275	4.13
ne.deu	389	1.99	244	3.67
ne.ambig	16	0.08	12	1.25
ne.lang3	112	0.57	95	1.43
mixed	231	1.18	183	2.75
deu tur	231	100.0	183	100.0

Table 5.2.: The frequency breakdown of tokens by language IDs in the German-Turkish dataset. *All*: the total number of tokens per tag, *%*: the percentage of them for the total number of tokens; *Unique*: the number of unique word types, and *Unique %*: the percentage of them with respect to the total number of unique word types.

Spanish–Wixarika Our second dataset consists of 985 sentences and 8K tokens in Spanish and Wixarika. The data is collected from public postings and comments from Facebook accounts. To ensure the public character of these posts, we manually collect data that is accessible publicly without being logged in to Facebook to comply with the users' terms of use and privacy. These posts and comments are taken from 34 users: 14 women, 10 men, and the rest do not publicly reveal their gender. None of them have

[3] E.g., separating copular suffixes from the roots to which they are attached to, cf. Ö. Çetinoğlu & Çöltekin (2016) for details.

publicaly mentioned their age. To obtain a dataset that focuses on the LID task, we only consider threads where the CS phenomenon appears. As the Wixarika language has non-wide accepted normalization, we decided to avoid any normalization process, even for the Spanish tokens. We replace usernames with `@username` to preserve privacy. Afterward, we tokenize the text, segment the mixed words, and add language IDs to words and segments.

Tokens	All	%	Unique	Unique %
`esp`	4218	50.73	1527	45.76
`hch`	2019	24.28	1191	35.69
`eng`	24	0.29	21	0.63
`ambig`	28	0.34	25	0.75
`pther`	1664	20.01	288	8.63
`ne.esp`	96	1.15	85	2.55
`ne.hch`	77	0.93	49	1.47
`ne.en`	11	0.13	9	0.27
`mix`	177	2.13	142	4.26
esp hch	35	*19.77*	31	*21.83*
hch esp	122	*68.93*	93	*65.49*
hch esp hch	17	*9.60*	31	*10.56*
hch eng	1	*0.07*	1	*0.07*
eng esp	1	*0.07*	1	*0.07*

Table 5.3.: Number of tokens classified by language tags in the Spanish-Wixarika dataset. We show the total number of *Tokens* per tag, their proportion (*%*) with the total tokens, the *Unique* word types, and their proportion (*Unique %*) of them with the total number of unique word types.

Wixarika, like many other indigenous languages, has a robust oral tradition but lacks a widely adopted standardized form and usage. With the advent of the internet and mobile communications, previously isolated communities began using social media, such as Facebook, daily. The interaction with the Spanish environment has made using code-switching text in social media very common. `other` (punctuation, numbers, emoticons, etc), `ne.esp`, `ne.hch` and `ne.eng` (named entities). Mixed words are segmented, and each segment is labeled with its corresponding language (`esp`, `hch`, `eng`).

Table 5.3 shows a detailed dataset description. The percentage of mixed words is higher than in the `deu–tur`dataset: 3.13% of the tokens and 4.26% of the types. The most common combination is Spanish roots with Wixarika affixes. Furthermore, 16.55% of the sentences contain mixed words.

We split the `deu–tur` corpus and the `esp–hch` corpus into training and test sets of

	Train	Dev	Test	Total
ES–WI	616	154	216	985
DE–TR	640	160	228	1029

Table 5.4.: Overview of our datasets.

sizes 800:229 and 770:216, respectively. Error analysis and hyperparameter tuning are performed on the training set via 5-fold cross-validation. Finally, we present the results of the test sets. Both datasets are available at https://www.ims.uni-stuttgart.de/institut/mitarbeiter/ozlem/NAACL2019.html Table 5.4 shows the final data splitting.

	DE–TR						
	Segmentation			Tagging			Char
	P	R	F1	P	R	F1	Acc.
SegRNN	60.4	46.8	**53.0**	78.8	60.2	74.0	**72.9**
BiLSTM+Seq2Seq	46.1	33.4	38.7	84.3	66.8	**74.5**	67.7
BiLSTM+CRF	49.4	34.5	40.6	84.3	66.8	**74.5**	68.0
CRFTag+Seq2Seq	12.7	6.8	8.9	27.8	15.2	19.7	37.2
CRFTag+CRF	11.0	5.9	7.7	27.8	15.2	19.7	36.8
CharBiLSTM	19.1	26.6	22.2	32.9	45.7	38.2	61.1
	ES–WIX						
	Segmentation			Tagging			Char
	P	R	F1	P	R	F1	Acc.
SegRNN	75.6	62.7	**68.5**	85.3	70.5	**77.2**	**84.6**
BiLSTM+Seq2Seq	66.6	52.2	58.5	82.4	66.2	73.4	78.4
BiLSTM+CRF	61.1	48.4	54.0	82.4	66.3	73.4	76.7
CRFTag+Seq2Seq	47.2	31.9	38.1	63.5	43.0	51.3	69.6
CRFTag+CRF	47.6	32.4	38.6	63.5	43.0	51.3	69.4
CharBiLSTM	49.7	52.2	50.8	63.1	68.1	66.5	75.7

Table 5.5.: Segmentation and LID test results for mixed words only.

5.2.3. Proposed Model

SegRNN We suggest a SegRNN (Kong et al., 2016) would be the best fit for our task because it models a joint probability distribution of $p(y, \mathbf{E}|\mathbf{X})$ as it segments of a character sequence $x = \langle x_1, \ldots, x_{|x|} \rangle$ and labeling of the segments. Each segment has

a duration $e \in \mathbb{E}$ and a label $t \in T$. Note that this model is equivalent to a semi-Markov conditional random field (CRF Lafferty et al., 2001) that defines the conditional probability with the auxiliary segment labels **E**, such that

$$p(y, \mathbf{E}|\mathbf{X}) = \frac{1}{Z(x)} \tag{5.1}$$

where Z normalizes over the sum of all possible (t, E). The function f is

$$f(t_j, e_i, \mathbf{X}) = \mathbf{w}^T \phi(t_j, e_i, \mathbf{X}), \tag{5.2}$$

where **w** is a weight vector and ϕ is a non-linear activation function. We use a bidirectional long short term memory (LSTM) for the features: one LSTM computes the forward embeddings, and a second LSTM computes the backward embeddings. The resulting hidden states are then concatenated with embeddings representing the segment durations and tags, respectively. The output of g_t and g_e functions maps the segment durations and tags into a vector space, first representing them as a one-hot vector, which is then mapped into a continuous space by a linear embedding matrix. The inference is made via dynamic programing.

The model is trained to optimize the following objective, which corresponds to the joint log-likelihood of the segment lengths e and the language tags t:

$$\mathcal{L}(\theta) = sum_{(x,t,e)\in\mathcal{D}} - \log p(t, e|x) \tag{5.3}$$

$\mathcal{D}$ denotes the training data, θ is the set of model parameters, x the input, t the tag sequence and e is the sequence of the segment lengths.

Our inputs are single words.[4] As hyperparameters, we use 1 RNN layer, a 64-dimensional input layer, 32 dimensions for tags, 16 for segments, and 4 for lengths. For training, we use Adam (Kingma & Ba, 2014).

5.2.4. Experimental Setup

Baselines

Now, we will describe the baselines we use to evaluate the overall performance of our proposed model. First, table 5.6 shows input, process, and output examples for pipelines

[4] We also experimented with entire phrases as inputs, and the achieved scores were slightly worse than for word-based inputs.

(BiLSTM+Seq2Seq/BiLSTM+CRF, CRFTag+Seq2Seq/CRFTag+CRF) and CharBiLSTM.

BiLSTM+Seq2Seq/BiLSTM+CRF Our first baselines are pipelines. First, the input text is tagged with language IDs. [5] The language IDs of a mixed word are directly predicted as a combination of all the language ID tags of the word (i.e., `hch_esp`). Second, a subword-level model segment words with composed language ID tags. For word-level tagging, we use a hierarchical bidirectional LSTM (**BiLSTM**) that incorporates both token- and character-level information (Plank et al., 2016), similar to the winning system (Samih et al., 2016) of the Second Code-Switching Shared Task (Molina et al., 2016). [6]

For the subword level, we use two low resource supervised segmentation methods that had proven to achieve good results for the morphological surface segmentation task. The first method is a **CRF** segmenter proposed by Ruokolainen et al. (2013) that models segmentation as a labeling problem and a sequence-to-sequence (**Seq2Seq**) model trained with an auxiliary task as proposed by Kann et al. (2018).

CRFTag+Seq2Seq/CRFTag+CRF Since our datasets might be small for training neural networks, we substitute the BiLSTM with a CRF tagger (Müller et al., 2013, **CRFTag**) in the first step. We use the same two approaches as the previous baselines for segmentation.

CharBiLSTM We further employ a BiLSTM to tag each character with a language ID. Each character inherits the language ID of the word or segment to which it belongs for training. At the prediction time, if the characters of a word have different language IDs, the word is split. We experiment with single words as input (*Word*-CharBiLSTM) and the entire phrase as input (*Phrase*-CharBiLSTM).

Metrics

We use two metrics for evaluation. First, we follow Kong et al. (2016) and calculate precision (P), recall (R), and F1 using, segments as units (an unsegmented word cor-

[5] We choose LID first because the best tagger achieves a higher accuracy (95.11% and 91.22 %) than the best segmentation model (92.70% and 83.268%) on the development set for `esp–hch`and `deu–tur`, respectively.

[6] For all BiLSTM models the input dimension is 100 with a hidden layer size of 100. For training, we use a stochastic gradient descent Bottou (2010), 30 epochs, with a learning rate of 0.1. A 0.25 dropout factor is applied.

CharBiLSTM		
Input	*ne'iwa*	*pe cansado x+*
Character-level	*n e ' i w a*	*p e | c a n s a d o | x +*
Tag	hch hch hch hch hch hch	hch hch esp esp esp esp esp esp esp hch hch
Reconstruct	*ne'iwa* hch	*pe* hch | cansado esp | *r+* hch
Pipeline approach		
Input	*ne'iwa*	*pecansadox+*
LSTM/CRFs tag	hch	hch-esp-hch
seq2seq/CRFs split	*ne'iwa*	pe | cansado | x+
Reconstruct	*ne'iwa* hch	pe hch / cansado esp /r+ hch

Table 5.6.: Example that shows the input, process, and output approaches for our baselines. Pipeline approaches refer to either combination of CRFs and LSTM with Seq2Seq(BiLSTM + Seq2Seq/ BiLSTM + CRF and CRFTag + Seq2Seq/ CRFTag + CRF)

responds to one segment). The word-based evaluation is stricter and expects an exact match on the word level when calculating precision, recall, and F1. We also report a tagging accuracy (Char Acc.) by assigning a language ID to each character and calculating the ratio of correct language tags overall characters.

The overall evaluation considers each unsegmented word and each segment of a segmented word as a token and calculates F1 scores. Precision (P) is the ratio of the count of valid gold tokens observed in the hypothesis to the count of the hypothesis events. Recall (R) is the count of correct gold tokens observed in the hypothesis over the count of gold events. F1 is the harmonic mean of P and R. The same principle is applied to tagging. We further give a tagging accuracy (Char Acc.) by assigning a language ID to each character and calculating the ratio of correct language tags for all characters. Intra-word evaluation checks the exact match for only the segmented words. Precision is the ratio of the count of correct to matches over the count of segmented words according to the hypothesis. Recall is the ratio of the count of correct matches to the count of gold-segmented words.

We perform the evaluation on the resulting output, taking as equivalent words and word segments. We calculate the F1 score for each LID and output token, where precision (P) is the number of the correct counts of segments or tokens from the hypothesis observed in the gold standard over the total amount of hypothesis events. For recall (R), we used the correct counts of segments or tokens from the gold standard observed in the hypothesis over the total number of gold standard events. F1 is the harmonic mean of P and R. To get a bigger, more precise picture of the subword-level phenomenon; we

assigned to each character a corresponding LID tag. With this information, we calculated the overall accuracy (acc.).

	deu–tur			esp–hch		
	Seg. F1	Tag. F1	Char Acc.	Seg. F1	Tag. F1	Char Acc.
SegRNN	**98.7**	94.0	93.6	97.8	**92.5**	**92.4**
BiLSTM+Seq2Seq	98.6	**95.1**	94.3	**98.1**	90.9	90.7
BiLSTM+CRF	**98.7**	94.9	**94.4**	97.9	87.8	90.6
CRFTag+Seq2Seq	98.4	93.7	93.1	97.7	90.4	90.1
CRFTag+CRF	98.4	93.7	93.1	97.6	90.3	90.1
CharBiLSTM	87.7	88.0	92.5	89.7	87.9	91.3

Table 5.7.: Test set results for entire datasets.

5.2.5. Results and Discussion

Table 5.7 shows all test results for the entire dataset. We find the following: (i) For esp–hch, SegRNN performs slightly better for tagging than the best baseline in terms of F1 and character accuracy. For deu–tur, SegRNN and, BiLSTM+CRF are the best segmentation models, but the BiLSTM model has slightly better results for esp–hch(with a 1.42% higher accuracy than the the BiLSTM baseline), while the pipeline of BiLSTM tagging and splitting approach also achieved a slightly better score for the DE-TR data (.9% of acc. over the best SegRNN). (ii) The CRF pipelines perform slightly worse than the best word-level BiLSTM models for both datasets and all evaluations, with 93.12% acc. for deu–turand 90.11% acc. for esp–hch. (iii) In most cases, F1 and the accuracy scores are correlated. The exception is the CharBiLSTM models, whose F1 scores are relatively low compared to the accuracies. This is caused by the accuracy measured at the character level, while F1 scores evaluate the final segmentation and tags.

Table 5.5 shows the results of tagging and segmentation only for the mixed words in our datasets. Here, we can see that: (i) Our SegRNN model achieves the best performance for segmentation. Differences to the other approaches are $\geq 10\%$, showing clearly why these models are suitable for the task when the number of words belonging to two languages is high. (ii) The pipeline BiLSTM models work best for tagging the DE–TR data with a slight margin but underperform on the esp–hch dataset as compared to the SegRNN models. (iii) Both CRFTag models achieve very low results for both segmentation and tagging. (iv) CharBiLSTM performs better than the CRFTag models

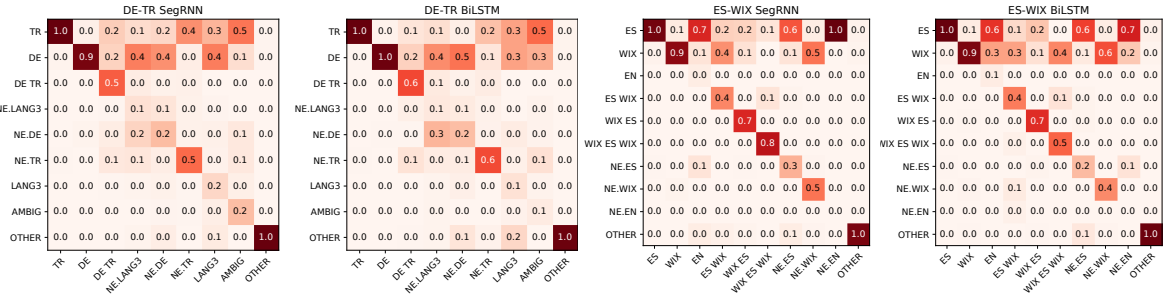

Figure 5.3.: Confusion matrices of the two best models on both datasets. The x axis represents the tags seen in the gold standard, and the y axis shows the corresponding predicted tags. Values are rounded up. Therefore, not all columns add up to 1.

on both tasks but is worse than all other approaches in our experiments.

More generally, we further observe that recall on mixed words for the `deu–tur` pair is low for all systems compared to `esp–hch`. This effect is especially strong for the CRFTag and CharBiLSTM models, which seem to be unable to correctly identify mixed words. While this tendency can also be seen for the `esp–hch` pair, it is less extreme. The models that are best at finding mixed words are the SegRNN models, followed by the BiLSTM pipeline models. We suggest that the better segmentation and tagging of mixed words for `esp–hch` might be due to two factors: the higher percentage of available examples of mixed words in the training set for `esp–hch`. The second finding is that SegRNN worked better overall for the `esp–hch` pair than the pipeline models. This is relevant because this language pair has a higher quantity of mixed words.

Overall, we conclude that SegRNN models seem to work better on language pairs that have more intra-word CS, while pipeline approaches might be as good for language pairs where the number of mixed words is lower since these models had a better performance tagging correctly monolingual words.

Error analysis. Figure 5.3 shows confusion matrices for SegRNN and BiLSTM+Seg2Seg. Both models achieve good results by assigning monolingual tags (`esp`, `hch`, `deu`, `tur`) and punctuation symbols (`others`). The hardest labels to classify are named entities (`ne`, `ne.tur`, `ne.tur`, `ne.hch`, `ne.esp`), as well as a third language and ambiguous tags (`lang3`, `eng`, `ambig`). Performance on multilingual tags (`deu tur`, `hch ESP`, `esp hch`, `hch esp hch`) is mixed. For `deu tur`, BiLSTM+Seq2Seq gets slightly better classifications, but for the `esp–hch`tags SegRNN achieves better results.

Regarding the over-segmentation problems, BiLSTM+Seq2Seq (0.8% for `deu–tur` and

2.0% for esp–hch) slightly underperforms SegRNN (0.7% for DE–TR and 1.13% for esp–hch). The BiLSTM+Seq2Seq (2.4%) makes fewer under-segmentation errors for deu–tur than SegRNN (2.7%). However, for esp–hch, SegRNN performs better with 3.81% under-segmentation errors compared to 4.2% of BiLSTM+Seq2Seq.

5.3. Code-switching with MT

In this section, we explore the impact of CS on the machine translation task (MT) when we translate from or into a CS text to a single of these languages. We also investigate the usefulness of CS to improve the traditional language MT task, as previous work has shown that synthetic CS data helps to improve the final system performance.

For this investigation, we work on two language pairs Wixarika (hch)–Spanish (esp) and Egyptian Arabic (arz)– English (eng). For hch–engWe use the existing parallel datasets (Mager et al., 2017) curated for the AmericasNLP shared task (A. Ebrahimi et al., 2022), and for arz–eng we use Hamed et al. (2020). We also introduce a CS parallel dataset (see table 5.8) for hch–esp.

esp	a mi no me gusta estar casado
hch	*nekwanetsinake* estar casado
CS	*nekwanetsinake nem+ne+krni*

Table 5.8.: Example of the Wixarika–Spanish parallel code-switching dataset. Italic tokens are Wixarika words. The + symbol is part of the Wixarika alphabet.

In this section, we aim to explore the following research questions: i) is real CS data helpful for MT machine translation?; ii) what is the difference between using real CS data and synthetic CS?; and iii) what is the best way to train a system with CS data? Our research questions are constrained to a highly low-resource environment, where a languages has extreme morphological reach and to a single language pair.

To answer these questions, we base our experiments on the state-of-the-art MT system from (Vázquez et al., 2021) for the hch–esp language pair. We test in all directions. We also perform the same experiments on arz and eng to compare them to a higher-resource (but still low-resource) setup.

1. **Synthetic CS data**. Following previous work, we first generate a synthetic CS dataset by substituting random words with their equivalent from a dictionary and pair them with the original text. Then, we train the parallel and synthetic CS data

as two different language pairs. A variant of this experiment is a simple merge of both datasets without any language token.

2. **Real CS data**. Then we substitute the synthetic data with real CS data and compare their performance.

3. **Multilingual system**. We also explore using all data together in a multilingual setup.

5.3.1. Previous Work

Previous work has mostly explored the usage of synthetic CS to improve MT models. Nevertheless, recent studies also explore the MT process from (and into) CS. K. Song, Zhang, et al. (2019) use synthetic CS to perform a data augmentation for MT. This data augmentation involves replacing words from a monolingual dataset with words from the second language with the help of a bilingual dictionary. In addition, pre-training models with synthetic CS help to improve MT systems (Z. Yang et al., 2020). Here, a random word in the monolingual dataset is swapped with a translated word in the second language. This word is sampled from accurate translations inferred from a probabilistic bilingual lexicon.

Hamed et al. (2022) investigates the usage of lexical replacement methods to generate synthetic code-switching data to improve Arabic–English CS to English MT. With the same language pair, we (Gaser, 2022) explore the influence of morphological segmentation on MT. Here, we search for the best segmentation method to boost the MT process when a translation has code-switching. Both previous studies use the ArzEn dataset (Hamed et al., 2020). The main difference between this dataset from ours is that it only contains the pair of Arabic–English CS to English the Arabic–English to Arabic direction.

Our work uses the synthetic code-switching generation approaches to compare its performance with real code-switching data when used for an MT task.

5.3.2. CS parallel dataset

To perform this study, we explore the Wixarika–Spanish language pair. We use the already existing dataset introduced in section 5.2.2, which already contains code-switching. These data, as explained, was collected from Facebook public comments and publications. To increase the number of data points, we also collect 511 instances from the

	CS	esp	hch
Tokens (N)	9254	9918	8714
N / setences	6.16	6.60	5.80
Vocabulary (V)	4218	2667	3599
V / N	45.58%	26.89%	41.30%
ED vs CS	-	18.42	21.80

Table 5.9.: Parallel Code-switching dataset description. We show the number of tokens (N) for each language, the number of tokens per sentence (N/sentences), and the number of unique tokens that appear per language (V). We also show the difference, using edit distance (ED), between the CS dataset with the Spanish and Wixarika translations.

	Train	Dev.	Test	Total
Phrases	297	400	800	1496

Table 5.10.: Splits and total size (phrases) of our parallel Code-switching dataset.

same platform that contain Code-Switching. Afterward, we asked one native Wixarika speaker[7] to translate each sentence into Spanish and Wixarika. Therefore, we could gather a three-way dataset containing CS, Wixarika, and Spanish.

Table 5.9 gives a better insight into the dataset. We noticed that the edit distance (ED) between the translation (`hch` and `esp`) to `CS` is not that large, with both not exceeding 21.80. We also noticed that, as expected, the vocabulary size increases for `CS`, and that the data sparsity of `hch` is higher than for `esp` (V).

Given the extreme scarcity of data for this task, we chose to split our dataset, giving preference to the test set that contains more 53.47% of the dataset. For finetuning, we used half, 400 instances, and the rest for training (297). The intuition behind this decision is that with a strong pre-trained system, we can finetune such a system with a small dataset and still understand the role of CS, for the MT task. Table 5.10 shows the final dataset splitting.

5.3.3. Artificial CS creation

Following previous work, we want to study the impact of synthetic code-switching text for the Egyptian–Engish pair. Hamed et al. (2022) shows that selecting random words

[7] His dialect is from Zoquipan, the Mexican state of Nayarit. This is the same dialect as our Wixarika-Spanish parallel dataset.

in a matrix language sentence and substituting that word with a closely related word in the second language can boost MT applications. This approach also outperforms a carefully designed method that includes morphosyntactic agreements. However, in the case of highly low-resourced languages, it is impossible to replicate their experiments completely: we have no substantial monolingual datasets for the Wixarika language, with only 511 monolingual sentences that we can use, collected from Facebook social media[8] (Mager, Carrillo, & Meza, 2018b). However, a small dictionary provided by Gómez & López (1999) contains 767 Spanish–Wixarika aligned words. As we use only the parallel Wixarika-Spanish corpora for MT training, performing a word alignment algorithm is also not an option. Therefore, the only word substitution method that can be used under these circumstances is a dictionary-based approach.

The approach that we use for creating of the synthetic dataset is as follows:

1. We first select a Spanish monolingual dataset as close as possible to our CS dataset. We found that using WhatsApp messages collected by Dorantes et al. (2018) matches the criteria of similarity regarding conversation topics, language usage, and noise. This means that the conversations are mostly daily casual conversations, with many word modifications, including spelling errors, emoticons, and creative language usage.

2. For the Wixarika side, we use the public Facebook posts collected by Mager, Carrillo, & Meza (2018b). These phrases match the same domain, language usage, and even social media platforms as the CS data.

3. The Spanish–Wixarika vocabulary included in Gómez & López (1999) is used as our substitution dictionary.

4. We iterate through the dataset until we find a word contained in the substitution dictionary. Given the low word substitution rate for, we decided to use substitutions in all cases where a word was found in the dictionary.

Table 5.11 shows the characteristics of the synthetically generated dataset. We divide the dataset into two: the `CS`generated from Spanish monolingual text and `CS`generated from Wixarika text. In both cases, we can see that the edit distance (ED) is considerably lower than in natural CS text: 7.19, and 6.76 from the artificial CS data, compared to

[8] https://github.com/pywirrarika/wixarikacorpora/blob/master/wixmonolingual/social.wix

	CS(from esp)	esp	CS(from hch)	hch
Phrases (Ph)	599		164	
Tokens (N)	5285	5094	1526	1523
N / (Pp)	8.82	8.50	9.30	9.28
Vocabulary (V)	1991	2002	941	942
V / N (%)	37.67	39.30	61.66	61.85
ED vs CS	7.19	-	6.79	-

Table 5.11.: Synthetic Parallel Code-switching dataset description.

Dataset	Language	Phrase
Natural CS	CS	pani nepeukwaim+k+ lo voy a comprar
	esp	quiero comer pan lo voy a comprar
	hch	pani nepeukwaim+k+ nepinaneni
Synth CSf esp	CS	kepaukwa lo veas sabrás cuál
	esp	cuándo lo veas sabrás cuál
Synth CSf hch	CS	Ukiratsi kepatme dinero pemeu
	hch	Ukiratsi kepatme tumini pemeu

Table 5.12.: Code-switching examples, comparing real code-switching phrases with our synthetically generated data. The examples are not aligned.

18.42 and 21.80 for the natural CS data. Because the dictionary-based substitution-based method relies on a constrained number of words, we decided to apply all possible substitutions. The percentage of sentences that could be modified from the original dataset was 22.06% for Spanish and 32.09%.

Table 5.12 shows examples of natural CS and synthetically generated CS for, both cases: from Spanish and Wixarika monolingual text. We can see that in both synthetically generated cases, the substitution is just applied in for one word, while in the Natural CS examples, we see more words from a second language. Even if these examples were chosen from the synthetic dataset, we found this pattern for most sentences during the manual data inspection.

5.3.4. Training MT models with CS

Transfer-learning in extremely low resource setups

For our experiments, we use as a base the ModelB setup of the winning system (Vázquez et al., 2021) of the AmericasNLP2021 shared task. This setup trains a multilingual model

for the 10 languages (Nahuatl, Wixarika, Hñahñu, Bribri, Shipibo-Konibio, Guaraní, Quechua, Ashaninka, Aymara, and Rarámirui) to Spanish (and in the reverse direction). The dataset used for training is the AmericasNLI dataset A. Ebrahimi et al. (2022) with a collection of additional datasets, such as the bible and constitutions (see the original paper to have a complete description). The collected dataset is filtered using OpusFilter (Aulamo et al., 2020) and cleaned from duplicate instances afterward. Text normalization scripts are also applied in the case that they are available for a particular language. For training, two steps are used: i) the languages are trained together with a large Spanish–English parallel dataset (90% `esp–eng` and 10% for the other languages). ii) afterward, the model is fine-tuned with the 10-language dataset for the → Spanish direction and fine-tuned just for the specific language in the opposite direction. For implementation, we use the same base code from the original result: as a machine translation framework, we use OpenNMT (Klein et al., 2017) with the same hyper-parameters defined by Vázquez et al. (2021). For all models, we do not use back translation data, as we use that data to perform CS modifications.

Adapting MT models for CS

To adapt the models, we add the code-switching data as a new language. For the experiments, we have the following training setups. As a reference, the reader can figure 5.4.

Additional language to mark CS (`addCS`) . In this setup, we add the code-switching data as an additional language pair instead of training our base model with the 10 languages of AmericasNLP (see 5.4.1) This means that it will have its language tag and can be translated into directly from the input.

Additional finetuning step (`ftCS`) . As seen in figure 5.4.2, we fine-tune the initial model to each CS data.

For all cases, we want to explore the models' performance for the Spanish → Wixarika (and inverse direction), and for translating from and into Code-switching. We also want to explore the possible benefits and problems of including CS in the equation.

5.3.5. Experimental Setup

In this section, we will present the results of our experiments and discuss them. We use the chrF Popović (2015) as our evaluation metric, in concordance with section 2.4.3.

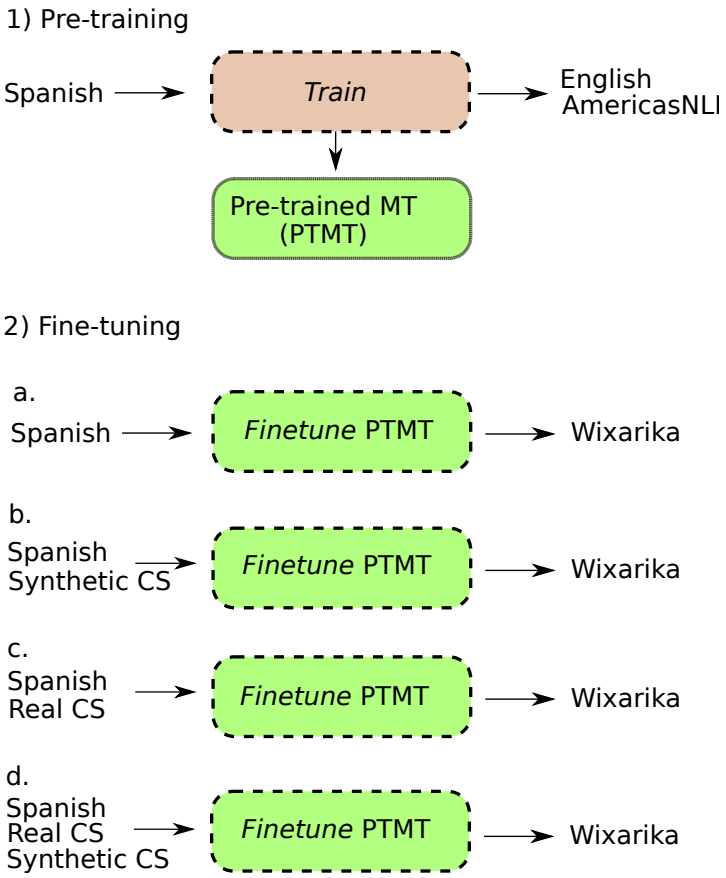

Figure 5.4.: Training and finetuning strategies to include CS data into the base Helsinki translation model.

5.3.6. Results

Table 5.13 shows the results when the inputs are monolingual, in both directions: from esp ⟷ hch. For all esp*to*hch cases, we see small gains over the bilingual baseline when adding synthetic code-switching (synthCS), with 0.83 chrF on average. Nevertheless, the opposite happens for the esp*to*hchdirection, where the addition of synthetic CS data downgraded the performance by an average of 0.35 crfF. On the other hand, we found that adding natural CS (naturalCS) can also sometimes hurt the performance of the bilingual model. In the case of esp⟷hch direction, we found that the performance decreases when using naturally occurring CS (0.83 for addCS). This drop can be mitigated by training the system in a multi-lingual fashion (−0.35). For the opposite direction (hch→esp), the results show that the scores for hch→esp when of the system when using naturalCS are not significantly different from the bilingual baseline at $P < 0.5$.

Method	**System**	esp→ hch	hch→ esp
	Bilingual Baseline	30.06	25.40
addCS	Fintetuned w/ naturalCS	♦29.22	♦25.81
	Fintetuned w/ synthCS	♦30.65	♦24.59
mlCS	Fintetuned w/ naturalCS	♦29.70	♦25.89
	Fintetuned w/ synthCS	♦31.14	25.31

Table 5.13.: Results for the translations with different training strategies for esp and hch translaiton. All scores are in chrF. The significant difference (at $P < .05$) concerning the bilingual baseline is the market with ♦.

When comparing the naturalCS with the syntheticCS systems, we notice that the performance is mixed. For the esp→hch direction, the systems trained on synthCS (30.65 addCS, and 31.14 mlCS) perform better than the ones trined on naturalCS (29.22 addCS, and 29.70 mlCS). However, the trend flips when it comes to the hch→espdirection: the naturalCS systems perform better (25.81 addCS, and 25.89 mlCS) than the ones trained on synthCS (24.59 addCS, and 25.31 mlCS).

Finally, we observe a general trend for all training corpora: the best CS integration model for the bilingual task is using the CS data as a separate language in a multilingual system.

Table 5.14 shows the results when translating a CS text into a monolingual esp or hch text. The results show that the systems trained without code-switching text have a problem handling CS text in the input. The bilingual baseline underperforms the average of systems with CS in 6.924 for CS→hch, and 18.01 to esp. We also notice

Method	System	CS→hch	CS→esp
	Bilingual Baseline	27.03	21.51
addCS	Fintetuned w/ naturalCS	32.87	41.26
	Fintetuned w/ synthCS	33.75	40.49
mlCS	Fintetuned w/ naturalCS	33.73	43.01
	Fintetuned w/ synthCS	35.47	33.33

Table 5.14.: Results with different training strategies translating from and into CS text from hch or hch. All scores are in chrF. The significant differences (at $P < .05$) with respect to the bilingual baseline are market with ♦.

that, as in the previous experiments, including CS text into the training process works better when adding it as a separate language (mlCS), rather than just appending the text (addCS).

When comparing the performance of naturalCS and synthCS, we can also find the same trend as in the previous experiments. For CS→hchthe performance increases when using synthCS (33.75 chrF for addCS and 35.47 crhF for mlCS), compared to the lower scores of naturalCS (32.87 chrF for addCS and 33.75 crhF for mlCS). The opposite happens when applying synthCS (32.87 chrF for addCS and 33.75 crhF for mlCS) when compared to the better scores of naturalCS (32.87 chrF for addCS and 33.75 crhF for mlCS).

5.3.7. Discussion

First, we discuss the best way to use code-switching text in a machine translation task. From the results presented in §5.3.6 we can see that, between our experimented strategies to incorporate CS data into a machine translation system, the multilingual approach is the best fit, with a consistent improvement over just appending the CS data to the bilingual corpus. This result also holds for translating from a CS input into a monolingual text (esp or hch).

For the second research question in this section, we found that having a direct comparison between natural and synthetic code-switching text is difficult to perform, given the extremely low-resource environment we are working on. There are three main factors that need to be considered when creating synthetic data: i) a possible domain missmatch between the natural CS text and the monolingual data that is a candidate to be modified. Finding a monolingual text for extremely low-resourced languages is difficult,

especially when we need to find text that matches a specific domain. ii) the resources available to generate synthetic data are not optimal: in our case, we have to rely on a static word list to perform 1-to-1 substitutions. Furthermore, iii) the small amount of data available to generate the data from the extremely low-resourced language is prohibitive to take real advantage of synthetic data.

Even with the issues listed, synthetic data is better suited to improve MT systems when the data is large enough than the naturally occurring CS. As we can see in the results, when the available synthetic data has the same size as the naturally occurring CS data, the synthetic has the best results. We hypothesize that the naturally occurring CS is noisier and introduces errors that a model in low resource conditions cannot handle correctly. This can be seen as adding natural CS can even hurt the performance of a system trained on bilingual paired text.

Finally, as expected, models trained on bilingual paired text only perform poorly when the input contains code-switching. In this case, all systems trained with synthetic or natural CS achieve better performance than the former.

5.3.8. Analysis

To better understand the results shown in §5.3.6, we sample 3 examples from the development set from naturalCS and synthCS when used with the better system `mlCS` setup and compare them to the bilingual baseline. The examples for the translation between the `esp`⟷`hch` are shown in table 5.15. The first aspect we notice is the overall semantically incorrect translations in all directions and models. This insight shows us that the automatic metrics can miss-lead about the quality of the translation. However, we can see the difficult problem of concepts created in a polysynthetic language like Wixarika. If we look at example number 3b, we can see that there is no explicit word in Wixarika for rainforest (Spanish *selva*). This is substituted with the notion of hot and green in Wixarika. The models with code-switching surprisingly could catch this literal translation: *era verde y soplaba caliente.* (it was green, and the hot air was blowing). On the other hand, in example 2b, surprisingly, the model introduced a negation when interpreting wrong the morpheme *ka*. Finally, in example 1b, the models without CS could have a correct translation, but the CS models failed.

Table 5.16 shows the examples when translating a CS text into `esp` or `hch`. In this case, we can observe more accurate translations in general. In example 1b, we can see that the CS model could correctly translate the Wixarika word *muxuri* into *guamuchil*. It might be possible that the word was seen during training in the extended

CS dataset, showing the advantages of using real CS data. No other system could find it. In example 3b, we also found that Wixarika words were just removed from the translation. When it comes to the translation from `CS` to `hch`, we see in example 2a the good capacity of the model to form verbs correctly. The models with CS can find partially correct conjugations of an alternative stem (*mie*) for the verb coming. However, as seen in previous work, the translation systems can generate a decent translation of simple sentences but fail regarding more complex constructions, as seen in example 3a.

5.4. Findings

In this chapter, we investigated the common appearing phenomenon of code-switching (CS) when handling indigenous languages of the Americas (ILA) text. We introduced the phenomenon and the most important challenges in our context. We performed experiments to answer research question 3 of this thesis: "How can we handle and take advantage of Code-Switching in these languages?". Our findings regarding this question are focused on the Wixarika language, as we could find and gather data for it. This language is paired with Spanish, and to a minor degree with English. All findings are restricted to this language pair and cannot be generalized to all indigenous languages of the Americas:

1. Wixarika, as other ILA, has a robust oral tradition but has recently been used in written form. Therefore, it is often found together with non-standard Spanish text. Also, Wixarika is not standardized, which adds to the challenge.

2. As Wixarika is a polysynthetic language, code-switching is also important on the morphological level, making it essential to include this factor when processing data. To take this into account, we introduce an extension to the language identification (LID) task: sub-word level language identification (SLID). This extension is a joint task of segmenting all words into parts, where each part belongs to a different language and the corresponding language tag for each segment. No segmentation is performed if a word belongs to a single language.

3. Even with a small amount of annotated data, it is possible to obtain subword segmentation and tagging using a segmentental RNN (SegRNN) or a pipeline approach (first segment, then tag).

4. For the MT task, a bilingual-only trained model is not capable of handling CS input, as well as if we included CS text into its training.

5. On the other hand, CS can be simulated with synthetic data from the monolingual text. This synthetic data is overall more helpful for training than natural CS.

6. Finally, we also found that the best way (from the explored methods) to include CS text into a multilingual machine translation model is to train it as a different language instead of just appending the code-switching text to the bilingual data.

esp→hch			
No.	CS Data	Input	Output
1a	Gold	Y en realidad no lo entendió.	yemek+ p+kei ti eniet+ya.
1a	-	Y en realidad no lo entendió.	meta nit+ma atsi kare'enax+.
1a	Natural	Y en realidad no lo entendió.	meta nit+ma kaheitima.
1a	Synth	Y en realidad no lo entendió.	meta atsi pukare'eri.
2a	Gold	Estoy haciendo lo que debo hacer hoy.	m+k+ k+ axapeure ka eriet+ yeika.
2a	-	Estoy haciendo lo que debo hacer hoy.	m+k+ 'ek+ m+pa+ pe'it+arieximeni.
2a	Natural	Estoy haciendo lo que debo hacer hoy.	m+k+ta kametini kuaim+k+.
2a	Synth	Estoy haciendo lo que debo hacer hoy.	m+k+ m+k+ kametini kuaim+k+.
3a	Gold	Era una selva tropical.	pe h+k+t+kai metsi+ra+ ye tsie nieka ti+x+kat+.
3a	-	Era una selva tropical.	peh+k+t+kai +tsita.
3a	Natural	Era una selva tropical.	peh+k+t+kai tupiriyari.
3a	Synth	Era una selva tropical.	peh+k+t+kai tupiriya tropical.
hch→esp			
No.	CS Data	Input	Output
1b	Gold	yemek+ p+kei ti eniet+ya.	Y en realidad no lo entendió.
1b	-	yemek+ p+kei ti eniet+ya.	Es muy difícil de entender.
1b	Natural	yemek+ p+kei ti eniet+ya.	No hay manera de que se encoja.
1b	Synth	yemek+ p+kei ti eniet+ya.	y fue una voz muy fuerte.
2b	Gold	m+k+ k+ axapeure ka eriet+ yeika.	Estoy haciendo lo que debo hacer hoy.
2b	-	m+k+ k+ axapeure ka eriet+ yeika.	por lo que ella no se sintió satisfecha.
2b	Natural	m+k+ k+ axapeure ka eriet+ yeika.	Él no se sintió seguro.
2b	Synth	m+k+ k+ axapeure ka eriet+ yeika.	y ella no se sintió satisfecha.
3b	Gold	pe h+k+t+kai metsi+ra+ ye tsie nieka ti+x+kat+.	Era una selva tropical.
3b	-	pe h+k+t+kai metsi+ra+ ye tsie nieka ti+x+kat+.	sino que era una botella de aceite sobre su hombro.
3b	Natural	pe h+k+t+kai metsi+ra+ ye tsie nieka ti+x+kat+.	era verde y caliente.
3b	Synth	pe h+k+t+kai metsi+ra+ ye tsie nieka ti+x+kat+.	era verde y soplaba caliente.

Table 5.15.: Machine translation for the Spanish (esp) – Wixarika (hch) language pair when, the systems are trained together with the CS text.

CS→hch			
No.	CS Data	Input	Output
1a	Gold	en mi casa puxuawe muxuri	ne kie pexuawe muxuri
1a	-	en mi casa puxuawe muxuri	ne kitsie puxuawe muxuri
1a	Natural	en mi casa puxuawe muxuri	neki puxuawe muxuri
1a	Synth	en mi casa puxuawe muxuri	nekie puxuawe muxuri
2a	Gold	hik+tá a donde vas	hik+tá hakewa pepe+t+a
2a	-	hik+tá a donde vas	hik+ +xa hakewa pepet+a
2a	Natural	hik+tá a donde vas	hik+ta hakewa pemeyemie
2a	Synth	hik+tá a donde vas	hik+ta hakewa kenemie
3a	Gold	Así fue como se fueron dando las cosas!! tam+Katiy+we	m+pa+ p+ ane kwitsi mautsi+ra+ye.
3a	Gold	Así fue como se fueron dando las cosas!! tam+Katiy+we	ipa+ p+ti+y+ muyuku xexe+riwakai pinite!! Tam+ katiy+we.
3a	Natural	Así fue como se fueron dando las cosas!! tam+Katiy+we	m+pa+ta pureyu tuaxa!! tam+ +Katiy+we
3a	Synth	Así fue como se fueron dando las cosas!! tam+Katiy+we	m+pa+ta pey+ wa+ka pini!! tam+ +Katiy+we

CS→esp			
No.	CS Data	Input	Output
1b	Gold	en mi casa puxuawe muxuri	en mi casa hay guamuchil
1b	-	en mi casa puxuawe muxuri	En mi casa hay muxuri
1b	Natural	en mi casa puxuawe muxuri	En mi casa hay guamuchil
1b	Synth	en mi casa puxuawe muxuri	En mi casa hay muxur
2b	Gold	hik+tá a donde vas	y ahora a donde vas
2b	-	hik+tá a donde vas	hik+ta a hakewa pepeyewie
2b	Natural	hik+tá a donde vas	Hoy en día , una varita .
2b	Synth	hik+tá a donde vas	Ahora tenemos una casa vas
3b	Gold	Así fue como se fueron dando las cosas!! tam+Katiy+we	asi fue como se fueron dando las cosas !! Yo creo que si se puede.
3b	-	Así fue como se fueron dando las cosas!! tam+Katiy+we	¡ Aix+ aix+ p+h+k+ aix+ p+katiy+we ! tam+ katiy+we
3b	Natural	Así fue como se fueron dando las cosas!! tam+Katiy+we	¡ Asi fue como se fuen dando las cosas !
3b	Synth	Así fue como se fueron dando las cosas!! tam+Katiy+we	¡ Así fue como se fueron dando las cosas!!

Table 5.16.: Machine translation examples, where the source is a code-switching text, to Spanish (esp) and Wixarika (hch).

6. Conclusions

Much of the success of NLP relies on vast amounts of monolingual data and a decent amount of annotated data. However, these resources are not available for most of the world's languages. While ISO codes account for 6359 languages,[1] less than 80 have more than 1 million paired sentences in publicly available datasets (Arivazhagan, Bapna, Firat, Lepikhin, et al., 2019), mostly with English. Only 67 languages have more than 100 thousand articles in Wikipedia, and only 166 languages have a web crawl in OSCAR (Ortiz Suárez et al., 2019). Moreover, only 282 languages have an official Wikipedia page. Thus, the languages with extensive monolingual datasets only correspond to about 4.5% of all known languages. This problem is not trivial when we try to build high-quality models for the indigenous languages of the Americas. Most of the languages from the continent are part of the excluded languages, with only Guaraní, Quechua, Nahuatl, Inuktitut, Nēhiyawēwin, Aymara, Navajo, and Cree. However, only Inuktitut, Nēhiyawēwin, and Aymara have more than one hundred articles.

In addition to this scenario of extreme resources, most languages are also endangered. This leads us to the problem that the speaker community is small and has no young generations that use it daily.

In this context, in this thesis, we study eight indigenous languages of the Americas (Mexicanero, Nahuatl, Wixarika, Yorem Nokki, Shipibo-Konibo, Rarámuri, Popoluca, and Tepehua). All of these languages are polysynthetic. As the morphology in these languages has an important role, we first explore in chapter 3 the modeling of the morphology of these languages, with two NLP tasks: surface segmentation for languages that have a concatenative morphology (Mexicanero, Nahuatl, Wixarika, Yorem Nokki, Rarámuri), and canonical segmentation for Popoluca and Tepehua, which have a fusional morphology. To answer our **research question 1** "Is it possible to model the complex polysynthetic morphology with neural networks in low resource scenarios?", we modeled the morphology with a wide array of neural network architectures and found that the sequence-to-sequence architecture with Recurrent Neural Networks is able to outperform

[1] Ethnologue (M. P. Lewis, 2009) counts 6909 languages.

the best non-neural models, even in extreme low-resource scenarios. The best variations for this task are seq-to-seq models with explicit or induced copy capabilities. To allow neural networks to take advantage of the not annotated text, we also introduced a semi-supervised approach that uses unsupervised morphological segmentation as silver data for pre-training our models. To make all our experiments possible, we introduced new segmentation datasets for Popoluca, Tepehua, Shipibo-Konibo, and Rarámuri.

Being able to translate from and into a language is vital for various reasons. It allows native speakers to access their language texts from all around the world and share their texts; in general, it is a tool for learning and access for future generations in case the language is lost. In Chapter 4, we introduce the general concepts of machine translation (MT) and its applications to low-resource languages, and in particular to the indigenous languages of the Americas. In order to answer our **research question 2** "Is morphological segmentation useful for Machine Translation of polysynthetic languages?", we apply morphological segmentation to the polysynthetic language side of the MT systems. We study a comprehensive array of segmentation methods: supervised, unsupervised, semi-supervised, neural, and non-neuronal. We compare all these systems with the standard frequency-based byte pair encoding (BPE) algorithm. We show that unsupervised morphological segmentation achieves the best results, even outperforming significantly BPE. However, we also found no direct correlation between the segmentation quality with the translation quality. This is especially noticeable with the supervised methods, which perform best for the morphological segmentation task but have the worst scores in machine translation. We found that our semi-supervised sequence-to-sequence segmentation method that includes in-domain monolingual text can close the gap.

As all of the indigenous languages of the Americas are minority languages in their countries, contact with Spanish, English, French, or Portuguese is common. This also provokes a strong bilingual phenomenon in the native-speaker communities. A consequence of this is the common presence of code-switching (CS) in the daily usage of the languages. In chapter 5, we study how to model CS for two tasks: language identification and machine translation. This allows us to tackle our **research question 3** "How can we handle and take advantage of Code-Switching in these languages?". We first expand the LID task and propose sub-word level language identification (SLID), which not only identifies the language of each work but also segments words that contain multiple languages and assigns a language tag to each part. We then experiment with a set of models and pipelines in extremely low-resources. We found that for the Wixarika–Spanish pair, the most efficient method is a segmentental RNN (SegRNN).

We further contrast our results with the agglutinative language Turkish, which is paired with German, confirming our findings. We then use code-switching to enhance the machine translation of the Wixarika–English language pair. For this, we use a multi-lingual model trained on 10 indigenous languages from the AmericasNLP shared task (Mager et al., 2021), and finetune it, adding the CS corpora as an additional language. To compare it with previous findings, a synthetic CS dataset is also created. When comparing the performance of both CS techniques, we notice that synthetic CS helps better boost the translation performance than naturally occurring CS. The downside is the lack of sufficient in-domain monolingual data for Wixarika that we can use to generate sufficient synthetic data. Finally, our experiments also show that CS data is needed to be able to translate CS data into Spanish or Wixarika.

We conclude that the space for improvements is large even with the advances in NLP methods for indigenous languages of the Americas presented in this thesis. When examining the outputs of our systems (i.e., machine translation systems), the quality is far from usable. The dialectical variations of the languages, the lack of orthographic standards, the challenging morphology, and the lack of data are not easily solvable. Only with a strong and long-term community effort that includes (not mutually exclusive) native speakers, NLP researchers, linguists, language activists, and sponsors will we see usable systems. With this in mind, we have started the AmericasNLP workshop and shared a task that has gathered researchers worldwide.

The main research in this thesis has focused on text processing and generation. However, the strong oral tradition of these languages also makes us call attention to speech processing. If we aim to create technologies that fit the needs of native speakers and communities, automatic speech recognition (ASR) and speech synthesis is a priority. As an ongoing work, we have centered the AmericasNLP shared task of 2022 on speech-to-text translation. Nevertheless, this topic will be a part of future work.

References

Abdulmumin, I., Galadanci, B. S., & Garba, A. (2019). Tag-less back-translation. *arXiv preprint arXiv:1912.10514*.

Adelani, D., Alabi, J., Fan, A., Kreutzer, J., Shen, X., Reid, M., ... Manthalu, S. (2022, July). A few thousand translations go a long way! leveraging pre-trained models for African news translation. In *Proceedings of the 2022 conference of the north american chapter of the association for computational linguistics: Human language technologies* (pp. 3053–3070). Seattle, United States: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2022.naacl-main.223 doi: 10.18653/v1/2022.naacl-main.223

Agić, Ž., & 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* (pp. 3204–3210).

Aguilar-Gil, Y. E. (2020, December). *A modest proposal to save the world.* https://restofworld.org/2020/saving-the-world-through-tequiology/.

Aharoni, R., & Goldberg, Y. (2017, July). Morphological inflection generation with hard monotonic attention. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 2004–2015). Vancouver, Canada: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/P17-1183 doi: 10.18653/v1/P17-1183

Aharoni, R., Johnson, M., & 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)* (pp. 3874–3884).

Ahmadnia, B., & Dorr, B. J. (2019). Augmenting neural machine translation through round-trip training approach. *Open Computer Science*, *9*(1), 268–278.

Al-Badrashiny, M., & Diab, M. (2016a). Lili: A simple language independent approach for language identification. In *Coling.*

Al-Badrashiny, M., & Diab, M. (2016b, December). LILI: A simple language independent approach for language identification. In *Proceedings of COLING 2016, the 26th international conference on computational linguistics: Technical papers* (pp. 1211–1219). Osaka, Japan: The COLING 2016 Organizing Committee. Retrieved from https://www.aclweb.org/anthology/C16-1115

Alfredo, T., & Alberto, G. R. (1978). Teoría, métodos y técnicas en la investigación social. *México. Ediciones de Cultura Popular*, 11–74.

Anastasopoulos, A., & Chiang, D. (2018, June). Tied multitask learning for neural speech translation. In *Proceedings of the 2018 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long papers)* (pp. 82–91). New Orleans, Louisiana: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/N18-1008 doi: 10.18653/v1/N18-1008

Anderson, S. R., & Anderson, S. (2012). *Languages: A very short introduction* (Vol. 320). Oxford University Press.

Arivazhagan, N., Bapna, A., Firat, O., Aharoni, R., Johnson, M., & Macherey, W. (2019). The missing ingredient in zero-shot neural machine translation. *arXiv preprint arXiv:1903.07091*.

Arivazhagan, N., Bapna, A., Firat, O., Lepikhin, D., Johnson, M., Krikun, M., ... others (2019). Massively multilingual neural machine translation in the wild: Findings and challenges. *arXiv preprint arXiv:1907.05019*.

Arppe, A., Junker, M.-O., & Torkornoo, D. (2017). Converting a comprehensive lexical database into a computational model: The case of East Cree verb inflection. *ComputEL-2*, 52.

Artetxe, M., Labaka, G., & Agirre, E. (2017, July). Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 451–462). Vancouver, Canada: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/P17-1042 doi: 10.18653/v1/P17-1042

Artetxe, M., Labaka, G., & Agirre, E. (2019). An effective approach to unsupervised machine translation. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 194–203).

Artetxe, M., Labaka, G., Agirre, E., & Cho, K. (2018). Unsupervised neural machine translation. In *6th international conference on learning representations, iclr 2018.*

Artetxe, M., Ruder, S., Yogatama, D., Labaka, G., & Agirre, E. (2020). A call for more rigor in unsupervised cross-lingual learning. *arXiv preprint arXiv:2004.14958*.

Artetxe, M., & Schwenk, H. (2019). Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, *7*, 597–610.

Assini, A. A. (2014). *Natural language processing and the mohawk language: creating a finite state morphological parser of mohawk formal nouns.* Scholars' Press.

Ataman, D., & Federico, M. (2018a). Compositional representation of morphologically-rich input for neural machine translation. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 2: Short papers)* (pp. 305–311).

Ataman, D., & Federico, M. (2018b, March). An evaluation of two vocabulary reduction methods for neural machine translation. In *Proceedings of the 13th conference of the association for machine translation in the Americas (volume 1: Research papers)* (pp. 97–110). Boston, MA: Association for Machine Translation in the Americas. Retrieved from https://www.aclweb.org/anthology/W18-1810

Ataman, D., Negri, M., Turchi, M., & Federico, M. (2017). Linguistically motivated vocabulary reduction for neural machine translation from turkish to english. *arXiv preprint arXiv:1707.09879*.

Aulamo, M., Virpioja, S., Tiedemann, J., et al. (2020). Opusfilter: A configurable parallel corpus filtering toolbox. In *58th annual meeting of the association for computational linguistics (acl 2020): System demonstrations system demonstrations.*

Austin, P. K., & Sallabank, J. (2013). *Endangered languages: An introduction* (Vol. 34) (No. 4). Taylor & Francis.

Avelino, H. (2021). Nuevas perspectivas en documentaciÓn lingÜÍstica. *Ciencias Antropologicas*, 93–119.

Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). Layer normalization. *stat*, *1050*, 21.

Bahdanau, D., Cho, K., & Bengio, Y. (2015a). Neural machine translation by jointly learning to align and translate. In Y. Bengio & Y. LeCun (Eds.), *3rd international conference on learning representations, ICLR 2015, san diego, ca, usa, may 7-9, 2015, conference track proceedings.* Retrieved from http://arxiv.org/abs/1409.0473

Bahdanau, D., Cho, K. H., & Bengio, Y. (2015b). Neural machine translation by jointly learning to align and translate. In *3rd international conference on learning representations, iclr 2015.*

Baisa, V. (2009). Problems of machine translation evaluation. *RASLAN 2009 Recent Advances in Slavonic Natural Language Processing*, 121.

Baker, M. (2018). *In other words: A coursebook on translation.* Routledge.

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

Barman, U., Das, A., Wagner, J., & Foster, J. (2014). Code mixing: A challenge for language identification in the language of social media. *EMNLP*.

Barrault, L., Biesialska, M., Bojar, O., Costa-jussà, M. R., Federmann, C., Graham, Y., ... Zampieri, M. (2020, November). Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the fifth conference on machine translation* (pp. 1–55). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.wmt-1.1

Barron, A. R. (1993). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information theory*, *39*(3), 930–945.

Barron-Romero, C., Manuel, M., & Fernando, R. A. (2016). Richard feynman, los alfabetos y los lenguajes.

Baziotis, C., Haddow, B., & Birch, A. (2020). Language model prior for low-resource neural machine translation. *arXiv preprint arXiv:2004.14928*.

Beauclair, N. (2010). Éticas andinas y discursos de reivindicaciones indígenas: asociando tradición y alter-mundialización. *Tinkuy: Boletín de investigación y debate*(12), 9–34.

Belinkov, Y., & Bisk, Y. (2018). Synthetic and natural noise both break neural machine translation. In *International conference on learning representations.*

Bender, E. M. (2011). On achieving and evaluating language-independence in nlp. *Linguistic Issues in Language Technology*, *6*(3), 1–26.

Bentahila, A., & Davies, E. E. (1983). The syntax of arabic-french code-switching. *Lingua*, *59*(4), 301–330.

Bickel, B., & Nichols, J. (2005). Inflectional morphology. In T. Shopen (Ed.), *Language typology and syntactic description.* Cambridge: Cambridge University Press. (2nd edition)

Bird, S. (2020, December). Decolonising speech and language technology. In *Proceedings of the 28th international conference on computational linguistics* (pp. 3504–3519). Barcelona, Spain (Online): International Committee on Computational Linguistics. Retrieved from https://aclanthology.org/2020.coling-main.313 doi: 10.18653/v1/2020.coling-main.313

Blackwood, G., Ballesteros, M., & Ward, T. (2018). Multilingual neural machine translation with task-specific attention. In *Proceedings of the 27th international conference on computational linguistics* (pp. 3112–3122).

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146.

Bollmann, M., Aralikatte, R., Murrieta Bello, H., Hershcovich, D., de Lhoneux, M., & Søgaard, A. (2021, June). Moses and the character-based random babbling baseline: CoAStaL at AmericasNLP 2021 shared task. In *Proceedings of the first workshop on natural language processing for indigenous languages of the americas* (pp. 248–254). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.americasnlp-1.28 doi: 10.18653/v1/2021.americasnlp-1.28

Botha, J. A., & Blunsom, P. (2013). Adaptor grammars for learning non- concatenative morphology..

Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In *Compstat.* Springer.

Bowers, D., Arppe, A., Lachler, J., Moshagen, S., & Trosterud, T. (2017). A morphological parser for Odawa. *ComputEL-2*, 1.

Brambila, D. (1976). *Diccionario rarámuri-castellano (tarahumar).* Obra Nacional de la buena Prensa.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... Amodei, D. (2020). Language models are few-shot learners.

Bustamante, G., Oncevay, A., & Zariquiey, R. (2020, May). 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* (pp. 2914–2923). Marseille, France: European Language Resources Association. Retrieved from https://aclanthology.org/2020.lrec-1.356

Caballero, G. (2008). Choguita rarámuri (tarahumara) phonology and morphology.

Caballero, G. (2010). Scope, phonology and morphology in an agglutinating language: Choguita rarámuri (tarahumara) variable suffix ordering. *Morphology*, *20*(1), 165–204.

Campbell, L. (2000). *American indian languages: the historical linguistics of native america* (Vol. 4). Oxford University Press on Demand.

Can, B., & Manandhar, S. (2018). Tree structured dirichlet processes for hierarchical morphological segmentation. *Computational Linguistics*, *44*(2), 349–374. Retrieved from http://aclweb.org/anthology/J18-2005 doi: 10.1162/COLI_a_00318

Canger, U. (2001). *Mexicanero de la sierra madre occidental.* El Colegio de México.

Carey, H. M. (2010). Lancelot threlkeld, biraban, and the colonial bible in australia. *Comparative Studies in Society and History*, *52*(2), 447–478.

Caruana, R. (1997). Multitask learning. *Machine learning*, *28*(1), 41–75.

Caswell, I., Chelba, C., & Grangier, D. (2019). Tagged back-translation. In *Proceedings of the fourth conference on machine translation (volume 1: Research papers)* (pp. 53–63).

Çetinoğlu, O. (2016). A Turkish-German code-switching corpus. In *Lrec.*

Çetinoğlu, Ö., & Çöltekin, Ç. (2016). Part of speech annotation of a Turkish-German code-switching corpus. In *Law-x.* Berlin, Germany.

Celikyilmaz, A., Clark, E., & Gao, J. (2020). Evaluation of text generation: A survey. *arXiv preprint arXiv:2006.14799* .

Çetinoğlu, Ö. (2017, April). A code-switching corpus of Turkish-German conversations. In *Proceedings of the 11th linguistic annotation workshop* (pp. 34–40). Valencia, Spain: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/W17-0804 doi: 10.18653/v1/W17-0804

Çetinoğlu, Ö., Schulz, S., & Vu, N. T. (2016, November). Challenges of computational processing of code-switching. In *Proceedings of the second workshop on computational approaches to code switching* (pp. 1–11). Austin, Texas: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/W16-5801 doi: 10.18653/v1/W16-5801

Chakravarthi, B. R., Priyadharshini, R., Banerjee, S., Saldanha, R., McCrae, J. P., M, A. K., ... Johnson, M. (2021, April). Findings of the shared task on machine translation in Dravidian languages. In *Proceedings of the first workshop on speech and language technologies for dravidian languages* (pp. 119–125). Kyiv: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/2021.dravidianlangtech-1.15

Chatzikoumi, E. (2020). How to evaluate machine translation: A review of automated and human metrics. *Natural Language Engineering*, *26*(2), 137–161.

Chen, G., Ma, S., Chen, Y., Dong, L., Zhang, D., Pan, J., ... Wei, F. (2021, November). Zero-shot cross-lingual transfer of neural machine translation with multilingual pretrained encoders. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 15–26). Online and Punta Cana, Dominican Republic: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.emnlp-main.2

Cheng, Y. (2019). Joint training for pivot-based neural machine translation. In *Joint training for neural machine translation* (pp. 41–54). Springer.

Cheng, Y., Jiang, L., & Macherey, W. (2019). Robust neural machine translation with doubly adversarial inputs. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 4324–4333).

Cheng, Y., Jiang, L., Macherey, W., & Eisenstein, J. (2020, July). AdvAug: Robust adversarial augmentation for neural machine translation. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 5961–5970). Online: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/2020.acl-main.529 doi: 10.18653/v1/2020.acl-main.529

Cheng, Y., Tu, Z., Meng, F., Zhai, J., & Liu, Y. (2018). Towards robust neural machine translation. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 1756–1766).

Chesterman, A. (2001). Proposal for a hieronymic oath. *The translator*, *7*(2), 139–154.

Cheyfitz, E. (1997). *The poetics of imperialism: Translation and colonization from the tempest to tarzan*. University of Pennsylvania Press.

Chiruzzo, L., Amarilla, P., Ríos, A., & Giménez Lugo, G. (2020, May). Development of a Guarani - Spanish parallel corpus. In *Proceedings of the 12th language resources and evaluation conference* (pp. 2629–2633). Marseille, France: European Language Resources Association. Retrieved from https://aclanthology.org/2020.lrec-1.320

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & 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)* (pp. 1724–1734).

Christodouloupoulos, C., & Steedman, M. (2015). A massively parallel corpus: the bible in 100 languages. *Language resources and evaluation*, *49*(2), 375–395.

Cohn, T., & Lapata, M. (2007). Machine translation by triangulation: Making effective use of multi-parallel corpora. In *Proceedings of the 45th annual meeting of the association of computational linguistics* (pp. 728–735).

Conforti, C., Huck, M., & Fraser, A. (2018). Neural morphological tagging of lemma sequences for machine translation. In *Amta* (Vol. 1, pp. 39–53).

Conneau, A., & Lample, G. (2019). Cross-lingual language model pretraining. In *Advances in neural information processing systems* (pp. 7057–7067).

Conneau, A., Lample, G., Ranzato, M., Denoyer, L., & Jégou, H. (2017). Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., ... others (2022). No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Cotterell, R., Kirov, C., Sylak-Glassman, J., Walther, G., Vylomova, E., Xia, P., ... others (2017). CoNLL-SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection in 52 Languages. *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*, 1–30.

Cotterell, R., Kirov, C., Sylak-Glassman, J., Yarowsky, D., Eisner, J., & Hulden, M. (2016). The SIGMORPHON 2016 shared task—morphological reinflection. In *Proceedings of the 14th sigmorphon workshop on computational research in phonetics, phonology, and morphology* (pp. 10–22).

Cotterell, R., Müller, T., Fraser, A., & Schütze, H. (2015a). Labeled morphological segmentation with semi-markov models. In *Proceedings of the nineteenth conference on computational natural language learning* (pp. 164–174).

Cotterell, R., Müller, T., Fraser, A., & Schütze, H. (2015b, July). Labeled morphological segmentation with semi-Markov models. In *Proceedings of the nineteenth conference on computational natural language learning* (pp. 164–174). Beijing, China: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/K15-1017 doi: 10.18653/v1/K15-1017

Cotterell, R., Peng, N., & Eisner, J. (2014). Stochastic contextual edit distance and probabilistic fsts. In *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 2: Short papers)* (pp. 625–630).

Cotterell, R., Schütze, H., & Eisner, J. (2016, August). Morphological smoothing and extrapolation of word embeddings. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 1651–1660). Berlin, Germany: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/P16-1156 doi: 10.18653/v1/P16-1156

Cotterell, R., Vieira, T., & Schütze, H. (2016). A joint model of orthography and morphological segmentation. In *Naacl-hlt* (pp. 664–669).

Couillault, A., Fort, K., Adda, G., & de Mazancourt, H. (2014, May). Evaluating corpora documentation with regards to the ethics and big data charter. In *Proceedings of the ninth international conference on language resources and evaluation (LREC-2014)* (pp. 4225–4229). Reykjavik, Iceland: European Languages Resources Association (ELRA). Retrieved from http://www.lrec-conf.org/proceedings/lrec2014/pdf/424_Paper.pdf

Creutz, M., & Lagus, K. (2002). Unsupervised discovery of morphemes. In *Conll–sigmorphon*. Retrieved from http://aclweb.org/anthology/W02-0603

Creutz, M., & Lagus, K. (2007). Unsupervised models for morpheme segmentation and morphology learning. *ACM TSLP*, *4*(1), 3.

Cruz, E. (2011). *Phonology, tone and the functions of tone in san juan quiahije chatino* (Unpublished doctoral dissertation).

Crystal, D. (1987). *The cambridge encyclopedia of language* (Vol. 2). Cambridge University Press Cambridge.

Czaykowska-Higgins, E. (2009). Research models, community engagement, and linguistic fieldwork: Reflections on working within canadian indigenous communities.

Dabre, R., Chu, C., & Kunchukuttan, A. (2019). A survey of multilingual neural machine translation. *arXiv preprint arXiv:1905.05395*.

Daes, E.-I. A. (1993). *Discrimination against indigenous peoples: Study on the protection of the cultural and intellectual property of indigenous peoples*. United Nations, Economic and Social Council.

Daumé, H., Langford, J., & Marcu, D. (2009). Search-based structured prediction. *Machine learning*, *75*(3), 297–325.

DelValls, T. (1978). El instituto lingüístico de verano, instrumento del imperialismo. *Nueva Antropología*, *3*(9), 117–142.

Denisov, P., Mager, M., & Vu, N. T. (2021, August). IMS' systems for the IWSLT 2021 low-resource speech translation task. In *Proceedings of the 18th international conference on spoken language translation (iwslt 2021)* (pp. 175–181). Bangkok, Thailand (online): Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.iwslt-1.21 doi: 10.18653/v1/2021.iwslt-1.21

Denkowski, M., & Lavie, A. (2011). Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the sixth workshop on statistical machine translation* (pp. 85–91).

Devlin, J., Chang, M.-W., Lee, K., & 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)* (pp. 4171–4186).

Dobrin, L. M. (2009). Sil international and the disciplinary culture of linguistics: Introduction. *Language*, *85*(3), 618–619.

Doğruöz, A. S., Sitaram, S., Bullock, B., & Toribio, A. J. (2021). A survey of code-switching: Linguistic and social perspectives for language technologies. 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)* (pp. 1654–1666).

Doherty, S. (2016). Translations| the impact of translation technologies on the process and product of translation. *International journal of communication*, *10*, 23.

Dorantes, A., Sierra, G., Pérez, T. Y. D., Bel-Enguix, G., & Rosales, M. J. (2018). Sociolinguistic corpus of whatsapp chats in spanish among college students. In *Proceedings of the sixth international workshop on natural language processing for social media* (pp. 1–6).

Dryer, M. S., & Haspelmath, M. (Eds.). (2013). *Wals online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. Retrieved from https://wals.info/

Duan, X., Ji, B., Jia, H., Tan, M., Zhang, M., Chen, B., ... Zhang, Y. (2020). Bilingual dictionary based neural machine translation without using parallel sentences. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 1570–1579).

Dwyer, A. M. (2006). Ethics and practicalities of cooperative fieldwork and analysis. *Essentials of language documentation*, *178*.

Ebrahimi, A., Mager, M., Oncevay, A., Chaudhary, V., Chiruzzo, L., Fan, A., ... Kann, K. (2022, May). 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)* (pp. 6279–6299). Dublin, Ireland: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2022.acl-long.435 doi: 10.18653/v1/2022.acl-long.435

Ebrahimi, A., Mager, M., Rijhwani, S., Rice, E., Oncevay, A., Baltazar, C., ... Kann, K. (2023, July). Findings of the AmericasNLP 2023 shared task on machine translation into indigenous languages. In *Proceedings of the workshop on natural language processing for indigenous languages of the americas (americasnlp)* (pp. 206–219). Toronto, Canada: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2023.americasnlp-1.23 doi: 10.18653/v1/2023.americasnlp-1.23

Ebrahimi, J., Lowd, D., & Dou, D. (2018, August). On adversarial examples for character-level neural machine translation. In *Proceedings of the 27th international conference on computational linguistics* (pp. 653–663). Santa Fe, New Mexico, USA: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/C18-1055

Edunov, S., Baevski, A., & Auli, M. (2019). Pre-trained language model representations for language generation. 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)* (pp. 4052–4059).

Edunov, S., Ott, M., Auli, M., & Grangier, D. (2018). Understanding back-translation at scale. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 489–500).

Elman, J. L. (1990). Finding structure in time. *Cognitive science*, *14*(2), 179–211.

Enríquez, P. (2019). El rol de la lengua kichwa en la construcción de la identidad en la población indígena de cañar. *Lenguas en contacto: desafíos en la diversidad*, 267.

Errington, J. (2001). Colonial linguistics. *Annual review of anthropology*, *30*(1), 19–39.

Eskander, R., Callejas, F., Nichols, E., Klavans, J., & Muresan, S. (2020, May). MorphAGram, evaluation and framework for unsupervised morphological segmentation.

In *Proceedings of the twelfth language resources and evaluation conference* (pp. 7112–7122). Marseille, France: European Language Resources Association. Retrieved from https://aclanthology.org/2020.lrec-1.879

Eskander, R., Rambow, O., & Yang, T. (2016). Extending the use of adaptor grammars for unsupervised morphological segmentation of unseen languages. In *Proceedings of coling 2016, the 26th international conference on computational linguistics: Technical papers* (pp. 900–910).

Fadaee, M., Bisazza, A., & Monz, C. (2017). Data augmentation for low-resource neural machine translation. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 2: Short papers)* (pp. 567–573).

Faust, N. (1973). *Lecciones para el aprendizaje del idioma shipibo-conibo* (Vol. 1). Yarinacocha: Instituto Lingüístico de Verano.

Feldman, I., & Coto-Solano, R. (2020, December). Neural machine translation models with back-translation for the extremely low-resource indigenous language Bribri. In *Proceedings of the 28th international conference on computational linguistics* (pp. 3965–3976). Barcelona, Spain (Online): International Committee on Computational Linguistics. Retrieved from https://aclanthology.org/2020.coling-main.351 doi: 10.18653/v1/2020.coling-main.351

Feng, X., Feng, Z., Zhao, W., Qin, B., & Liu, T. (2020). Enhanced neural machine translation by joint decoding with word and pos-tagging sequences. *Mobile Networks and Applications*, 1–7.

Feng, X., Huang, L., Tang, D., Ji, H., Qin, B., & Liu, T. (2016, August). A language-independent neural network for event detection. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers)* (pp. 66–71). Berlin, Germany: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/P16-2011 doi: 10.18653/v1/P16-2011

Fernández, J. L. S. (2022). La libertad religiosa en el pueblo huichol religious freedom in the huichol community. *Revista de Garantismo y Derechos Humanos*, 43.

Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *The journal of machine learning research*, *15*(1), 3133–3181.

Fort, K., Adda, G., Sagot, B., Mariani, J., & Couillault, A. (2011). Crowdsourcing for language resource development: criticisms about amazon mechanical turk overpowering use. In *Language and technology conference* (pp. 303–314).

Fraser, A. (2020, November). Findings of the WMT 2020 shared tasks in unsupervised MT and very low resource supervised MT. In *Proceedings of the fifth conference on machine translation* (pp. 765–771). Online: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/2020.wmt-1.80

Freeze, R. A. (1989). *Mayo de Los Capomos, Sinaloa*. El Colegio de México.

Furcy, D., & Koenig, S. (2005). Limited discrepancy beam search. In *Ijcai*.

Galarreta, A.-P., Melgar, A., & Oncevay, A. (2017, September). 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* (pp. 238–244). Varna, Bulgaria: INCOMA Ltd. Retrieved from https://doi.org/10.26615/978-954-452-049-6_033 doi: 10.26615/978-954-452-049-6_033

Garcia, X., Foret, P., Sellam, T., & Parikh, A. P. (2020). A multilingual view of unsupervised machine translation. *arXiv preprint arXiv:2002.02955*.

Gardner-Chloros, P., & Edwards, M. (2004). Assumptions behind grammatical approaches to code-switching: when the blueprint is a red herring. *Transactions of the Philological Society*, *102*(1), 103–129.

Gardner-Chloros, P., et al. (2009). *Code-switching*. Cambridge university press.

Gaser, M. (2022). *Subword-level segmentation for neural machine translation of code-switched dialectal egyptian arabic - english text* (Master's thesis). doi: 10.13140/RG.2.2.10302.77123

Gaser, M., Mager, M., Hamed, I., Habash, N., Abdennadher, S., & Vu, N. T. (2023, May). Exploring segmentation approaches for neural machine translation of code-switched egyptian arabic-english text. In *Proceedings of eacl*. Berlin, Germany: Association for Computational Linguistics.

Gehring, J., Auli, M., Grangier, D., & Dauphin, Y. (2017, July). A convolutional encoder model for neural machine translation. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp.

123–135). Vancouver, Canada: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/P17-1012 doi: 10.18653/v1/P17-1012

Gers, F. A., Schmidhuber, J., & Cummins, F. (2000). Learning to forget: Continual prediction with lstm. *Neural computation*, *12*(10), 2451–2471.

Gerz, D., Vulić, I., Ponti, E. M., Reichart, R., & Korhonen, A. (2018). On the relation between linguistic typology and (limitations of) multilingual language modeling. In *Emnlp* (pp. 316–327).

Gheini, M., Ren, X., & May, J. (2021, November). Cross-attention is all you need: Adapting pretrained Transformers for machine translation. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 1754–1765). Online and Punta Cana, Dominican Republic: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.emnlp-main.132

Gibadullin, I., Valeev, A., Khusainova, A., & Khan, A. (2019). A survey of methods to leverage monolingual data in low-resource neural machine translation. *arXiv preprint arXiv:1910.00373*.

Gilmour, R. (2007). Missionaries, colonialism and language in nineteenth-century south africa. *History Compass*, *5*(6), 1761–1777.

Gobierno, F. (2022). Constitución política de los estados unidos mexicanos. *Diario Oficial de la Federación*, *10*.

Goddard, I. (1996). Introduccion. In W. C. Sturtevant (Ed.), *Handbook of north american indians (vol. 17)* (p. 1-6). University of Texas.

Goldsmith, J. (2001). Unsupervised learning of the morphology of a natural language. *Computational Linguistics*, *27*(2), 153–198.

Gómez, P., & López, P. G. (1999). *Huichol de San Andrés Cohamiata, Jalisco.* El Colegio de México.

Góngora, S., Giossa, N., & Chiruzzo, L. (2021, June). Experiments on a Guarani corpus of news and social media. In *Proceedings of the first workshop on natural language processing for indigenous languages of the americas* (pp. 153–158). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.americasnlp-1.16 doi: 10.18653/v1/2021.americasnlp-1.16

Góngora, S., Giossa, N., & Chiruzzo, L. (2022, May). Can we use word embeddings for enhancing Guarani-Spanish machine translation? In *Proceedings of the fifth workshop on the use of computational methods in the study of endangered languages* (pp. 127–132). Dublin, Ireland: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2022.computel-1.16 doi: 10.18653/v1/2022.computel-1.16

Gordon Jr, R. G. (2005). Ethnologue, languages of the world. *http://www. ethnologue. com/*.

Graham, Y., Haddow, B., & Koehn, P. (2020, November). Statistical power and translationese in machine translation evaluation. In *Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp)* (pp. 72–81). Online: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/2020.emnlp-main.6

Graves, A., Fernández, S., & Schmidhuber, J. (2005). Bidirectional lstm networks for improved phoneme classification and recognition. In *International conference on artificial neural networks* (pp. 799–804).

Greenberg, J. H. (1963). Universals of language.

Grönroos, S.-A., Virpioja, S., & Kurimo, M. (2019). North sámi morphological segmentation with low-resource semi-supervised sequence labeling. In *International workshop on computational linguistics for uralic languages* (pp. 15–26).

Grönroos, S.-A., Virpioja, S., Smit, P., & Kurimo, M. (2014). Morfessor flatcat: An hmm-based method for unsupervised and semi-supervised learning of morphology. In *Coling* (pp. 1177–1185).

Grosjean, F. (2010). *Bilingual: Life and reality*. Harvard University Press.

Gu, J., Bradbury, J., Xiong, C., Li, V. O., & Socher, R. (2018). Non-autoregressive neural machine translation. In *Iclr*.

Gu, J., Wang, Y., Cho, K., & Li, V. O. (2019). Improved zero-shot neural machine translation via ignoring spurious correlations. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 1258–1268).

Gulcehre, C., Firat, O., Xu, K., Cho, K., Barrault, L., Lin, H.-C., ... Bengio, Y. (2015). On using monolingual corpora in neural machine translation. *arXiv preprint arXiv:1503.03535*.

Guo, J., Xu, L., & Chen, E. (2020). Jointly masked sequence-to-sequence model for non-autoregressive neural machine translation. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 376–385).

Gutierrez-Vasques, X. (2017). Exploring bilingual lexicon extraction for Spanish-Nahuatl. In *Acl workshop in women and underrepresenting minorities in natural language processing.*

Gutierrez-Vasques, X., Sierra, G., & Pompa, I. H. (2016, May). Axolotl: a web accessible parallel corpus for Spanish-Nahuatl. In *Proceedings of the tenth international conference on language resources and evaluation (LREC 2016)* (pp. 4210–4214). Portorož, Slovenia: European Language Resources Association (ELRA). Retrieved from https://www.aclweb.org/anthology/L16-1666

Guzmán, F., Chen, P.-J., Ott, M., Pino, J., Lample, G., Koehn, P., ... 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)* (pp. 6100–6113).

Haddow, B., Bawden, R., Miceli Barone, A. V., Helcl, J., & Birch, A. (2022). Survey of low-resource machine translation. *Computational Linguistics*, 1–67.

Hale, K. (1992). Endangered languages: On endangered languages and the safeguarding of diversity. *language*, *68*(1), 1–42.

Hämäläinen, M. (2021). Endangered languages are not low-resourced! *arXiv preprint arXiv:2103.09567*.

Hamed, I., Habash, N., Abdennadher, S., & Vu, N. T. (2022). Investigating lexical replacements for arabic-english code-switched data augmentation. *arXiv preprint arXiv:2205.12649*.

Hamed, I., Vu, N. T., & Abdennadher, S. (2020, May). ArzEn: A speech corpus for code-switched Egyptian Arabic-English. In *Proceedings of the 12th language resources*

and evaluation conference (pp. 4237–4246). Marseille, France: European Language Resources Association. Retrieved from https://aclanthology.org/2020.lrec-1.523

Han, L. (2016). Machine translation evaluation resources and methods: A survey. *arXiv preprint arXiv:1605.04515*.

Harrigan, A. G., Schmirler, K., Arppe, A., Antonsen, L., Trosterud, T., & Wolvengrey, A. (2017). Learning from the computational modelling of Plains Cree verbs. *Morphology*, *27*(4), 565–598.

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

Hart, P. E., Nilsson, N. J., & Raphael, B. (1968). A formal basis for the heuristic determination of minimum cost paths. *IEEE transactions on Systems Science and Cybernetics*, *4*(2), 100–107.

Hawkins, J. A. (2015). *A comparative typology of english and german: Unifying the contrasts*. Routledge.

Hernandez, F., & Nguyen, V. (2020, November). The ubiqus English-Inuktitut system for WMT20. In *Proceedings of the fifth conference on machine translation* (pp. 213–217). Online: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/2020.wmt-1.21

Hershcovich, D., Frank, S., Lent, H., de Lhoneux, M., Abdou, M., Brandl, S., ... others (2022). Challenges and strategies in cross-cultural nlp. *arXiv preprint arXiv:2203.10020*.

Hidayatullah, A. F., Qazi, A., Lai, D. T. C., & Apong, R. A. (2022). A systematic review on language identification of code-mixed text: techniques, data availability, challenges, and framework development. *IEEE access*.

Hill, K. C. (2002). On publishing the hopi dictionary. *Making dictionaries: Preserving Indigenous languages of the Americas*, 299–311.

Himmelmann, N. P. (2008). Language documentation: What is it and what is it good for? In *Essentials of language documentation* (pp. 1–30). De Gruyter Mouton.

Hiroki Nomoto, D. M., Hannah Choi, & Bond, F. (2018, may). Malindo morph: Morphological dictionary and analyser for malay/indonesian. In K. Shirai (Ed.), *Proceedings of the eleventh international conference on language resources and evaluation (lrec 2018).* Paris, France: ELRA.

Hoang, V. C. D., Koehn, P., Haffari, G., & Cohn, T. (2018). Iterative back-translation for neural machine translation. In *Proceedings of the 2nd workshop on neural machine translation and generation* (pp. 18–24).

Hochreiter, S. (1991). Untersuchungen zu dynamischen neuronalen netzen. *Diploma, Technische Universität München*, *91*(1).

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

Horváth, C., Szilágyi, N., Vincze, V., & Nagy, Á. (2017). Language technology resources and tools for mansi: an overview. In *Proceedings of the third workshop on computational linguistics for uralic languages* (pp. 56–65).

Hu, J. E., Khayrallah, H., Culkin, R., Xia, P., Chen, T., Post, M., & Van Durme, B. (2019). Improved lexically constrained decoding for translation and monolingual rewriting. 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)* (pp. 839–850).

Hu, Z., Tan, B., Salakhutdinov, R. R., Mitchell, T. M., & Xing, E. P. (2019). Learning data manipulation for augmentation and weighting. In *Advances in neural information processing systems* (pp. 15738–15749).

Hulden, M. (2009). Foma: a finite-state compiler and library. In *Proceedings of the 12th conference of the european chapter of the association for computational linguistics: Demonstrations session* (pp. 29–32).

Ide, N. (2017). Introduction: The handbook of linguistic annotation. In *Handbook of linguistic annotation* (pp. 1–18). Springer.

Ilo, I. (1989). C169-indigenous and tribal peoples convention. In *Convention concerning indigenous and tribal peoples in independent countries. geneva: Ilo.*

INALI. (2017). *Etnografía del pueblo tarahumara (rarámuri).*

INALI. (2022). *CatÁlogo de las lenguas indÍgenas nacionales.* https://www.inali.gob.mx/clin-inali/. (Accessed: 2022-10-10)

INEGI. (2008a). Catálogo de las lenguas indígenas nacionales: Variantes lingüísticas de méxico con sus autodenominaciones y referencias geoestadísticas. *Diario Oficial*, 31–108.

INEGI. (2008b). *Programa de revitalización, fortalecimiento y desarrollo de las lenguas indígenas nacionales, 2008-2012.* Instituto Nacional de Lenguas Indígenas.

INEGI. (2020). *Censo de población y vivienda.*

Jean, S., Cho, K., Memisevic, R., & Bengio, Y. (2015, July). On using very large target vocabulary for neural machine translation. In *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 1: Long papers)* (pp. 1–10). Beijing, China: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/P15-1001 doi: 10.3115/v1/P15-1001

Jiménez, A. T., & Ramos, A. G. (1985). *Teoría, métodos y técnicas en la investigación social.* Eds. Taller Abierto.

Jin, D., Jin, Z., Zhou, J. T., & Szolovits, P. (2020). Unsupervised domain adaptation for neural machine translation with iterative back translation. *arXiv preprint arXiv:2001.08140*.

Joanis, E., Knowles, R., Kuhn, R., Larkin, S., Littell, P., Lo, C.-k., ... Micher, J. (2020, May). The Nunavut hansard Inuktitut–English parallel corpus 3.0 with preliminary machine translation results. In *Proceedings of the 12th language resources and evaluation conference* (pp. 2562–2572). Marseille, France: European Language Resources Association. Retrieved from https://www.aclweb.org/anthology/2020.lrec-1.312

Johan, K., Eddy, R., & Zwarts, J. (2001). *Lexicon of linguistics.* Utrecht institute of Linguistics OTS, Utrecht University.

Johnson, M. (2008). Using adaptor grammars to identify synergies in the unsupervised acquisition of linguistic structure. In *Proceedings of acl-08: Hlt* (pp. 398–406).

Johnson, M., Griffiths, T., & Goldwater, S. (2006). Adaptor grammars: A framework for specifying compositional nonparametric bayesian models. *Advances in neural information processing systems*, *19*.

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., ... others (2017). Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, *5*, 339–351.

Jordan, M. I. (1986). Serial order: A parallel distributed processing approach, ics report 8604. *Institute for Cognitive Science, UCSD, La Jolla*.

Jose, N., Chakravarthi, B. R., Suryawanshi, S., Sherly, E., & McCrae, J. P. (2020). A survey of current datasets for code-switching research. In *2020 6th international conference on advanced computing and communication systems (icaccs)* (pp. 136–141).

Junczys-Dowmunt, M. (2019). Microsoft translator at wmt 2019: Towards large-scale document-level neural machine translation. *arXiv preprint arXiv:1907.06170*.

Kamholz, D., Pool, J., & Colowick, S. M. (2014). Panlex: Building a resource for panlingual lexical translation. In *Lrec* (pp. 3145–3150).

Kann, K., Cotterell, R., & Schütze, H. (2016a). Neural morphological analysis: Encoding-decoding canonical segments. In *Emnlp* (pp. 961–967).

Kann, K., Cotterell, R., & Schütze, H. (2016b, November). Neural morphological analysis: Encoding-decoding canonical segments. In *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 961–967). Austin, Texas: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/D16-1097 doi: 10.18653/v1/D16-1097

Kann, K., Mager Hois, J. M., Meza-Ruiz, I. V., & Schütze, H. (2018, June). 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)* (pp. 47–57). New Orleans, Louisiana: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/N18-1005 doi: 10.18653/v1/N18-1005

Kawaguchi, K., Kaelbling, L. P., & Bengio, Y. (2017). Generalization in deep learning. *arXiv preprint arXiv:1710.05468*.

Kay, M. (1977). Morphological and syntactic analysis. *Linguistic structures processing*, *5*, 131.

Khayrallah, H., Thompson, B., Post, M., & Koehn, P. (2020). Simulated multiple reference training improves low-resource machine translation. *arXiv preprint arXiv:2004.14524*.

Kim, Y., Graça, M., & Ney, H. (2020, November). When and why is unsupervised neural machine translation useless? In *Proceedings of the 22nd annual conference of the european association for machine translation* (pp. 35–44). Lisboa, Portugal: European Association for Machine Translation. Retrieved from https://www.aclweb.org/anthology/2020.eamt-1.5

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv:1412.6980*.

Klavans, J. L. (2018, August). Computational challenges for polysynthetic languages. In *Proceedings of the workshop on computational modeling of polysynthetic languages* (pp. 1–11). Santa Fe, New Mexico, USA: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/W18-4801

Klein, G., Kim, Y., Deng, Y., Senellart, J., & Rush, A. M. (2017). Opennmt: Open-source toolkit for neural machine translation. In *Proceedings of acl 2017, system demonstrations* (pp. 67–72).

Knowles, R., Stewart, D., Larkin, S., & Littell, P. (2020, November). NRC systems for the 2020 Inuktitut-English news translation task. In *Proceedings of the fifth conference on machine translation* (pp. 156–170). Online: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/2020.wmt-1.13

Knowles, R., Stewart, D., Larkin, S., & Littell, P. (2021, June). NRC-CNRC machine translation systems for the 2021 AmericasNLP shared task. In *Proceedings of the first workshop on natural language processing for indigenous languages of the americas* (pp. 224–233). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.americasnlp-1.25 doi: 10.18653/v1/2021.americasnlp-1.25

Kobayashi, S. (2018). Contextual augmentation: Data augmentation by words with paradigmatic relations. In *Proceedings of the 2018 conference of the north american chapter of the association for computational linguistics: Human language technologies, volume 2 (short papers)* (pp. 452–457).

Kocmi, T. (2020, November). CUNI submission for the Inuktitut language in WMT news 2020. In *Proceedings of the fifth conference on machine translation* (pp. 171–174). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.wmt-1.14

Koehn, P., et al. (2005). Europarl: A parallel corpus for statistical machine translation. In *Mt summit* (Vol. 5, pp. 79–86).

Kohonen, O., Virpioja, S., & Lagus, K. (2010). Semi-supervised learning of concatenative morphology. *ACL*, 78.

Kong, L., Dyer, C., & Smith, N. A. (2016). Segmental recurrent neural networks. *Interspeech*.

Konig, E., & Van der Auwera, J. (2013). *The germanic languages*. Routledge.

Krubiński, M., Chochowski, M., Boczek, B., Koszowski, M., Dobrowolski, A., Szymański, M., & Przybysz, P. (2020, November). Samsung R&D institute Poland submission to WMT20 news translation task. In *Proceedings of the fifth conference on machine translation* (pp. 181–190). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.wmt-1.16

Kurimo, M., Creutz, M., Varjokallio, M., Arisoy, E., & Saraçlar, M. (2006). Unsupervised segmentation of words into morphemes–challenge 2005: An introduction and evaluation report. In *Proceedings of the pascal challenge workshop on unsupervised segmentation of words into morphemes* (pp. 1–11).

Kurimo, M., Virpioja, S., Turunen, V., & Lagus, K. (2010, July). Morpho challenge 2005-2010: Evaluations and results. In *Proceedings of the 11th meeting of the ACL special interest group on computational morphology and phonology* (pp. 87–95). Uppsala, Sweden: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/W10-2211

Kusampudi, S. S. V., Chaluvadi, A., & Mamidi, R. (2021). Corpus creation and language identification in low-resource code-mixed telugu-english text. In *Proceedings of the international conference on recent advances in natural language processing (ranlp 2021)* (pp. 744–752).

Lafferty, J., McCallum, A., & Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Icml*.

Lakew, S. M., Lotito, Q. F., Negri, M., Turchi, M., & Federico, M. (2018). Improving zero-shot translation of low-resource languages. In *Proceedings of the 14h iwslt* (pp. 113–119).

Lample, G., Conneau, A., Denoyer, L., & Ranzato, M. (2017). Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*.

Laskar, S. R., Khilji, A. F. U. R., Pakray, P., & Bandyopadhyay, S. (2020, December). Zero-shot neural machine translation: Russian-Hindi @LoResMT 2020. In *Proceedings of the 3rd workshop on technologies for mt of low resource languages* (pp. 38–42). Suzhou, China: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/2020.loresmt-1.5

Lastra de Suárez, Y. (1980). *Náhuatl de Acaxochitlán (Hidalgo)*. Colegio de México.

Leidner, J. L., & Plachouras, V. (2017, April). Ethical by design: Ethics best practices for natural language processing. In *Proceedings of the first ACL workshop on ethics in natural language processing* (pp. 30–40). Valencia, Spain: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/W17-1604 doi: 10.18653/v1/W17-1604

Leng, Y., Tan, X., Qin, T., Li, X.-Y., & Liu, T.-Y. (2019). Unsupervised pivot translation for distant languages. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 175–183).

Leonard, W. (2020). Producing language reclamation by decolonising'language'.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady* (Vol. 10, pp. 707–710).

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Lewis, M. P. (2009). *Ethnologue: Languages of the world*. SIL international.

Leza, J. L. I., & López, P. G. (2006). *Gramática wixarika* (Vol. 1). Lincom Europa.

Li, Z., Wang, R., Chen, K., Utiyama, M., Sumita, E., Zhang, Z., & Zhao, H. (2020). Data-dependent gaussian prior objective for language generation. In *International conference on learning representations*.

Lin, Z., Pan, X., Wang, M., Qiu, X., Feng, J., Zhou, H., & Li, L. (2020, November). Pre-training multilingual neural machine translation by leveraging alignment information. In *Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp)* (pp. 2649–2663). Online: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/2020.emnlp-main.210 doi: 10.18653/v1/2020.emnlp-main.210

Lindén, K., Axelson, E., Hardwick, S., Pirinen, T. A., & Silfverberg, M. (2011). Hfst—framework for compiling and applying morphologies. In *International workshop on systems and frameworks for computational morphology* (pp. 67–85).

Ling, W., Trancoso, I., Dyer, C., & Black, A. W. (2015). Character-based neural machine translation. *arXiv preprint arXiv:1511.04586*.

Littell, P., Kazantseva, A., Kuhn, R., Pine, A., Arppe, A., Cox, C., & Junker, M.-O. (2018, August). Indigenous language technologies in canada: Assessment, challenges, and successes. In *Proceedings of the 27th international conference on computational linguistics* (pp. 2620–2632). Santa Fe, New Mexico, USA: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/C18-1222

Liu, C., Dominé, L., Chavez, K., & Socher, R. (2020). Central yup'ik and machine translation of low-resource polysynthetic languages. *arXiv preprint arXiv:2009.04087*.

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., ... Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, *8*, 726–742.

Liu, Z., Richardson, C., Hatcher Jr, R., & Prud'hommeaux, E. (2022). Not always about you: Prioritizing community needs when developing endangered language technology. *arXiv preprint arXiv:2204.05541*.

Liu, Z., Xu, Y., Winata, G. I., & Fung, P. (2019). Incorporating word and subword units in unsupervised machine translation using language model rescoring. In *Proceedings of the fourth conference on machine translation (volume 2: Shared task papers, day 1)* (pp. 275–282).

Lo, C.-k. (2020, November). Extended study on using pretrained language models and YiSi-1 for machine translation evaluation. In *Proceedings of the fifth conference on machine translation* (pp. 895–902). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.wmt-1.99

Lörscher, W. (2005). The translation process: Methods and problems of its investigation. *Meta: journal des traducteurs/Meta: Translators' Journal*, *50*(2), 597–608.

Lowerre, B. T. (1976). *The harpy speech recognition system* (Tech. Rep.). CARNEGIE-MELLON UNIV PITTSBURGH PA DEPT OF COMPUTER SCIENCE.

Lu, Y., Keung, P., Ladhak, F., Bhardwaj, V., Zhang, S., & Sun, J. (2018). A neural interlingua for multilingual machine translation. In *Proceedings of the third conference on machine translation: Research papers* (pp. 84–92).

Ludescher, M. (2001). Instituciones y prácticas coloniales en la amazonía peruana: pasado y presente. *Indiana*, 313–359.

Luong, T., Sutskever, I., Le, Q., Vinyals, O., & Zaremba, W. (2015, July). Addressing the rare word problem in neural machine translation. In *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 1: Long papers)* (pp. 11–19). Beijing, China: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/P15-1002 doi: 10.3115/v1/P15-1002

MacKay, C. J., & Trechsel, F. R. (2010). *Tepehua de pisaflores, veracruz.* El Colegio de México, Centro de Estudios Lingüísticos y Literarios.

Mackey, B., Johnson, R. J., & Wood, T. (1995). Cognitive and affective outcomes in a multi-age language arts program. *Journal of Research in childhood Education*, *10*(1), 49–61.

Mager, M., Bhatnagar, R., Neubig, G., Kann, K., & Vu, N. T. (2023, July). Neural machine translation for the indigenous languages of the americas: An introduction. In *Proceedings of the third workshop on natural language processing for indigenous languages of the americas.* Toronto: Association for Computational Linguistics.

Mager, M., Carrillo, D., & Meza, I. (2018a). Probabilistic finite-state morphological segmenter for the Wixarika (Huichol) language. *Journal of Intelligent & Fuzzy Systems*, *34*(5), 3081–3087.

Mager, M., Carrillo, D., & Meza, I. (2018b). Probabilistic finite-state morphological segmenter for wixarika (huichol) language. *Journal of Intelligent & Fuzzy Systems*, *34*(5), 3081–3087.

Mager, M., Çetinoğlu, Ö., & Kann, K. (2019, June). Subword-level language identification for intra-word code-switching. 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)* (pp. 2005–2011). Minneapolis, Minnesota: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/N19-1201 doi: 10.18653/v1/N19-1201

Mager, M., Çetinoğlu, Ö., & Kann, K. (2020, November). Tackling the low-resource challenge for canonical segmentation. In *Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp)* (pp. 5237–5250). Online: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/2020.emnlp-main.423 doi: 10.18653/v1/2020.emnlp-main.423

Mager, M., Gonzalez, D., & Meza, I. (2017, 11). Probabilistic finite-state morphological segmenter for wixarika (huichol).
doi: 10.13140/RG.2.2.29286.11845

Mager, M., Gutierrez-Vasques, X., Sierra, G., & 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* (pp. 55–69). ACL. Retrieved from http://aclweb.org/anthology/C18-1006

Mager, M., Gutierrez-Vasques, X., Sierra, G., & Meza-Ruiz, I. (2018b, August). Challenges of language technologies for the indigenous languages of the Americas. In *Proceedings of the 27th international conference on computational linguistics* (pp. 55–69). Santa Fe, New Mexico, USA: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/C18-1006

Mager, M., & Kann, K. (2020, July). The IMS–CUBoulder system for the SIGMORPHON 2020 shared task on unsupervised morphological paradigm completion. In *Proceedings of the 17th sigmorphon workshop on computational research in phonetics, phonology, and morphology* (pp. 99–105). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.sigmorphon-1.9 doi: 10.18653/v1/2020.sigmorphon-1.9

Mager, M., Mager, E., Kann, K., & Vu, N. T. (2023). Ethical considerations for machine translation of indigenous languages: Giving a voice to the speakers. *arXiv preprint arXiv:2305.19474*.

Mager, M., Mager, E., Medina-Urrea, A., Meza Ruiz, I. V., & Kann, K. (2018, August). 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* (pp. 73–83). Santa Fe, New Mexico, USA: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/W18-4808

Mager, M., & Meza, I. (2021a). Challenges in the construction of automatic translators for indigenous languages of mexico. *DIGITAL SCHOLARSHIP IN THE HUMANITIES*, *36*, 37–42.

Mager, M., & Meza, I. (2021b, 05). Retos en construcción de traductores automáticos para lenguas indígenas de México. *Digital Scholarship in the Humanities*, *36*(Supplement 1), 43-48. Retrieved from https://doi.org/10.1093/llc/fqz093

Mager, M., Oncevay, A., Ebrahimi, A., Ortega, J., Rios, A., Fan, A., ... Kann, K. (2021, June). Findings of the AmericasNLP 2021 Shared Task on Open Machine Translation for Indigenous Languages of the Americas. In *Proceedings of thethe first workshop on nlp for indigenous languages of the americas.* Online: Association for Computational Linguistics.

Mager, M., Oncevay, A., Mager, E., Kann, K., & Vu, T. (2022, May). BPE vs. morphological segmentation: A case study on machine translation of four polysynthetic languages. In *Findings of the association for computational linguistics: Acl 2022* (pp. 961–971). Dublin, Ireland: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2022.findings-acl.78 doi: 10.18653/v1/2022.findings-acl.78

Mager, M., Rosales, M. J., Çetinoğlu, Ö., & Meza, I. (2019). Low-resource neural character-based noisy text normalization. *Journal of Intelligent & Fuzzy Systems*, *36*(5), 4921–4929.

Maharjan, S., Blair, E., Bethard, S., & Solorio, T. (2015). Developing language-tagged corpora for code-switching tweets. In *Proceedings of the 9th linguistic annotation workshop* (pp. 72–84).

Makarov, P., & Clematide, S. (2018a). Imitation learning for neural morphological string transduction. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 2877–2882).

Makarov, P., & Clematide, S. (2018b, October-November). Imitation learning for neural morphological string transduction. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 2877–2882). Brussels, Belgium: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/D18-1314 doi: 10.18653/v1/D18-1314

Makarov, P., & Clematide, S. (2018c, August). Neural transition-based string transduction for limited-resource setting in morphology. In *Proceedings of the 27th international conference on computational linguistics* (pp. 83–93). Santa Fe, New Mexico, USA: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/C18-1008

Makarov, P., Ruzsics, T., & Clematide, S. (2017). Align and copy: Uzh at sigmorphon 2017 shared task for morphological reinflection. In *Proceedings of the conll sigmorphon 2017 shared task: Universal morphological reinflection* (pp. 49–57).

Malaviya, C., Neubig, G., & Littell, P. (2017). Learning language representations for typology prediction. In *Proceedings of the 2017 conference on empirical methods in natural language processing* (pp. 2529–2535).

Mandal, S., Das, S. D., & Das, D. (2018). Language identification of bengali-english code-mixed data using character & phonetic based lstm models. *arXiv preprint arXiv:1803.03859* .

Marie, B., Rubino, R., & Fujita, A. (2020, July). Tagged back-translation revisited: Why does it really work? In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 5990–5997). Online: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/2020.acl-main.532 doi: 10.18653/v1/2020.acl-main.532

Marie, B., Sun, H., Wang, R., Chen, K., Fujita, A., Utiyama, M., & Sumita, E. (2019). Nict's unsupervised neural and statistical machine translation systems for the wmt19 news translation task. In *Proceedings of the fourth conference on machine translation (volume 2: Shared task papers, day 1)* (pp. 294–301).

Martin, J., Johnson, H., Farley, B., & Maclachlan, A. (2003). Aligning and using an english-inuktitut parallel corpus. In *Proceedings of the hlt-naacl 2003 workshop on building and using parallel texts: data driven machine translation and beyond-volume 3* (pp. 115–118).

Mayer, T., & Cysouw, M. (2014). Creating a massively parallel bible corpus. *Oceania*, *135*(273), 40.

MC. (2022). *Base de datos de pueblos indígenas u originarios.* https://bdpi.cultura.gob.pe/pueblos/shipibo-konibo. (Accessed: 2022-10-10)

McCarthy, A. D., Wicks, R., Lewis, D., Mueller, A., Wu, W., Adams, O., ... Yarowsky, D. (2020a). The johns hopkins university bible corpus: 1600+ tongues for typological exploration. In *Proceedings of the 12th language resources and evaluation conference* (pp. 2884–2892).

McCarthy, A. D., Wicks, R., Lewis, D., Mueller, A., Wu, W., Adams, O., ... Yarowsky, D. (2020b, May). The johns hopkins university bible corpus: 1600+ tongues for typological exploration. In *Proceedings of the 12th language resources and evaluation conference* (pp. 2884–2892). Marseille, France: European Language Resources Association. Retrieved from https://www.aclweb.org/anthology/2020.lrec-1.352

McNemar, Q. (1947, June). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, *12*(2), 153–157. Retrieved from https://doi.org/10.1007/bf02295996 doi: 10.1007/bf02295996

McQuown, N. A. (1955). The indigenous languages of latin america. *American Anthropologist*, *57*(3), 501–570.

Medina-Urrea, A. (2007). Affix discovery by means of corpora: Experiments for Spanish, Czech, Ralámuli and Chuj. In *Aspects of automatic text analysis* (pp. 277–299). Springer.

Medina-Urrea, A. (2008). Affix discovery based on entropy and economy measurements. *Texas Linguistics Society*, 99–112.

Medina Urrea, A., & García, M. A. (2006). Un experimento de reconocimiento automático de la derivación léxica en el ralámuli. *La lengua y la antropología para un conocimiento global del hombre*.

Meza Salcedo, G. (2017). Ética de la investigación desde el pensamiento indígena: derechos colectivos y el principio de la comunalidad. *Revista de Bioética y Derecho*(41), 141–159.

Miceli-Barone, A. V., Helcl, J., Sennrich, R., Haddow, B., & Birch, A. (2017). Deep architectures for neural machine translation. In *Proceedings of the second conference on machine translation* (pp. 99–107).

Micher, J. (2017). Improving coverage of an inuktitut morphological analyzer using a segmental recurrent neural network. In *Proceedings of the 2nd workshop on the use of computational methods in the study of endangered languages* (pp. 101–106).

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* .

Milroy, J., et al. (1995). *One speaker, two languages: Cross-disciplinary perspectives on code-switching* (Vol. 10). Cambridge University Press.

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

Moeng, T., Reay, S., Daniels, A., & Buys, J. (2022). Canonical and surface morphological segmentation for nguni languages. In *Artificial intelligence research: Second southern african conference, sacair 2021, durban, south africa, december 6–10, 2021, proceedings* (p. 125).

Moira, Y. (2002). *Cambridge textbooks in linguistics*. Cambridge: Cambridge University Press.

Molina, G., AlGhamdi, F., Ghoneim, M., Hawwari, A., Rey-Villamizar, N., Diab, M., & Solorio, T. (2016). Overview for the second shared task on language identification in code-switched data. In *Cs workshop.*

Monson, C., Llitjós, A. F., Aranovich, R., Levin, L., Brown, R., Peterson, E., ... Lavie, A. (2006). Building NLP systems for two resource-scarce indigenous languages: mapudungun and quechua. *Strategies for developing machine translation for minority languages*, 15.

Moreno, O. (2021, June). The REPU CS' Spanish–Quechua submission to the AmericasNLP 2021 shared task on open machine translation. In *Proceedings of the first workshop on natural language processing for indigenous languages of the americas* (pp. 241–247). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.americasnlp-1.27 doi: 10.18653/v1/2021.americasnlp-1.27

Moseley, C. (2010). *Atlas of the world's languages in danger*. Unesco.

Müller, T., Schmid, H., & Schütze, H. (2013). Efficient higher-order CRFs for morphological tagging. In *Emnlp*.

Myers-Scotton, C. (1993). Common and uncommon ground: Social and structural factors in codeswitching. *Language in society*, *22*(4), 475–503.

Nagoudi, E. M. B., Chen, W.-R., Abdul-Mageed, M., & Cavusoglu, H. (2021, June). IndT5: A text-to-text transformer for 10 indigenous languages. In *Proceedings of the first workshop on natural language processing for indigenous languages of the americas* (pp. 265–271). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.americasnlp-1.30 doi: 10.18653/v1/2021.americasnlp-1.30

Naradowsky, J., & Goldwater, S. (2009). Improving morphology induction by learning spelling rules. In *Twenty-first international joint conference on artificial intelligence*.

Nekoto, W., Marivate, V., Matsila, T., Fasubaa, T., Fagbohungbe, T., Akinola, S. O., ... Bashir, A. (2020, November). Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the association for computational linguistics: Emnlp 2020* (pp. 2144–2160). Online: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/2020.findings-emnlp.195 doi: 10.18653/v1/2020.findings-emnlp.195

Neubig, G. (2017). Neural machine translation and sequence-to-sequence models: A tutorial. *arXiv preprint arXiv:1703.01619*.

Neubig, G., & 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* (pp. 875–880).

Ngoc Le, T., & Sadat, F. (2020, December). Revitalization of indigenous languages through pre-processing and neural machine translation: The case of Inuktitut. In *Proceedings of the 28th international conference on computational linguistics* (pp. 4661–4666). Barcelona, Spain (Online): International Committee on Computational Linguistics. Retrieved from https://aclanthology.org/2020.coling-main.410 doi: 10.18653/v1/2020.coling-main.410

Nguyen, D., & Cornips, L. (2016). Automatic detection of intra-word code-switching. In *Sigmorphon.*

Nguyen, D., Trieschnigg, D., & Cornips, L. (2015). Audience and the use of minority languages on twitter. In *Icwsm.*

Nguyen, D. Q., Sirts, K., & Johnson, M. (2015, December). Improving topic coherence with latent feature word representations in MAP estimation for topic modeling. In *Proceedings of the australasian language technology association workshop 2015* (pp. 116–121). Parramatta, Australia. Retrieved from https://www.aclweb.org/anthology/U15-1014

Nguyen, T. Q., & Chiang, D. (2017). Transfer learning across low-resource, related languages for neural machine translation. In *Proceedings of the eighth international joint conference on natural language processing (volume 2: Short papers)* (pp. 296–301).

Nichols, J. (1986). Head-marking and dependent-marking grammar. *Language*, *62*(1), 56–119.

Nida, E. (1945). Linguistics and ethnology in translation-problems. *Word*, *1*(2), 194–208.

Niehues, J., & Cho, E. (2017, September). Exploiting linguistic resources for neural machine translation using multi-task learning. In *Proceedings of the second conference on machine translation* (pp. 80–89). Copenhagen, Denmark: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/W17-4708 doi: 10.18653/v1/W17-4708

Niehues, J., Cho, E., Ha, T.-L., & Waibel, A. (2016, December). Pre-translation for neural machine translation. In *Proceedings of COLING 2016, the 26th international conference on computational linguistics: Technical papers* (pp. 1828–1836). Osaka, Japan: The COLING 2016 Organizing Committee. Retrieved from https://www.aclweb.org/anthology/C16-1172

Niranjana, T. (1990). Translation, colonialism and rise of english. *Economic and Political Weekly*, 773–779.

Niranjana, T. (1992). Siting translation: History, post-structuralism and the colonial context. *URL: https://doi. org/10.1525/9780520911369 [2021-04-01]*.

Nissing, H.-G., & Müller, J. (2009). *Grundpositionen philosophischer ethik: von aristoteles bis jürgen habermas*. WBG.

Nooralahzadeh, F., Bekoulis, G., Bjerva, J., & Augenstein, I. (2020). Zero-shot cross-lingual transfer with meta learning. *arXiv preprint arXiv:2003.02739*.

Nortier, J. M. (1995). Code switching in moroccan arabic/dutch vs. moroccan arabic/french language contact. *International journal of the sociology of language*, *1995*(112), 81–96.

Ojha, A. K., Malykh, V., Karakanta, A., & Liu, C.-H. (2020, December). Findings of the LoResMT 2020 shared task on zero-shot for low-resource languages. In *Proceedings of the 3rd workshop on technologies for mt of low resource languages* (pp. 33–37). Suzhou, China: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/2020.loresmt-1.4

Olson, M., Wyner, A., & Berk, R. (2018). Modern neural networks generalize on small data sets. In *Advances in neural information processing systems* (pp. 3619–3628).

Oncevay, A. (2021, June). 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* (pp. 194–201). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.americasnlp-1.22 doi: 10.18653/v1/2021.americasnlp-1.22

OpenAI. (2023). Gpt-4 technical report. *ArXiv*, *abs/2303.08774*. Retrieved from https://api.semanticscholar.org/CorpusID:257532815

Orife, I., Kreutzer, J., Sibanda, B., Whitenack, D., Siminyu, K., Martinus, L., ... others (2020). Masakhane–machine translation for africa. *arXiv preprint arXiv:2003.11529*.

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

Ortiz Suárez, P. J., Sagot, B., & Romary, L. (2019, July). Asynchronous Pipeline for Processing Huge Corpora on Medium to Low Resource Infrastructures. In *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*. Cardiff, United Kingdom. Retrieved from https://hal.inria.fr/hal-02148693

Palacios, A. (2008). La lengua como instrumento de identidad y diferenciación: más allá de la influencia de las lenguas amerindias. *De moneda nunca usada*.

Pan, Y., Li, X., Yang, Y., & Dong, R. (2020, July). Multi-task neural model for agglutinative language translation. In *Proceedings of the 58th annual meeting of the association for computational linguistics: Student research workshop* (pp. 103–110). Online: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/2020.acl-srw.15

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002, July). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the association for computational linguistics* (pp. 311–318). Philadelphia, Pennsylvania, USA: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/P02-1040 doi: 10.3115/1073083.1073135

Parida, S., Panda, S., Dash, A., Villatoro-Tello, E., Doğruöz, A. S., Ortega-Mendoza, R. M., ... Motlicek, P. (2021, June). Open machine translation for low resource South American languages (AmericasNLP 2021 shared task contribution). In *Proceedings of the first workshop on natural language processing for indigenous languages of the americas* (pp. 218–223). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.americasnlp-1.24 doi: 10.18653/v1/2021.americasnlp-1.24

Patro, J., Samanta, B., Singh, S., Basu, A., Mukherjee, P., Choudhury, M., & Mukherjee, A. (2017). All that is English may be Hindi: Enhancing language identification through automatic ranking of the likeliness of word borrowing in social media. In *Emnlp*.

Pavlick, E., Post, M., Irvine, A., Kachaev, D., & Callison-Burch, C. (2014). The language demographics of amazon mechanical turk. *Transactions of the Association for Computational Linguistics*, *2*, 79–92.

Pennington, J., Socher, R., & Manning, C. (2014, October). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543). Doha, Qatar: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/D14-1162 doi: 10.3115/v1/D14-1162

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 conference of the north american chapter of the association for computational linguistics: Human language technologies, volume 1 (long papers)* (pp. 2227–2237).

Pintado-Cortina, A. P. (2004). *Tarahumaras*. CDI, Comisión Nacional para el Desarrollo de los Pueblos Indígenas.

Plank, B., Søgaard, A., & Goldberg, Y. (2016). Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss. In *Acl*.

Platanios, E. A., Sachan, M., Neubig, G., & Mitchell, T. (2018). Contextual parameter generation for universal neural machine translation. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 425–435).

Poncelas, A., Popović, M., Shterionov, D., de Buy Wenniger, G. M., & Way, A. (2019). Combining pbsmt and nmt back-translated data for efficient nmt. In *Proceedings of the international conference on recent advances in natural language processing (ranlp 2019)* (pp. 922–931).

Poon, H., Cherry, C., & Toutanova, K. (2009). Unsupervised morphological segmentation with log-linear models. In *Naacl-hlt* (pp. 209–217).

Poplack, S. (1980). Sometimes i'll start a sentence in spanish y termino en espanol: toward a typology of code-switching1.

Poplack, S. (2008). Code-switching. In *Volume 1* (pp. 589–596). De Gruyter Mouton.

Poplack, S., & Meechan, M. (1995). Patterns of language mixture: Nominal structure in wolof-french and fongbe-french bilingual discourse. Cambridge University Press.

Popović, M. (2015). chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation* (pp. 392–395).

Porta, A. O. (2010). The use of formal language models in the typology of the morphology of amerindian languages. In *Proceedings of the acl 2010 student research workshop* (pp. 109–113).

Post, M. (2018, October). A call for clarity in reporting BLEU scores. In *Proceedings of the third conference on machine translation: Research papers* (pp. 186–191). Belgium,

Brussels: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/W18-6319 doi: 10.18653/v1/W18-6319

Pottier, B. (2002). *The mexicanero dialect of the western sierra-madre.* SOC LINGUISTIQUE ROMANE UNIV SCI HUMAINES 25 RUE DU MARECHAL-JUIN, F-67084 ....

Pourdamghani, N., & Knight, K. (2019). Neighbors helping the poor: improving low-resource machine translation using related languages. *Machine Translation*, *33*(3), 239–258.

Pratapa, A., Bhat, G., Choudhury, M., Sitaram, S., Dandapat, S., & Bali, K. (2018). Language modeling for code-mixing: The role of linguistic theory based synthetic data. In *Acl* (Vol. 1, pp. 1543–1553).

Premsrirat, S., & Malone, D. (2003). Language development and language revitalization in asia. *SIL International, Mahidol University*.

Press, O., & Wolf, L. (2017). Using the output embedding to improve language models. In *Proceedings of the 15th conference of the european chapter of the association for computational linguistics: Volume 2, short papers* (pp. 157–163).

Rachels, J., & Rachels, S. (1986). *The elements of moral philosophy*. Temple University Press Philadelphia.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, *21*, 1–67.

Raganato, A., Vázquez, R., Creutz, M., & Tiedemann, J. (2021, November). An empirical investigation of word alignment supervision for zero-shot multilingual neural machine translation. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 8449–8456). Online and Punta Cana, Dominican Republic: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.emnlp-main.664

Ranathunga, S., Lee, E.-S. A., Skenduli, M. P., Shekhar, R., Alam, M., & Kaur, R. (2021). Neural machine translation for low-resource languages: A survey. *arXiv preprint arXiv:2106.15115*.

Ren, S., Wu, Y., Liu, S., Zhou, M., & Ma, S. (2020, July). A retrieve-and-rewrite initialization method for unsupervised machine translation. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 3498–3504). Online: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/2020.acl-main.320 doi: 10.18653/v1/2020.acl-main.320

Ren, S., Zhang, Z., Liu, S., Zhou, M., & Ma, S. (2019). Unsupervised neural machine translation with smt as posterior regularization. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 33, pp. 241–248).

Rijhwani, S., Sequiera, R., Choudhury, M., Bali, K., & Maddila, C. S. (2017). Estimating code-switching on twitter with a novel generalized word-level language detection technique. In *Acl*.

Riley, P., Caswell, I., Freitag, M., & Grangier, D. (2020, July). Translationese as a language in "multilingual" NMT. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 7737–7746). Online: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/2020.acl-main.691 doi: 10.18653/v1/2020.acl-main.691

Rios, A. (2016). A basic language technology toolkit for quechua.

Roest, C., Edman, L., Minnema, G., Kelly, K., Spenader, J., & Toral, A. (2020, November). Machine translation for English–Inuktitut with segmentation, data acquisition and pre-training. In *Proceedings of the fifth conference on machine translation* (pp. 274–281). Online: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/2020.wmt-1.29

Rolnick, D., Veit, A., Belongie, S., & Shavit, N. (2017). Deep learning is robust to massive label noise. *arXiv preprint arXiv:1705.10694*.

Romero, C. B., Hois, J. M. M., & Avilés, F. R. (2016). Richard feynman, los alfabetos y los lenguajes. *Relingüística aplicada*(19), 2.

Rosales, M. J., Mager, M., & Ruız, I. V. M. (2019). Towards a twitter corpus of the indigenous languages of the americas. In *Opencor*.

Ruder, S. (2020). *Why You Should Do NLP Beyond English.* http://ruder.io/nlp-beyond-english.

Rudra, K., Rijhwani, S., Begum, R., Bali, K., Choudhury, M., & Ganguly, N. (2016). Understanding language preference for expression of opinion and sentiment: What do Hindi-English speakers do on twitter? In *Emnlp.*

Ruokolainen, T., Kohonen, O., Sirts, K., Grönroos, S.-A., Kurimo, M., & Virpioja, S. (2016). A comparative study of minimally supervised morphological segmentation. *Computational Linguistics*, *42*(1), 91–120.

Ruokolainen, T., Kohonen, O., Virpioja, S., & Kurimo, M. (2013). Supervised morphological segmentation in a low-resource learning setting using conditional random fields. In *Conll* (pp. 29–37).

Ruokolainen, T., Kohonen, O., Virpioja, S., & Kurimo, M. (2014, April). Painless semi-supervised morphological segmentation using conditional random fields. In *Proceedings of the 14th conference of the European chapter of the association for computational linguistics, volume 2: Short papers* (pp. 84–89). Gothenburg, Sweden: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/E14-4017 doi: 10.3115/v1/E14-4017

Ruzsics, T., & Samardzic, T. (2017). Neural sequence-to-sequence learning of internal word structure. In *Proceedings of the 21st conference on computational natural language learning (conll 2017)* (pp. 184–194).

Sabty, C., Mesabah, I., Çetinoğlu, Ö., & Abdennadher, S. (2021). Language identification of intra-word code-switching for arabic–english. *Array*, *12*, 100104.

Sachan, D. S., & Neubig, G. (2018). Parameter sharing methods for multilingual self-attentional translation models. *arXiv preprint arXiv:1809.00252*.

Sahin, G. G., & Steedman, M. (2018). Character-level models versus morphology in semantic role labeling. In *Acl* (Vol. 1, pp. 386–396).

Sakakini, T., Bhat, S., & Viswanath, P. (2017). Morse: Semantic-ally drive-n morpheme segment-er. In *Acl* (pp. 552–561). ACL. Retrieved from http://aclweb.org/anthology/P17-1051 doi: 10.18653/v1/P17-1051

Salcedo, G.-M. (2016). El 'vivir nosotros' amerindio vs 'decir nosotros' de la globalización. *Cuadernos de filosofía latinoamericana*, *37*(114), 151–166.

Salesky, E., Etter, D., & Post, M. (2021). Robust open-vocabulary translation from visual text representations. *arXiv preprint arXiv:2104.08211*.

Saleva, J., & Lignos, C. (2021, April). The effectiveness of morphology-aware segmentation in low-resource neural machine translation. In *Proceedings of the 16th conference of the european chapter of the association for computational linguistics: Student research workshop* (pp. 164–174). Online: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/2021.eacl-srw.22

Samih, Y., Maharjan, S., Attia, M., Kallmeyer, L., & Solorio, T. (2016). Multilingual code-switching identification via LSTM recurrent neural networks. In *Cs workshop.*

Sánchez-Cartagena, V. M., Esplà-Gomis, M., Pérez-Ortiz, J. A., & Sánchez-Martínez, F. (2021, November). Rethinking data augmentation for low-resource neural machine translation: A multi-task learning approach. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 8502–8516). Online and Punta Cana, Dominican Republic: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.emnlp-main.669

Sano, M., Suzuki, J., & Kiyono, S. (2019). Effective adversarial regularization for neural machine translation. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 204–210).

Sarawagi, S., & Cohen, W. W. (2005). Semi-markov conditional random fields for information extraction. In *Advances in neural information processing systems* (pp. 1185–1192).

Scherrer, Y., Grönroos, S.-A., & Virpioja, S. (2020, November). The University of Helsinki and aalto university submissions to the WMT 2020 news and low-resource translation tasks. In *Proceedings of the fifth conference on machine translation* (pp. 1129–1138). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.wmt-1.134

Schmidt, F. A. (2013). The good, the bad and the ugly: Why crowdsourcing needs ethics. In *2013 international conference on cloud and green computing* (pp. 531–535).

Schmidt, R. W. (2011). American indian identity and blood quantum in the 21st century: A critical review. *Journal of Anthropology*, *2011*.

Schwartz, L. (2022). Primum non nocere: Before working with indigenous data, the acl must confront ongoing colonialism.

Schwartz, L., Tyers, F., Levin, L., Kirov, C., Littell, P., Lo, C.-k., ... others (2020). Neural polysynthetic language modelling. *arXiv preprint arXiv:2005.05477*.

Sechrest, L., Fay, T. L., & Zaidi, S. H. (1972). Problems of translation in cross-cultural research. *Journal of cross-cultural psychology*, *3*(1), 41–56.

See, A., Liu, P. J., & Manning, C. D. (2017, July). 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)* (pp. 1073–1083). Vancouver, Canada: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/P17-1099 doi: 10.18653/v1/P17-1099

Seeker, W., & Çetinoğlu, Ö. (2015). A graph-based lattice dependency parser for joint morphological segmentation and syntactic analysis. *TACL*, *3*(1), 359–373.

SEGOB. (2020). *Sistema de Información Cultural - Lenguas indígenas: Huichol* . https://sic.gob.mx/ficha.php?table=inali_li.

Sennrich, R., Haddow, B., & 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)* (pp. 86–96).

Sennrich, R., Haddow, B., & Birch, A. (2016b, August). 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)* (pp. 86–96). Berlin, Germany: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/P16-1009 doi: 10.18653/v1/P16-1009

Sennrich, R., Haddow, B., & Birch, A. (2016c). 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)* (pp. 1715–1725).

Sennrich, R., Haddow, B., & Birch, A. (2016d, August). 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) (pp. 1715–1725). Berlin, Germany: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/P16-1162 doi: 10.18653/v1/P16-1162

Sennrich, R., & Zhang, B. (2019). Revisiting low-resource neural machine translation: A case study. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 211–221).

Serra, J., Suris, D., Miron, M., & Karatzoglou, A. (2018). Overcoming catastrophic forgetting with hard attention to the task. In *International conference on machine learning* (pp. 4548–4557).

Sharma, A., Katrapati, G., & Sharma, D. M. (2018a, October). IIT(BHU)–IIITH at CoNLL–SIGMORPHON 2018 shared task on universal morphological reinflection. In *Proceedings of the CoNLL–SIGMORPHON 2018 shared task: Universal morphological reinflection* (pp. 105–111). Brussels: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/K18-3013 doi: 10.18653/v1/K18-3013

Sharma, A., Katrapati, G., & Sharma, D. M. (2018b, October). IIT(BHU)–IIITH at CoNLL–SIGMORPHON 2018 shared task on universal morphological reinflection. In *Proceedings of the CoNLL–SIGMORPHON 2018 shared task: Universal morphological reinflection* (pp. 105–111). Brussels: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/K18-3013 doi: 10.18653/v1/K18-3013

Siddhant, A., Bapna, A., Cao, Y., Firat, O., Chen, M., Kudugunta, S., ... Wu, Y. (2020, July). Leveraging monolingual data with self-supervision for multilingual neural machine translation. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 2827–2835). Online: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/2020.acl-main.252

Sierra Martínez, G., Montaño, C., Bel-Enguix, G., Córdova, D., & Mota Montoya, M. (2020, May). CPLM, a parallel corpus for Mexican languages: Development and interface. In *Proceedings of the 12th language resources and evaluation conference* (pp. 2947–2952). Marseille, France: European Language Resources Association. Retrieved from https://aclanthology.org/2020.lrec-1.360

Sirts, K., & Goldwater, S. (2013). Minimally-supervised morphological segmentation using adaptor grammars. *Transactions of the Association for Computational Linguistics*, *1*, 255–266.

Sitaram, S., Chandu, K. R., Rallabandi, S. K., & Black, A. W. (2019). A survey of code-switched speech and language processing. *arXiv preprint arXiv:1904.00784*.

Skiba, R. (1997). Code switching as a countenance of language interference. *The internet TESL journal*, *3*(10), 1–6.

Smit, P., Virpioja, S., Grönroos, S.-A., & Kurimo, M. (2014, April). 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* (pp. 21–24). Gothenburg, Sweden: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/E14-2006 doi: 10.3115/v1/E14-2006

Smit, P., Virpioja, S., Grönroos, S.-A., Kurimo, M., et al. (2014). Morfessor 2.0: Toolkit for statistical morphological segmentation. In *The 14th conference of the european chapter of the association for computational linguistics (eacl), gothenburg, sweden, april 26-30, 2014*.

Smith, L. T. (2021). *Decolonizing methodologies: Research and indigenous peoples*. Bloomsbury Publishing.

Snoek, C., Thunder, D., Loo, K., Arppe, A., Lachler, J., Moshagen, S., & Trosterud, T. (2014). Modeling the noun morphology of Plains Cree. In *Proceedings of the 2014 workshop on the use of computational methods in the study of endangered languages* (pp. 34–42).

Søgaard, A., Ruder, S., & Vulić, I. (2018). On the limitations of unsupervised bilingual dictionary induction. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 778–788).

Solorio, T., Blair, E., Maharjan, S., Bethard, S., Diab, M., Gohneim, M., ... others (2014). Overview for the first shared task on language identification in code-switched data. *CS Workshop*.

Solorio, T., & Liu, Y. (2008, October). Learning to predict code-switching points. In *Proceedings of the 2008 conference on empirical methods in natural language processing* (pp. 973–981). Honolulu, Hawaii: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/D08-1102

Solórzano, S. F. (2017). *Desarrollo de un analizador automático de estados finitos para la lengua bribri*. http://morphology.bribri.net/.

Song, H., Dabre, R., Mao, Z., Cheng, F., Kurohashi, S., & Sumita, E. (2020). Pre-training via leveraging assisting languages and data selection for neural machine translation. *arXiv preprint arXiv:2001.08353*.

Song, K., Tan, X., Qin, T., Lu, J., & Liu, T.-Y. (2019). Mass: Masked sequence to sequence pre-training for language generation. In *International conference on machine learning* (pp. 5926–5936).

Song, K., Zhang, Y., Yu, H., Luo, W., Wang, K., & Zhang, M. (2019, June). Code-switching for enhancing NMT with pre-specified 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)* (pp. 449–459). Minneapolis, Minnesota: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/N19-1044 doi: 10.18653/v1/N19-1044

Soto, X., Shterionov, D., Poncelas, A., & Way, A. (2020, July). Selecting backtranslated data from multiple sources for improved neural machine translation. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 3898–3908). Online: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/2020.acl-main.359 doi: 10.18653/v1/2020.acl-main.359

Spiegler, S., & Monson, C. (2010, August). EMMA: A novel evaluation metric for morphological analysis. In *Proceedings of the 23rd international conference on computational linguistics (coling 2010)* (pp. 1029–1037). Beijing, China: Coling 2010 Organizing Committee. Retrieved from https://www.aclweb.org/anthology/C10-1116

Srinivasan, T., Sanabria, R., & Metze, F. (2019). Multitask learning for different subword segmentations in neural machine translation. *arXiv preprint arXiv:1910.12368*.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, *15*(1), 1929–1958.

Stanovsky, G., Smith, N. A., & Zettlemoyer, L. (2019). Evaluating gender bias in machine translation. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 1679–1684).

Stojanovski, D., Hangya, V., Huck, M., & Fraser, A. (2019). The lmu munich unsupervised machine translation system for wmt19. In *Proceedings of the fourth conference on machine translation (volume 2: Shared task papers, day 1)* (pp. 393–399).

Sun, H., Wang, R., Chen, K., Utiyama, M., Sumita, E., & Zhao, T. (2020). Robust unsupervised neural machine translation with adversarial training. *arXiv preprint arXiv:2002.12549*.

Syawkany, Y. (2022). *What is code-switching.* Retrieved from https://www.academia.edu/23264194/What_is_Code_Switching

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 2818–2826).

Taguchi, C., Iwata, S., & Watanabe, T. (2022). Universal dependencies treebank for tatar: Incorporating intra-word code-switching information. In *Proceedings of the workshop on resources and technologies for indigenous, endangered and lesser-resourced languages in eurasia within the 13th language resources and evaluation conference* (pp. 95–104).

Taguchi, C., Sakai, Y., & Watanabe, T. (2021). Transliteration for low-resource code-switching texts: Building an automatic cyrillic-to-latin converter for tatar. In *Proceedings of the fifth workshop on computational approaches to linguistic code-switching* (pp. 133–140).

Tan, X., Leng, Y., Chen, J., Ren, Y., Qin, T., & Liu, T.-Y. (2019). A study of multilingual neural machine translation. *arXiv preprint arXiv:1912.11625*.

Team, N. (2022). No language left behind: Scaling human-centered machine translation.

Thouvenot, M. (2011). Chachalaca en cen, juntamente. In *Compendio enciclopedico del nahuatl, dvd.*

Tiedemann, J. (2016). Opus–parallel corpora for everyone. *Baltic Journal of Modern Computing*, 384.

Tiedemann, J. (2018). Emerging language spaces learned from massively multilingual corpora. *arXiv preprint arXiv:1802.00273*.

Toral, A., Castilho, S., Hu, K., & Way, A. (2018). Attaining the unattainable? reassessing claims of human parity in neural machine translation. In *Proceedings of the third conference on machine translation: Research papers* (pp. 113–123).

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... others (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Trieu, H.-L., Tran, D.-V., Ittoo, A., & Nguyen, L.-M. (2019, June). Leveraging additional resources for improving statistical machine translation on asian low-resource languages. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, *18*(3). Retrieved from https://doi.org/10.1145/3314936 doi: 10.1145/3314936

Tymoczko, M. (2006). Translation: Ethics, ideology, action. *The Massachusetts Review*, *47*(3), 442–461.

UN. (2022). *Who are indigenous peoples?* Retrieved from https://www.un.org/esa/socdev/unpfii/documents/5session_factsheet1.pdf

Utiyama, M., & Isahara, H. (2007). A comparison of pivot methods for phrase-based statistical machine translation. In *Human language technologies 2007: The conference of the north american chapter of the association for computational linguistics; proceedings of the main conference* (pp. 484–491).

Valenzuela, P. (2003). *Transitivity in Shipibo-Konibo grammar* (Unpublished doctoral dissertation). University of Oregon.

Vanhove, M., Stolz, T., Urdze, A., & Otsuka, H. (2012). *Morphologies in contact*. Walter de Gruyter.

Vania, C., Grivas, A., & Lopez, A. (2018). What do character-level models learn about morphology? the case of dependency parsing. In *Emnlp* (pp. 2573–2583).

Vania, C., & Lopez, A. (2017a). 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)* (pp. 2016–2027).

Vania, C., & Lopez, A. (2017b, July). 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)* (pp. 2016–2027). Vancouver, Canada: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/P17-1184 doi: 10.18653/v1/P17-1184

Vasquez, A., Ego Aguirre, R., Angulo, C., Miller, J., Villanueva, C., Agić, Ž., ... Oncevay, A. (2018, November). Toward universal dependencies for Shipibo-konibo. In *Proceedings of the second workshop on universal dependencies (UDW 2018)* (pp. 151–161). Brussels, Belgium: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/W18-6018 doi: 10.18653/v1/W18-6018

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. In *Nips*.

Vázquez, R., Scherrer, Y., Virpioja, S., & Tiedemann, J. (2021, June). The Helsinki submission to the AmericasNLP shared task. In *Proceedings of the first workshop on natural language processing for indigenous languages of the americas* (pp. 255–264). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.americasnlp-1.29 doi: 10.18653/v1/2021.americasnlp-1.29

Vilca, C., David, H., Mariñó, C., Cagniy, F., & Mamani Calderon, E. F. (2012). Analizador morfológico de la lengua quechua basado en software libre helsinkifinite-statetransducer (hfst). *COMTEL*.

Vinyals, O., Fortunato, M., & Jaitly, N. (2015). Pointer networks. In *Neurips* (pp. 2692–2700).

Virpioja, S., Väyrynen, J. J., Creutz, M., & Sadeniemi, M. (2007). Morphology-aware statistical machine translation based on morphs induced in an unsupervised manner. In *Proceedings of machine translation summit xi: Papers*.

Vulić, I., Glavaš, G., Reichart, R., & Korhonen, A. (2019, November). Do we really need fully unsupervised cross-lingual embeddings? 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) (pp. 4407–4418). Hong Kong, China: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/D19-1449 doi: 10.18653/v1/D19-1449

Vulić, I., Ruder, S., & Søgaard, A. (2020). Are all good word vector spaces isomorphic? *arXiv preprint arXiv:2004.04070*. Retrieved from https://arxiv.org/pdf/2004.04070.pdf

Wagner, C. (2016). Las lenguas indígenas de américa (lenguas amerindias). *Revista Documentos Lingüísticos y Literarios UACh*(17), 30–37.

Wang, H., Ma, S., Dong, L., Huang, S., Zhang, D., & Wei, F. (2022). Deepnet: Scaling transformers to 1,000 layers. *arXiv preprint arXiv:2203.00555*.

Wang, L., Cao, Z., Xia, Y., & de Melo, G. (2016). Morphological segmentation with window LSTM neural networks. In *Aaai*.

Wang, L., Zhao, W., Jia, R., Li, S., & Liu, J. (2019). Denoising based sequence-to-sequence pre-training for text generation. 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)* (pp. 3994–4006).

Wang, Q., Li, B., Xiao, T., Zhu, J., Li, C., Wong, D. F., & Chao, L. S. (2019, July). Learning deep transformer models for machine translation. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 1810–1822). Florence, Italy: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/P19-1176 doi: 10.18653/v1/P19-1176

Wang, R., Tan, X., Luo, R., Qin, T., & Liu, T.-Y. (2021, 8). A survey on low-resource neural machine translation. In Z.-H. Zhou (Ed.), *Proceedings of the thirtieth international joint conference on artificial intelligence, IJCAI-21* (pp. 4636–4643). International Joint Conferences on Artificial Intelligence Organization. Retrieved from https://doi.org/10.24963/ijcai.2021/629 (Survey Track) doi: 10.24963/ijcai.2021/629

Wang, X., Pham, H., Arthur, P., & Neubig, G. (2019, May). Multilingual neural machine translation with soft decoupled encoding. In *International conference on*

learning representations (iclr). New Orleans, LA, USA. Retrieved from https://arxiv.org/abs/1902.03499

Wang, X., Tsvetkov, Y., & Neubig, G. (2020, July). Balancing training for multilingual neural machine translation. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 8526–8537). Online: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/2020.acl-main.754 doi: 10.18653/v1/2020.acl-main.754

Wang, Y., Tian, F., He, D., Qin, T., Zhai, C., & Liu, T.-Y. (2019). Non-autoregressive machine translation with auxiliary regularization. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 33, pp. 5377–5384).

Weber, D. (1989). *A grammar of huallaga (huánuco) quechua* (Vol. 112). Univ of California Press.

Weiss, K., Khoshgoftaar, T. M., & Wang, D. (2016). A survey of transfer learning. *Journal of Big Data*, *3*(1), 9.

Wichmann, S. (2007). *Popoluca de texistepec* (Vol. 27). El Colegio de Mexico AC.

Wiemerslage, A., Silfverberg, M., Yang, C., McCarthy, A., Nicolai, G., Colunga, E., & Kann, K. (2022, May). Morphological processing of low-resource languages: Where we are and what's next. In *Findings of the association for computational linguistics: Acl 2022* (pp. 988–1007). Dublin, Ireland: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2022.findings-acl.80 doi: 10.18653/v1/2022.findings-acl.80

Wieting, J., Berg-Kirkpatrick, T., Gimpel, K., & Neubig, G. (2019). Beyond bleu: Training neural machine translation with semantic similarity. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 4344–4355).

Winata, G. I., Aji, A. F., Yong, Z.-X., & Solorio, T. (2022). The decades progress on code-switching research in nlp: A systematic survey on trends and challenges. *arXiv preprint arXiv:2212.09660*.

Wolfart, H. C., & Pardo, F. (1973). Computer-assisted linguistic analysis. *University of Manitoba Anthropology Papers*(6).

Wu, H., & Wang, H. (2007). Pivot language approach for phrase-based statistical machine translation. *Machine Translation*, *21*(3), 165–181.

Wu, X., Lv, S., Zang, L., Han, J., & Hu, S. (2019). Conditional bert contextual augmentation. In *International conference on computational science* (pp. 84–95).

Xu, H., Van Durme, B., & Murray, K. (2021, November). BERT, mBERT, or BiBERT? a study on contextualized embeddings for neural machine translation. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 6663–6675). Online and Punta Cana, Dominican Republic: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.emnlp-main.534

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., ... Raffel, C. (2021). mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 conference of the north american chapter of the association for computational linguistics: Human language technologies* (pp. 483–498).

Yang, Y., Li, S., Zhang, Y., & Zhang, H.-P. (2019). Point the point: Uyghur morphological segmentation using pointernetwork with gru. In *China national conference on chinese computational linguistics* (pp. 371–381).

Yang, Z., Hu, B., Han, A., Huang, S., & Ju, Q. (2020, November). CSP:code-switching pre-training for neural machine translation. In *Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp)* (pp. 2624–2636). Online: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/2020.emnlp-main.208 doi: 10.18653/v1/2020.emnlp-main.208

Zaremoodi, P., Buntine, W., & Haffari, G. (2018). Adaptive knowledge sharing in multi-task learning: Improving low-resource neural machine translation. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 2: Short papers)* (pp. 656–661).

Zeiler, M. D. (2012). Adadelta: an adaptive learning rate method. *arXiv:1212.5701*.

Zevallos, R., Ortega, J., Chen, W., Castro, R., Bel, N., Toshio, C., ... Nelsi Melgarejo, H. (2022, July). Introducing QuBERT: A large monolingual corpus and BERT model for Southern Quechua. In *Proceedings of the third workshop on deep learning for low-resource natural language processing* (pp. 1–13). Hybrid: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2022.deeplo-1.1

Zhang, B., Williams, P., Titov, I., & Sennrich, R. (2020a). Improving massively multilingual neural machine translation and zero-shot translation. *arXiv preprint arXiv:2004.11867*.

Zhang, B., Williams, P., Titov, I., & Sennrich, R. (2020b, July). Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 1628–1639). Online: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/2020.acl-main.148 doi: 10.18653/v1/2020.acl-main.148

Zhang, S., Frey, B., & Bansal, M. (2020, November). ChrEn: Cherokee-English machine translation for endangered language revitalization. In *Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp)* (pp. 577–595). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.emnlp-main.43 doi: 10.18653/v1/2020.emnlp-main.43

Zhang, S., Frey, B., & Bansal, M. (2021, August). ChrEnTranslate: Cherokee-English machine translation demo with quality estimation and corrective feedback. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing: System demonstrations* (pp. 272–279). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.acl-demo.33 doi: 10.18653/v1/2021.acl-demo.33

Zhang, S., Frey, B., & Bansal, M. (2022). How can nlp help revitalize endangered languages? a case study and roadmap for the cherokee language. *arXiv preprint arXiv:2204.11909*.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). Bertscore: Evaluating text generation with bert. In *Iclr*.

Zhang, Y., Riesa, J., Gillick, D., Bakalov, A., Baldridge, J., & Weiss, D. (2018). A fast, compact, accurate model for language identification of codemixed text. In *Emnlp*.

Zhang, Y., Wang, Z., Cao, R., Wei, B., Shan, W., Zhou, S., ... Zhu, J. (2020, November). The NiuTrans machine translation systems for WMT20. In *Proceedings of the fifth conference on machine translation* (pp. 338–345). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.wmt-1.37

Zhang, Z., Huang, Y., & Zhao, H. (2019). Open vocabulary learning for neural chinese pinyin ime. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 1584–1594).

Zheng, F., Reid, M., Marrese-Taylor, E., & Matsuo, Y. (2021, June). Low-resource machine translation using cross-lingual language model pretraining. In *Proceedings of the first workshop on natural language processing for indigenous languages of the americas* (pp. 234–240). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.americasnlp-1.26 doi: 10.18653/v1/2021.americasnlp-1.26

Zheng, H., Cheng, Y., & Liu, Y. (2017). Maximum expected likelihood estimation for zero-resource neural machine translation. In *Ijcai* (pp. 4251–4257).

Zhou, R., & Hansen, E. A. (2005). Beam-stack search: Integrating backtracking with beam search. In *Icaps* (pp. 90–98).

Zhou, S., Zeng, X., Zhou, Y., Anastasopoulos, A., & Neubig, G. (2019, August). Improving robustness of neural machine translation with multi-task learning. In *Proceedings of the fourth conference on machine translation (volume 2: Shared task papers, day 1)* (pp. 565–571). Florence, Italy: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/W19-5368 doi: 10.18653/v1/W19-5368

Zhu, C., Yu, H., Cheng, S., & Luo, W. (2020, July). Language-aware interlingua for multilingual neural machine translation. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 1650–1655). Online: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/2020.acl-main.150 doi: 10.18653/v1/2020.acl-main.150

Zhu, J., Gao, F., Wu, L., Xia, Y., Qin, T., Zhou, W., ... Liu, T.-Y. (2019). Soft contextual data augmentation for neural machine translation. *arXiv preprint arXiv:1905.10523*.

Zhu, J., Xia, Y., Wu, L., He, D., Qin, T., Zhou, W., ... Liu, T.-Y. (2020). Incorporating bert into neural machine translation. *arXiv preprint arXiv:2002.06823*.

Zoph, B., Yuret, D., May, J., & Knight, K. (2016). Transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1604.02201*.

Zou, W. Y., Socher, R., Cer, D., & Manning, C. D. (2013). Bilingual word embeddings for phrase-based machine translation. In *Proceedings of the 2013 conference on empirical methods in natural language processing* (pp. 1393–1398).

A. Appendix

A.1. Complementary Low-Resource discussion on Machine Translation

In this appendix, we expand the information regarding current work on MT for LRL.

A.1.1. Expanded LR work on Multilingual supervised training

Arivazhagan, Bapna, Firat, Aharoni, et al. (2019) introduce a representational invariance training objective across languages that achieve comparable results with pivoting methods. Promising results of multilingual models have encouraged experiments with models trained on a massive amount of language pairs, resulting in large multilingual models: Aharoni et al. (2019) train a single model on 102 languages to and from English in contrast to the 58 languages used by Neubig & Hu (2018). The negative aspect of this approach is the size of the network. Arivazhagan, Bapna, Firat, Lepikhin, et al. (2019) perform an extensive study on 102 language pairs to explore different settings and training setups and achieve good results for LRLs, while maintaining good performance for high-resource languages. Related massively multilingual NMT systems have been trained for analytic proposes (Tiedemann, 2018; Malaviya et al., 2017) and general zero-shot transfer learning (Artetxe & Schwenk, 2019). mRASP (Lin et al., 2020) use for pretraining of the multilingual model and add a randomly aligned substitution loss that aims to bring words and phrases closer in the cross-lingual space.

B. Zhang et al. (2020a) explores the main problems that arise for such models: multilingual NMT usually underperforms bilingual models (Arivazhagan, Bapna, Firat, Lepikhin, et al., 2019), the larger the number of languages gets, the more the performance drops (Aharoni et al., 2019), languages in datasets used for multilingual training are unbalanced in size, and poor zero-shot performance compared to pivot models (cf. §4.3.3). B. Zhang et al. (2020a) addresses these problems with a language-aware input layer, a deep transformer architecture (Q. Wang et al., 2019), and an online back-translation

approach. These modifications boost zero-shot translation performance for multilingual models.

To improve the problem of imbalanced and linguistically diverse training data, mostly heuristic methods have been proposed: Arivazhagan, Bapna, Firat, Lepikhin, et al. (2019) samples training data from different languages based on a data size scaled by temperature term. These heuristics impact performance and ignore other factors that are not size. Oversampling of data is used by Johnson et al. (2017); Neubig & Hu (2018); Conneau & Lample (2019). X. Wang et al. (2020) proposes a differentiable data selection method that automatically learns to weight training data, optimizing translation in all languages.

Multilingual modeling Sharing all parameters except for the attention mechanism shows improvements compared with sharing everything in an RNN NMT model (Blackwood et al., 2018). Sachan & Neubig (2018) explores parameter sharing in the transformer architecture for the decoder in the one-to-many translation setting and shows that transformers are more suitable than RNNs for this task. Also, parameter sharing in the decoder and embedding layer further improves performance. Lu et al. (2018) proposes a shared layer to capture interlingua knowledge and extend the typical RNN network with multiple blocks and a trainable routing network. The routing network enables adaptive collaboration by dynamic sharing of blocks conditioned on the task at hand, input, and model state (Zaremoodi et al., 2018). Furthermore, a contextual parameter generator is proposed to accept source and target language embeddings as input and generate the encoder and decoder parameters, respectively (Platanios et al., 2018). B. Zhang et al. (2020a) proposes a language-aware layer to improve such architectures further. With a similar idea, C. Zhu et al. (2020) incorporates two special language embeddings into the self-attention mechanism. The first encodes the unique characteristics of each language, while the second captures common semantics across languages.

One problem in multilingual NMT systems is the translation into the wrong language. To address this problem, B. Zhang et al. (2020b) add a language-aware layer normalization and a linear transformation that is inserted between the encoder and the decoder to induce a language-specific translation. Raganato et al. (2021) explore to weight the target language label by jointly training one cross-attention head with word alignments.

Other modifications of NMT model architectures to improve their performance on low-resource languages include label smoothing (Szegedy et al., 2016), deep RNNs (Miceli-Barone et al., 2017), normalization layers (Ba et al., 2016), direct lexical connections

(D. Q. Nguyen et al., 2015), word embedding layers conducive to lexical sharing (X. Wang et al., 2019), and pointer-generator layers (Z. Yang et al., 2020).

A.1.2. Extended Multi-task training

S. Zhou et al. (2019) uses this approach but extends it with a cascade architecture: the first decoder reads the encoder, the second decoder reads the encoder, and the first decoder (Niehues et al., 2016; Anastasopoulos & Chiang, 2018). The auxiliary task (first decoder) is a denoising decoder. With RNN NMT architectures, one can further decide if the attention mechanism should be shared among tasks (Niehues & Cho, 2017). The authors compare all architectures and find that they perform similarly, with only sharing the encoder being slightly better.

Using linguistic information as an auxiliary task has yet to be explored exhaustively. Niehues & Cho (2017) studies the usage of part-of-speech (POS) and named entity (NE) tags, finding that training on named entity recognition (NER), POS, tagging, and MT together improves performance the most. (Feng et al., 2020) explores POS tagging and chunking as auxiliary tasks with three variants: independent decoders, sharing partial gated units of two decoders, and a two-head model with fully shared parameters. For agglutinative languages, auxiliary morphological tasks can be beneficial: Pan et al. (2020) uses stemming with fully shared parameters.

As an alternative to linguistically informed auxiliary tasks Srinivasan et al. (2019) uses multiple BPE vocabulary sizes to generate different segmentations. Each segmentation is treated as an individual task.

A.1.3. Data augmentation

Back-translation Caswell et al. (2019) shows that adding a special tag to the synthetic data improves performance. A technique that exploits this idea is training an initial translation model with synthetic data generated via BT and then finetuning it with gold data (Abdulmumin et al., 2019). This simple yet effective training algorithm improves NMT for LRLs; however, it can also degrade performance on HRLs if trained without a tagging strategy (Marie et al., 2020).

Multiple improvements of BT have been proposed. Repeating BT multiple times can improve the resulting system's quality Hoang et al. (2018). (Edunov et al., 2018) shows that sampling or noisy beam search can generate more effective pseudo-parallel data. However, for LRLs an optimal beam search and greedy decoding are better. A

factor influencing BT's effectiveness is the quality of the initial MT systems (Hoang et al., 2018). Using back-translated data from multiple sources (Poncelas et al., 2019) or optimizing the ranking of back-translated data yields further gains (Soto et al., 2020).

BT results in gains when the parallel corpora are naturally occurring text and not translationese, as the latter would only improve automatic n metrics (Toral et al., 2018; Graham et al., 2020). Additionally, translationese and original data can be modeled as separate languages in a multilingual model (Riley et al., 2020). BT is also a central part of unsupervised MT (UMT; cf. §4.3.4) and zero-shot MT (Gu et al., 2019).

Sentence modification J. Zhu et al. (2019) proposes to replace a randomly chosen word in a sentence with a *soft-word*. That means that, instead of sampling a word from the lexical distribution of a LM like Kobayashi (2018), the authors use the hidden state vector of the LM directly. X. Wu et al. (2019) substitutes the RNN LMs from previous work and use BERT (Devlin et al., 2019) – a transformer trained with a masked language modeling objective – instead. The authors finetune BERT with a conditional masked language modeling objective that tries to avoid the prediction of words that do not correspond to the original sentence meaning. Such a task performs masks randomly on some of the tokens from a labeled sentence, and the objective is to predict the original vocabulary index of the masked word based on both its context and its label. However, there are two drawbacks to such an approach: we must have enough labeled data to perform the fine-tuning step, hindering it from being applicable in the low-resource setting, and an LM that is not connected to the final task, leaving it fixed and sub-optimal. (Z. Hu et al., 2019) proposes joint training on MT and language modeling. All these approaches require either existing pretrained models or sufficient monolingual data to generate a good LM. This is not always possible for resource-poor or endangered languages. (Sánchez-Cartagena et al., 2021) found that changing the word order and feeding this information as multi-task training improves low-resource MT.

Another way to augment MT data is by paraphrasing. A good paraphrase system can increase the number of training instances (J. E. Hu et al., 2019). Paraphrasing can also be used at training time by sampling paraphrases of the reference sentence from a paraphrase and training the MT model to predict the distribution of the paraphraser (Khayrallah et al., 2020). Again, this helps the model to generalize. Wieting et al. (2019) propose a similar approach, using minimum risk training to optimize BLEU. In addition, they use paraphrasing to diversify the given reference to avoid BLEU's constraints to a specific reference.

Finally, existing data can be augmented by adding noise. This noise can be continuous or discrete. When applying continuous noise, noise vectors are added to the word embeddings (Cheng et al., 2018; Sano et al., 2019). Discrete noise is realized by inserting, deleting, or replacing words, BPE tokens, or characters to expand the training set in an adversarial fashion (Belinkov & Bisk, 2018; J. Ebrahimi et al., 2018; Cheng et al., 2019, 2020).

Pivoting While it is simple to implement and effective, pivot-based approaches suffer from error propagation. To overcome that for NMT, joint training H. Zheng et al. (2017); Cheng (2019) and round-trip training (Ahmadnia & Dorr, 2019) have been proposed.

Pivoting with NMT systems has been used for translating Japanese, Indonesian, and Malay into Vietnamese (Trieu et al., 2019), translation of related languages (Pourdamghani & Knight, 2019), multilingual zero-shot MT (Lakew et al., 2018), and UMT (cf. §4.3.4) between distant language pairs (Leng et al., 2019).

A.1.4. Recent low-resource Shared Tasks

First, the LoResMT 2020 shared task (Ojha et al., 2020) explores the case of language pairs that have no parallel data between them (Hindi–Bhojpuri, Hindi–Magahi, and Russian–Hindi). The winning system (Laskar et al., 2020) uses a MASS model in a zero-shot fashion with additional monolingual data (see §4.3.4). Second, the WMT 2020 shared tasks on UMT and very low-resource supervised MT (Fraser, 2020) provide text and 60k aligned phrases for German–Upper Sorbian., The most important technique in all tracks is transfer learning, achieving surprisingly good results. For the AmericasNLP 2021 shared task on open MT (Mager et al., 2021), 10 indigenous language languages were paired with Spanish, resulting in an extremely low-resource setting (4k to 125k paired sentences), with challenges out as domain, dialectical, and orthographic mismatches between splits and datasets. The best systems show that data cleaning and collection (§4.2) as well as multilingual approaches (§4.3.1) result in the best performance in these conditions. Finally, the shared task on MT in Dravidian languages (Chakravarthi et al., 2021) features 3 languages paired with English as well as Tamil–Telugu. Again, the winning system uses a multilingual approach. The best performing systems use BT (§4.3.3) and BPE word segmentation (§2.4.1).

The results from these challenges indicate that the optimal selection and combination of methods differ between cases (i.e., amount of monolingual, parallel data, cleanness of data, domain mismatch, the linguistic closeness of languages). This implies that data

analysis and linguistic knowledge are needed to improve a final system's performance.

A.1.5. Transfer learning

This helps low-resource tasks as less data can be used for training. One application of transfer learning to MT is the usage of a pretrained RNN LM (Gulcehre et al., 2015) as the decoder in an NMT system. Zoph et al. (2016) is the first work that uses pretrained models to improve NMT systems. The authors perform two experiments with an RNN encoder–decoder architecture with an attention mechanism: the model is first pretrained on a high-resource language pair – French–English or German–English – and then finetuned on LRLs. The resulting model is known as a child model. This works even better if related languages are used during pretraining (T. Q. Nguyen & Chiang, 2017). Using pretrained LMs at the decoding time and as priors at training time also improves vanilla models (Baziotis et al., 2020). With the advancement of multilingual MT models, the initial seed can be one of those models, which is then finetuned on the low-resource languages Neubig & Hu (2018). To avoid overfitting, models can be finetuned on both an HRLs pair and an LRLs pair in a multi-task fashion (Neubig & Hu, 2018). The authors report a 1.7 BLEU improvement over the zero-shot 15.5 BLEU of the multilingual model.

However, how can we represent the vocabulary best? Zoph et al. (2016) use separate embeddings for the source and the target language. However, using tied embeddings has been shown to yield better results (Press & Wolf, 2017). Edunov et al. (2019) employs ELMO (Peters et al., 2018) representations as pretrained features in the encoder of a transformer model. A pretrained LM is used as a decoder, and the entire system is finetuned on parallel data, achieving important gains in the low-resource setting. One of the challenges for low-resource languages is the lack, in many cases, of a large amount of monolingual data that can be used for training. H. Song et al. (2020) shows that it is possible to improve performance by combining monolingual texts from linguistically related languages and performing a script mapping. It is also possible to extract features from a BERT model in the source language and combine these with an NMT system (J. Zhu et al., 2020), but using a BERT model pretrained with mixed sentences from source and target languages leads to even better results (Xu et al., 2021).

Encoder-decoder pretrained models have gained popularity in the last few years for low-resource MT. Conneau & Lample (2019) proposes training the encoder and the decoder separately in order to get cross-language representations (XLM). The method includes a shared sub-word level vocabulary and training masked language models for

both languages. This can be finetuned to a translation language model, where parallel data can be fitted together with a boundary token. This idea has further been extended by K. Song, Tan, et al. (2019, MASS) to mask a *sequence* of tokens from the input. Training MASS in a multilingual fashion and using monolingual data for pretraining helps to improve NMT for low-resource languages and zero-shot translation (Siddhant et al., 2020). Another approach is to train the entire transformer model as a denoising autoencoder (BART; M. Lewis et al., 2019). The multilingual version of BART (mBART) is more suitable for NMT tasks and yields important gains (Y. Liu et al., 2020). It is also possible to pretrain a transformer in a multi-task, text-to-text fashion, where one of the tasks is MT (T5; Raffel et al., 2020). One advantage is that the model does not rely on a single masking function so various techniques can be used. All four models can be finetuned for MT or used in an unsupervised fashion. Improvements to BART can be obtained by augmenting the maximum likelihood objective with an additional objective, a data-dependent Gaussian prior distribution (Li et al., 2020). Huge LMs can improve zero-shot and few-shot learning even further (Brown et al., 2020), but at a high computational cost. Pursuing another direction, (L. Wang et al., 2019) develops a hybrid architecture between a transformer and a pointer-generator network. At training time, the authors jointly train the encoder and the decoder in a denoising auto-encoding fashion. Also, forcing the pretraining step to be multilingual shows some gains. Finally, Z. Yang et al. (2020) uses a synthetic code-switching dataset for pretraining to improve their model's cross-lingual capabilities.

One crucial problem for transfer-learning is minimizing catastrophic forgetting (Serra et al., 2018). Chen et al. (2021) show that it is possible to combine a pre-trained multilingual model by fine-tuining it with one single language pair, to improve zero-shot machine translation. Another way to handle this problem is by reducing the number of parameters to be updated. Gheini et al. (2021) propose to only update the cross attention parameters.

Applications of these methods on non-NMT tasks include summarization (L. Wang et al., 2019) where they propose a hybrid architecture of a transformer seq2seq and a pointer-generator network. BART has also been applied for domain adaption (Jin et al., 2020) using back-translation.

A.1.6. Unsupervised MT

The addition of other components, such as masked LMs and denoising auto-encoding, has also been tried Stojanovski et al. (2019). Unsupervised methods are vulnerable

to adversarial attacks of word substitution and order change in the input. Adversarial training can improve performance in such situations (Sun et al., 2020). Since the initialization step is crucial for UMT, (Ren et al., 2020) aligns semantically similar sentences from two monolingual corpora with the help of cross-lingual embeddings. Then they train an alignment model that is used to delete non-aligned words between those sentences. At the end, a rewriting encoder–decoder transformer architecture is trained with monolingual text, using word dropouts and random swaps. Finally, these models improve the synthetic parallel data, rewriting the retrieved sentences. With these, an SMT system is trained to warm up an NMT system.

However, UMT still has to overcome a set of challenges. Søgaard et al. (2018) shows that performance decays dramatically for languages with different typological features, such as morphology or word order, and monolingual texts from different domains: translation from English to Finnish achieves only 0.09 BLEU, compared to 32.71 BLEU for English to Turkish. since, in such situations, bilingual word embeddings (Conneau et al., 2017) are far from isomorphic.

Vulić et al. (2020) finds that isomorphism is also less likely if small amounts of monolingual data are used for training bilingual word embeddings. Nooralahzadeh et al. (2020) discovers that performance quickly deteriorates for a source and target domain mismatch and that the initialization of word embeddings can affect MT performance. All of this makes UMT for LRLs or endangered languages challenging. But the perhaps most important question for UMT is if real-world cases exist where a low-resource language has a huge amount of monolingual data and no parallel corpora Artetxe et al. (2020). Linguistic dissimilarity and domain mismatch are common in low-resource scenarios. Therefore, for resource-poor languages, supervised and semi-supervised models perform better than UMT Kim et al. (2020). This is especially important because, thanks to the Bible, the only existing monolingual available for many languages is one that we can find in parallel corpora.

Some of the described issues have been addressed: Z. Liu et al. (2019) proposes combining word- and subword-level embeddings to account for morphological complexity. It uses BPE for one NMT system and a second NMT system with word-based embeddings. Then a language model re-ranks the output of both systems. The experiments were performed in Czech and German. For the problem of distant language pairs, Leng et al. (2019) proposes pivoting (cf. §4.3.3). Isomorphism of bilingual word embeddings can be improved with semi-supervised methods (Vulić et al., 2019).

Garcia et al. (2020) introduces multilingual UMT systems. The main idea consists

of generalizing UMT into a multilingual approach using the multi-way back-translation objective. Also, the usage of an auxiliary translation task uses parallel data for one direction but not for the other. This is done by proposing a cross-translation loss term from the initial probabilistic framework that enforces cross-language pair consistency cross-translation of the loss term. Recently, pretrained multilingual transformer networks have been used to improve further UMT (cf. §4.3.4).

B. Complete survey and answers of the MT ethics study

B.1. Questionnaire

The following questions, together with their answers are listed in this section.

1. **How do you consider yourself**
   a) Leader of your town or community
   b) Language activist
   c) None of the above
   d) Other, explain
2. **To what degree do you speak your mother tongue (indigenous / native / native)**
   a) Perfectly
   b) Fairly good, but with shortcomings
   c) I speak with difficulties
   d) I don't speak it but I do understand
   e) I am not a speaker of the language, but I know some words
3. **What people / tribe / nation do you belong to?**
   a) Open Question
4. **In your town / nation / tribe there is some restriction to share your language with people outside your community**
   a) Yes
   b) No

c) It depends

Comment the question

5. **On a scale from 1-5, how dangerous do you think an MT system would be?**
   - 1 - Not dangerous
   - 5 - Dangerous

   Comment the question

6. **What advantages would you see with an automatic translation system**
   - Open ended

   Comment the question

7. **What dangers would you see in a machine translation system:**
   - Open ended

   Comment the question

8. **What topics would you see as positive in which automatic translation can be carried out (there may be several options)**
   a) Medicine and health
   b) Medicinal plants
   c) Religion and sacred songs
   d) Tales and literature and non-sacred songs
   e) Laws
   f) Commercial matters
   g) Talk everyday
   h) Science and education
   i) Culture and traditions
   j) others

9. **What would you see as damaging topics that should not be machine translated?**

- Open question

10. **What would you think of a system with poor translation quality?**
    - Not to be used or made available to the public
    - It would be good, to see the progress and be able to help correct it
11. **To create translation systems we need data (text in the native language aligned with text in the foreign language). These data should:**
    a) Be public and without restrictions
    b) Be owned by an organ of the community or town.
    c) Be the property of the research group that compiled it
    d) Be the property of the speakers who made the translation
12. **In a machine translation project**
    a) Any team of researchers in the world can produce systems
    b) It is preferable to integrate community members
    c) Only community members should work on it
    d) Only with express permission of the community body
    e) These investigations should not be carried out

B.2. Complete answers of the open questions

All questions refer to the ones enumerated in appendix B.1. In table B.4 we present the complete results for the open question 6. In table B.1 we present the complete results for the open question 7. In table B.2 we present the complete results for the open question 9.

	Original	English Translation
	Qué ventajas tendría un sistema de traducción automática para su idioma?	**What dangers would you see in a machine translation system (for indigenous languages):**
1	Sería practico pero confuso, por las variantes (aprox. mas de 30 variantes)	It would be practical but confusing, because of the variants (approx. more than 30 variants)
2	Facilitaría el aprendizaje para los ajenos a la lengua y entender cada uno de los nativos la forma estructura de su escritura.	It would facilitate learning for those foreign to the language and for each of the natives to understand the structure of their writing.
3	Fácil acceso	Easy access
4	La unificacion de la lengua	The unification of the language
5	La ventaja de que las personas que no lo hablan, puedan comunicarse con la gente monolingüe	The advantage that people who do not speak it can communicate with monolingual people
6	Ninguna.	None
7	La dispersión del aprendizaje de este	The dispersion of the learning of this
8	Tal vez un aporte de modernidad a la comunidad, preservacion de la lengua nativa.	Perhaps a contribution of modernity to the community, preservation of the native language.
9	Con respeto a los derechos lingüísticos su difusión pertinencia	With respect to linguistic rights its dissemination relevance
10	Facilitaría la enseñanza porque se tendría muchas herramientas de apoyo.	It would facilitate teaching because it would have many support tools.
11	Conocer una aproximación de los significados	Know an approximation of the meanings
12	Ayudaría a las personas que no hablan o ya no hablan el idioma Popti', ayudaría a los niños en las escuelas para aprender el idioma.	It would help people who don't speak or no longer speak the Popti' language, it would help children in schools to learn the language.
13	Contribuiría al uso social de nuestro idioma	It would contribute to the social use of our language
14	El uso de traductores automáticos en espacios como hospitales, oficinas de gobierno, etc.	The use of automatic translators in spaces such as hospitals, government offices, etc.
15	Contribuiría al Status de las lenguas indígenas y quizá también Corpus.	It would contribute to the Status of indigenous languages and perhaps also Corpus.
16	Solo seria útil para conocer vocabulario y pequeñas frases quizás.	It would only be useful to learn vocabulary and small phrases perhaps.
17	hacer la lengua mas visible y más accesible a personas fuera de la comunidad de habla	make the language more visible and more accessible to people outside the speaking community
18	Mantener la lengua a través del tiempo y la distancia	hold the tongue through time and distance
19	Mucha ventaja para las personas que han estado interesados en aprender este idioma.	Much advantage for people who have been interested in learning this language.
20	nos ayudaría a normalizar la presencia del kichwa en un medio actual de comunicación.	It would help us to normalize the presence of Kichwa in a current means of communication.
21	These are great for interview videos or audio. I don't see anything wrong with this.	
22	Ayudar aprender nuestra lengua	Help learn our language

Table B.1.: Answers for the open question 6 (see §B.1). All translations are automatic translations with human editing. If the original is in English, we leave the translation field blank.

	Original	English Translation
	Podría usted commentar sobre algún posible problema que podría generar un sistema de traducción automática para su idioma?	**What dangers would you see in a machine translation system**
1	Por la variedad de la Lengua, algunas palabras significan diferente, no es recomendable un diccionario estandar para la lengua, si por variante.	Due to the variety of the language, some words mean differently, a standard dictionary for the language is not recommended, if by variant.
2	Los conceptos culturales. En su defecto hay conceptos lingüísticos culturales que solo se entienden en la lengua nativa. De igual forma, encontraremos conceptos ajenos a nuestra lengua que no son traducibles.	Cultural concepts. Failing that, there are cultural linguistic concepts that are only understood in the native language. In the same way, we will find concepts foreign to our language that are not translatable.
3	Falta de recursos de información	Lack of information resources
4	La gente de cada región se aferra a su propia variante	The people of each region cling to their own variant
5	No se puede entender estrictamente como peligroso, pero si un poco difícil porque existen muchas variantes	It cannot be strictly understood as dangerous, but it is a bit difficult because there are many variants
6	La desnaturalización del idioma. Al tratarse de un idioma donde no imperan los sustantivos, se carece de artículos y de algunas preposiciones, además de ser aglutinante, el trabajo es mucho más complicado que en otros idiomas. Además, la existencia de tantas variantes haría que el proyecto fuera poco o nada rentable y llevaría a los "expertos" a un intento por homologar el habla, lo cual sería un error tremendo.	The denaturation of the language. Being a language where nouns do not prevail, articles and some prepositions are lacking, in addition to being agglutinative, the work is much more complicated than in other languages. In addition, the existence of so many variants would make the project unprofitable or unprofitable and would lead the "experts" to attempt to standardize speech, which would be a tremendous mistake.
7	Ayudaría a el aprendizaje de la lengua.	It would help to learn the language.
8	Ninguna, de hecho tenemos avances y una comunidad internacional(Peru, Bolivia, Argentina y Chile) de aymara formada con este objetivo.	None, in fact we have progress and an international community (Peru, Bolivia, Argentina and Chile) of Aymara formed with this objective.
9	Su comercialización	the commercialization
10	Hay algunas palabras que no se pueden traducir, porque al hacerlo pierde el sentido, además de que esta lengua tiene otras grafías que están ausentes en el sistema consonántico y vocálico del español (que es como la referencia de todo lo que se hace). Además de que la lengua O'dam es aglutinante, las correspondencias entre palabras u oraciones no siempre será uno a uno porque las palabras en O'dam son largas y puede significar toda una frase. Es una lengua aglutinante pues.	There are some words that cannot be translated, because by doing so they lose their meaning, in addition to the fact that this language has other spellings that are absent in the consonantal and vowel system of Spanish (which is like the reference for everything that is done). In addition to the fact that the O'dam language is agglutinative, the correspondences between words or sentences will not always be one to one because the words in O'dam are long and can mean a whole sentence. It is an agglutinative language.
11	Tomar en cuenta la cultura, existen elementos culturales que deben tomarse en cuenta	Take culture into account, there are cultural elements that must be taken into account
12	Ninguno.	None
13	Que no sea pertinente cultural y lingüísticamente	Not culturally and linguistically adequate

Table B.2.: Answers for the open question 7 (see §B.1). All translations are automatic translations with human editing. If the original is in English, we leave the translation field blank.

14	Que por "practicidad" se priorice el uso de traductores automáticos en lugar de aprender la lengua.	That for "practicality" the use of automatic translators is prioritized instead of learning the language.
15	Podrían sin lugar a dudas desvirtuar el adecuado uso del idioma. El quechua es un idioma contextual.	They could undoubtedly distort the proper use of the language. Quechua is a contextual language.
16	la pregunta anterior es muy ambigua, peligroso en qué sentido, para sus hablantes? para el linguista? ... un problema que podría generar es que no toda la comunidad de habla esté de acuerdo en poner la langua al alcance de todos	the previous question is very ambiguous, dangerous in what sense, for its speakers? for the linguist? ... a problem that could arise is that not the entire speech community agrees to make the language available to everyone
17	Lo primero es que no hay un teclado especial	The first thing is that there is no special keyboard
18	Sería un poco difícil de realizar un translate automático del chol, ya que es un lengua compleja y a veces solamente se entiende viendo, estando o conociendo el contexto del habla.	It would be a bit difficult to translate Chol automatically, since it is a complex language and sometimes it is only understood by seeing, being or knowing the context of the speech.
19	ninguno	none
20	siempre y cuando la comunidad este involucrada.	as long as the community is involved.
21	As long as the community owns the language corpus then I think it would be helpful. I have concerns of other entities, academic or commercial, being involved in handling the corpus.	
22	Entes externos podrían aprovecharse económicamente	External entities could take advantage economically

Table B.3.: Answers for the open question 7 (see §B.1). All translations are automatic translations with human editing. If the original is in English, we leave the translation field blank.

	Original	English Translation
	Qué temas vería usted cómo perjudiciales en los que no se debería realizar traducción automática para su lengua?	**What would you see as damaging topics that should not be machine translated?**
1	Ninguno	None
2	Leyes, medicina y salud, ciencia, cuestiones mercantiles, religión y canciones sagradas,	Law, medicine and health, science, business matters, religion and sacred songs,
3	Derecho	Laws
4	Las groserías	the bad words
5	No veo problema alguno en la traducción de todos lados temas	I do not see any problem in the translation of all sides topics
6	ninguno	None
7	Ninguno.	None
8	Ninguna, pero hay muchas versiones de aymara y con influyencia de castellano.	None, but there are many versions of Aymara and with Castilian influence.
9	Temas que atenten contra la vida orgánica	Issues that threaten organic life
10	Lo sagrado (discursos rituales, cantos, etc.) por eso de que las palabras en el caso del Mucha fuerza al pronunciarse razón por la que no deben decirse fuera de su contexto.	The sacred (ritual speeches, songs, etc.) That is why the words in the case of Much force when pronounced, reason why they should not be said out of context.
11	Significados culturales	cultural meanings
12	Religión occidenta	western religion
13	Religión	Religion
14	Conocimientos considerados sagrados.	Knowledge considered sacred.
15	Las traducciones automáticas no son completamente válidas ni aún en los idiomas modernos.	Machine translations are not completely valid even in modern languages.
16	ningun tema	no topic
17	Situaciones políticas y religiones a menos que sea del interés de la persona	Political situations and religions unless it is in the interest of the person
18	Los cantos sagrados, como los de un curandero.	The sacred songs, like those of a healer.
19	ninguno	none
21	Medicina y Salud, Plantas medicinales, Religión y canciones sagradas	Medicine and Health, Medicinal plants, Religion and sacred songs
20	Anything ceremonial	
22	curaciones y cuestiones personales	cures and personal issues

Table B.4.: Answers for the open question 9 (see §B.1). All translations are automatic translations with human editing. If the original is in English, we leave the translation field blank.

List of Figures

List of Tables