Proceedings of the 1st Joint SLTU and CCURL Workshop (SLTU-CCURL 2020), pages 79–87
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Investigating Language Impact in Bilingual Approaches for Computational Language Documentation

Marcely Zanon Boito[1], **Aline Villavicencio**[2,3], **Laurent Besacier**[1]
(1) Laboratoire d'Informatique de Grenoble (LIG), UGA, G-INP, CNRS, INRIA, France
(2) Department of Computer Science, University of Sheffield, England
(3) Institute of Informatics (INF), UFRGS, Brazil
contact: marcely.zanon-boito@univ-grenoble-alpes.fr

Abstract

For endangered languages, data collection campaigns have to accommodate the challenge that many of them are from oral tradition, and producing transcriptions is costly. Therefore, it is fundamental to translate them into a widely spoken language to ensure interpretability of the recordings. In this paper we investigate how the choice of translation language affects the posterior documentation work and potential automatic approaches which will work on top of the produced bilingual corpus. For answering this question, we use the MaSS multilingual speech corpus (Boito et al., 2020) for creating 56 bilingual pairs that we apply to the task of low-resource unsupervised word segmentation and alignment. Our results highlight that the choice of language for translation influences the word segmentation performance, and that different lexicons are learned by using different aligned translations. Lastly, this paper proposes a *hybrid* approach for bilingual word segmentation, combining *boundary clues* extracted from a non-parametric Bayesian model (Goldwater et al., 2009a) with the attentional word segmentation neural model from Godard et al. (2018). Our results suggest that incorporating these clues into the neural models' input representation increases their translation and alignment quality, specially for challenging language pairs.

Keywords: word segmentation, sequence-to-sequence models, computational language documentation, attention mechanism

1. Introduction

Computational Language Documentation (CLD) is an emerging research field whose focus lies on helping to automate the manual steps performed by linguists during language documentation. The need for this support is ever more crucial given predictions that more than 50% of all currently spoken languages will vanish before 2100 (Austin and Sallabank, 2011). For these very low-resource scenarios, transcription is very time-consuming: one minute of audio is estimated to take one hour and a half on average of a linguist's work (Austin and Sallabank, 2013).

This *transcription bottleneck* problem (Brinckmann, 2009), combined with a lack of human resources and time for documenting all these endangered languages, can be attenuated by translating into a widely spoken language to ensure subsequent interpretability of the collected recordings. Such parallel corpora have been recently created by aligning the collected audio with translations in a well-resourced language (Adda et al., 2016; Godard et al., 2017; Boito et al., 2018), and some linguists even suggested that more than one translation should be collected to capture deeper layers of meaning (Evans and Sasse, 2004). However, in this documentation scenario, the impact of the language chosen for translation rests understudied, and it is unclear if similarities among languages have a significant impact in the automatic bilingual methods used for information extraction (these include word segmentation, word alignment, and translation).

Recent work in CLD includes the use of aligned translation for improving transcription quality (Anastasopoulos and Chiang, 2018), and for obtaining bilingual-rooted word segmentation (Duong et al., 2016; Boito et al., 2017). There are pipelines for obtaining manual (Foley et al., 2018) and automatic (Michaud et al., 2018) transcriptions, and for aligning transcription and audio (Strunk et al., 2014). Other examples are methods for low-resource segmentation (Lignos and Yang, 2010; Goldwater et al., 2009b), and for lexical unit discovery without textual resources (Bartels et al., 2016). Moreover, direct speech-to-speech (Tjandra et al., 2019) and speech-to-text (Besacier et al., 2006; Bérard et al., 2016) architectures could be an option for the lack of transcription, but there is no investigation yet about how exploitable these architectures can be in low-resource settings. Finally, previous work also showed that Neural Machine Translation models at the textual level are able to provide exploitable soft-alignments between sentences by using only 5,130 training examples (Boito et al., 2019).

In this work, we investigate the existence of language impact in bilingual approaches for CLD, tackling word segmentation,[1] one of the first tasks performed by linguists after data collection. More precisely, the task consists in detecting word boundaries in an unsegmented phoneme sequence in the language to document, supported by the translation available at the sentence-level. The phonemes in the language to document can be manually obtained, or produced automatically as in Godard et al. (2018).

For our experiments, we use the eight languages from the multilingual speech-to-speech MaSS dataset (Boito et al., 2020): Basque (EU), English (EN), Finnish (FI), French (FR), Hungarian (HU), Romanian (RO), Russian (RU) and Spanish (ES). We create 56 bilingual models, seven per language, simulating the documentation of each language supported by different sentence-level aligned translations. This setup allows us to investigate how having the same content, but translated in different languages, affects bilingual word segmentation. We highlight that in

[1]Here, word is defined as a sequence of phones that build a minimal unit of meaning.

this work we use a dataset of well-resourced languages due to the lack of multilingual resources in documentation languages that could be used to investigate this hypothesis. Thus, for keeping results coherent and generalizable for CLD, we down-sample our corpus, running our experiments using only 5k aligned sentences as a way to simulate a low-resource setting. We train bilingual models based on the segmentation and alignment method from Godard et al. (2018), investigating the language-related impact in the quality of segmentation, translation and alignment.

Our results confirm that the language chosen for translation has a significant impact on word segmentation performance, what aligns with Haspelmath (2011) who suggests that the notion of word cannot always be meaningfully defined cross-linguistically. We also verify that joint segmentation and alignment is not equally challenging across different languages: while we obtain good results for EN, the same method fails to segment the language-isolate EU. Moreover, we verify that the bilingual models trained with different aligned translations learn to focus on different structures, what suggests that having more than one translation could enrich computational approaches for language documentation. Lastly, the models' performance is improved by the introduction of a *hybrid* approach, which leverages the *boundary clues* obtained by a monolingual non-parametric Bayesian model (Goldwater et al., 2009b) into the bilingual models. This type of intermediate annotation is often produced by linguists during documentation, and its incorporation into the neural model can be seen as a form of validating word-hypotheses.

This paper is organized as follows. Section 2. presents the models investigated for performing word segmentation. Section 3. presents the experimental settings, and Section 4. the results and discussion. Section 5. concludes the work.

2. Models for Word Segmentation

2.1. Monolingual Bayesian Approach

Non-parametric Bayesian models (Goldwater, 2007; Johnson and Goldwater, 2009) are statistical approaches that can be used for word segmentation and morphological analysis, being known as very robust in low-resource settings (Godard et al., 2016; Goldwater et al., 2009a). In these monolingual models, words are generated by a uni or bigram model over a non-finite inventory, through the use of a Dirichlet process. Although providing reliable segmentation in low-resource settings, these monolingual models are incapable of automatically producing alignments with a foreign language, and therefore the discovered pseudo-word segments can be seen as "meaningless". Godard et al. (2018) also showed that `dpseg`[2] (Goldwater et al., 2006; Goldwater et al., 2009a) behaves poorly on pseudo-phone units discovered from speech, which limits its application. Here, we investigate its use as an intermediate monolingual-rooted segmentation system, whose discovered boundaries are used as clues by bilingual models.

[2]Available at `http://homepages.inf.ed.ac.uk/sgwater/resources.html`

2.2. Bilingual Attention-based Approach

We reproduce the approach from Godard et al. (2018) who train Neural Machine Translation (NMT) models between language pairs, using as source language the translation (word-level) and as target, the language to document (unsegmented phoneme sequence). Due to the attention mechanism present in these networks (Bahdanau et al., 2014), posterior to training, it is possible to retrieve *soft-alignment probability matrices* between source and target sentences.

The soft-alignment probability matrix for a given sentence pair is a collection of context vectors. Formally, a context vector for a decoder step t is computed using the set of source annotations H and the last state of the decoder network (s_{t-1}, the translation context). The attention is the result of the weighted sum of the source annotations H (with $H = h_1, ..., h_A$) and their α probabilities (Eq. 1). Finally, these are obtained through a feed-forward network $align$, jointly trained, and followed by a softmax operation (Eq. 2).

$$c_t = Att(H, s_{t-1}) = \sum_{i=1}^{A} \alpha_i^t h_i \quad (1)$$

$$\alpha_i^t = \mathrm{softmax}(align(h_i, s_{t-1})) \quad (2)$$

The authors show that these soft-alignment probability matrices can be used to produce segmentation over phoneme (or grapheme) sequences. This is done by segmenting neighbor phonemes whose probability distribution (over the words in the aligned source translation) peaks at different words. The result is a pair of phoneme sequences and translation words, as illustrated on the bottom half of Figure 1. In this work we refer to this type of model simply as **neural model**.

2.3. Bilingual Hybrid Approach

The monolingual approach (§2.1.) has the disadvantage of not producing bilingual alignment, but it segments better than the bilingual approach (§2.2.) when the phonemic input is used (Godard et al., 2018). In this work we investigate a simple way of combining both approaches by creating a *hybrid* model which takes advantage of the Bayesian method's ability to correctly segment from small data while jointly producing translation alignments.

We augment the original unsegmented phoneme sequence with the `dpseg` output boundaries. In this augmented input representation, illustrated in Figure 1, a boundary is denoted by a special token which separates the words identified by `dpseg`. We call this *soft-boundary insertion*, since the `dpseg` boundaries inserted into the phoneme sequence can be ignored by the NMT model, and new boundaries can be inserted as well. For instance, in Figure 1 `aintrat` becomes `a intrat` (boundary insertion), and `urat debine` becomes `uratdebine` (soft-boundary removal).

3. Experimental Settings

Multilingual Dataset: For our experiments we use the MaSS dataset (Boito et al., 2020), a fully aligned and multilingual dataset containing 8,130 sentences extracted

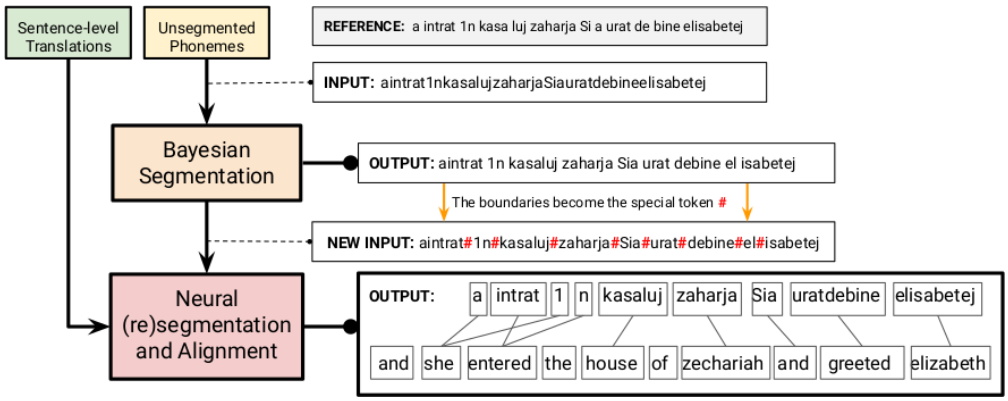

Figure 1: An illustration of the hybrid pipeline for the EN>RO language pair. The Bayesian model receives the unsegmented phonemes, outputing segmentation. The discovered boundaries are then replaced by a special token, and bilingual re-segmentation and alignment are jointly performed.

from the Bible. The dataset provides multilingual speech and text alignment between all the available languages: English (EN), Spanish (ES), Basque (EU), Finnish (FI), French (FR), Hungarian (HU), Romanian (RO), Russian (RU). As sentences in documentation settings tend to be short, we used RO as the pivot language for removing sentences longer (in terms of number of tokens) than 100 symbols. The resulting corpus contains 5,324 sentences, a size which is compatible with real language documentation scenarios. Table 1 presents some statistics. For the phonemic transcription of the speech (target side of the bilingual segmentation pipeline), we use the automatic phonemization from *Maus forced aligner* (Kisler et al., 2017), which results in an average vocabulary reduction of 835 types, the smallest being for RO (396), and the most expressive being for FR (1,708). This difference depends on the distance between phonemic and graphemic forms for each language. The phonemizations present an average number of unique phonemes of 42.5. Table 2 presents the statistic for the phonemic representation.

Training and Evaluation: For monolingual segmentation, we use `dpseg`'s unigram model with the same hyperparameters as Godard et al. (2016). The bilingual neural models were trained using a one-layer encoder (embeddings of 64), and a two-layers decoder (embeddings of 16). The remaining parameters come from Godard et al. (2018). From this work, we also reproduced the *multiple runs averaging*: for every language pair, we trained two networks, averaging the soft-alignment probability matrices produced. This averaging can be seen as *agreement* between the alignment learned with different parameters initialization. Regarding the data, 10% of the multilingual ids were randomly selected for validation, and the remaining were used for training. We report BLEU scores (Papineni et al., 2002) over the validation set for assessing translation quality. For hybrid setups, the soft-boundary special token is removed from the output before scoring, so results are comparable. Finally, for the reported word discovery results, the totality of the corpus is considered for evaluation.

	#Types	#Tokens	Token Length	Token/ Sentence
EN	5,232	90,716	3.98	17.04
ES	8,766	85,724	4.37	16.10
EU	11,048	67,012	5.91	12.59
FI	12,605	70,226	5.94	13.19
FR	7,226	94,527	4.12	17.75
HU	13,770	69,755	5.37	13.10
RO	7,191	88,512	4.06	16.63
RU	11,448	67,233	4.66	12.63

Table 1: Statistics for the textual portion of the corpus. The last two columns bring the average of the named metrics.

	#Types	#Tokens	Token Length	Phonemes/ Sentence
EN	4,730	90,657	3.86	56.18
ES	7,980	85,724	4.30	68.52
EU	9,880	67,012	6.94	71.13
FI	12,088	70,226	5.97	72.37
FR	5,518	93,038	3.21	52.86
HU	12,993	69,755	5.86	65.52
RO	6,795	84,613	4.50	68.04
RU	10,624	67,176	6.19	59.26

Table 2: Statistics for the phonemic portion of the corpus. The last two columns bring the average of the named metrics.

4. Bilingual Experiments

Word segmentation boundary F-scores are presented in Table 3. For the bilingual methods, Table 4 presents the averaged BLEU scores. We observe that, similar to the trend observed in Table 3, hybrid models are in average superior in terms of BLEU scores.[3] Moreover, we observe that segmentation and translation scores are strongly correlated for six of the eight languages, with an average ρ-value of 0.76

[3] We find an average BLEU scores difference between best hybrid and neural setups of 1.50 points after removing the outlier (RO). For this particular case, hybrid setups have inferior translation performance (average BLEU reduction of 11.44).

		EN	ES	EU	FI	FR	HU	RO	RU
neural	EN	-	51.8	36.1	53.8	**65.8**	47.7	57.5	50.3
	ES	60.1	-	**38.4**	46.3	63.4	45.9	53.5	46.3
	EU	48.3	44.2	-	42.5	46.4	41.2	44.7	41.8
	FI	60.0	46.8	36.5	-	53.7	**50.1**	51.5	**53.5**
	FR	**69.1**	**57.7**	37.0	53.7	-	47.4	**62.8**	49.8
	HU	53.3	46.0	36.5	52.9	48.7	-	48.7	49.8
	RO	60.9	51.5	37.9	51.1	63.9	47.6	-	51.6
	RU	58.7	47.6	35.6	**54.7**	54.0	49.3	53.9	-
hybrid	EN	-	57.9	43.5	57.5	**69.6**	52.9	64.2	58.1
	ES	66.4	-	**47.3**	54.3	68.8	51.7	63.4	56.1
	EU	58.6	53.1	-	50.1	58.1	49.2	55.1	50.1
	FI	66.5	55.6	45.7	-	62.7	**58.5**	60.7	**62.6**
	FR	**73.3**	**62.1**	45.6	56.9	-	54.2	**70.0**	59.5
	HU	62.6	54.2	45.0	59.7	60.0	-	58.8	59.3
	RO	68.2	57.6	46.9	56.2	69.3	53.8	-	60.1
	RU	66.8	56.1	44.6	**60.7**	63.0	55.3	63.6	-
dpseg		82.4	79.2	81.0	80.0	78.1	75.5	82.0	78.3

Table 3: Word Segmentation Boundary F-score results for neural (top), hybrid (middle) and `dpseg` (bottom). The columns represent the target of the segmentation, while the rows represented the translation language used. For bilingual models, darker squares represent higher scores. Better visualized in color.

(significant to $p < 0.01$). The exceptions were EU (0.46) and RO (-0.06). While for EU we believe the general lack of performance of the systems could explain the results, the profile of RO hybrid setups was surprising. It highlights that the relationship between BLEU score and segmentation performance is not always clearly observed. In summary, we find that the addition of soft-boundaries will increase word segmentation results, but its impact to translation performance needs further investigation.

Looking at the segmentation results, we verify that, given the same amount of data and supervision, the segmentation performance for different target languages vary: EN seems to be the easiest to segment (neural: 69.1, hybrid: 73.3), while EU is the most challenging to segment with the neural approach (neural: 38.4, hybrid: 47.3). The following subsections will explore the relationship between segmentation, alignment performance and linguistic properties.

4.1. Source Language Impact

Bilingual Baseline Comparison: The results confirm that there is an impact related to using different source languages for generating the segmentations, and we identify interesting language pairs emerging as the most efficient, such as FI>HU (Uralic Family), FR>RO and FR>ES (Romance family).[4] In order to consolidate these results, we investigate if the language ranking obtained (in terms of *best translation languages for segmenting a target language*) is due to a similar profile of the source and target languages in terms of word length and tokens per sentence. Since translation words are used to cluster the phoneme sequences into words (bilingual-rooted word segmentation), having more or less translation words could be a determining aspect in the bilingual segmentation performed (more details about this in Section 4.3.). For this investigation, we use a naive bilingual baseline called proportional (Godard et al., 2018). It performs segmentation by distributing phonemes equally between the words of the aligned translation, insuring that words that have more letters, receive more phonemes (hence *proportional*). The average difference between the best hybrid (Table 3) and proportional (Table 5) results is of 25.92 points. This highlights not only the challenge of the task, but that the alignments learned by the bilingual models are not trivial.

We compute Pearson's correlation between bilingual hybrid and proportional segmentation scores, observing that no language presents a significant correlation for $p < 0.01$. However, when all languages pairs are considered together ($N = 56$), a significant positive correlation (0.74) is observed. Our interpretation is that the token ratio between the number of tokens in source and the number of tokens

[4] We denote L1>L2 as using L1 for segmenting L2. L1<>L2 means L1>L2 and L2>L1.

		EN	ES	EU	FI	FR	HU	RO	RU
neural	EN	-	39.7	35.2	**45.1**	40.5	**36.0**	43.3	37.3
	ES	37.3	-	**37.7**	37.9	37.6	32.7	39.2	32.8
	EU	28.3	32.8	-	33.8	26.8	28.0	31.1	27.6
	FI	40.4	36.0	34.6	-	35.2	35.5	39.2	37.5
	FR	**45.7**	**42.2**	35.9	43.9	-	34.7	**50.8**	37.5
	HU	35.4	34.7	33.5	41.0	31.4	-	36.3	36.0
	RO	40.7	39.7	36.0	42.4	**44.3**	34.8	-	**37.8**
	RU	38.9	36.7	32.8	43.0	35.2	34.6	40.4	-
hybrid	EN	-	40.5	35.5	**47.0**	42.3	37.2	30.9	39.5
	ES	38.3	-	**38.4**	39.2	37.3	33.7	28.6	34.9
	EU	28.8	33.5	-	34.4	26.8	29.6	22.6	28.2
	FI	41.5	36.7	35.4	-	36.3	**37.4**	27.9	38.6
	FR	**46.6**	**43.6**	36.2	45.3	-	36.2	**34.9**	39.1
	HU	35.6	35.8	35.0	41.8	32.4	-	26.5	36.2
	RO	42.7	41.4	36.3	43.8	**46.1**	36.6	-	**40.2**
	RU	39.7	37.2	34.6	44.4	36.4	36.2	28.8	-

Table 4: BLEU 4 average results for neural (top) and hybrid (bottom) bilingual models. The columns represent the target of the segmentation. Darker squares represent higher scores. Better visualized in color.

	EN	ES	EU	FI	FR	HU	RO	RU
EN	-	36.0	32.5	37.1	**41.4**	34.2	36.6	36.6
ES	37.6	-	32.3	36.9	41.0	34.0	36.7	36.8
EU	35.5	36.1	-	38.0	38.8	34.5	36.2	**37.3**
FI	36.1	36.1	32.9	-	39.3	34.3	36.5	37.1
FR	**38.4**	**36.4**	32.2	36.4	-	33.9	**36.9**	36.5
HU	35.9	35.9	**33.0**	37.8	39.3	-	36.4	37.2
RO	37.6	36.3	32.6	36.8	40.9	34.0	-	36.8
RU	34.8	35.9	32.9	**38.5**	38.2	**34.8**	36.2	-

Table 5: Proportional segmentation results. The columns represent the target of the segmentation. Darker squares represent higher word boundary F-scores. Better visualized in color.

in target sentences have a significant impact on bilingual segmentation difficulty. However, it does not, by itself, dictates the best choice of translation language for a documentation scenario. For instance, the proportional baseline results indicate that EU is the best choice for segmenting RU. This choice is not only linguistically incoherent, but bilingual models reached their worst segmentation and translation results by using this language. This highlights that while statistical features might impact greatly low-resource alignment and should be taken into account, relying only on them might result in sub-optimal models.

Language Ranking: Looking into the quality of the segmentation results and their relationship with the language ranking, our intuition was that languages from the same family would perform the best. For instance, we expected ES<>FR, ES<>RO, FR<>RO (Romance family) and FI<>HU (Uralic family) to be strong language pairs. While some results confirm this hypothesis (FR>ES, FI>HU, FR>RO), the exceptions are: EN>FR, RU<>FI and ES>EU. For EN>FR, we argue that EN was ranked high for almost all languages, which could be due to some convenient statistic features. Table 1 shows that EN presents a very reduced vocabulary in comparison to the other languages. This could result in an easier language modeling scenario, which could then reflect in a better alignment capacity of the trained model. Moreover, for this and for RU<>FI scenarios, results seemed to reproduce the trend from the proportional baseline, in which these pairs were also found to be the best. This could be the result of a low syntactic divergence between languages of these pairs. Finally, the language isolate EU is not a good choice for segmenting any language (worst result for all languages). If we consider that this language has no relation to any other in this dataset, this result could be an indication that documentation should favor languages somehow related to the language they are trying to document. In fact, results for EU segmentation are both low (F-score and BLEU) and very close to the proportional baseline (average difference of 4.23 for neural and 13.10 for hybrid), which suggests that these models were not able to learn meaningful bilingual alignment.

4.2. Hybrid Setups

Looking at the hybrid results, we verify that the these models outperform their neural counterparts. Moreover, the impact of having the *soft-boundaries* is larger for the languages whose bilingual segmentation seems to be more challenging, hinting that the network is learning to leverage the *soft-boundaries* for generating a better-quality alignment between challenging language pairs. Table 6 presents the intersection between the correct types discovered by both monolingual and hybrid models. Results show that while the monolingual baseline *informs* the bilingual models, it is not completely responsible for the increase in performance. This hints that giving boundary clues to the network will not simply force some pre-established segmentation, but instead it will enrich the network's internal representation. Moreover, it is interesting to observe that the degree of overlap between the vocabulary generated will depend on the language target of segmentation,

	EN	ES	EU	FI	FR	HU	RO	RU
EN	-	**0.60**	**0.74**	**0.64**	0.68	**0.59**	**0.69**	**0.51**
ES	0.76	-	0.67	0.45	0.59	0.43	0.59	0.43
EU	**0.81**	0.57	-	0.49	**0.70**	0.48	0.68	0.50
FI	0.72	0.46	0.58	-	0.61	0.34	0.57	0.34
FR	0.72	0.44	0.68	0.48	-	0.48	0.56	0.41
HU	0.76	0.47	0.57	0.34	0.64	-	0.59	0.37
RO	0.76	0.56	0.70	0.51	0.62	0.48	-	0.43
RU	0.74	0.48	0.60	0.35	0.61	0.39	0.56	-

Table 6: Intersection between the correct types discovered by both monolingual and hybrid models. We notice that the target language of the segmentation (columns) has an impact in the acceptance of soft-boundaries by the neural model.

hinting that some languages might *accept* more easily the *soft-boundaries* proposed by the monolingual approach.

Nonetheless, compared to the monolingual segmentation (Table 3), even if the hybrid approach improves over the base neural one, it deteriorates considerably the performance with respect to `dpseg` (average difference of 16.54 points between the best hybrid result and its equivalent monolingual segmentation). However, this deterioration is necessary in order to discover semantically meaningful structures (joint bilingual segmentation and alignment), which is a harder task than monolingual segmentation. In this scenario, the monolingual results should be interpreted as an intermediate, good quality, segmentation/word-hypotheses created by linguists, which might be validated or not in light of the system's bilingual output.

4.3. Analysis of the Discovered Vocabulary

Next we study the characteristics of the vocabulary outputed by the bilingual models focusing on the impact caused by the aligned translation. For this investigation, we report results for hybrid models only, since their neural equivalents present the same trend. We refer as *token* the collection of phonemes segmented into word-like units. *Types* are defined as the set of distinct tokens. Table 7 brings the hybrid model's total number of types.

Looking at the rows, we see that EN, ES, FR, RO, which are all fusional languages, generated in average the smallest vocabularies. We also notice that HU and FI are the languages that tend to create the largest vocabularies when used as translation language. This could be due to both languages accepting a flexible word order, thus creating a difficult alignment scenario for low-resource settings. Moreover, these languages, together with EU, are agglutinative languages. This might be an explanation for the lack of performance in general for setups using these languages as target. In these conditions, the network must learn to align many translation words to the same structure in order to achieve the expected segmentation. However, sometimes over-segmentation might be the result of the network favoring alignment content instead of phoneme clustering.

Notwithstanding, the models for agglutinative languages are not the only ones over-segmenting. Looking at the average token length of the segmentations produced in Figure 2,

	EN	ES	EU	FI	FR	HU	RO	RU
EN	-	12,170	9,173	17,532	13,658	17,029	15,830	15,844
ES	13,732	-	13,249	12,965	10,984	13,283	13,073	13,247
EU	13,942	16,202	-	17,106	15,996	17,931	16,138	17,904
FI	16,201	18,349	16,540	-	17,478	**19,993**	17,470	17,938
FR	10,886	13,985	13,737	15,217	-	15,574	13,609	14,531
HU	**17,086**	**18,398**	**17,218**	**21,097**	**18,472**	-	**18,861**	**18,728**
RO	12,063	13,948	12,768	15,226	12,094	15,764	-	14,811
RU	14,973	16,856	16,027	18,805	16,515	18,349	16,595	-

Table 7: Number of types produced by the hybrid models.

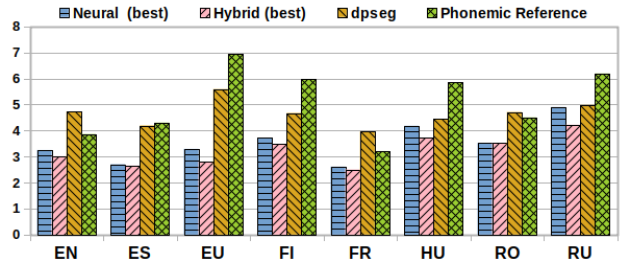

Figure 2: Average token length of the reference, monolingual `dpseg`, and best neural and hybrid setups from Table 3.

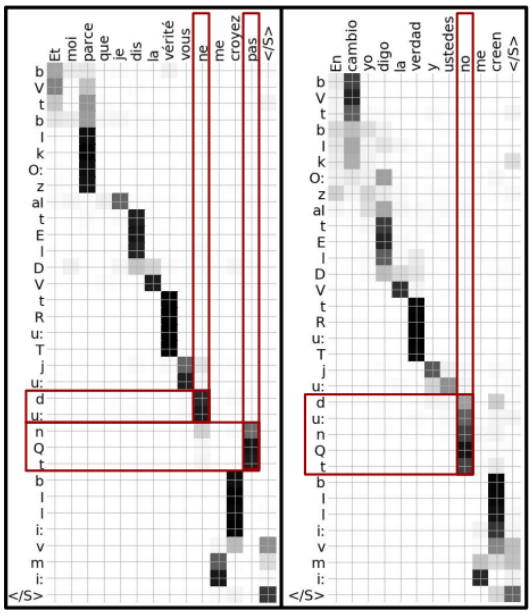

Figure 3: EN attention matrices generated by neural FR (left) and ES (right) bilingual models. The squares represent alignment probabilities (the darker the square, the higher the probability). The EN phonemization (rows) correspond to the following sentence: "But because I tell the truth, you do not believe me".

and supported by the size of the vocabularies, we verify that bilingual approaches tend to over-segment the output independent of the language targeted. This over-segmentation tends to be more accentuated in hybrid setups, with the exception of EN, FR and RO. This is probably due to the challenge of clustering the very long sequence of phonemes into the many available source words (see statistics for words and phonemes per sentence in Tables 1 and 2).

Furthermore, the very definition of a word might be difficult to define cross-linguistically, as discussed by Haspelmath (2011), and different languages might encourage a more fine-grained segmentation. For instance, in Figure 3 we see the EN alignment generated by the FR and ES neural models for the same sentence. Focusing at the *do not* (`du:nQt`) at the end of the sentence, we see that the ES model does not segment it, aligning everything to the ES translation `no`. Meanwhile the FR model segments the structure in order to align it to the translation `ne pas`. In both cases the discovered alignments are correct however, the ES segmentation is considered wrong. This highlights that the use of a segmentation task for evaluating the learned alignment might be sub-optimal, and that a more in-depth evaluation of source-to-target correspondences should be considered. In Section 4.4. we showcase a method for filtering the alignments generated by the bilingual models.

Concluding, in this work we study the alignment *implicitly* optimized by a neural model. An interesting direction would be the investigation of explicit alignment optimization for translation models, such as performed in Godard et al. (2019), where the authors consider the segmentation length generated by the bilingual alignments as part of their loss during training.

4.4. Alignment Confidence

The neural approach used here for bilingual-rooted word segmentation produces alignments between source and target languages. In this section we investigate how these alignments vary in models trained using different translation (source) languages. This extends the results from the previous section, that showed that models trained on different languages will present different lexicon sizes. We aim to show that this difference in segmentation behavior comes from the different alignments that are discovered by the models with access to different languages.

We use the approach from Boito et al. (2019) for extracting the alignments the bilingual models are more *confident about*. For performing such a task, *Average Normalized Entropy*, as defined in Boito et al. (2019), is computed for every (segmentation, aligned translation) pair. The scores are used for ranking the alignments in terms of confidence, with low-entropy scores representing the high-confidence automatically generated alignments. In previous work, we showed that this approach allow us to increase type retrieval scores by filtering the good from the bad quality alignments discovered. For this investigation, we chose to present results applied to the target language FR.

Table 8 presents the top 10 low-entropy (high-confidence) pairs from 3 different translation languages (from Table 3, FR column). The phoneme sequences are accompanied by their grapheme equivalents to increase readability, but all presented results were computed over phoneme sequences. The other translation languages were also omitted for readability purpose.

We observe a different set of discovered types depending on the language used, but all languages learn a fair amount

	EN			ES			RU		
1	Galates	galat	Galatians	N-A-W	Jo	Cordero	Jean	Za~	Иохан
2	Femmes	fam	Wives	Jeanne	Zan	Juana	les+huissiers	leHisie	Служители
3	Jude	Zyd	Jude	guéri	geRi	recuperará	Galates	galat	Галатам
4	Kaïnan	kaj	Cainan	Galates	galat	Gálatas	neuf	n2f	9
5	Philippiens	filipje~	Philippians	onze	?o~z	11	Marc	maRk	Марк
6	N-A-W	tR	treacherous	Hébreux	ebR2	Hebreos	Matthieu	matj2	Матай
7	Luc	lyk	Luke	manne	man	maná	sachez	saSe	Знайте
8	car	kaR	main	douze	duz	12	déclare	deklaR	Проповедуй
9	Seth	sEt	Seth	N-A-W	afliZ	afligidos	asa	aza	Аса
10	boue	bu	mud	treize	tREz	13	amis	ami	друзья

Table 8: Top low-entropy/high-confidence (graphemization, phonemic segmentation, aligned translation) results for EN, ES and RU models for segmenting FR. The output of the system is the phonemic segmentation, and graphemization is provided only for readability purpose. N-A-W identify unknown/incorrect generated types.

of biblical names and numbers, very frequent due to the nature of the dataset.[5] This highlights that very frequent types might be captured independently of the language used, but other structures might be more dependent on the chosen language. We also notice the presence of incorrect alignments (the word `car` (because) aligned to the word `main`), concatenations (the words `les huissiers` (the ushers) became a single word) and incorrect types (`N-A-W` in the table). This is to be expected, as these are automatic alignments.

Confirming the intuition that the models are focused on different information depending on the language they are trained on, we studied the vocabulary intersection of the FR bilingual models for the top 200 correct discovered types ranked by alignment confidence. We observed that the amount of shared lexicon for the sets is fairly small: the smallest intersection being of 20% (between EU and RO) and the largest one of 35.5% (between RU and FI). In other words, this means that the high-confidence alignments learned by distinct bilingual models differ considerably. Even for models that shared most structures, such as FI and RU (35.5%), and HU and RU (34%), this intersection is still limited. This shows that the bilingual models will discover different structures, depending on the supervision available. This is particularly interesting considering that the content of the aligned information remains the same, and the only difference between the bilingual models is the language in which the information is expressed. It highlights how collecting data in *multilingual settings* (that is, in more than one translation language) could enrich approaches for CLD. Lastly, we leave as future work a more generalizable study of the distinctions in the bilingual alignments, including the evaluation of the word-level alignments discovered by the models.

5. Conclusion

In language documentation scenarios, transcriptions are most of the time costly and difficult to obtain. In order to ensure the interpretability of the recordings, a popular solution is to replace manual transcriptions by translations of the recordings in well-resourced languages (Adda et al., 2016). However, while some work suggests that translations in multiple languages may capture deeper layers of meaning (Evans and Sasse, 2004), most of the produced corpora from documentation initiatives are bilingual. Also, there is a lack of discussion about the impact of the language chosen for these translations in posterior automatic methods.

In this paper we investigated the existence of language-dependent behavior in a bilingual method for unsupervised word segmentation, one of the first tasks performed in post-collection documentation settings. We simulated such a scenario by using the MaSS dataset (Boito et al., 2020) for training 56 bilingual models, the combination of all the available languages in the dataset. Our results show that in very low-resource scenarios (only 5,324 aligned sentences), the impact of language can be great, with a large margin between best and worst results for every target language. We also verify that the languages are not all equally difficult to segment. Moreover, while some of our *language rankings*, in terms of best translation languages for segmenting a target language, could be explained by the linguistic family of the languages (FR>ES, FI>HU, FR>RO), we found some surprising results such as ES>EU and EN>FR. We believe these are mostly due to the impact of existing statistic features (e.g. token length ratio between source and target sentences, and vocabulary size), related to the corpus, and not to the language features.

Additionally, we investigated providing a different form of supervision to the bilingual models. We used the monolingual-rooted segmentation generated by `dpseg` for augmenting the phoneme sequence representation that the neural models learn from at training time. We observed that the networks learned to leverage `dpseg`'s *soft-boundaries* as hints of alignment break (boundary insertion). Nonetheless, the networks are still robust enough to ignore this information when necessary. This suggests that, in a documentation scenario, `dpseg` could be replaced by early annotations of potential words done by a linguist, for instance. The linguist could then validate the output of the neural system, and review their word hypotheses considering the generated bilingual alignment.

In summary, our results highlight the existence of a relationship between language features and performance in

[5]The chapter names and numbers (e.g. "Revelation 2") are included in the dataset, totalizing 260 examples of *"name, number"*.

(neural) bilingual segmentation. We verify that languages close in phonology and linguistic family score better, while less similar languages yield lower scores. While we find that our results are rooted in linguistic features, we also believe there is a non-negligible relationship with corpus statistic features which can impact greatly neural approaches in low-resource settings.

6. Bibliographical References

Adda, G., Stüker, S., Adda-Decker, M., Ambouroue, O., Besacier, L., Blachon, D., Bonneau-Maynard, H., Godard, P., Hamlaoui, F., Idiatov, D., Kouarata, G.-N., Lamel, L., Makasso, E.-M., Rialland, A., de Velde, M. V., Yvon, F., and Zerbian, S. (2016). Breaking the unwritten language barrier: The BULB project. *Procedia Computer Science*, 81:8–14.

Anastasopoulos, A. and Chiang, D. (2018). Leveraging translations for speech transcription in low-resource settings. In *Proc. Interspeech 2018*, pages 1279–1283.

Austin, P. K. and Sallabank, J. (2011). *The Cambridge handbook of endangered languages*. Cambridge University Press.

Austin, P. K. and Sallabank, J. (2013). *Endangered languages*. Taylor & Francis.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Bartels, C., Wang, W., Mitra, V., Richey, C., Kathol, A., Vergyri, D., Bratt, H., and Hung, C. (2016). Toward human-assisted lexical unit discovery without text resources. In *Spoken Language Technology Workshop (SLT), 2016 IEEE*, pages 64–70. IEEE.

Bérard, A., Pietquin, O., Servan, C., and Besacier, L. (2016). Listen and translate: A proof of concept for end-to-end speech-to-text translation. *arXiv preprint arXiv:1612.01744*.

Besacier, L., Zhou, B., and Gao, Y. (2006). Towards speech translation of non written languages. In *Spoken Language Technology Workshop, 2006. IEEE*, pages 222–225. IEEE.

Boito, M. Z., Bérard, A., Villavicencio, A., and Besacier, L. (2017). Unwritten languages demand attention too! word discovery with encoder-decoder models. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 458–465. IEEE.

Boito, M. Z., Anastasopoulos, A., Lekakou, M., Villavicencio, A., and Besacier, L. (2018). A small griko-italian speech translation corpus. *arXiv preprint arXiv:1807.10740*.

Boito, M. Z., Villavicencio, A., and Besacier, L. (2019). Empirical evaluation of sequence-to-sequence models for word discovery in low-resource settings. *arXiv preprint arXiv:1907.00184*.

Boito, M. Z., Havard, W. N., Garnerin, M., Ferrand, É. L., and Besacier, L. (2020). Mass: A large and clean multilingual corpus of sentence-aligned spoken utterances extracted from the bible. *Language Resources and Evaluation Conference (LREC)*.

Brinckmann, C. (2009). Transcription bottleneck of speech corpus exploitation.

Duong, L., Anastasopoulos, A., Chiang, D., Bird, S., and Cohn, T. (2016). An attentional model for speech translation without transcription. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 949–959.

Evans, N. and Sasse, H.-J. (2004). In *Searching for meaning in the Library of Babel: field semantics and problems of digital archiving*. Open Conference Systems, University of Sydney, Faculty of Arts.

Foley, B., Arnold, J. T., Coto-Solano, R., Durantin, G., Ellison, T. M., van Esch, D., Heath, S., Kratochvil, F., Maxwell-Smith, Z., Nash, D., et al. (2018). Building speech recognition systems for language documentation: The coedl endangered language pipeline and inference system (elpis). In *SLTU*, pages 205–209.

Godard, P., Adda, G., Adda-Decker, M., Allauzen, A., Besacier, L., Bonneau-Maynard, H., Kouarata, G.-N., Löser, K., Rialland, A., and Yvon, F. (2016). Preliminary experiments on unsupervised word discovery in mboshi. In *Proc. Interspeech*.

Godard, P., Adda, G., Adda-Decker, M., Benjumea, J., Besacier, L., Cooper-Leavitt, J., Kouarata, G.-N., Lamel, L., Maynard, H., Müller, M., et al. (2017). A very low resource language speech corpus for computational language documentation experiments. *arXiv preprint arXiv:1710.03501*.

Godard, P., Zanon Boito, M., Ondel, L., Berard, A., Yvon, F., Villavicencio, A., and Besacier, L. (2018). Unsupervised word segmentation from speech with attention. In *Interspeech*.

Godard, P., Besacier, L., and Yvon, F. (2019). Controlling utterance length in nmt-based word segmentation with attention. *arXiv preprint arXiv:1910.08418*.

Goldwater, S., Griffiths, T. L., and Johnson, M. (2006). Contextual dependencies in unsupervised word segmentation. In *Proc. International Conference on Computational Linguistics*, pages 673–680, Sydney, Australia.

Goldwater, S., Griffiths, T. L., and Johnson, M. (2009a). A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, 112(1):21–54.

Goldwater, S., Griffiths, T. L., and Johnson, M. (2009b). A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, 112(1):21–54.

Goldwater, S. J. (2007). *Nonparametric Bayesian models of lexical acquisition*. Ph.D. thesis, Citeseer.

Haspelmath, M. (2011). The indeterminacy of word segmentation and the nature of morphology and syntax. *Folia linguistica*, 45(1):31–80.

Johnson, M. and Goldwater, S. (2009). Improving nonparameteric bayesian inference: experiments on unsupervised word segmentation with adaptor grammars. In *Proc. NAACL-HLT*, pages 317–325. Association for Computational Linguistics.

Kisler, T., Reichel, U., and Schiel, F. (2017). Multilingual processing of speech via web services. *Computer Speech & Language*, 45:326 – 347.

Lignos, C. and Yang, C. (2010). Recession segmentation: simpler online word segmentation using limited re-

sources. In *Proceedings of the fourteenth conference on computational natural language learning*, pages 88–97. Association for Computational Linguistics.

Michaud, A., Adams, O., Cohn, T. A., Neubig, G., and Guillaume, S. (2018). Integrating automatic transcription into the language documentation workflow: Experiments with na data and the persephone toolkit.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Strunk, J., Schiel, F., Seifart, F., et al. (2014). Untrained forced alignment of transcriptions and audio for language documentation corpora using webmaus. In *LREC*, pages 3940–3947.

Tjandra, A., Sakti, S., and Nakamura, S. (2019). Speech-to-speech translation between untranscribed unknown languages. *arXiv preprint arXiv:1910.00795*.