Corpus Creation and Initial SMT Experiments between Spanish and Shipibo-konibo

Ana-Paula Galarreta[1], **Andrés Melgar**[2] and **Arturo Oncevay-Marcos**[2]
[1]Departamento de Ciencias, [2]Departamento de Ingeniería
Grupo de Reconocimiento de Patrones e Inteligencia Artificial Aplicada
Pontificia Universidad Católica del Perú, Lima, Perú
{a.galarreta,amelgar,arturo.oncevay}@pucp.edu.pe

Abstract

In this paper, we present the first attempts to develop a machine translation (MT) system between Spanish and Shipibo-konibo (es-shp). There are very few digital texts written in Shipibo-konibo and even less bilingual texts that can be aligned, hence we had to create a parallel corpus using both bilingual and monolingual texts. We will describe how this corpus was made, as well as the process we followed to improve the quality of the sentences used to build a statistical MT model or SMT. The results obtained surpassed the baseline proposed (dictionary based) and made a promising result for further development considering the size of corpus used. Finally, it is expected that this MT system can be reinforced with the use of additional linguistic rules and automatic language processing functions that are being implemented.

1 Introduction

In Perú, there are around 47 indigenous or original languages according to the database of the Ministry of Culture (Ministerio de Cultura, Perú, 2016). Besides the official Spanish (Castillan) language, which is spoken by the majority of the population, there are other native languages that have remained throughout history in the different regions of the country. For instance, in the highlands, there are a lot of speakers from the Quechua family, followed by Aymara. Furthermost, in the Amazon region, there is a high density presence of many linguistic families between the different native communities, where Shipibo-konibo is one of the most studied languages by linguists at Perú (Valenzuela, 2003). Besides, nowadays there are a lot of efforts from the government in order to accelerate the integration of the communities that don't speak Spanish as the main language.

In this context, there is a need for computational resources to facilitate the accomplishment of tasks for the social inclusion of this native communities. As first steps, there are basic natural language processing (NLP) tools required for morphological or syntax analysis that are under development (Pereira et al., 2017). However one of the most important applications needed is a machine translation (MT) system, at least a prototype one.

MT systems may be rule (RBMT) or corpus driven (SMT). Also, there is an hybrid approach that integrates both of them (Costa-Jussa and Fonollosa, 2015). The RBMT system is based on the linguistic knowledge and could be very expensive to build in time and effort, while the SMT relies on the amount of parallel corpus available in order to obtain consistent results. Besides, there is a novel corpus-driven approach called Neural Machine Translation (NMT), but it could be less effective in low-resource scenarios unless there is another (rich-resourced) language involved as a pivot (Zoph et al., 2016).

In this context, it may seem that a SMT is not an appropriate approach for minority languages or with scarce digital resources. However, there have been different studies trying to take advantage of any small corpus they could. For instance, there were attempts for many pair of languages (involving a low resourced one) such as English-Lao, Myanmar and Thai (Pa et al., 2016); English-Estonian, Hungarian, Latvian, Lithuanian and Polish (Skadiņš et al., 2014) and also Persian-English (Salami et al., 2016).

While those studies use corpus with more than 50 000 sentences (for example, Skadiņš et al. (2014) obtained a BLEU score of 59.70 using a 0.5M en-hu corpus), our work will use at most the fifth part. Nevertheless, this does not stop

Proceedings of Recent Advances in Natural Language Processing, pages 238–244,
Varna, Bulgaria, Sep 4–6 2017.
https://doi.org/10.26615/978-954-452-049-6_033

this study and experiment, since the final goal is to be able to integrate the SMT system to a platform linked with more linguistic resources, taking advantage of rules, and other NLP functionalities such as POS-tagging or dependency parsing (Costa-Jussa and Fonollosa, 2015). In this sense, the individual results obtained were very promising considering the size of corpus used.

This study is organized as follows. Section 2 presents the Shipibo-konibo language. After that, Section 3 details the procedure followed to build our parallel corpus. Then, the experiment design is described in Section 4 and the obtained results are discussed in Section 5. Finally, we conclude the paper in Section 6, including future work proposals.

2 The Shipibo-Konibo Language

Shipibo-konibo (shp) is one of the most representative native languages in Perú, behind the Quechua language family, Aymara and Ashaninkas. It belongs to the Panoan family, which is an important subject of study of many linguist researchers in Perú (Adelaar et al., 2011; Zariquiey, 2006). This language is spoken by around 22k people in 150 communities, and is taught in almost 300 public schools (Ministerio de Educación, Perú, 2013b).

Shipibo-konibo is an agglomerative language, with a high use of common suffixes (130) plus some prefixes (13) for its words formation process. Thereby, this language has a very rich morphology, which increase the difficulty in an SMT task, because each variation of a word (flexed by an affix) would be taken into account as a completely new word. Also, the basic sentence order construction is SOV (subject-object-verb) unlike Spanish (SVO) (Valenzuela, 2003). An example of a translation (to English) could be seen in Figure 1.

3 Corpus Creation Procedure

The followed procedure consists mainly in the selection, preparation (that could includes digitalization, manual correction and translation) and alignment of a parallel corpus from different sources. We take into account two main domains for the corpus: religious and educational, and it is available in the project site[1]. The next subsections will

[1] chana.inf.pucp.edu.pe/resources

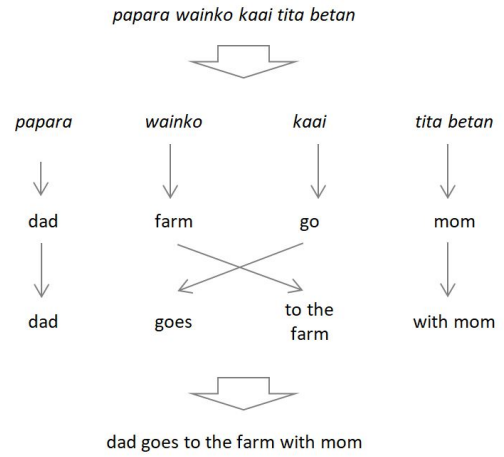

Figure 1: A translation sample between Shipibo-Konibo and English

detail the generation of the corpus and its later analysis at word level.

3.1 Generation of Parallel Corpus

In order to develop a SMT model, a large amount of parallel corpus is needed (Koehn, 2009). This kind of corpus is a very valuable resource for machine translation tasks but at the same time is a prohibitive element for low-resourced languages. In our case, Shipibo-konibo, as a minority language in Perú, has a very poor collection of digital text corpus, even monolingual ones. So it has been very difficult to identify texts translated between that language and Spanish.

Despite that, we decided to carry out an exhaustive search as a first step. In this context, we obtained a Shipibo-konibo to Spanish dictionary (which included translated sentences), laws, law proposals and the catholic Bible in both languages. However, the amount of parallel sentences obtained from the legal domain were too limited (1 142) and it was very difficult to increase the corpus size of that domain, due to the length of sentences and complexity of words used to write laws in Spanish.

The dictionary (James et al., 1993) included 5 143 translated sentences used to illustrate the meaning of the different terminology entries. Since each entry is different, we expected to have several out of vocabulary words (OOVW) when splitting the corpus into train-tune-test. For this reason, instead of using the dictionary examples to create a parallel corpus, we decided to elaborate rules from the entries in Shipibo-konibo and their

Table 1: Religious domain corpus: Token count per sentence

	shp		es	
max number of words	number of sentences	% of sentences	number of sentences	% of sentences
5	1 723	12.7	1 953	14.4
10	4 672	34.4	4 428	32.6
15	7 713	56.8	7 225	53.2
20	10 101	74.3	9 715	71.5
25	11 692	86.1	11 530	84.9
30	12 570	92.5	12 594	92.7
35	13 108	96.5	13 156	96.8

Table 2: Educational domain corpus (kindergarten book): Token count per sentence

	shp		es	
max number of words	number of sentences	% of sentences	number of sentences	% of sentences
5	983	49.9	523	14.4
10	1 595	84.3	1 224	64.7
15	1 742	92.1	1 565	82.8
20	1 798	95.1	1 691	89.4
25	1 828	96.7	1 754	92.8
30	1 848	97.7	1 793	94.8
35	1 862	98.5	1 826	96.6

correspondent translation. This procedure was tedious because only a scanned version of the dictionary was available.

In order to obtain the rules mentioned above, an OCR function was applied to the dictionary. However, the obtained text had several errors, which meant we had to correct it manually (with help of linguistic and engineering students). Finally, using information about the structure of the dictionary, we were able to obtain 13 783 rules to translate words directly from Shipibo-konibo to Spanish.

Regarding the religious domain, we were able to automatically obtain more than 10 000 phrases. Even though this corpus can not be enlarged or renewed we decided to run tests with it, since no one has analyzed the Bible in Shipibo-konibo before.

On the other side, because one of the main motivations to implement the translator is to use it in an educational context (to support the generation of bilingual education textbooks), we decided to translate educational text from Shipibo-konibo to Spanish and vice versa.

In the next subsections, we will detail the contents of each domain.

3.1.1 Religious Domain

The Bible is a common parallel corpus used in many experiments, regarding its limited size, due to the availability of many languages translations (Christodouloupoulos and Steedman, 2015). Using both Shipibo-konibo and Spanish Bibles, we semi-automatically aligned 9 804 versicles.

It is important to mention that not all of the Bible's books have been translated to Shipibo-konibo and that this translation may differ from the Spanish one. We identified several differences and had to make manual corrections in order to align the versicles. This is the case of chapter 3 of Joel's book (es), which was aligned with the last versicles of chapter 2 (shp).

Then, each versicle was split using specific punctuation signs (dots, colons, question and exclamation marks). For example, when Apocalypse 15:3 was split, 3 sentences were automatically obtained per language (Shipibo-konibo and Spanish):

shp: *[1] Jatianra ja Diossen yonoti Moiséssen bewá, itan *Corderon bewá bewai neskákana iki: [2] "Non Ibo Dios, jatíbi atipana koshi, itan ratéti jawékibo riki jatíbi min akábo; ponté, itan ikon riki min ikábo. [3] Mia riki jatíbiainoa joni-*

baon Apo

es: *[1] Estos cantan el cántico de Moisés, servidor de Dios, y el cántico del Cordero: [2] Grandes y maravillosas son tus obras,Señor Dios, Todopoderoso. [3] Justicia y verdad guían tus pasos, oh Rey de las naciones.*

en: *[1] And they sing the song of Moses the servant of God, and the song of the Lamb, saying, [2] Great and marvellous are thy works, Lord God Almighty; [3] just and true are thy ways, thou King of saints.*

Finally, we cleaned the text by removing all remaining punctuation signs (such as dashes and commas) and 13 587 sentences were obtained.

3.1.2 Educational Domain

An educational book used for kindergarten Shipibo-konibo students (Ministerio de Educación, Perú, 2013a) was translated by a human translator certified by the Ministry of Culture. As a result, 1 466 of the book's paragraphs were translated. Then, we split several paragraphs using punctuation signs. To split a paragraph, it had to fulfill both of the following conditions:

- The number of obtained phrases is the same for both languages.
- For all obtained phrases, the $ratio_{shp-es}$ of tokens was in a range value between 0.5 and 2 (see subsection 3.2.2).

After duplicated sentences were removed, we were left with 1 891 aligned sentences. For example, the following paragraph was split in 2 sentences:

shp: *[1] ¿Jawe janerin min jeman jane?, [2] ¿Jawekeskamein min jema peokotai bena ika iki?*

es: *[1] ¿Cómo se llama tu comunidad?, [2] ¿Cómo se formó tu comunidad al principio?*

en: *[1] What is the name of your community?, [2] How was your community formed in the beginning?*

3.2 Analysis of Parallel Corpora

Since manual validation of the human translations or the automatic alignments was not performed, we decided to obtain some descriptive information of the parallel corpus. Note that for the educational domain, only results from the kindergarten book are shown.

3.2.1 Token Count per Sentence

SMT systems are usually trained with short sentences (Dimeo, 2014), such as 10 tokens at most. In this case, in order to analyze the size of our corpus to know its usefulness, we decided to calculate the number of tokens for both Spanish and Shipibo-konibo sentences with a five-step variation. Results are shown on Table 1 and Table 2.

As we can observe, if we filter our corpus to limit sentence length to 10 tokens in Shipibo-konibo or Spanish, we would be left with less than 35% of total sentences. For this reason, we decided not to limit the sentence length to 10 or less tokens.

3.2.2 Ratio shp-es

We also calculated the Shipibo-konibo to Spanish ratio per sentence, using the following formula:

$$ratio_{shp-es} = \frac{tokens_{shipibo-konibo}}{tokens_{spanish}}$$

Since Shipibo-konibo is an agglutinate language, a sentence in Spanish usually presents fewer words in the Shipibo-konibo translated counterpart. Hence, we expected this ratio to be lower than 1 in most cases and this is true for the educational corpus. However, as seen in Figure 2, in the religious domain that statement does not follow the expected trend. We think this is due to the way the Bible was translated (from Spanish to Shipibo-konibo) and because of the nature of the texts: the missioners that translated it wanted to preserve the meaning of each sentence, so they probably used more words than usual.

3.2.3 Token Frequency

We determined the frequency of each token in both domains and observed that several tokens are *hapax legomenon* (HL) or words that appear only once within the corpus. We can see that information in Table 3.

Table 3: Token-level corpus stats: T = number of tokens; $|\mathcal{V}|$ = word vocabulary size or unique tokens; HLT = number of HL tokens (*hapax legomenon*).

	Religious		Educational	
	es	shp	es	shp
T	215 818	210 828	21 150	14 225
$\|\mathcal{V}\|$	14 386	20 500	2 629	2 793
HLT	6 706	11 898	1 324	1 642
$\%HLT$	3.1	5.6	6.3	11.5

We also calculated the percentage of sentences that contain HL tokens, and those results are presented in Table 4. It was important to identify these HL tokens, since they would only appear in

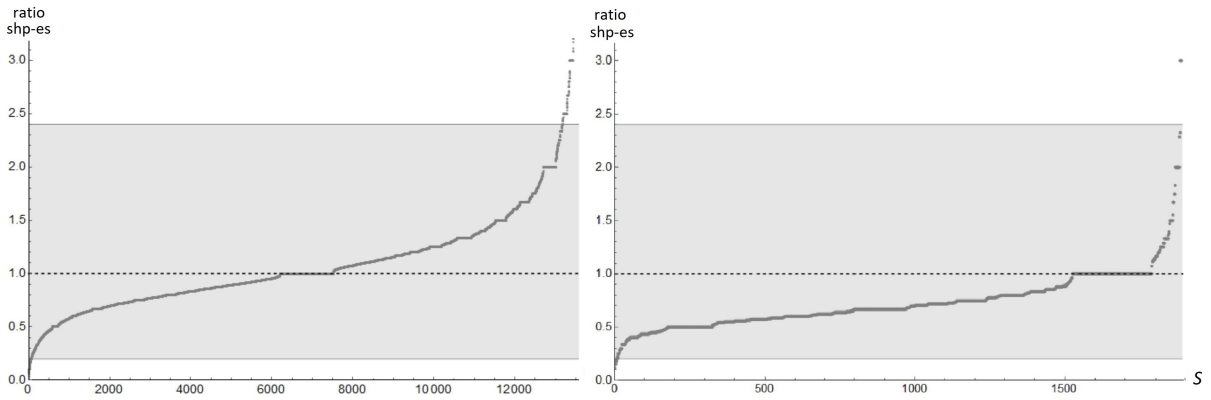

Figure 2: Sentence number (S) versus $ratio_{shp-es}$ for both the religious (left) and educational domain (right) corpus. The 96.6% of religious domain sentences have $ratio_{shp-es}$ values between 0.2 and 2.4, and 99.3% of educational domain sentences have $ratio_{shp-es}$ values between 0.4 and 2.4

one of the sub-datasets (train, validation or test). This means HL tokens present in the train subset would be the OOVW in the test subset, which can affect the overall result of the SMT system.

Table 4: Sentence-level corpus stats: S = total number of sentences or phrases; $SHLT$ = number of sentences or phrases that contains at least one HL token.

	Religious		Educational	
	es	shp	es	shp
S	13 587	13 587	1 891	1 891
$SHLT$	13 411	12 945	1 174	1 732
$\%SHLT$	98.7	95.3	62.1	91.6

4 Experiment

The corpus was split into train, tune/validation and test subsets, using a 80-10-10 proportion and an algorithm that avoids having too many unique tokens in a single subset. The number of sentences for each set is shown in Table 5. We decided not to remove sentences that contain this kind of tokens since, as seen in Table 4, we would be left with less than 10% of the original corpus.

Table 5: Number of sentences in train-tune-test subsets

	Train	Tune	Test	Total
Religious	10 021	1 263	1 263	12 547
Educational	1 429	184	184	1 797

Also, since we could not do a manual verification of all the aligned sentences (in both domains), we decided to remove sentences that might not be correctly aligned, basing this decision in the token count and the $ratio_{shp-es}$.

A small sample of aligned sentences was taken and we observed that long sentences were usually misaligned. So, for both domains, we established a token count threshold of 35, considering that when this value is more strict, the corpus is smaller.

Using the same criteria, we analyzed the quality of the alignment using the $ratio_{shp-es}$. For the religious domain, we only selected sentences with $ratio_{shp-es}$ values between 0.2 and 2.4, and limited the token count to 35. For the educational domain the ratios were between 0.4 and 2.4, and the threshold was set to 35 also.

Then, we trained the SMT using the MOSES platform (Koehn et al., 2007). The texts were translated from Shipibo-konibo to Spanish and the language model was trained using 3-grams followed by some additional operations:

- **Test 1:** Initial experiment with 3-grams.
- **Test 2:** The unknown words (words that were not translated using MOSES) are directly translated (replaced) by the entry in the dictionary.
- **Test 3:** In the translation output, the consecutive duplicated words were removed.
- **Baseline:** Direct application of the dictionary rules in the Shipibo-konibo test file (replace of terms with their direct translation).

Table 6: BLEU scores obtained

Domain	min ratio	max ratio	max length	BLEU-1 score Baseline	BLEU score Test 1	BLEU score Test 2	BLEU score Test 3
Religious	0.2	2.4	35	4.90	4.31	4.32	**4.42**
Educational	0.4	2.4	35	6.99	13.86	13.90	**13.92**

Baseline: Original text translated using rules from dictionary.

BLEU score obtained in both domains is zero. We show BLEU-1 scores.

Test 1: Initial experiment

Test 2: OOVW were automatically translated using the dictionary

Test 3: Duplicated consecutive words were removed

5 Results and Discussion

As we can see in Table 6, the BLEU (Papineni et al., 2002) results obtained using the SMT system in the Educational domain greatly exceed the ones we got using only rules from the dictionary (baseline). A BLEU score of 4.31 was obtained for the religious domain. Besides, despite being trained with fewer sentences, the BLEU score for the educational domain was 13.86.

Also, replacing unknown tokens using dictionary rules (Test 2) and removing duplicated consecutive ones (Test 3) slightly improve the results.

The different scores between both domains may be explained by the sentence length of each corpus. As shown in Tables 1 and 2, 84.3% Shipibo-konibo sentences from the educational domain have 10 or less tokens and only 34.4% of Shipibo-konibo sentences from the religious domain have the same token count.

Another reason may be the simplicity of the educational corpus, since it was obtained from a kindergarten book. This means that all the sentences contains very simple words and belong to a very closed domain. On the other hand, the Bible contains several proper names and its different books were written by several authors in different periods of time, which adds the complexity level of the text.

6 Conclusions and Future Work

In this paper, we attempted the first steps of the implementation of a new language automatic translation pair between Spanish and Shipibo-konibo, a highly agglomerative native language from the peruvian Amazon. The main limitation in this study was the size of the parallel corpus available, and the great difference in the linguistic features between the language targets. Nevertheless, the results obtained surpassed the baseline proposed (dictionary based) for one of the domains analyzed (educational) and made a promising result for further development, since this corpus can be enlarged. We obtained a BLEU score of 4.42 for the religious domain corpus and 13.92 for the educational one.

As future work, this SMT implementation could be used as a part of an hybrid MT system guided mainly by rules, and also could incorporate other information related to the morphology (lemmas), POS-tags and syntax (dependency relations), since those language processing tools are currently under development. Regarding the morphological analysis, there could be test comparisons between the integration of a supervised (Pereira et al., 2017) or an unsupervised (Creutz and Lagus, 2005) word segmentation to the corpus in the Shipibo-konibo language, in order to identify which approach could improve further the results of the SMT system. In addition, NMT approaches may be tested if a third language could be added as a pivot for transfer learning (Zoph et al., 2016) or with the application of a data augmentation process for the parallel corpus (Fadaee et al., 2017). Finally, it is expected to achieve similar results with other close languages of the same family, in order to develop future pivots MT systems.

Acknowledgments

For this study, the authors acknowledge the support of the “Concejo Nacional de Ciencia, Tecnología e Innovación Tecnológica” (CONCYTEC Perú) under the contract 225-2015-FONDECYT, and the PAIP research program from the Vicerrectorado de Investigación, PUCP.

References

Willem Frederik Hendrik Adelaar, Pilar Valenzuela Bismarck, Roberto Zariquiey, and Rodolfo Marcial Cerrón-Palomino. 2011. *Estudios sobre lenguas andinas y amazónicas: homenaje a Rodolfo Cerrón-Palomino*. Pontificia Universidad Católica del Perú, Fondo Editorial.

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

Marta R Costa-Jussa and José AR Fonollosa. 2015. Latest trends in hybrid machine translation and its applications. *Computer Speech & Language* 32(1):3–10.

Mathias Creutz and Krista Lagus. 2005. *Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor 1.0*. Helsinki University of Technology.

Claire Dimeo. 2014. *Building an Automatic Translation System from English to Scots*. Master's thesis, University of Edinburgh.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2017. Data augmentation for low-resource neural machine translation. *arXiv preprint arXiv:1705.00440* .

Loriot James, Erwin Lauriault, and Dwight Day. 1993. *Diccionario Shipibo-Castellano*. Instituto Lingüístico de Verano.

Philipp Koehn. 2009. *Statistical machine translation*. Cambridge University Press.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*. Association for Computational Linguistics, pages 177–180.

Ministerio de Cultura, Perú. 2016. Base de datos de pueblos indígenas u originarios - Pueblos indígenas del Perú. Available in: http://bdpi.cultura.gob.pe.

Ministerio de Educación, Perú. 2013a. *Axeti kirika - Tsanas 4 Baritiayabaona. Cuaderno de Inicial Shipibo 4 años*. Ministerio de Educación.

Ministerio de Educación, Perú. 2013b. *Documento nacional de lenguas originarias del Perú*. Ministerio de Educación. URI: http://repositorio.minedu.gob.pe/handle/123456789/3549.

Win Pa Pa, Ye Kyaw Thu, Andrew Finch, and Eiichiro Sumita. 2016. A study of statistical machine translation methods for under resourced languages. *Procedia Computer Science* 81:250–257.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, pages 311–318.

Jose Pereira, Rodolfo Mercado, Andres Melgar, Marco Sobrevilla-Cabezudo, and Arturo Oncevay-Marcos. 2017. Ship-lemmatagger: building an NLP toolkit for a peruvian native language. In *Text, Speech, and Dialogue: 20th International Conference, TSD 2017*. Springer. In-press.

Shahram Salami, Mehrnoush Shamsfard, and Shahram Khadivi. 2016. Phrase-boundary model for statistical machine translation. *Computer Speech & Language* 38:13–27.

Raivis Skadiņš, I Skadiņa, M Pinnis, A Vasiļjevs, and Tomas Hudik. 2014. Application of machine translation in localization into low-resourced languages. In *Proceedings of the Seventeenth Annual Conference of the European Association for Machine Translation (EAMT 2014)*. pages 209–216.

Pilar Valenzuela. 2003. *Transitivity in shipibo-konibo grammar*. Ph.D. thesis, University of Oregon.

Roberto Zariquiey. 2006. Reinterpretación fonológica de los préstamos léxicos de base hispana en la lengua shipibo-conibo. *Boletín de la Academia Peruana de la Lengua* 41.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 1568–1575.