Models and resources for attention-based unsupervised word segmentation : an application to computational language documentation

Marcely Zanon Boito

Université Grenoble Alpes

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

Spécialité : **Informatique**

Arrêté ministériel : **25 mai 2016**

Présentée par

Marcely ZANON BOITO

Thèse dirigée par **M. Laurent BESACIER**
et codirigée par **Mme. Aline VILLAVICENCIO**

préparée au sein du **Laboratoire d'Informatique de Grenoble (LIG)**
et de l'**École Doctorale Mathematiques, Sciences et Technologies de l'Information, Informatique (ED-MSTII)**

Models and Resources for Attention-based Unsupervised Word Segmentation

An Application to Computational Language Documentation

Thèse soutenue publiquement le **July, 9, 2021**,
devant le jury composé de :

M. François PORTET
Professor, Université Grenoble Alpes, Président

M. Thierry POIBEAU
Research Director, CNRS, ENS/PSL and Université Sorbonne Nouvelle, Rapporteur

Mme. Karen LIVESCU
Associate Professor, Toyota Technological Institute of Chicago, Rapportrice

Mme. Claire GARDENT
Research Director, CNRS and Université de Lorraine, Examinatrice

M. Laurent BESACIER
Professor, Naver Labs Europe, Directeur de thèse

Mme. Aline VILLAVICENCIO
Associate Professor, Sheffield University, Co-Directrice de thèse

Models and Resources for Attention-based Unsupervised Word Segmentation

Abstract:

Computational Language Documentation (CLD) is a research field interested in proposing methodologies capable of speeding up language documentation, helping linguists to efficiently collect and process data from many dialects, some of which are expected to vanish before the end of this century (Austin and Sallabank, 2013). In order to achieve that, the proposed methods need to be robust to low-resource data processing, as corpora from documentation initiatives lack size, and they must operate from speech, as many of these languages are from oral tradition, meaning that there is a lack of standard written form.

In this thesis we investigate the task of Unsupervised Word Segmentation (UWS) from speech. The goal of this approach is to segment utterances into smaller chunks corresponding to the words in that language, without access to any written transcription. Here we propose to ground the word segmentation process in aligned bilingual information. This is inspired by the possible availability of translations, often collected by linguists during documentation (Adda et al., 2016).

Thus, using bilingual corpora made of speech utterances and sentence-aligned translations, we propose the use of attention-based Neural Machine Translation (NMT) models in order to align and segment. Since speech processing is known for requiring considerable amounts of data, we split this approach in two steps. We first perform Speech Discretization (SD), transforming input utterances into sequences of discrete speech units. We then train NMT models, which output soft-alignment probability matrices between units and word translations. This *attention-based* soft-alignment is used for segmenting the units with respect to the bilingual alignment obtained, and the final segmentation is carried to the speech signal. Throughout this work, we investigate the use of different models for these two tasks.

For the SD task, we compare five different approaches: three Bayesian HMM-based models (Ondel et al., 2016, 2019; Yusuf et al., 2020), and two Vector Quantization (VQ) neural models (van den Oord et al., 2017; Baevski et al., 2020a). We find that the Bayesian SD models, in particular the SHMM (Ondel et al., 2019) and H-SHMM (Yusuf et al., 2020), are the most exploitable for direct application in text-based UWS in our documentation setting. For the alignment and segmentation task, we compare three attention-based NMT

models: RNN (Bahdanau et al., 2015), 2D-CNN (Elbayad et al., 2018), and Transformer (Vaswani et al., 2017). We find that the attention mechanism is still exploitable in our limited setting (5,130 aligned sentences only), but that the soft-alignment probability matrices from novel NMT approaches (2D-CNN, Transformer) are inferior to the ones from the simpler RNN model.

Finally, our attention-based UWS approach is evaluated in topline conditions using the true phones (Boito et al., 2019a), and in realistic conditions using the output of SD models (Godard et al., 2018c). We use eight languages and fifty six language pairs for verifying the language-related impact caused by grounding segmentation in bilingual information (Boito et al., 2020b), and we present extensions for increasing the quality of the produced soft-alignment probability matrices (Boito et al., 2021).

Overall we find our method to be generalizable. In realistic settings and across different languages, attention-based UWS is competitive against the nonparametric Bayesian model (`dpseg`) from Goldwater et al. (2009). Moreover, ours has the advantage of retrieving bilingual annotation for the word segments it produces. Lastly, in this work we also present two corpora for CLD studies (Godard et al., 2018a; Boito et al., 2018), and a dataset for low-resource speech processing with diverse language pairs (Boito et al., 2020a).

Keywords: unsupervised word segmentation, neural machine translation, speech discretization, low-resource approaches, computational language documentation

Modèles et Ressources pour la Segmentation Non Supervisée des Mots basée sur l'Attention

Résumé:

La documentation computationnelle des langues (CLD) est un domaine de recherche qui vise à proposer des méthodologies capables d'accélérer la documentation des langues, en aidant les linguistes à collecter et à traiter efficacement les données de nombreux dialectes, dont certains devraient disparaître d'ici 2100 (Austin and Sallabank, 2013). Pour y parvenir, les méthodes proposées doivent être robustes au traitement de données disponibles en faible quantité, car les corpus issus des initiatives de documentation manquent de volume, et elles sont basées sur la parole, car beaucoup de ces langues sont de tradition orale, sans forme écrite standard.

Dans cette thèse, nous étudions la tâche de segmentation non supervisée en mots (UWS) à partir de la parole. Le but de cette approche est de segmenter la parole en petits morceaux correspondant aux mots de cette langue, sans avoir accès à une transcription écrite. Nous proposons ici de baser le processus de segmentation des mots sur des informations bilingues alignées. Ceci est inspiré par la potentielle disponibilité de traductions, souvent collectées par les linguistes lors de la documentation (Adda et al., 2016).

Ainsi, à l'aide de corpus bilingues composés d'énoncés vocaux et de traductions alignées au niveau des phrases, nous proposons l'utilisation de modèles de traduction automatique neuronale (NMT) basés sur l'attention afin d'aligner et de segmenter. Le traitement de la parole nécessitant des quantités considérables de données, nous divisons cette approche en deux étapes. Nous effectuons d'abord une discrétisation de la parole (SD), en transformant les énoncés d'entrée en séquences d'unités de parole discrètes. Nous entraînons ensuite des modèles NMT, qui produisent des matrices de probabilité d'alignement entre les unités et les traductions de mots. Cette probabilité d'alignement bilingue est utilisée pour segmenter les unités, et la segmentation finale est appliquée au signal vocal.

Pour la tâche de SD, nous comparons 5 approches : 3 modèles bayésiens basés sur les HMM (Ondel et al., 2016, 2019; Yusuf et al., 2020), et 2 modèles neuronaux à quantification vectorielle (van den Oord et al., 2017; Baevski et al., 2020a). Nous constatons que les modèles bayésiens, en particulier le SHMM (Ondel et al., 2019) et le H-SHMM (Yusuf et al., 2020), sont les plus exploitables pour l'UWS basée sur le texte dans notre cadre de documentation. Pour l'alignement et la segmentation, nous comparons 3 modèles NMT

basés sur l'attention : RNN (Bahdanau et al., 2015), 2D-CNN (Elbayad et al., 2018), and Transformer (Vaswani et al., 2017). Nous constatons que le mécanisme d'attention est toujours exploitable dans notre cadre limité (5130 phrases alignées uniquement), mais que les matrices produites par les modèles NMT récents (2D-CNN, Transformer) sont inférieures à celles du modèle RNN, plus simple.

Enfin, notre approche UWS basée sur l'attention est évaluée dans des conditions optimales en utilisant les phonèmes (Boito et al., 2019a), et dans des conditions réalistes en utilisant la sortie des modèles de SD (Godard et al., 2018c). Nous utilisons 8 langues et 56 paires de langues pour vérifier l'impact linguistique de la segmentation basée sur l'information bilingue (Boito et al., 2020b), et nous présentons des extensions pour augmenter la qualité des matrices de probabilité d'alignement produites (Boito et al., 2021).

Dans des contextes réalistes et en utilisant différentes langues, l'UWS basé sur l'attention est compétitif par rapport au modèle bayésien non-paramétrique de Goldwater et al. (2009). De plus, le nôtre a l'avantage de récupérer des annotations bilingues pour les segments de mots qu'elle produit. Enfin, dans ce travail, nous présentons également 2 corpus pour les études de CLD (Godard et al., 2018a; Boito et al., 2018), et un corpus pour le traitement de la parole à faibles ressources avec des paires de langues diverses (Boito et al., 2020a).

Mots-clés: segmentation non supervisée des mots, traduction automatique neuronale, discrétisation de la parole, approches à faibles ressources, documentation computationnelle des langues

---

Contents

Publications 1

Introduction **11**
I Introduction 13
1 Language Documentation 13
2 Unsupervised Word Segmentation (UWS) 16
3 Thesis Contribution 17
4 Thesis Outline 19

State of the Art **23**
II State of the Art 25
1 Monolingual Unsupervised Word Segmentation 25
1.1 Text-based Approaches for UWS 26
1.2 Speech UWS and Clustering 28
2 Towards Bilingual Supervision 29
2.1 Attention-based NMT Models 30
2.2 NMT for Low-resource Languages 41
3 Learning Representations from Speech 43
3.1 Neural Networks for Vector Quantization 44
3.2 NP Bayesian Generative Models 46
4 Discussion 49

Contributions **51**
III Resources 53
1 Mboshi-French Parallel Corpus 53
2 Griko-Italian Parallel Corpus 55
3 MaSS: Multilingual corpus of Sentence-aligned Spoken Utterances 57
4 Contributions Overview 61
IV A Bilingual Attention-based Unsupervised Word Segmentation Model 63
1 Attention-based UWS Model 65
2 Model Evaluation 67
2.1 Average Normalized Entropy 67
2.2 Segmentation Evaluation on the Speech Domain 68
3 Empirical Evaluation of NMT Models for UWS in Low-resource Settings 69
3.1 Experimental Setup 70
3.2 UWS Results 72

3.3 Robustness to Low-resource Settings 74
3.4 Correlation Between ANE and Boundary Scores 74
3.5 ANE and Syntactic Divergence 75
3.6 ANE for Vocabulary Filtering 77
3.7 Discussion 80
4 Investigating Language Impact in the Bilingual UWS model . . 80
4.1 Experimental Setup 81
4.2 UWS Results 83
4.3 Source Language Impact 84
4.4 Analysis of the Discovered Vocabulary 84
4.5 Alignment Confidence 86
4.6 Discussion 89
5 Conclusions 90
V Model Extensions for Attention-based UWS 93
1 Monolingual Data Leveraging 93
1.1 Experimental Setup 94
1.2 Results 96
2 Hybrid Bayesian-Neural Model 97
2.1 Experimental Setup 98
2.2 Results 98
3 Word-length biased NMT Training 102
3.1 Model Definition 103
3.2 Experimental Setup 103
3.3 Results 103
4 Multilingual Supervision for UWS 104
4.1 Experimental Setup 105
4.2 Results 105
5 Discussion 106
VI Attention-based UWS for Speech 109
1 Comparing SD Approaches in Low-resource Settings 110
1.1 Experimental Protocol 110
1.2 Resulting Representation 114
2 Bilingual Attention-based UWS from Speech 116
3 Hybrid Bayesian-Neural Model for Speech 119
4 Discussion 120

Conclusion **123**
VII Conclusion 125
1 Contributions 126
2 Limitations 129

3 Future Work 130
I Appendix 133
A Experiments from Chapter IV 135
1 Investigating Language Impact: Bilingual Baseline Comparison . 136
B Experiments from Chapter V 139
1 Hybrid Model for the MaSS Corpus 139
2 Multilingual Selection 140
C Experiments from Chapter VI 143
II French Translation 149
A Introduction 151
B Résumé des Chapitres 155
1 Chapitre 3 : Les Resources 155
2 Chapitre 4 : Un modèle bilingue de segmentation de mots non supervisé basé sur l'attention 155
3 Chapitre 5 : Extensions du modèle de UWS basé sur l'attention 158
4 Chapitre 6 : UWS basé sur l'attention au niveau de la parole . . 159
C Conclusion 161
1 Contributions 162
2 Limitations 166
Bibliography 167

List of Publications

The following papers were published as part of this thesis.

International Journal Proceedings

- **Boito, M. Z.**, Villavicencio, A., and Besacier, L. (2021) **Investigating alignment interpretability for low-resource NMT.** *Machine Translation Journal: Special Issue on Machine Translation for Low-Resource Languages. Springer.*

International Conference Proceedings

- Godard, P., Adda, G., Adda-Decker, M., Benjumea, J., Besacier, L., Cooper-Leavitt, J., Kouarata, G.-N., Lamel, L., Maynard, H., Mueller, M., Rialland,A., Stueker, S., Yvon, F., and **Boito, M. Z.** (2018). **A very low-resource language speech corpus for computational language documentation experiments.** *International Conference on Language Resources and Evaluation (LREC 2018).*

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

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

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

- **Boito, M. Z.**, Yusuf, B., Ondel, L., Villavicencio, A., and Besacier, L. (2021). **Unsupervised Word Segmentation from Discrete Speech Units in Low-Resource Settings.** Under review. *Submitted to Interspeech 2021.*

French and International Workshops

- **Boito, M. Z.**, Anastasopoulos, A., Lekakou, M., Villavicencio, A., and Besacier, L. (2018). **A Small Griko-Italian Speech Translation Corpus.** *International Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU 2018).*

- **Boito, M. Z.**, Villavicencio, A., and Besacier, L. (2019). **How does language influence documentation workflow? Unsupervised word discovery using translations in multiple languages.** *Scientific Meeting of the Computational, Formal and Field Linguistics Research Group (LIFT 2019).*

- Nguyen, H., Tomashenko, N., **Boito, M. Z.**, Caubrière, A., Bougares, F., Rouvier, M., Besacier, L., and Estève, Y. **ON-TRAC consortium end-to-end speech translation systems for the IWSLT 2019 shared task.** *International Workshop on Spoken Language Translation (IWSLT 2019).*

- **Boito, M. Z.**, Villavicencio, A., and Besacier, L. (2020). **Investigating language impact in bilingual approaches for computational language documentation.** *Joint SLTU and CCURL Workshop (SLTU-CCURL 2020).*

List of Figures

1.1 The differences between the general pipeline for classic (monolingual) unsupervised word segmentation compared to the bilingual approach we propose. The former produces a textual resource, while the latter produces speech segments. 17

2.1 Soft-alignment probability heatmaps from an English-French NMT model. Brighter squares correspond to higher source-to-target probabilities. Target corresponds to rows, and source to columns. Figure taken from Bahdanau et al. (2015). 30
2.2 The general scheme for an attention-based encoder-decoder NMT model. 32
2.3 The general scheme for an encoder and decoder Transformer layer inside the encoder and decoder stacks from Figure 2.2. . . 35
2.4 The general scheme for an 2D-CNN NMT model. 37
2.5 The general structure of a DenseNet Block (top), and the computation flow within each block (bottom). Figure extracted from Elbayad et al. (2018). 37

3.1 A tokenized and lower-cased sentence pair example in our Mboshi-French parallel corpus. 54
3.2 A tokenized and lower-cased sentence pair example in our Griko-Italian parallel corpus. 57
3.3 The pipeline for a given language in the *bible.is* website. 58
3.4 A tokenized multilingual parallel verse from our dataset (Hebrews 10, verse 31). 60

4.1 The general bilingual speech UWS pipeline. It requires as input a parallel dataset made of speech and sentence-level aligned translations. The system outputs word-level segmentation over the speech utterances. *Units* at the end of the first step correspond to the discrete speech units. 64
4.2 Detailed pipeline for the second step of our bilingual UWS pipeline from Figure 6.1. The discrete speech units, or phones, are fed to the decoder network. Soft-alignment probability matrices are extracted for each sentence after training, and transformed in speech segmentation. 65

4.3 Soft-alignment probability matrix heatmaps between speech labels (rows) and french words (columns) for three sentences from the Mboshi-French Parallel Corpus (Chapter III). The higher the soft-alignment probability for a given pair, the darker is the square color. 66

4.4 Soft-alignment probability matrices from the UWS task between English true phones (rows) and French words (columns). ANE values (from left to right) are 0.11, 0.64 and 0.83. The gold segmentation is "BAH1T MAA1MAH0 PAA1PAH0 IH0Z AW1T", which corresponds to the English sentence "But mama, papa is out". 69

4.5 An illustration of the apparent correlation between Sentence ANE and soft-alignment quality. The heatmaps displayed correspond to random sentences sampled from the RNN model trained on the MB-FR language pair. This tendency is observed for all NMT models. 75

4.6 Soft-alignment probability matrices from the alignment complexity buckets 1 to 4 (left to right) for examples with same source length. Darker squares correspond to higher probabilities. The sentence ANE scores are, from left to right, 0.26, 0.40, 0.47 and 0.53. The language pair is French (words) and Mboshi (phonemes). 77

4.7 Average token length of the best bilingual UWS models from Table 4.8, `dpseg`, and reference. 86

4.8 EN soft-alignment probability matrices generated by 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". 87

5.1 An example of the same target sequence, with its monolingual (<`mono`>) and bilingual (<`bi`>) aligned source information. Note that the tags are inserted at the decoder (target). 95

5.2 An illustration of the hybrid model using a sentence from the EN-RO language pair from the MaSS Corpus. The NP Bayesian model (`dpseg`) receives the unsegmented phonemes, producing segmentation. The discovered boundaries are then replaced by a special token (`#`), and bilingual alignment and re-segmentation are jointly performed. 99

5.3 Soft-alignment probability matrix heatmaps for hybrid models trained on sentences from the MB-FR corpus. Darker squares correspond to higher pair alignment probability. The examples are ordered, from left to right, by alignment complexity buckets. The # is the soft-boundary symbol. 100

6.1 The general bilingual speech UWS pipeline. It requires as input a parallel dataset made of speech and sentence-level aligned translations. The system outputs word-level segmentation over the speech utterances. *Units* at the end of the first step correspond to the discrete speech units. 110

6.2 Discrete speech unit segmentation for the same Mboshi utterance by each HMM-based SD system. The black lines denote the true boundaries, and the dashed white lines denote the discrete speech units' boundaries discovered by each system. 111

6.3 Discrete speech unit segmentation for the same Mboshi utterance by each VQ-based SD system. The black lines denote the true boundaries, and the dashed white lines denote the discrete speech units' boundaries discovered by each system. 113

3.1 The soft-alignment probability matrices produced for the same sentence pair, but using different HMM-based SD approaches: HMM (left), SHMM (middle) and H-SHMM (right). Darker squares correspond to higher soft-alignment probabilities. . . . 144

3.2 The soft-alignment probability matrices produced for the same sentence pair, but using different VQ-based SD approaches: VQ-VAE (left), VQ-WAV2VEC-V16 (middle) and VQ-WAV2VEC-V36 (right). Darker squares correspond to higher soft-alignment probabilities. 145

3.3 The soft-alignment probability matrices produced for the same sentence pair, but using different languages to train the **HMM SD model**: Finnish (left), Hungarian (left-to-center), Romanian (center-to-right), and Russian (right). Darker squares correspond to higher soft-alignment probabilities. 146

3.4 The soft-alignment probability matrices produced for the same sentence pair, but using different languages to train the **SHMM SD model**: Finnish (left), Hungarian (left-to-center), Romanian (center-to-right), and Russian (right). Darker squares correspond to higher soft-alignment probabilities. 147

3.5 The soft-alignment probability matrices produced for the same sentence pair, but using different languages to train the **H-SHMM SD model**: Finnish (left), Hungarian (left-to-center), Romanian (center-to-right), and Russian (right). Darker squares correspond to higher soft-alignment probabilities. 148

List of Tables

3.1 Number of sentences, tokens and types for the different sets, for both languages. For Mboshi, audio duration is also detailed. 55
3.2 General statistics for the languages present in the dataset. The metrics were computed on the totality of the corpus (5,130 sentences). 55
3.3 General statistics for the 330 sentence-long Griko-Italian parallel corpus. 56
3.4 A comparison between CMU's multilingual alignment and ours. Text in italic presents alignment mismatches between English and French. We used a slightly different (*non-drama*) version of the Bible, hence the small differences in the displayed texts. 59
3.5 Statistics of the MASS corpus. 60

4.1 Statistics for the three source-target datasets. 71
4.2 Precision (P), Recall (R), and F-score (F) UWS boundary results for the NMT models (RNN, 2D-CNN, Transformer) trained on the three corpora (EN 33k and 5k, MB 5k) in bilingual (real) and monolingual (topline) settings. Best results for each setting presented in bold. 72
4.3 Average Corpus ANE scores over the 5 runs for the different models we trained. Scores $\in [0, 1]$, smaller values being better (lower entropy). 75
4.4 Type discovery precision scores for the alignment complexity buckets, and for the totality of the corpus (All buckets). Results in each of the rows are cumulative and use the Alignment ANE thresholds indicated in the first column. 78
4.5 Type discovery recall scores using Alignment ANE for keeping the most confident `(type, translation)` pairs. Results in each row are cumulative and use the Alignment ANE thresholds indicated in the first column. 79
4.6 Top 10 low and high ANE ranking for the discovered types (EN 5k), with gold transcription and aligned information between parentheses (respectively). "INV" means incorrect type. . . . 79
4.7 Statistics for the subset of 5,324 sentences of the MaSS corpus. 82

4.8 UWS Boundary F-score (left) and BLEU score (right) results for all language pairs using the RNN model. The columns represent the target of the segmentation, while the rows represented the translation language used. Darker squares represent higher **column** scores. Best scores presented in bold. Better visualized in color. 83

4.9 Type discovery recall scores for the bilingual-rooted UWS models. The columns represent the target of the segmentation, while the rows represented the translation language used. Darker squares represent higher **row** scores. Best (column) scores presented in bold. Better visualized in color. 85

4.10 Top 10 low Alignment ANE ranking for FR models trained with EN, ES and RU supervision. Each column brings the discovered types with gold transcription and aligned information between parentheses (respectively). "INV" means incorrect type. 88

5.1 Statistics for the MB transcriptions for both sets, as well as for the totality of the corpus (All). The monolingual and bilingual sets have a type intersection of 1,338. 94

5.2 UWS Precision (P), Recall (R), and F-score (F) for the discovered boundaries and types. The Base model corresponds to the RNN result obtained in Chapter IV. All segmentations are scored over the totality of the corpus (5,130 sentences), including type results for the 1st step. 96

5.3 UWS boundary and type discovery scores for the RNN (base) and `dpseg` models from Chapter IV, and for the hybrid model. 99

5.4 General statistics for the produced segmentations. 101

5.5 Precision type discovery scores for the alignment complexity buckets, and for the totality of the corpus (All buckets), by using the matrices produced by the hybrid model. Results are cumulative and use the Alignment ANE thresholds indicated in the first column. The difference between the obtained scores and the ones from the base model (Table 4.4, Chapter IV) is displayed between parentheses. The buckets from 1 to 4 correspond to increasing alignment complexity scenarios (4 is the hardest). 102

5.6 UWS Boundary and Type scores for the RNN (Base) and word-length biased model. 104

5.7 General statistics for the produced segmentations. 104

5.8 UWS Boundary F-score results for bilingual (left) and multilingual (right) UWS models. The bilingual results are the same from Table 4.8. Darker squares represent higher **column** scores. Best scores presented in bold. Better visualized in color. . . . 106
5.9 Boundary and Type UWS F-scores for base model (1), `dpseg` segmentation baseline (2), and the proposed model extensions (3-5). 107

6.1 Word Boundary Recall for the sequences generated by the 5 SD models before (left) and after (right) the silence post-processing. Results use the Mboshi utterances as input. For VQ-WAV2VEC, **VX** corresponds to the version of the model with a **V**ocabulary of **X** units. 115
6.2 Statistics for the produced discretization (unsegmented) using Mboshi utterances, and before (left) and after (right) the silence post-processing. For VQ-WAV2VEC, **VX** corresponds to the version of the model with a **V**ocabulary of **X** units. 115
6.3 Boundary F-scores results for the UWS models (`dpseg` and attention-based) using the SD models (1-6) and true phones (7, from Chapter IV), and applied to the Mboshi-French parallel corpus. Best results presented in bold. 117
6.4 Statistics for the Mboshi-French (MB-FR) and MaSS datasets computed over the text (FR), or over the audio and phonetic representation (MB, FI, HU, RO and RU). 118
6.5 UWS Boundary F-scores for the MaSS dataset using HMM-based models (+SIL only) and true phones (Chapter IV). Best results for each language and SD model presented in bold. . . 119
6.6 Precision (P), Recall (R), and F-score (F) boundary UWS results for the Mboshi-French parallel corpus using the HMM-based models (+SIL only). Best results presented in bold. . . 120

7.1 Examples of two translation word clusters, and the discovered types within. Extracted from the EN-FR 5k model from Chapter IV. The cluster on the left is *assassin* (murderer), the one on the right is *nourriture* (food). For ANE scores, lower is better. 131

1.1 Number of trainable parameters inside the models trained on different datasets (English (EN) 33k and 5k, Mboshi (MB) 5k) for both monolingual and bilingual settings from Section 3, Chapter IV. The amount of trainable parameters depend on the vocabulary size, due to the embedding layer and the softmax projection inside the decoder network. 135

1.2 Type discovery precision, recall and F-score results for the bilingual models from Section 4, Chapter IV. The columns represent the target of the segmentation, while the rows represented the translation language used. Darker squares represent higher **column** scores. Best scores presented in bold. 135
1.3 UWS Boundary F-score results for the proportional baseline. The columns represent the target of the segmentation, while the rows represented the translation language used. Darker squares represent higher **column** scores. Best scores presented in bold. Better visualized in color. 136

2.1 UWS 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 **column** scores. Best results in bold. Better visualized in color. 139
2.2 Intersection between the correct types discovered by both `dpseg` and hybrid models. We notice that the target language of the segmentation (columns) has an impact in the acceptance of soft-boundaries by the NMT model. 140
2.3 Boundary UWS F-scores for the multilingual selection approaches applied to the RO language, from Table 5.8. The bilingual baseline scored 62.8. 141

3.1 Statistics for the segmentation produced by our UWS models for the Mboshi corpus, and by using the different SD approaches. TTR corresponds to Type-Token Ratio. 143

Introduction

Chapter I
Introduction

In the scope of computational approaches for language documentation, in this thesis we propose a bilingual unsupervised word segmentation approach from speech. This proposed model grounds speech segmentation in the sentence-aligned word translations, solving the task without the use of manual transcriptions, and in low-resource settings.

This chapter is organized as follows. Section 1 presents the language documentation field, and Section 2 defines the task of unsupervised word segmentation. Section 3 summarizes the contribution of our work, and Section 4 outlines this dissertation's chapters.

1 Language Documentation

Language documentation, as defined by Austin (2012), is the subfield of linguistics that deals with creating multipurpose records of languages through audio and video recording of speakers and signers. It includes annotation, translation, preservation, and distribution of the resulting material (e.g. grammars, dictionaries, text collections).

The goal of this process is to *document* the languages studied, in other words, to preserve them through the creation of well-organized, long-lasting corpora and resources. These can be posteriorly explored for subsequent research in the target language, or they can be used for practical technological applications such as machine translation and speech recognition. This data can also be the starting point for *language revitalization* initiatives (Pine and Turin, 2017).

One of the main targets of language documentation are the *endangered languages*. These are defined as a subset of existing languages whose number of speakers have been significantly decreasing, leaving them at risk of falling out of use as their speakers perish or shift to different languages. In *The Cambridge Handbook of Endangered Languages*, Austin and Sallabank (2011) estimated that, from the approximately 7,000 currently spoken languages, at least 50% of them will go extinct by 2100.

Between the many reasons that are causing this language shift and the homogenization of the spoken languages across the globe, it is notable the

impact of neocolonialism and globalization (Austin and Sallabank, 2011). Endangered languages are spoken in isolated communities across the globe. As these communities start getting integrated into economic pipelines, the language spoken in larger economic centers is carried into these places. Rural exodus also causes an impact, as the younger generations migrate to larger cities in search of better job opportunities, significantly reducing their contact with their native language.

Some argue that language extinction in its core is a natural phenomenon (Ladefoged, 1992). Even so, the impact it causes on communities is widely recognized. Languages embody unique world-views, value systems, philosophies and particular cultural features. Their extinction results in irrecoverable loss of unique cultural, historical, spiritual and ecological knowledge, useful not only for the community, but for countless others (Drude et al., 2003; Bird, 2018; UNESCO, 2020). Moreover, the loss of languages also represents a scientific problem, as future linguits will only have access to a fraction of the world's linguistic diversity available for study (Austin and Sallabank, 2011; Grenoble and Whaley, 1996; Nettle et al., 2000).

In this context, it does not help that most of the world's languages are not actively written, even the ones with an official writing system (Bird, 2011). For documenting these oral languages, audio recordings are usually collected, and then transcribed. However, this 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).

Moreover, the documentation process is iterative, and the transcriptions are expected to be revised several times before the final product (Crowley, 2007). Because of that, field linguists spend a large amount of their time transcribing and revising materials, and this makes documentation very human expensive and slow. Brinckmann (2009) defines this as the *transcription bottleneck problem* of documentation initiatives.

For attenuating this bottleneck, recent work suggested replacing transcriptions by multilingual links, added to the audio recordings. These can come in the form of sentence or word-level translations (Adda et al., 2016), or in the form of overlapping labels over the audio's time frames (Bird, 2021). These approaches highlight the content present in the audios, instead of creating extensive transcriptions. By doing so, they treat *transcription as an observation* (Cucchiarini, 1993), instead of considering it the ultimate goal of documentation.

However, in order to process and extract information from this new form of corpora, technology needs to step in, providing robust computational methods able to deal with this data that is: low-resource, multilingual, and sometimes multimodal (e.g. images, videos). The recent emergence of the **Com-**

putational Language Documentation (CLD) field tries to propose answers to that. It brings together linguists and technology experts in order to provide methodologies and models for automatically processing data and for assisting linguists, attenuating the human resources and the time needed for documenting languages.

The following are examples of recent work for CLD. Focusing in the production of transcriptions from speech, there are pipelines for obtaining manual (Foley et al., 2018) and automatic (Michaud et al., 2018; Matsuura et al., 2020) ones, for aligning existing transcriptions to audio (Strunk et al., 2014), and for automatically increasing transcription quality by using aligned translations (Anastasopoulos and Chiang, 2018a). Focusing in the information present in these transcriptions, there are methods for monolingual (Lignos and Yang, 2010; Goldwater et al., 2009; Godard, 2019) and bilingual (Duong et al., 2016; Boito et al., 2017) low-resource unsupervised word segmentation, and for lexical unit discovery without textual resources (Bartels et al., 2016).

Nonetheless, as this recent research field thrives by proposing methods for processing speech and text in extreme low-resource settings, Bird (2020) denounces the lack of real application of proposed approaches in the targeted communities. In *Decolonising Speech and Language Technology*, he says the following:

> *"For a fraction of the world's languages – perhaps no more than 10% – the dominant ideology is that a language is a communication tool, a public corpus, readily interchangeable with others, raw data for commercial exploitation by algorithms, (...). For the remaining 90%, language tends to be oral, emergent, untranslatable, tightly coupled to a place. Representatives from the former may approach the latter with a sense of entitlement: to project, to save, to know, to mine. They may be unwilling to hear local aspirations, unable to see how differently language functions in each place. It is simply a given that language loss must be halted, and that technology is up to the task."*

Indeed the mainstream vision of language considers it as a commodity: from data gathering procedures, to the classification of languages ranging from low to high *resource*. If a language is to embody its community of speakers' culture and value system, documentation should not exist in isolation from this community. Instead, it should be performed in collaboration with them, and respecting their wishes for any developed technology and for their own language. In other words, the end goal for CLD should be to develop *for the communities*, and not only from their data.

2 Unsupervised Word Segmentation (UWS)

In this thesis, our investigation covers one of the first tasks performed during documentation: word segmentation. Commonly, this task occurs together or right after the transcription of the audio data. It consists of joining a sequence of phones[1] into larger units representing words, and thus, providing a *segmentation* of the transcription at the word level.

There are some very successful pipelines for high-quality low-resource unsupervised[2] word segmentation (Goldwater, 2007; Johnson and Goldwater, 2009). However, for discovering boundaries between phones, these approaches require the existence of an extensive transcription of the audio data. As the transcription process is the bottleneck of documentation initiatives (Brinckmann, 2009), this results in word segmentation being a very difficult resource to obtain.

Meanwhile, Adda et al. (2016) and Bird (2021) highlight translations in high-resource languages as an inexpensive way of labeling the information present in the audios collected during the documentation process. These translations are considered *inexpensive* because they are usually collected together with the audio by linguists for organizational purposes, such as labeling or indexing the content of the audios.

Based on that, in this thesis we defend a more realistic approach for word segmentation, which takes advantage of the audios and their translations, instead of being rooted in the audio transcription only. The goal of our approach is to segment directly from the speech signal, while using the translations as a weak form of supervision. The expected output is then a collection of speech segments, corresponding to words, aligned to bilingual annotation.

The task we propose is more challenging than segmenting from transcriptions: it combines low-resource audio processing with the weak supervision of translations. However, as it does not require the manual transcription of the audio, it has the potential for reaching a larger number of low-resource languages than the classic approach. Figure 1.1 highlights the difference between our approach, which we refer to as *bilingual*, and the classic pipeline for word segmentation, referred as *monolingual*.

[1] Phones are language-agnostic representations of any distinct speech sound or gesture.

[2] The task is defined as *unsupervised* because it does not require a dictionary or language priors as input for guiding the segmentation.

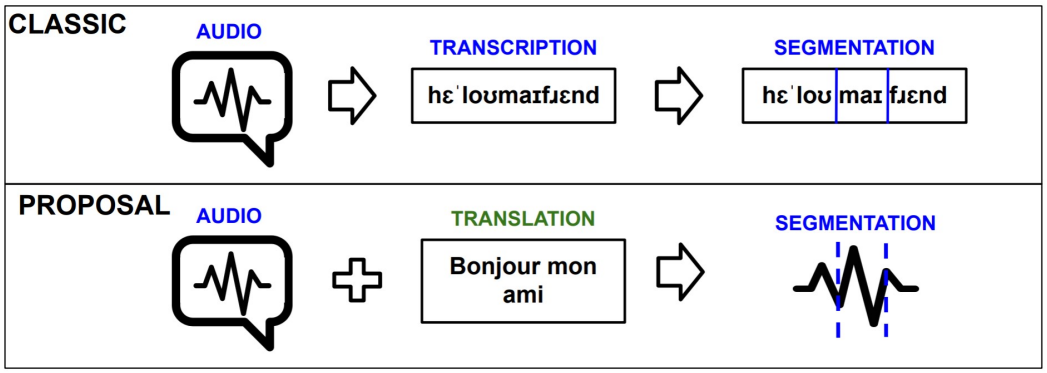


Figure 1.1: The differences between the general pipeline for classic (monolingual) unsupervised word segmentation compared to the bilingual approach we propose. The former produces a textual resource, while the latter produces speech segments.

3 Thesis Contribution

This thesis is one of the many computational language documentation approaches which aim to produce technology useful for processing data in the context of language documentation. In particular, we propose an approach for unsupervised word segmentation from speech. Solving such a task from the speech signal, instead of segmenting in the textual domain, is motivated by extensive transcriptions being a known bottleneck of data collection processes (Brinckmann, 2009).

Moreover, considering translations as an inexpensive process for data labeling (Adda et al., 2016), we chose to include these as weak supervision for our utterances during segmentation. Thus, we consider our segmentation process to be *bilingually grounded*, and during this thesis we discuss how language impacts the quality of the segments discovered.

Our model is made of two components: (1) speech discretization, and (2) text-based alignment and speech segmentation. This separation is necessary in order to attenuate the challenge of speech processing in very low-resource settings. The goal of the first component is to produce sequences of discrete speech units (phones), exploitable in low-resource settings, using only a few hours of speech. Consequently, in this thesis we investigate the quality and exploitability of speech discretization models in our documentation setting.

For the second component, we use neural networks for creating alignment probability matrices between the speech discretization and their sentence-level translation. This is performed by a special layer inside neural machine translation models called *attention*, whose output can be seen as bilingual

soft-alignment. This soft-alignment is used for producing segmentation over the discrete speech units, which is then carried to the original speech signal. Thus, in this work we extensively investigate the quality and exploitability of the attention mechanism in our setting, and we also introduce a task-agnostic metric for assessing the *alignment confidence* of soft-alignment probability matrices (Boito et al., 2019a).

This unsupervised word segmentation pipeline in two steps that we propose is compared against a well-established baseline (Goldwater, 2007), and across different languages (Godard et al., 2018c; Boito et al., 2019a, 2020b). Focusing on documentation scenarios, we propose an extension which considers the availability of partial transcriptions, and a model which leverages preexisting segmentation into the bilingual alignment model (Boito et al., 2021).

Finally, the model we propose requires a bilingual corpus made of speech utterances and aligned sentence translations. In order to realistically test our models and allow the research community to do the same, we gathered and published three datasets, which we present in this work (Godard et al., 2018a; Boito et al., 2018, 2020a).

Research Questions: The proposed model results in the following research questions, which we investigate throughout this thesis.

- **Q1:** Focusing on the first step of our pipeline, can we use low-resource speech discretization approaches for producing an exploitable discrete representation for direct application to text-based UWS approaches?

- **Q2:** Focusing on the second step of our pipeline, is the attention mechanism from neural machine translation approaches directly interpretable in low-resource settings? Can we use it for segmenting a sequence of phones with respect to the aligned translation words?

- **Q3:** What is the performance of the proposed approach compared to a strong baseline (Goldwater et al., 2009)?

- **Q4:** Considering that we propose to ground segmentation in bilingual information, how does this supervision impact the quality of the segmentation?

- **Q5:** Considering that partial transcription or intermediate segmentation from documentation initiatives might exist, can we include these into our UWS pipeline?

Contributions: For answering these research questions, we produce the following contributions. Research questions and chapters are presented between parentheses.

- **C1:** A thorough comparison of recent speech discretization approaches for low-resource speech processing, focusing on their direct applicability to text-based UWS. **(Q1)**
  (Chapter VI)
- **C2:** A study of the direct interpretability of the attention mechanism in neural machine translation models, and in low-resource settings. **(Q2)**
  (Chapter IV)
- **C3:** A comparison between unsupervised word segmentation approaches: our attention-based model and two baselines(the well-established model from Goldwater et al. (2009), and a *proportional* bilingual model). **(Q3)**
  (Chapter IV and VI)
- **C4:** The investigation of language-related impact in our pipeline, focusing on the quality of the segmentation discovered by using different languages for grounding the segmentation of a target language. **(Q4)**
  (Chapter IV and VI)
- **C5:** The proposal of pipeline extensions for incorporating extra information (transcriptions, segmentation) into the segmentation model. **(Q5)**
  (Chapter V)
- **C6:** The gathering and publishing of three datasets useful for low-resource and computational language documentation approaches.
  (Chapter III)

4 Thesis Outline

This thesis is organized as follows.

Chapter II. We present the state of the art for unsupervised word segmentation in the textual domain. We also discuss *neural machine translation* models as a proxy for obtaining bilingual alignment between sentence-level aligned text. Targeting speech, we present models for *Speech Discretization*, which are able to produce *discrete speech units* from speech without the use of any transcription. We finish this chapter by linking these different components, proposing a model which operates in two steps: (1) speech discretization, and (2) text-based alignment and speech segmentation.

Chapter III. We present three datasets we published during this thesis, and that we use for our experiments. Two of them are from oral and potentially endangered languages: the *Mboshi-French Parallel Corpus* (Godard et al., 2018a), and the *Griko-Italian Parallel Corpus* (Boito et al., 2018). Both represent a low-resource setting (respectively 5,130 and 330 aligned sentences), and present speech-to-text alignments. We also detail the collection and processing of a third dataset: the *MaSS dataset* (Boito et al., 2020a). This dataset is a multilingual speech-to-speech and speech-to-text collection with 56 language pairs. We finish this chapter with a quick overview of how the community has been using these datasets.

Chapter IV. We detail the second step of our pipeline that, from a given speech discretization and its sentence-level aligned translation, retrieves bilingual soft-alignment from neural machine translation models. This soft-alignment is then used for producing attention-based speech segmentation. For assessing the feasibility of our proposal, we evaluate models trained using a perfect discretization, which corresponds to the true phones in the language we want to segment. We present results across three different neural machine translation models, and by using Average Normalized Entropy for assessing alignment quality (Boito et al., 2019a). Moreover, we also showcase the language-related impact in our bilingual models by training 56 bilingual models from 8 different languages (Boito et al., 2020b).

Chapter V. We study three extensions for the best segmentation model from Chapter IV. We investigate (1) using partial annotations for pretraining the model, (2) the incorporation of pre-existing segmentation into training by using them as *soft-boundaries*, and (3) the biasing of the attention layer for reducing over-segmentation. We present the comparison between these models (Boito et al., 2021), and some less-successful experiments for grounding our segmentation using multilingual annotations.

Chapter VI. We compare five speech discretization approaches in low-resource settings, focusing on their direct exploitability to our task. From these, we investigate three Bayesian and two neural approaches. We then present our complete pipeline for unsupervised word segmentation from speech (Godard et al., 2018c), in which we compare the different discretization models for training segmentation models in five different languages.

Chapter VII. We conclude our work by reviewing and summarizing the findings of the investigations presented from Chapter IV to VI. We then discuss

benefits and limitations of the proposed pipeline, and possible extensions.

State of the Art

Chapter II

State of the Art

In the last chapter we introduced the goal of this thesis: a bilingual approach for unsupervised word segmentation from speech, and in low-resource settings. In order to contextualize the reader, in this chapter we review past work on monolingual unsupervised word segmentation (Session 1), bilingual approaches for processing text (Session 2), and approaches for extracting information from speech (Session 3). We end this chapter with a discussion about how these different fields relate to the approach developed in this thesis (Section 4).

1 Monolingual Unsupervised Word Segmentation

Previously we defined Unsupervised Word Segmentation (UWS) as a language documentation task. However, this task has also been extensively investigated by the language acquisition field (Saffran et al., 1996; Brent, 1999; Goldwater, 2007; Johnson and Goldwater, 2009; Johnson et al., 2014; Lignos and Yang, 2016; Larsen et al., 2017), which is interested in understanding and mimicking how infants learn language. Based on the observation that children learn to speak without the aid of written words or large amounts of supervision, this field aims towards methods for extracting meaningful information from multimodal data using a limited number of examples. They argue that approaches should emulate human learning, which is in nature multimodal (interaction between vision, speech, gestures), instead of relying on very large datasets of labeled data.

Because of that, many parallels can be drawn between approaches for language acquisition and the ones for language documentation, especially since both aim to extract information from small amounts of data, which can be of multimodal nature. Their input also presents similar characteristics: small sentences, and tailored vocabulary. For documentation, vocabulary is tailored in order to isolate specific phenomena being studied. For acquisition, it is due to the target demographic's age.

In this session we review work in UWS from both fields, without distinction. Instead, we separate the work in terms of their input representation:

segmentation for text is presented in Section 1.1, and speech-based approaches are presented in Section 1.2.

1.1 Text-based Approaches for UWS

Based on the Bayesian properties from early computational models for word segmentation (Saffran et al., 1996; Brent, 1999), Nonparametric Bayesian (NB) models for UWS and morphological analysis were introduced by Goldwater (2007). They are able to achieve very competitive UWS results using small quantities of data (Goldwater et al., 2009), although parameter optimization is considerably hard since there is no objective criterion to find hyperparameters in a fully unsupervised manner (Kawakami et al., 2019).

For these *nonparametric* models, the number of parameters grows together with the size of the corpus, which makes them very efficient even when working with only a few examples. Moreover, their structure makes them very flexible. They are defined by two components: a lexicon generator and an adaptor. We now describe these two components as defined in Goldwater et al. (2009).

The *lexicon generator*, which is task-dependent, models the lexicon items, and it can be unigram or bigram-based. For the former, words are considered statistically independent events,[1] while the latter considers every word dependent on a single previous word of context. The probability of novel lexical items is defined as the product of the probability of each of its phonemes, which ensures very long words will be dispreferred. Novel lexical items have a high generation probability at first, and this probability decreases as more word tokens are generated, which makes the model penalize large vocabularies. Moreover, the probability of a lexical item depends on the number of times it already occurred in the lexicon. This pushes the model towards a power-law distribution behavior, where only a few words are very frequently used, such as the behavior of natural languages (Powers, 1998).

The *lexicon adaptor* assigns frequencies to the lexical items from the lexicon generator. Assuming that the hypotheses under consideration by the model are possible segmentations over the word sequences, consistent word sequences, in respect to the corpus, receive maximum prior probability. This makes the posterior probability of a sequence determined by its prior, which is computed by considering that every word sequence from a segmentation hypothesis is created according to a particular probabilistic generative process.

This definition of the word segmentation task in two parts, one modeling the construction of words (generator), and the other assigning frequencies to these words (adaptor), makes the unigram and bigram models instances of

[1] In their study, Goldwater et al. (2009) states that the unigram model tends to undersegment.

the Dirichlet and Hierarchical Dirichlet Processes. Inference is performed by the use of Gibbs sampling for sampling from the posterior distribution over segmentations for both models. Throughout this work, we use this model, referred as `dpseg`,[2] as a **segmentation baseline.**

Since its introduction, many works have extended and improved `dpseg`. Johnson and Goldwater (2009) introduced adaptor grammars for inference, which Godard et al. (2018b) applied for taking into account the expertise of linguists for studying word hypotheses during language documentation. Mochihashi et al. (2009) proposed a nested hierarchical Pitman-Yor Process (PYP) language model for modeling spelling inside the word model, and Neubig (2014) replaced `dpseg`'s Dirichlet Process by a PYP, allowing the parallelization of the sampling process by blocked sampling, which made the resulting model faster. The former was used as part of a joint segmentation and translation pipeline in Nguyen et al. (2010). The latter was exploited by Adams et al. (2015) for inducing a bilingual lexicon in language documentation scenarios. Godard et al. (2016) compared `dpseg` against these two PYP-based models, finding that, while all models tend to over-segment the input, `dpseg` still led to better segmentation results in true low-resource settings (less than two thousand sentences).

Another branch of successful statistical models for word segmentation are the *generic* unsupervised models (Liang and Klein, 2009; Berg-Kirkpatrick et al., 2010). These can be seen as *generic* because they are not designed considering UWS as the end goal. Instead, they separate modeling from optimizing, generating a model which can be applied to different tasks by changing the optimization objective (e.g. document classification, word alignment, word segmentation). Liang and Klein (2009) use online stepwise EM optimization for UWS, obtaining promising results. Berg-Kirkpatrick et al. (2010) then propose models based on locally normalized generative decisions with a feature-enhanced EM optimization algorithm. Their model surpasses strong baselines (Liang and Klein, 2009; Johnson and Goldwater, 2009), including `dpseg`, in terms of accuracy using the Bernstein-Ratner Corpus (Flokstra, 1987). Moreover, they do so while proposing a model which is considerably simpler to optimize than NB models.

Working on a phonemic level and also aiming to develop a simple model for UWS, Lignos (2011, 2012) introduced an online bootstrapping algorithm for modeling word segmentation in language acquisition settings. Their segmentation model has no access to previous sentences when segmenting, only keeping the produced lexicon, and segmenting on a left-to-right fashion. They were able to achieve very good results segmenting adult utterances from the

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

CHILDES database (MacWhinney, 2004), while keeping the model simple and relatable to the way infants acquire language. Moreover, optionally their algorithm can include stress patterns, which allows the model to generalize to different languages. Doyle and Levy (2013) also investigates the use of stress patterns. Using phonemes as input, they treat stress patterns and word boundaries as a joint inference task, verifying, as in Lignos (2011), that these stress cues increase segmentation performance. Elsner et al. (2013) used NB models to study the benefits of executing the tasks of word segmentation, lexical acquisition and phonetic variability together. Their UWS model was able to slightly improve upon their baseline, while better relating to the way children learn language.

On a different trend, recently Recurrent Neural Networks (RNN) were discovered to be very good tools for modeling long-range dependencies, a characteristic that makes them ideal for processing language (Mikolov et al., 2010). Moreover, the block-like nature of neural networks, which easily allows for the addition and removal of different processing layers (blocks), makes them considerably easy to implement and deploy.

Kawakami et al. (2019) is an example of a monolingual neural model for UWS. They differentiate themselves from the statistical models mentioned above by unifying the segmentation with language modeling, while also allowing for multimodal information, in the form of pictures, for grounding word meaning. They achieve new state-of-the-art results in UWS, comparing their work with Goldwater et al. (2009) and Berg-Kirkpatrick et al. (2010). Moreover, the authors highlight how the previous models might contain English-specific design considerations, which might be a limitation when applying these models to different languages.

1.2 Speech UWS and Clustering

All the approaches mentioned so far focus on textual representation, having sequences of characters or phones as their input. However, there is an increasing interest from the community in creating and adapting models to deal with speech signals. This is because, directly segmenting words from speech not only helps when transcriptions are not available, but it is also closer to the way humans learn languages (Lignos and Yang, 2016).

The Zero Resource Speech (ZRC) Challenge is a campaign that instigates scientists to develop unsupervised methods for processing and recognizing structures in speech from scratch (no supervision, limited amounts of data). The challenges from 2015 (Versteegh et al., 2015, 2016) and 2017 (Dunbar et al., 2017) presented tracks for Unsupervised Term Discovery (UTD), which falls very close to UWS. This task's aim is to segment utterances into word-like

segments, differing from UWS by not necessarily producing a full segmentation of the target speech signal. Many of the works mentioned in this session are entries from these ZRC campaigns.

Lee et al. (2015a) present a probabilistic framework for jointly inferring word segmentation and discovering lexicon from acoustic signals. Their findings suggest that modeling phonetic variability is critical for inferring lexical units from speech. Räsänen et al. (2015) proposes a *cognitively-motivated* syllable-based pipeline: they start by extracting syllable-like units from the signal, which are clustered considering the features' similarity across segments. The potential patterns are extracted by searching for recurring combinations of the segments. Their model was very effective in recovering speech segments corresponding to lexical words.

Kamper et al. (2016) present a novel Bayesian model for segmenting fixed-dimensional speech embeddings. They segmented and clustered unlabelled speech utterances into word hypotheses units by using a Bayesian Gaussian Mixture Model (GMM). This resulted in roughly 10% of performance improvement in terms of Word Error Rate (WER), compared to a traditional HMM-based baseline. This model is extended in Kamper et al. (2017), where they replace the clustering component by the Embedded Segmental K-means algorithm, which makes the model considerably faster and lighter in terms of hyper-parameters. The resulting approach is a trade-off between pure Bayesian models, which always converge but are very heavy, and the cognitively-motivated model from Räsänen et al. (2015), which has no convergence guarantees but is very fast to train. This system was the best submission on track 2 for the Zero Resource Speech Challenge 2017.[3]

Lastly, Lyzinski et al. (2015) evaluates graph-clustering methods, finding that modularity-based clustering results in better UTD performance. They also test supervised deep-learning bottleneck features trained in English, in order to improve their model's performance in a low-resource language. The addition of the neural features allowed them to perform near on par with a high-resource UTD system by using considerably less data during training.

2 Towards Bilingual Supervision

Recently, encoder-decoder architectures equipped with *attention mechanisms* emerged as a popular solution for addressing sequence-to-sequence (seq2seq) problems for a variety of tasks.[4] These include Automatic Speech Recognition

[3]Available at: `http://www.zerospeech.com/2017/results.html`

[4]Surveys on different attention mechanisms for NLP tasks are presented in (Hu, 2019; Galassi et al., 2019).

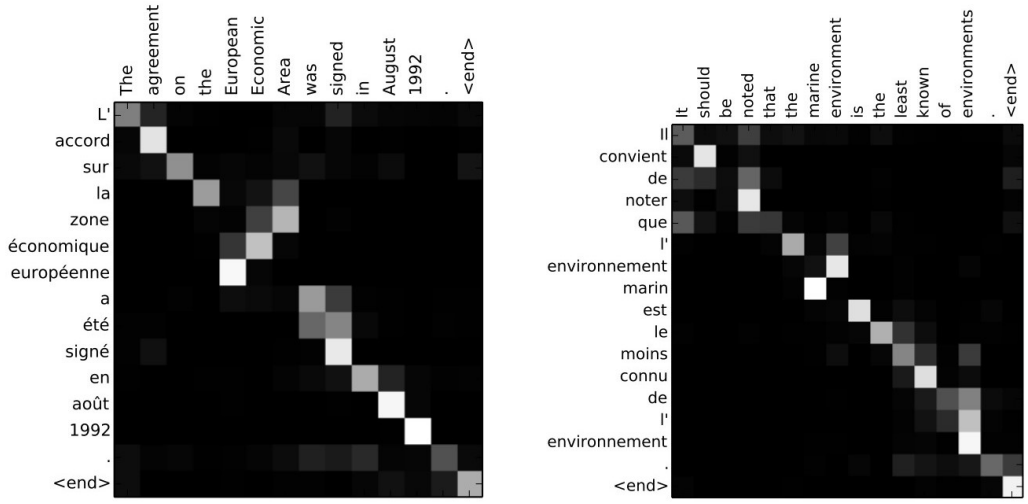


Figure 2.1: Soft-alignment probability heatmaps from an English-French NMT model. Brighter squares correspond to higher source-to-target probabilities. Target corresponds to rows, and source to columns. Figure taken from Bahdanau et al. (2015).

(Watanabe et al., 2017; Chorowski et al., 2015), Text-to-Speech Synthesis (Wang et al., 2017; Shen et al., 2018), and Neural Machine Translation (NMT) (Bahdanau et al., 2015; Elbayad et al., 2018; Vaswani et al., 2017; Gehring et al., 2017; Sutskever et al., 2014). For the latter, popular leaderboards, such as WMT 2014 and IWSLT 2015, have been dominated by these attention-based approaches for years now.[5]

In the scope of this work we are interested in methods for integrating translations into a UWS pipeline in low-resource settings. Considering that translation models are by nature bilingual, we find inspiration in the work on attention-based NMT. In the next section we review some of the work in this field (Section 2.1), comparing attention mechanisms and discussing methods for exploiting and analysing them. We then present literature on low-resource NMT models (Section 2.2), discussing the challenge of adapting neural networks for low-resource settings.

2.1 Attention-based NMT Models

The attention mechanism in seq2seq models provides a dynamic bridge between source and target representations in the form of the weighting of the source sequences. For NMT, it was first introduced in Bahdanau et al. (2015),

[5]Leaderboards available at: `https://paperswithcode.com/task/machine-translation`

and it replaced the fixed length vectors which were used prior (Sutskever et al., 2014), and that limited the performance of the resulting translation model when dealing with long sentences.

The attention layer weighting can be seen as a query searching problem, in which the target is the obtained result, and the goal is to search in the source input for the key, query pair which satisfies the obtained value. In practice, this layer is implemented as a combination of projections and non-linearities which consume source and (masked) target sequences, producing weights over the source sequence. These are then transformed into probabilities by a softmax projection.

An interesting feature of attention is the possibility of, posterior to training, visualizing the learned weights between source and target sentences in the form of probability matrices. Such is the example in Figure 2.1, taken from Bahdanau et al. (2015). There, the probabilities generated by their English-French NMT model for two random sentences from their training dataset are presented in the form of heatmaps. Because these source-to-target probabilities, in the context of translation, might be a good representation of what the bilingual alignment between the languages looks like, these visualizations are referred to as *soft-alignment* probability matrices.

In the sections that follow (2.1.1, 2.1.2 and 2.1.3) we present the three different architectures for attention-based NMT we use in our work, focusing on their attention implementation.[6] In Section 2.1.4 we present work focused on the *interpretability* of the attention mechanism.

2.1.1 Basic Encoder-Decoder Attention

The general scheme for attention-based encoder-decoder NMT architectures is illustrated in Figure 2.2. The input for these systems is a parallel dataset of sentence-level aligned sentences. These are first projected into an embedding layer, and then fed into their respective stacks (step 1). In the encoder stack, source sequences are reduced into a sequence of source annotations, which are sent to the attention layer (step 2). For every target token, this layer weights the source annotations, outputting a context vector (step 3). This vector captures the *importance* of every source token for the generation of each target token. This is used, together with the context given by the last token generated by the decoder stack, for generating the next target token (step 4). This process is repeated until the End-Of-Sentence (EOS) token is produced (step 5).

Finally, from the model's predictions, the cross-entropy loss is computed, as in Equation 2.1. There, $|S|$ is the sentence length, and $|V|$ is the target

[6] An extensive survey on attention-based NMT is available at Yang et al. (2020).

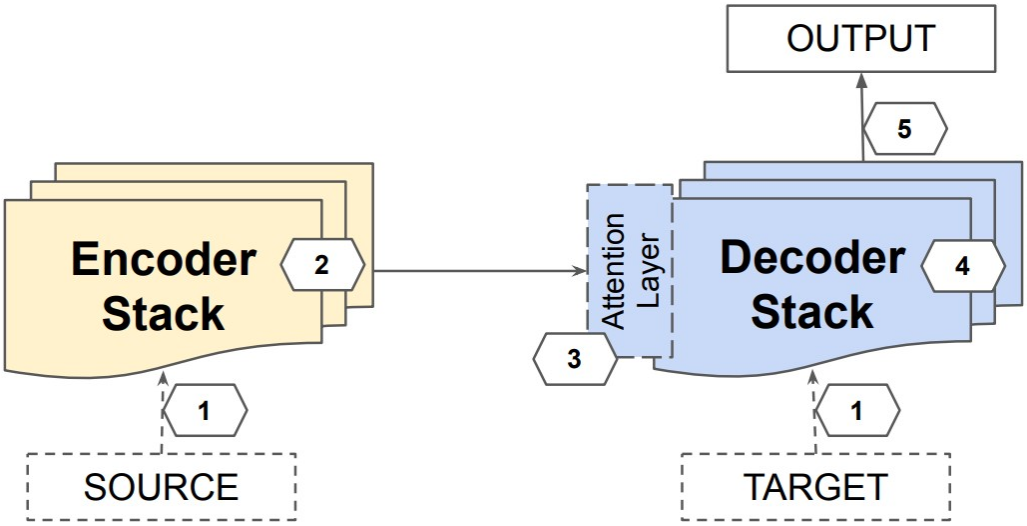


Figure 2.2: The general scheme for an attention-based encoder-decoder NMT model.

vocabulary length. In summary, this loss function sums over the negative log likelihoods that the model gives to the correct translation word ($f(i,j) = 1$) at each position of the output sentence.

$$\mathcal{L}_{NLL} = -\sum_{i=0}^{|S|}\sum_{j=0}^{|V|} f(i,j) \times log(y'_{i,j}) \tag{2.1}$$

$$f(i,j) = \begin{cases} 1 & \text{if } i = j, \text{ (predicted and correct words match)} \\ 0 & \text{otherwise} \end{cases} \tag{2.2}$$

From this class of models, we highlight the attention-based RNN encoder-decoder model from Bahdanau et al. (2015). It combines a bidirectional LSTM encoder with an unidirectional LSTM decoder. In this model, 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 (translation context). The attention is the result of the weighted sum of the source annotations H (with $H = \{h_1, ..., h_{|s|}\}$) and their probabilities α (Equation 2.3) obtained through a feed-forward network *align* (Equation 2.4). Throughout this work, we will refer to this model as the **RNN model**.

$$c_t = Att(H, s_{t-1}) = \sum_{j=1}^{|s|} \alpha_{t,j} h_j \tag{2.3}$$

$$\alpha_{t,j} = \text{softmax}(align(h_j, s_{t-1})) \tag{2.4}$$

Subsequently, Luong et al. (2015) extend this definition of attention, creating the concept of local attention, which differs from *global attention* by using only a subset of source annotation in the computation. They also propose a simpler computation path in comparison to Bahdanau et al. (2015), and experiment with two different implementations of the attention mechanism, achieving very competitive results against existing NMT literature.

The standard translation model described uses word-level for source and target representations. However, the existing issue with word-level translation is that learning from the word form limits the translation capacity of the network to the vocabulary present in the training set. At inference time, the network is then incapable of producing translation for unseen tokens. This is called the *Out-of-Vocabulary (OOV) problem.*

A radical solution for this problem is to learn translation directly from a sequence of characters. This way, the network is able to produce a good *guess* for unknown words, by deducing its meaning from the characters composition. However, character-level NMT models are costly to train (Lee et al., 2017; Kreutzer and Sokolov, 2018; Ataman et al., 2019). Because of that, a popular compromise is the use of sub-word units for training the networks. These can be morpheme-based (Belinkov et al., 2020) or statistically-based, such as the Byte Pair Encoding (BPE) approach (Sennrich et al., 2016).

Kreutzer and Sokolov (2018) investigate which of the mentioned representations a network would *choose* if it could change the input representation level during training time. They add a dynamic embedding layer in the encoder and decoder stacks, which can decide towards a character-level representation or a more clustered one (sub-words, words) at training time. Comparing their model, which can dynamically change the representation level, with static representation models (character-level, BPE and word-level), they discovered that they reach comparable results, and that their model had preference for character-level encoding.

On the same trend, Hahn and Baroni (2019) trained neural language models, tracking the units' activation. They discover that character-level LSTMs are capable of working with unsegmented text, learning to specialize some of the cells for tracking boundaries, and thus learning boundaries and words' dependencies. Based on their findings, they question the necessity of an explicit rigid word lexicon for language learning.

Belinkov et al. (2020) perform an extensive investigation of the linguistic representational power the described NMT model captures within its layers. Their experiments are performed by extracting the layers activation, then using these for training classifiers on the following domains: syntactic, semantic

and morphological. They find that word morphology is learned at lower layer levels of encoder-decoder architectures, while non-local linguistic phenomena in syntax and semantics are better represented at higher layers.

They also highlight character-level models are able to better capture morphology features, resulting in a better translation model for morphologically rich languages. In contrast, sub-word units models were better for capturing syntactic and semantic information, which require learning non-local dependencies. Finally, they mention that a character-based representation might be a poor choice for handling long-range dependencies, making the resulting NMT models inferior when translating syntactically divergent language pairs.

Bisazza and Tump (2018) also tackles morphology, finding that the amount of information encoded by the NMT encoder varies and that it depends on the target language. Moreover, the encoder has a *lazy* tendency, only learning grammatical features which are directly transferable to their target equivalents.

2.1.2 Multi-head Encoder-Decoder Attention

Recently, Vaswani et al. (2017) proposed **Transformer**, a fully attentional encoder-decoder architecture, which has obtained state-of-the-art results for several NMT shared tasks. This model keeps the general architecture structure from previous work, but it replaces the use of sequential cell units (such as LSTM) by Multi-Head Attention (MHA) operations, which make the architecture considerably faster.

Figure 2.3 illustrates a Transformer encoder (top) and decoder (bottom) layer inside the stacks from Figure 2.2. Both encoder and decoder networks are stacked layers sets that receive source and target sequences, embedded and concatenated with positional encoding. An encoder layer is made of two sub-layers: a *Self-Attention* MHA and a feed-forward sub-layer. A decoder layer is made of three sub-layers: a *masked Self-Attention* MHA; an *Encoder-Decoder* MHA; and a feed-forward sub-layer. The mask in the decoder's first MHA is necessary to avoid attending to subsequent positions. The Encoder-Decoder MHA operates over the encoder stack's final output and the decoder's first sub-layer output (translation context). Dropout and residual connections are applied between all sub-layers. Final output probabilities are generated by applying a linear projection over the decoder stack's output, followed by a softmax operation. We now detail the computation of the attention in Transformer models.

Multi-Head Attention mechanism: attention is seen as a mapping problem in which, given a pair of key-value vectors and a query vector, the task is the computation of the weighted sum of the given values (output). In this

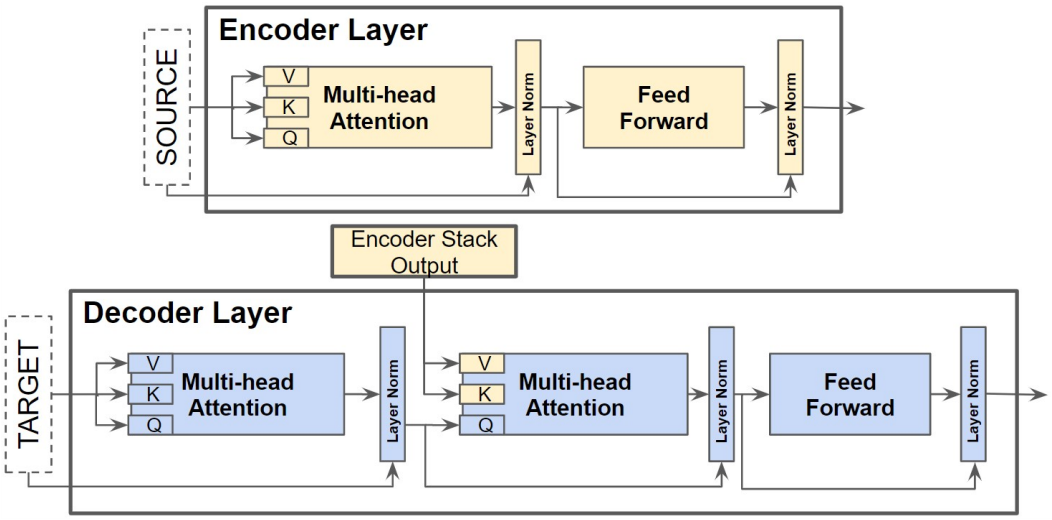


Figure 2.3: The general scheme for an encoder and decoder Transformer layer inside the encoder and decoder stacks from Figure 2.2.

setup, weights are learned by compatibility functions between key-query pairs of dimension d_k. For a given set of query (Q), keys (K) and values (V), the *Scaled Dot-Product* (SDP) Attention function is computed as in Equation 2.5.

$$Att(V, K, Q) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{2.5}$$

In practice, several *attentions* are computed for a given QKV set: the set is first projected into h different spaces (multiple heads of dimensionality d_{model}/h each), where the scaled dot-product attention is computed in parallel. Resulting values for all heads are then concatenated and once again projected, yielding the layer's output. Equation 2.6 and Equation 2.7 illustrate the process, in which H is the set of n heads ($H = \{h_1, ..., h_n\}$) and f is a linear projection. **Self-Attention** defines the case where query and values come from the same source (learning compatibility functions within the same sequence of elements). Throughout this work, we will refer to this model as the **Transformer model**.

$$MultiHead(V, K, Q) = f(Concat(H)) \tag{2.6}$$

$$h_i = Att(f_i(V), f_i(K), f_i(Q)) \tag{2.7}$$

Since its introduction, the Transformer's unique attention mechanism became a popular investigation target. The presence of several heads, instead of a single one, results in several source-to-target soft-alignment probability matrices. This flexibility potentially makes the alignment more dispersed across

heads, and in general the relationship between source and target sequences is less directly interpretable. In the original paper, the authors argue that this happens because different heads can capture different source-to-target syntactic and semantic relationships.

Aiming to understand how the heads inside the MHA mechanism contribute to the final translation model, Voita et al. (2019) perform an extensive study, weighting and classifying head importance. Using a standard transformer configuration, they removed 38 from the 48 existing heads during decoding stage, verifying that this resulted in negligible loss in translation performance. Based on this finding, they argue most of the model's heads are replaceable, and that just a few *specialized* heads are necessary after training. Michel et al. (2019) performs a similar study, verifying that many MHA layers inside a Transformer can be reduced to a single head during decoding, without any statistical significant drop in performance. Both studies find that the layer which benefits the most from the *multi-headness* is the Encoder-Decoder MHA, and they both highlight that in other cases the MHA might not be needed in order to achieve good translation performance during the decoding stage.

Focusing on the Encoder-Decoder MHA layer, the following works investigate methods for achieving better source-to-target alignments. Alkhouli et al. (2018) add one supervised extra head to this layer, giving maximal weight to the lexical items present in its dictionary. By doing so, they push the weights towards alignment, without explicitly forcing it. They show this approach's effectiveness for dictionary-guided translation. Chen et al. (2020) argues that the transformer model is able to capture good source-to-target alignment, and the challenge rests on finding the good information in the middle of the many heads and layers the model presents. With this goal, they present approaches for finding the best head and decoder step for extracting the soft-alignment probability matrices. Moreover, similar to Garg et al. (2019), they propose the inclusion of unsupervised alignments into the NMT training, jointly optimizing alignment and translation. By doing so, both (Garg et al., 2019; Chen et al., 2020) were able to generate more interpretable source-to-target alignments using Transformer.

2.1.3 Pervasive Attention

Different from the previous models, which are based on encoder-decoder structures interfaced by attention mechanisms, Elbayad et al. (2018) proposes the use of a single 2D-CNN for jointly encoding source and target sequences. Using masked convolutions, an auto-regressive model predicts the next output symbol based on a joint representation of both input and partial output se-

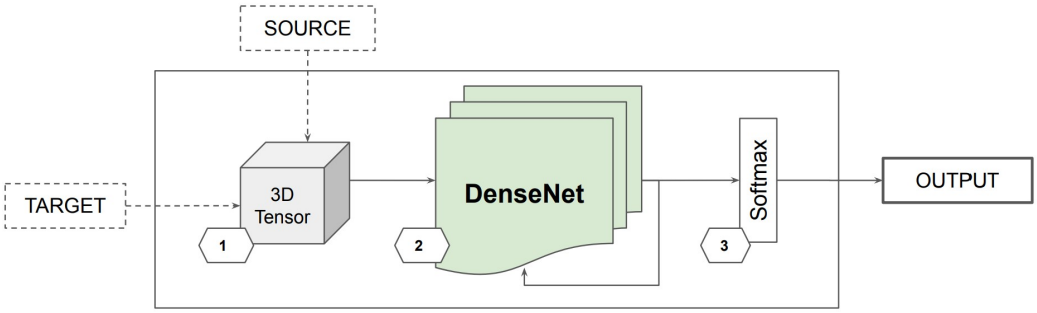


Figure 2.4: The general scheme for an 2D-CNN NMT model.

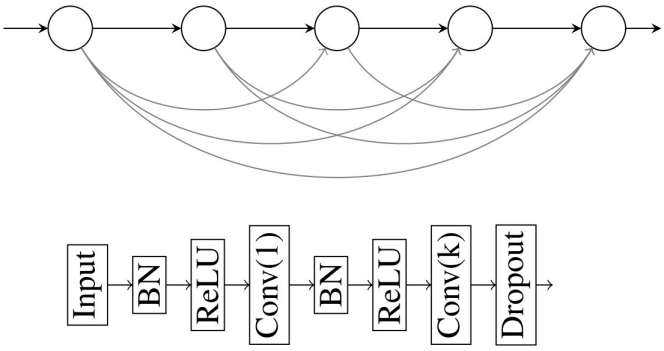


Figure 2.5: The general structure of a DenseNet Block (top), and the computation flow within each block (bottom). Figure extracted from Elbayad et al. (2018).

quences. Figure 2.2 illustrates the process.

Given a source-target pair $(\boldsymbol{s}, \boldsymbol{t})$ of lengths $|\boldsymbol{s}|$ and $|\boldsymbol{t}|$ respectively, tokens are first embedded in d_s and d_t dimensional spaces via look-up tables. Token embeddings $\{x_1, \ldots, x_{|\boldsymbol{s}|}\}$ and $\{y_1, \ldots, y_{|\boldsymbol{t}|}\}$ are then concatenated to form a 3D tensor $X \in \mathbb{R}^{|\boldsymbol{t}|\times|\boldsymbol{s}|\times f_0}$, with $f_0 = d_t + d_s$, where $X_{ij} = [y_i \quad x_j]$ (step 1). Each convolutional layer $l \in \{1, \ldots, L\}$ of the model is implemented as a DenseNet (Huang et al., 2017), illustrated in the Figure 2.5 extracted from Elbayad et al. (2018). It produces a tensor H_l of size $|\boldsymbol{t}|\times|\boldsymbol{s}|\times f_l$, where f_l is the number of output channels for that layer (step 2). To compute a distribution over the tokens in the output vocabulary, the second dimension of the tensor is used. This dimension is of variable length (given by the input sequence) and it is collapsed by max or average pooling to obtain the tensor H_L^{Pool} of size $|\boldsymbol{t}|\times f_L$. Finally, 1×1 convolution followed by a softmax operation are applied, resulting in the distribution over the target vocabulary for the next output token (step 3). We now describe the attention mechanism of this approach.

Pervasive Attention mechanism: joint encoding acts as an attention-like

mechanism, since individual source elements are re-encoded as the output is generated. The self-attention approach of Lin et al. (2017) is applied. It computes the attention weight tensor α, of size $|\boldsymbol{t}| \times |\boldsymbol{s}|$, from the last activation tensor H_L, to pool the elements of the same tensor along the source dimension, as in Equations 2.8 and 2.9. There, $W_1 \in \mathbb{R}^{f_a}$ and $W_2 \in \mathbb{R}^{f_a \times f_L}$ are weight tensors that map the f_L dimensional features in H_L to the attention weights via an f_a dimensional intermediate representation. Throughout this work, we will refer to this work as the **2D-CNN model**.

$$\alpha = softmax(W_1 \tanh(H_L W_2)) \tag{2.8}$$

$$H_L^{\text{Att}} = \alpha H_L \tag{2.9}$$

Gehring et al. (2017) is another example of a competitive CNN architecture for NMT. They differ from the work above by presenting a CNN encoder-decoder architecture, tied by a *Multi-step* attention mechanism, which allows the decoder to access the attention product from a given number of previous steps.

2.1.4 Attention as an Explanation

Recently, a group effort towards interpretability in neural networks emerged in the Natural Language Processing (NLP) community. Motivated by the fact that the neural networks are not directly *understandable* as statistical approaches are, recent work on interpretability aims to shed light into the internal processes of neural networks, investigating how learning is performed. An example of this is the Black box NLP Workshop (Alishahi et al., 2019), whose aim is to investigate the *black box processes* in deep learning approaches.[7]

Focusing on NMT, a target of interpretability studies has been the attention mechanism. As we started before, the source-to-target probabilities learned during training can be interpreted as alignment. Supported by this assessment, many works on NMT use the visualization of these probability matrices as a form of attesting translation quality. However, these matrices are just a by-product of translation, and the network is not optimized towards alignment.

Ghader and Monz (2017) highlight that there are differences between direct alignment and the attention's output. They use Luong et al. (2015)'s NMT architecture, investigating how its output relates to traditional alignment, focusing on the words Part-Of-Speech (POS). They verify that attention agrees with traditional alignment to a certain extent, especially for nouns. However,

[7]Available at: `https://blackboxnlp.github.io/cfp.html`

for other POS they show the attention might develop different patterns, which do not necessarily translate as source-to-target alignment.

Motivated by this, Ding et al. (2019) investigate methods for post-processing attention-based NMT models in order to retrieve better word alignment. They propose the use of *word saliency* for interpreting word alignments from NMT predictions in both online and offline fashion. Their evaluation shows that their model-agnostic approach is able to produce higher quality alignments compared to the raw product of the attention model. The authors perform experiments on Transformer (Vaswani et al., 2017), RNN (Bahdanau et al., 2015), and CNN (Gehring et al., 2017) models, highlighting that some of these architectures (CNN and RNN) already present good quality word alignments without post-processing.

He et al. (2019) introduce invariant gradients to compute the concept of *word importance* in NMT. In their experiments with the Transformer and RNN architectures, they show that their method for assessing the contribution of source tokens was superior to attention and other black box metrics for evaluating NMT quality on sentence-level. Moreover, their analysis showed that depending on the language pair, different syntactic categories of words receive more importance. They argue this highlights the importance of introducing an inductive bias into the model design.

For different NLP tasks, attention has also been a target of investigation. Focusing on RNN architectures for question answering, binary text classification and natural language inference, Jain and Wallace (2019) investigates the correlation between the attention model's weights and the final output yielded by the system. They perturb attention weights for assessing the impact in the output tokens, finding that only minimal changes occur. Thus, they argue *attention is not explanation* in the sense that, even if sometimes the attention layer's output correlates with the produced token (output), these weights are not directly responsible for the prediction, and therefore its visualization should not be used as a form explaining the systems' choices.

Extending this study, Serrano and Smith (2019) investigates the correlation between the attention layer's weighting of the input elements and the importance ranking obtained in topic classification models. They also find a lack of correlation able to justify the use of attention as a visualization tool for network learning, mentioning that attention might still be interpretable in other ways different from direct visualization.

Wiegreffe and Pinter (2019) challenge these works, arguing that *attention is not not explanation*. They explain that the difference lies in the definition of explanation itself: between plausible and faithful explanation. While attention might fail to provide faithful explanation for a set of NLP tasks, it still presents a plausible relationship between input and output tokens. In their

work, the authors use adversarial training methods for obtaining alternative attention distributions, showing that these perform quite poorly compared to the original ones. Supported by that, they say that attention does capture a meaningful relationship between input and output, and therefore it can be used for investigating network learning. They then suggest researchers to be careful when assessing model's quality through the use of their attention mechanism's visualization, and they provide an experimental test suite with the goal of making this investigation more sound.

Focusing on interpretability for NMT models, Moradi et al. (2019) investigated if the findings from Jain and Wallace (2019) hold in the case of sequence-to-sequence models. They separated words between function and content classes, and investigated the impact of using counterfactual attention weights during the decoding stage of a RNN model (Luong et al., 2015). In their experiments they noticed that it is harder to perturb the generation of function words, compared to content ones. They argue that this happens because function words depend more on the decoder context, while the content words depend mostly on the weighting of the encoder annotations performed by the attention layer. In summary, while preliminary, their results show that several counterfactual attention matrices can result in the same tokens being generated by the translation model. Based on that, the authors conclude that the interpretability of the attention layer is still an open research topic, and that people should refrain from using it for *explaining* the output of their NMT models.

Brunner et al. (2019) investigate the validity of self-attention as explanation in Transformer NMT models. They argue that a problem with interpretability studies of the attention mechanism is the assumption that the weights are relative to words, instead of their embeddings, which can be a mixture of several words present in the sentence. Investigating the Transformer architecture, they question methods accumulating attention weights over layers, since the embeddings are layer-dependent, and therefore the attention is not being computed over the same information. In their experiments, they observe that as they go deeper into the Transformer's layers, the relationship between words and their embeddings gets blurrier. However, by classifying words by their POS, they notice that for some core content classes in English, the contribution of a given word for its corresponding embedding stays high even in deeper layers. They conclude by saying that researchers need to be careful when using attention visualizations beyond the first layer to draw conclusions about word importance and translation quality.

Lastly, Vashishth et al. (2019) provide an extensive assessment of the impact of attention in NLP tasks. They argue that both view-points, *Attention is not explanation* (Jain and Wallace, 2019) and *Attention is not not explana-*

tion (Wiegreffe and Pinter, 2019), are in fact correct, but the lack of general vision made the distinct conclusions. They start by highlighting that the attention layer has different roles in different NLP models, which limit the generality of the claims from previous work. They classify attention models into: single sequence tasks (input consist on a single text sequence, i.e. sentiment analysis), pair sequence tasks (input consists on a pair of text sequences, i.e. question answering), and generation sequence tasks (consists in generating a sequence based on the input sequence, i.e. NMT). Throughout their work, they study the impact of perturbing attention weights on models from these different classes of NLP tasks.

They notice that the behavior in attention for single sequence tasks is different from the one observed in pair sequence and generation sequence tasks. While for the former, perturbing attention results on a marginal impact in models prediction, for the other tasks, there is a significant decrease in performance. Based on that, they propose a different way of seeing the attention mechanism. They argue that for single sequence tasks, the models depend less on their attention layer, which behaves as a gating mechanism and therefore, the impact in the generated output is limited. For the other cases, authors state that the dependency between the attention mechanism and the systems performance is higher and thus, in these cases attention takes the role of the *explainer* of the model.

2.2 NMT for Low-resource Languages

The superior abstraction power of neural networks comes with a heavy price in terms of data needs. These models demand considerable amounts of examples in order to train the large quantity of parameters inside their many layers. Because of that, its applicability stays narrowed to the subset of languages for which big datasets are commonly available (Maxwell and Hughes, 2006). For instance, the original Transformer NMT model (Vaswani et al., 2017) was trained on 4.5 million English-German parallel sentences.[8]

However, these data hungry approaches are not incapable of scaling down and performing reasonable well in scenarios with less data. Even so, the minimal amount they usually demand is not compatible with the available resources for many languages. Sennrich and Zhang (2019) searched for this minimum data amount for training effective NMT models in low-resource settings. They discovered that their baseline was only able to reach over 20 BLEU score by having 10^6 English words, which in their case meant having 40,000 aligned sentences. They then illustrated how targeted optimization

[8]Shared vocabulary of 37,000 types after BPE encoding.

can help reduce the need for data, and their final model was able to largely outperform this baseline. This highlights that existing approaches cannot be expected to work in an out-of-the-box fashion in low-resource settings. Instead it is necessary to find ways of adapting them.

On this topic, Kann et al. (2019) discuss realistic approaches for NLP in low-resource settings, focusing on the validation set. They argue that in these settings, separating some of the available data for validation represents a considerable toll in the amount of information available for training. They then propose to train low-resource neural models without validation sets. For doing so, they first train their neural models in different (high-resource) languages, averaging the number of epochs necessary for these to finish training. This average is then used to determine the training duration for the low-resource neural models. They evaluate neural models for three tasks (historical text normalization, morphological inflection, transliteration), showing that using all available data for training can result in as much as 18% of accuracy gain compared to models trained using validation sets.

Focusing on regularization, Rekabsaz et al. (2019) propose a multilingual Language Model (LM) trained on several low-resource languages as a form to counter the lack of data. This LM shares two layers between the different languages, which allows them to capture language specific features. Then, a shared third layer captures the common features present in the corpora. They show how this regularization strategy achieves better results compared to monolingual setups (one LM per low-resource language).

Gibadullin et al. (2019) presents a survey of methods for leveraging monolingual data into training for reducing the amount of parallel sentences necessary for creating NMT models in low-resource settings. The authors separate the works into two categories: architecture dependent and independent. The former refers to methods exploiting specific architectural features from NMT models in order to include the monolingual data into the NMT training pipeline. The latter refers to methods for data augmentation (creating a *pseudo-parallel* corpus from monolingual data), or to methods using a separate target-side LM for enriching the model. We highlight some methods from these categories.

Gulcehre et al. (2015) and Stahlberg et al. (2018) present architecture independent models for fusing a pretrained LM with a low-resource NMT model. The appeal of using an LM comes from the fact that monolingual data is easier to acquire than bilingual (sentence-aligned) datasets. They train their NMT models with up to 200,000 parallel sentences only, and their LMs with more than 3 million examples. They both verify a slight performance improvement in translation compared to pure NMT models.

Popular architecture dependent models are the ones which perform *trans-*

fer learning (Imankulova et al., 2019; Lin et al., 2019; Zoph et al., 2016). They work by pretraining models with considerable amounts of data in a high-resource language. After that, the networks are fine-tuned with small quantities of data from the target language. In this stage, some of the layers remain frozen (parameters are not updated), and new layers can also be added to the network. The resulting models tend to perform better than directly training on the low-resource language, as the pretraining stage provides a good *guess* for the layers' parameters in the target language.

Lin et al. (2019) raises an important aspect of transfer learning by questioning the impact of the language chosen for pretraining in this pipeline. They argue that *similar* languages should be preferred, as some syntactic and semantic information could be directly transferred from the high to the low-resource training stages, resulting in richer models. They then propose a toolkit for scoring from *which* language one should transfer from. This scoring uses the `lang2vec` resource (Littell et al., 2017) for investigating geographic proximity, phonological and syntactic similarities, data availability, and typology.

3 Learning Representations from Speech

In the last section we explained that neural models tend to require a considerable amount of examples in order to train their parameters. This becomes even more critical when text is replaced by speech utterances, since the dimensionality of the input increases drastically.[9] Learning from speech requires larger architectures, and consequently more examples in order to converge. The consequence of this is that recent models for speech processing depend on the availability of large amounts of speech data, which frequently need to be accompanied by extensive transcriptions.

However, learning supervised representations from speech differs from the unsupervised way infants learn language, hinting that it should be possible to develop more data-efficient, and unsupervised, speech processing models. Inspired by that, recent work suggested pretraining on large quantities of speech without supervision for application in downstream tasks (Chen et al., 2017; Chorowski et al., 2019; Schneider et al., 2019; Baevski et al., 2020b). While this reduces the amount of data transcription necessary for applying speech models, these still require some transcription for the downstream tasks.

More interesting for us are the models which provide *Speech Discretization* (SD) through unsupervised training (no access to transcriptions). Their

[9]For instance, a popular dataset for speech technologies is Librispeech (Panayotov et al., 2015), composed of 1,000 hours of recorded speech in English.

task consists in labeling the speech signal into discrete speech units, which can correspond or not to the language phonetic inventory. The advantage of this discretization is that it allows for the posterior application of text-based approaches, which are less data expensive.

Nowadays, there are two main approaches for SD. The first is the neural approach, in which models are typically made of an auto-encoder structure with a discretization layer (van den Oord et al., 2017; Chorowski et al., 2019; Baevski et al., 2020a). The second is the use of NP Bayesian generative models, which can be seen as infinite mixtures time series models (Lee and Glass, 2012; Ondel et al., 2016; Chen et al., 2017). Both have been recently investigated in the context of language acquisition and documentation (Versteegh et al., 2015; Dunbar et al., 2017, 2019).

The SD task can be formulated as the learning of a set of U discrete units with embeddings $\mathbf{H} = \{\boldsymbol{\eta}^1, \ldots, \boldsymbol{\eta}^U\}$ from a sequence of untranscribed acoustic features $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]$, as well as the assignment of frame to unit $\mathbf{z} = [z_1, \ldots, z_N]$. In simple terms, the network learns to summarize the speech using a number $|U|$ of units. These are used to label the speech frames. Depending on the approach, neural or Bayesian, the assumptions and the inference regarding these three quantities ($\mathbf{H}$, $\mathbf{X}$, $\mathbf{z}$) will differ. Section 3.1 describes two neural SD models, and Section 3.2 presents three Bayesian approaches.

3.1 Neural Networks for Vector Quantization

In this section we present two well-known neural networks for Vector Quantization (VQ) of unlabeled speech utterances. The first one, VQ-VAE (Section 3.1.1) is inspired by input dimensionality reduction architectures. The second model, VQ-WAV2VEC (Section 3.1.2), finds inspiration in self-supervised models trained with a context-prediction loss. We highlight that due to the size of these architectures, in terms of number of parameters and layers, models for VQ are usually trained in high-resource languages. Notwithstanding, fine-tuning methods could be an option for applying them to low-resource languages.

3.1.1 VQ-VAE

Variational Auto-Encoder (VAE) models (Kingma and Welling, 2013) are architectures for input dimensionality reduction. They are encoder-decoder networks, tied by a subspace given by the set of latent random variables $\mathbf{z}$. The encoder network parameterises a posterior distribution $q(\mathbf{z}|\mathbf{x_n})$ given the input data $\mathbf{X}$, a prior distribution $p(\mathbf{z})$, and a decoder with a distribution over

the input data $p(\mathbf{x_n}|\mathbf{z})$. In practice, the subspace provides a *summarization* of the input information captured by the encoder network. This summarization must be of enough quality in order to allow the decoder network to reconstruct the initial input.

The VQ Variational Auto-Encoder (VQ-VAE) models (van den Oord et al., 2017) are an extension of VAE models which output a discrete latent representation for the input. In order to reach this discrete representation, they apply Vector Quantization (VQ) training for circumventing a known problem of VAE models called "posterior collapse".

The VQ-VAE neural model comprises an encoder, a decoder and a set of unit-specific embeddings $\mathbf{H}$. The encoder is a neural network that transforms the data into a continuous latent representation $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_N)$. Each frame is then assigned to the closest embedding in the Euclidean sense, as in Equation 2.10. The decoder transforms the sequence of quantized vectors into parameters of the conditional log-likelihood of the data $p(\mathbf{x}_n|\mathbf{z})$ and the network is trained to maximize this likelihood.

$$z_n = \arg\min_u ||\mathbf{v}_n - \boldsymbol{\eta}^u||_2 \tag{2.10}$$

Since the quantization step is not differentiable, the encoder is trained with a straight through estimator (Bengio et al., 2013). In addition, a pair of ℓ_2 losses are used to minimize the quantization error, and the overall objective function that is maximized is presented in Equation 2.11. There, $\text{sg}[\cdot]$ is the stop-gradient operator. The likelihood $p(\mathbf{x}_n|z_n)$ is defined as $\mathcal{N}(\mathbf{x}_n; \boldsymbol{\mu}(\boldsymbol{\eta}^{z_n}), \mathbf{I})$. Under this assumption, the log-likelihood reduces to the mean-squared error $||\mathbf{x}_n - \boldsymbol{\mu}(\boldsymbol{\eta}^{z_n})||_2^2$.

$$\mathcal{L} = \frac{1}{N}\sum_{n=1}^{N}\Big(\ln p(\mathbf{x}_n|z_n) - k_1||\,\text{sg}[\boldsymbol{\eta}^{z_n}] - \mathbf{v}_n||_2^2 - k_2||\boldsymbol{\eta}^{z_n} - \text{sg}[\mathbf{v}_n]||_2^2\Big) \tag{2.11}$$

3.1.2 VQ-WAV2VEC

Baevski et al. (2020a) also propose a VQ-based SD model. It extends the self-supervised speech model `wav2vec` (Schneider et al., 2019), which solves a self-supervised context-prediction task with the same loss function from the classic `word2vec` (Mikolov et al., 2013). Different from VQ-VAE, the VQ-WAV2VEC model learns by using a future time-step prediction task, instead of input reconstruction.

This model is made of three components: encoder ($f : \mathbf{X} \rightarrow \mathbf{Z}$), quantizer ($q : \mathbf{Z} \rightarrow \hat{\mathbf{Z}}$) and aggregator ($g : \hat{\mathbf{Z}} \rightarrow \mathbf{C}$) networks. The encoder is a

CNN which maps the raw speech input $\mathbf{X}$ into the dense feature representation $\mathbf{Z}$. From this representation, the quantizer produces discrete labels $\hat{\mathbf{Z}}$ from a fixed size codebook $\mathbf{e} \in \mathbb{R}^{V \times d}$ with V representations of size d.

Since replacing an encoder feature vector $\mathbf{z}_i$ by a single entry in the codebook makes the method prone to model collapse (i.e. only some of the codebooks would actually be used), they independently quantize partitions of each feature vector. For achieving this, they generate multiple *groups* G, arranging the feature vector into a matrix form $\mathbf{z'} \in \mathbb{R}^{G \times (d/G)}$.

Considering each row by an integer index, they can thus represent the full feature vector by the indexes $\mathbf{i} \in [V]^G$, V being the possible number of *variables* for a given group, and each element $\mathbf{i}_j$ corresponding to a fixed codebook vector from a given group $j \in G$. For each of these G groups, the quantization is performed by using Gumbel-Softmax (Gumbel, 1948) or online k-means clustering.

Finally, the aggregator combines multiple quantized feature vector time steps into a new representation $\mathbf{c}_i$ for each time step i. Then, given this aggregated representation $\mathbf{c}_i$, the model is trained to distinguish a sample k steps in the future $\hat{\mathbf{z}}_{i+k}$ from *distractor* samples $\tilde{\mathbf{z}}$ drawn from a distribution p_n.

This is done by minimizing the contrastive loss for steps $k = \{1, \ldots, K\}$ as in Eq. 2.12, where T is the sequence length, $\sigma(x) = 1/(1 + exp(-x))$, $\sigma(\hat{\mathbf{z}}_{i+k}^{\intercal} h_k(\mathbf{c_i}))$ is the probability of $\hat{\mathbf{z}}_{i+k}$ being the true sample, and $h_k(\mathbf{c}_i)$ is the step-specific affine transformation $h_k(\mathbf{c}_i) = W_k \mathbf{c}_i + b_k$. Finally, this loss is accumulated over all k steps: $\mathcal{L} = \sum_{k=1}^{K} \mathcal{L}_k$.

$$\mathcal{L}_k = \sum_{i=1}^{T-k} \Big(\log \sigma(\hat{\mathbf{z}}_{i+k}^{\intercal} h_k(\mathbf{c_i})) + \lambda \mathbb{E}_{\tilde{\mathbf{z}} \sim p_n} [\log \sigma(-\tilde{\mathbf{z}}^{\intercal} h_k(\mathbf{c}_i))] \Big) \tag{2.12}$$

3.2 NP Bayesian Generative Models

For generative models, each acoustic unit embedding $\boldsymbol{\eta}_i$ represents the parameters of a probability distribution $p(\mathbf{x}_n | \boldsymbol{\eta}_{z_n}, z_n)$ with latent variables $\mathbf{z}$. Discovering the units amounts to estimating the posterior distribution over the embeddings $\mathbf{H}$ and the assignment variables $\mathbf{z}$ given by Equation 2.13.

$$p(\mathbf{z}, \mathbf{H} | \mathbf{X}) \propto p(\mathbf{X} | \mathbf{z}, \mathbf{H}) p(\mathbf{z} | \mathbf{H}) \prod_{u=1}^{U} p(\boldsymbol{\eta}^u) \tag{2.13}$$

From the definition above, in this section we describe three different generative models for performing SD: HMM (Section 3.2.1), SHMM (Section 3.2.2) and H-SHMM (Section 3.2.3). These models all share the feature of being

robust to low-resource settings. In Section 3.2.4 we detail their inference process.

3.2.1 HMM

This model, proposed by Ondel et al. (2016), solves the SD task using an infinite *phone-loop* architecture where each acoustic unit component is a 3-state left-to-right HMM/GMM[10] with parameters $\boldsymbol{\eta}^i$. It assigns a prior distribution over the units modeled by a Dirichlet process, and it incorporates a prior distribution over the parameters of the HMMs as well. These two features make this model *fully Bayesian*. We refer to this model as **HMM**, and we consider it as the NP Bayesian generative model baseline, serving as the backbone for the two subsequent models.

3.2.2 SHMM

The Subspace HMM (SHMM) model, proposed in Ondel et al. (2019), fixes a naive assumption of the HMM model. For the latter, the prior is defined as a combination of exponential family distributions forming a prior conjugate to the likelihood. While mathematically convenient, this prior does not incorporate any knowledge about phones: it considers all possible sounds as potential acoustic units. This means, for instance, that the sound of a car engine and the sound from the elicitation of the letter "a" are both equally considered by this model.

Looking back at the prior $p(\boldsymbol{\eta})$ in Equation 2.13, it corresponds to the probability that a sound, represented by an HMM with parameters $\boldsymbol{\eta}$, is an acoustic unit. In Ondel et al. (2019), they propose to remedy the mentioned shortcoming by defining the parameters of each unit u as in Equation 2.14, where $\mathbf{e}^u$ is a low-dimensional unit embedding, $\mathbf{W}$ and $\mathbf{b}$ are the parameters of the *phonetic subspace*, and the function $f(\cdot)$ ensures that the resulting vector $\boldsymbol{\eta}^u$ dwells in the HMM parameter space.

$$\boldsymbol{\eta}^u = f(\mathbf{W} \cdot \mathbf{e}^u + \mathbf{b}) \tag{2.14}$$

The subspace, defined by $\mathbf{W}$ and $\mathbf{b}$, is estimated from several labeled source languages. The prior $p(\boldsymbol{\eta})$ is defined over the low-dimensional embeddings $p(\mathbf{e})$ rather than $\boldsymbol{\eta}$ directly, therefore constraining the search of units in the relevant region of the parameter space.

[10]For simplicity, we refer to the HMM/GMM model simply as HMM from now on.

3.2.3 H-SHMM

While the SHMM model significantly improves over the HMM, it also suffers from an unrealistic assumption: it assumes that the phonetic subspace is the same for all languages. Yusuf et al. (2020) propose a model extension called Hierarchical SHMM (H-SHMM). In their work, they relax this assumption of a single phonetic multilingual subspace by proposing to adapt the subspace for each target language while learning the acoustic units. Formally, for a given language λ, the subspace and the acoustic units' parameters are constructed as in Equation 2.17.

$$\mathbf{W}^{\lambda} = \mathbf{M}_0 + \sum_{k=1}^{K} \alpha_k^{\lambda} \mathbf{M_k} \tag{2.15}$$

$$\mathbf{b}^{\lambda} = \mathbf{m}_0 + \sum_{k=1}^{K} \alpha_k^{\lambda} \mathbf{m_k} \tag{2.16}$$

$$\boldsymbol{\eta}^{\lambda,u} = f(\mathbf{W}^{\lambda} \cdot \mathbf{e}^{\lambda,u} + \mathbf{b}^{\lambda}) \tag{2.17}$$

The matrices $\mathbf{M}_0, \ldots, \mathbf{M}_K$ in Equation 2.15 and vectors $\mathbf{m}_0, \ldots, \mathbf{m}_K$ in Equation 2.16 represent a *template* phonetic subspace, linearly combined by a language embedding $\boldsymbol{\alpha}^{\lambda} = [\alpha_1^{\lambda}, \alpha_2^{\lambda}, \ldots, \alpha_K^{\lambda}]^{\top}$. The matrices $\mathbf{M}_i$ and the vectors $\mathbf{m}_i$ are estimated from labeled languages (i.e. multilingual transcribed speech). The acoustic units' low-dimensional embeddings $\{\mathbf{e}_i\}$ and the language embedding $\boldsymbol{\alpha}$ are learned on the target (unlabeled) speech data.

3.2.4 Inference of NP Bayesian Generative Models

Regarding inference, the posterior distribution is intractable and cannot be estimated. Instead, one seeks for an approximate posterior $q(\{\boldsymbol{\eta}_i\}, \mathbf{z}) = q(\{\boldsymbol{\eta}_i\})q(\mathbf{z})$ which maximizes the variational lower-bound $\mathcal{L}[q]$. For estimating $q(\mathbf{z})$, the *expectation* step is identical for all models and is achieved with a modified *forward-backward* algorithm described in Ondel et al. (2016).

The estimation of $q(\boldsymbol{\eta})$ (the *maximization* step) is model-specific and is described in Ondel et al. (2016) for the HMM, in Ondel et al. (2019) for SHMM, and in Yusuf et al. (2020) for the H-SHMM. Finally, the output of each Bayesian system is obtained from a modified Viterbi algorithm which uses the expectation of the log-likelihoods with respect to $q(\{\boldsymbol{\eta}_i\})$, instead of point estimates.

4 Discussion

As presented in this thesis introduction, here we propose a pipeline for *bilingual* UWS from speech, and in low-resource settings. We find inspiration in the fact that, in language documentation scenarios, linguists often write translations as a form of labeling the utterances they collect. We then propose the use of these translations for *grounding* the segmentation process.

For including the translation into the segmentation pipeline, we focus on attention-based NMT architectures. In Section 2 we showed that these models are by nature bilingual, and that their attention mechanisms produce *soft-alignment* between source and target sequences. Another aspect of these models that makes them very interesting for our task is that they can be extended for working directly from speech (Bérard et al., 2016; Weiss et al., 2017).

However, the use of neural networks presents its challenges. First of all, while world-level soft-alignment has been investigated for NMT, it remains to be seen if the soft-alignments produced remain exploitable when source and target sequences differ greatly (for instance, speech vectors and word translations). Moreover, there is the question of data scarcity robustness. While neural networks present state-of-the-art results for many different NLP tasks, they often require a considerable number of examples for training.

Another aspect we believe will impact our UWS approach is the nature of the supervision used for grounding. By using translations as a *guide* for segmentation, we might produce very distinct structures by varying the language. Haspelmath (2011) says that the very definition of a *word* might be difficult to define cross-linguistically.

Finally, there is the integration of speech input into the pipeline. Since end-to-end speech-to-translation training is unrealistic using datasets with just a couple of hours of labeled speech, in this thesis we propose a pipeline approach. It consists of first creating a sequence of discrete speech units from the speech utterances using the SD models presented in Section 3, and then training NMT models between this discretization and translation sentences.

Related to our work, Stahlberg et al. (2013) present a statistical pipeline model for segmentation and cross-lingual alignment between generated segmentations and translation words in low-resource settings. Working from manual transcriptions, they find that the translations improve segmentation performance.

Adams et al. (2015) use statistical alignment models for producing bilingual segmentation, and Duong et al. (2016) perform bilingual segmentation by using the soft-alignment learned by an attention-based NMT model. Both work from the phonetic transcription of the input.

Different from all the above mentioned, in this thesis we propose a pipeline

working directly from speech, by including a *speech discretization step*. We also provide an extensive investigation of our bilingual attention-based UWS model across different NMT architectures, input representations, translation languages, and dataset sizes. We also investigate including extra annotation into our pipeline, in the form of manual transcribed data and boundaries *clues*.

Contributions

Chapter III
Resources

In the last chapter we presented textual and speech-based approaches, discussing about the challenges of adapting models for processing language in scenarios with limited access to data. This need comes from the fact that most of the data resources freely available cover only a subset of languages, the so called *high-resource* languages (Maxwell and Hughes, 2006).

Furthermore, even when approaches scale to low-resource settings, we find a lack of realistic corpora for testing the generalization of the proposed models.[1] Thus, many works rely on sampling high-resource languages to *emulate* the expected behavior using low-resource languages. This methodology assumes that different languages are equally difficult to *learn*,[2] and more importantly, that they are learned in the same way. The result of this kind of assumption is the proposition of models which might be unintentionally language-biased towards a particular high-resource language, and that might not work well when applied to the real target (Kawakami et al., 2019).

The solution for this issue is then to thoroughly test proposed approaches on realistic settings and using many languages, which is not usually done due to a lack of data. Aiming to help fill this gap in available resources from low-resource languages, during this thesis we participated in three projects for releasing realistic low-resource speech corpora to the community, which we describe in this chapter.

We released two datasets from truly endangered languages (Sections 1 and 2); and one novel multilingual speech-to-speech dataset (Section 3) covering languages with interesting linguistic features. All the described datasets, together with evaluation references and scripts, are freely available online.

1 Mboshi-French Parallel Corpus

Mboshi (Bantu C25) is an oral language spoken in Congo-Brazzaville. It was one of the languages documented by the *Breaking the Unwritten Language*

[1] This is especially true for speech approaches. See Table 1 in Di Gangi et al. (2019) for an overview of available speech corpora.

[2] Cotterell et al. (2018) discuss this for the language modeling task.

Mboshi	`wáá ngá iwé léekundá ngá sá oyoá lendúma saa m ótéma`
French	`si je meurs enterrez-moi dans la forêt oyoa avec une guitare sur la poitrine`

Figure 3.1: A tokenized and lower-cased sentence pair example in our Mboshi-French parallel corpus.

Barrier (BULB) project (Adda et al., 2016; Stüker et al., 2016). Although mainly unwritten, linguists have defined a non-standard graphemic form for it, considered to be close to the language phonology.

The data was collected through the use of the `LIG-Aikuma` mobile app[3] (Blachon et al., 2016). This application is dedicated to fieldwork language documentation. Between many features, it allows linguists to capture oral and written translations, and elicitations from text or images. In the case of the Mboshi data, the corpus was built from two sources: a small dictionary (Beapami et al., 2000) and the reference sentences for oral language documentation (Bouquiaux and Thomas, 1976). Three speakers performed the elicitation of the sentences, resulting in a corpus of 5,130 sentences after post-processing. The translation language chosen was French.

The post-processing included manual correction of the translations, standardization of the characters encoding, and forced alignment between the audio and the transcriptions. The alignments were then used to create the reference files[4] for allowing researchers to evaluate and compare their spoken term discovery results obtained using the corpus. An example of the final written content of the corpus is presented in Figure 5.1.

Lastly, the corpus was split between *train* and *development* (or validation) sets. This was performed by first shuffling the data for ensuring comparable distributions in terms of speakers and origins.[5] There is no overlap for the transcriptions between the two sets, and no repeated sentences in the development set. General metrics for the resulting Mboshi-French parallel corpus[6] (Godard et al., 2018a) are presented in Table 3.1.

Multilingual Translations. Posterior to the release of the dataset, we extended it by adding translation in multiple languages (Boito et al., 2019b). This was possible by translating the original French text into four new languages using the `DeepL translator` tool.[7] The added languages are En-

[3]Available at: `https://lig-aikuma.imag.fr/`

[4]This reference works with the *Zero Resource Challenge* (Dunbar et al., 2017) 2017 track 2 evaluation track, for ensuring research reproducibility.

[5]Sentences come either from Bouquiaux and Thomas (1976) or Beapami et al. (2000).

[6]Available at: `https://github.com/besacier/mboshi-french-parallel-corpus`

[7]Available at: `https://www.deepl.com/translator`

language	set	#sentences	#tokens	#types	audio length(h)
Mboshi	train	4,616	27,563	6,196	4.02
	dev	514	2,993	1,146	0.26
French	train	4,616	38,481	4,921	-
	dev	514	4,234	1,174	-

Table 3.1: Number of sentences, tokens and types for the different sets, for both languages. For Mboshi, audio duration is also detailed.

	MB	FR	EN	ES	DE	PT
#tokens	30,556	42,715	37,379	37,428	37,515	37,095
#types	6,633	5,178	4,392	5,473	5,641	5,465
avg token length	4.18	4.41	4.19	4.36	4.91	4.40
avg #tokens per sentence	5.96	8.33	7.29	7.30	7.31	7.23

Table 3.2: General statistics for the languages present in the dataset. The metrics were computed on the totality of the corpus (5,130 sentences).

glish (EN), German (DE), Portuguese (PT) and Spanish (ES). Our motivation was to provide a version of this corpus that could be exploited for multilingual approaches. Since in documentation scenarios it is difficult to collect data, datasets tend to lack size. Our hope is that by relying on multilingual supervision, the effects of the lack of data for computational approaches could be attenuated. General metrics for the translated corpus[8] are presented in Table 3.2.

2 Griko-Italian Parallel Corpus

Griko is an endangered Greek dialect spoken in southern Italy, in the Grecìa Salentina area southeast of Lecce. It is one of the two Italo-Greek variety dialects in the region of Calabria. Less than 20,000 people (mostly over 60 years old) are believed to be native speakers (Horrocks, 2009; Douri and De Santis, 2015) but unfortunately, this number is quite likely an overestimation (Chatzikyriakidis, 2010).

The original corpus was collected during a field trip in Puglia, Italy, by two

[8]Available at: `https://github.com/mzboito/mmboshi`

language	#tokens	#types	avg token length	avg #tokens per sentence	audio length(h)
Griko	2,374	691	5.68	7.19	0.20
Italian	2,384	456	5.76	7.22	-

Table 3.3: General statistics for the 330 sentence-long Griko-Italian parallel corpus.

linguists, with a particular focus on the use of infinitive and verbal morphosyntax. It contains a total of 330 utterances from 9 different speakers (5 male, 4 female) from the 4 villages where native speakers could still be found (Calimera, Sternatia, Martano, Corigliano). The digitally collected audio files were manually segmented into utterances, transcribed, glossed in Italian, and annotated with extensive morphosyntactic tags by a trained linguist. The resulting dataset[9] (Lekakou et al., 2013) represents the only speech corpus for Griko available online.

In order to render the original corpus useful for speech-related computational research on Griko, new information was added by us to the corpus. First, the transcriptions were translated in Italian by a bilingual speaker. Gold-standard word-level alignment information, including silence marks, were added to the dataset, as well as gold-standard speech-to-translation alignments. We also automatically extracted pseudo-phones from the audio by using the Acoustic Unit Discovery (AUD) method from Ondel et al. (2016). Lastly, reference files allowing the use of the ZRC evaluation track for spoken term discovery, as in Session 1, were created.

The final Griko-Italian parallel corpus[10] (Boito et al., 2018) has several levels of information: speech, machine extracted pseudo-phones, transcriptions, translations and sentence alignment. We believe it can be an interesting resource for evaluating documentation techniques on (very) low-resource settings. Table 3.3 presents the general statistics, and Figure 3.2 illustrates a parallel sentence in the dataset.

[9]Available at: `http://griko.project.uoi.gr`
[10]Available at: `https://github.com/antonisa/griko-italian-parallel-corpus`

Griko	jatì ìche polemìsonta òli tin addomàda
Italian	perché aveva lavorato tutta la settimana

Figure 3.2: A tokenized and lower-cased sentence pair example in our Griko-Italian parallel corpus.

3 MaSS: Multilingual corpus of Sentence-aligned Spoken Utterances

Recently, a remarkable work introduced the *CMU Wilderness Multilingual Speech Dataset*[11] (Black, 2019). It provides data to build Automatic Speech Recognition (ASR) and Text-to-Speech (TTS) models for potentially 700 languages. Each language accounts for around 20 hours of data extracted from readings of the New Testament from the Bible. Segmentation was made at the punctuation level, and alignment between speech and corresponding text can be obtained with the pipeline provided along with the dataset. This pipeline, notably, can process a large amount of languages without using any extra resources such as acoustic models or pronunciation dictionaries. Such a resource allows the community to experiment and to develop speech technologies on an unprecedented number of languages.

Its source material, the New Testament from *The Faith Comes By Hearing* website[12] (or simply *bible.is*), is an online platform that provides audio-books of the Bible with transcriptions in 1,294 languages. On this website, the written content for a given language is always the same. However, there can be more than one audio-book available per language. Different versions might present different numbers of speakers, types of recording procedure, the presence of background music, and even the *dramatization*[13] of the text.

In their pipeline, Black (2019) extracted the soundtracks from the defaults links, and audio excerpts often contain music. It is also unknown if drama or non-drama versions were selected. Thus, although the quality of the alignment is good for many languages, it could be inaccurate (or noisy) for an unknown subset. Moreover, the final segmentation from chapters was obtained through the use of punctuation marks. While efficient for a speech-to-text monolingual scenario, this strategy does not allow accurate multilingual alignment, since different languages and translations may result in different sentence segmen-

[11] Available at: `http://www.festvox.org/cmu_wilderness/index.html`

[12] Available at: `https://www.bible.is`

[13] The *drama* version is an acted version of the text, corresponding to less tailored realizations.

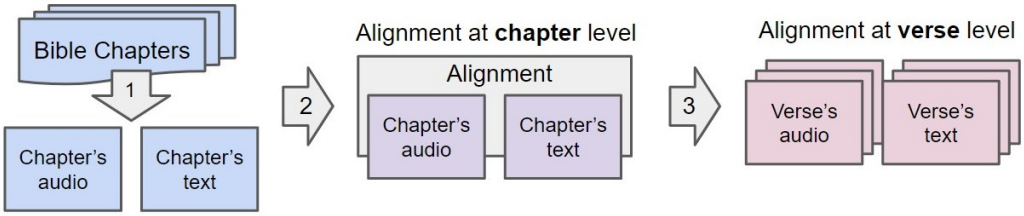


Figure 3.3: The pipeline for a given language in the *bible.is* website.

tation and ordering.

Inspired by the multilingual limitation of their approach, we proposed a new pipeline for extracting high-quality multilingual content from the *bible.is* website. Different from their pipeline, our method allows us to extract low-granularity multilingual speech segments. This is possible by taking advantage of the fact that the initial language material from the monolingual dataset (the Bible) is the same for all languages, thus constituting a multilingual and comparable[14] spoken corpus. Considering that for all languages, a chapter consists of the same set of *verses,*[15] the verse numbers give us a multilingual alignment between all language pairs.[16]

Our pipeline for a given language is described in Figure 3.3. We manually select and download the chapters on their non-drama version in the *Bible.is* website, and we also perform audio conversion (step 1). Note that the selection process does not require any language expertise, as we only verify that the audio-books selected do not present background music or dramatization.

Next we generate monolingual speech-to-text alignment by using the *Maus* forced aligner (Kisler et al., 2017) online platform[17] (step 2). We then use the generated alignment, together with the verse information present in the raw chapters version, to slice the chapter's audio into smaller chunks, identified by chapter and verse number (step 3). We highlight that this step works on any speech-to-text alignment generated (automatic or manual), as long as it is provided in a *TextGrid* file.

We applied our method to 8 languages (Basque, English, Finnish, French, Hungarian, Romanian, Russian and Spanish), resulting in 56 language pairs for which we obtain speech-to-speech, speech-to-text and text-to-text align-

[14]Our definition of a *comparable* corpus is the following: a non-sentence-aligned corpus, parallel at a broader granularity (e.g. chapter, document).

[15]A verse is the minimal segmentation unit used in the Bible and corresponds to a sentence, or more rarely to a phrase or a clause.

[16]This is mostly true, but for a small subset of chapters, due to different Bible versions and different translation approaches, the number of aligned speech verses will differ slightly.

[17]Available at: `https://clarin.phonetik.uni-muenchen.de/BASWebServices/interface/WebMAUSBasic`

Alignment from Black (2019)		
Files	**French**	**English**
00001	Matthieu	Matthew
00002	Jésus descend de la montagne et des foules nombreuses le suivent.	When he came down from the mountainside, large crowds followed him.
00003	Un lépreux s'approche, il se met à genoux devant Jésus et lui dit :	A man with leprosy came and knelt before him and said, *"Lord, if you are willing, you can make me clean."*
00004	Seigneur, si tu le veux, tu peux me guérir !	*Jesus reached out his hand and touched the man. "I am willing," he said. "Be clean!" Immediately he was cured of his leprosy.*
Our alignment		
Verses	**French**	**English**
00	Matthieu 8	Matthew 8
01	Lorsque Jésus fut descendu de la montagne une grande foule le suivit	When he came down from the mountain great crowds followed him
02	Et voici un lépreux s'étant approché se prosterna devant lui et dit : Seigneur si tu le veux tu peux me rendre pur	And behold a leper came to him and knelt before him saying Lord if you will you can make me clean
03	Jésus étendit la main le toucha et dit : Je le veux sois pur Aussitôt il fut purifié de sa lèpre	And Jesus stretched out his hand and touched him saying I will be clean And immediately his leprosy was cleansed

Table 3.4: A comparison between CMU's multilingual alignment and ours. Text in italic presents alignment mismatches between English and French. We used a slightly different (*non-drama*) version of the Bible, hence the small differences in the displayed texts.

ments. The output of our pipeline is a set of 8,160 audios segments, aligned at verse-level, in eight different languages, with an average of 20 hours of speech for each language. An example of an aligned verse is presented in Figure 3.4. Corpus statistics are presented in Table 3.5. Table 3.4 illustrates the difference between the multilingual alignment available on the CMU Wilderness Multilingual Speech dataset, compared to our approach.

The languages covered in our dataset present interesting linguistic features due to their different origins. We have five Indo-European languages, being three of them from the Romance family (French, Romanian, Spanish), one from the Germanic Family (English) and one from the Slavic Family (Russian). There are two entries for the Uralic languages: one from the Ugric family (Hungarian), and the other one from the Finnic (Finish). Finally, we also have a low-resource and isolated language: Basque.

We believe the obtained corpus can be useful in several applications, such as speech-to-speech retrieval (Lee et al., 2015b), multilingual speech representation learning (Harwath et al., 2018), and direct speech-to-speech translation (Tjandra et al., 2019; Zhang et al., 2020). Moreover, typological and dialectal fields could use such a corpus to solve some of the following novel tasks using parallel speech: word alignment, bilingual lexicon extraction, and semantic retrieval.

language	#tokens	#types	tokens per verse	types per verse	avg token length	audio length(h)	avg verse length(h)
English (EN)	176,461	6,471	21.52	18.03	3.82	18.50	8.27
Spanish (ES)	168,255	11,903	20.52	17.90	4.17	21.49	9.58
Basque (EU)	128,946	14,514	15.78	14.88	5.55	22.76	9.75
Finnish (FI)	134,827	18,824	16.44	15.04	5.66	23.16	10.21
French (FR)	183,786	10,080	22.36	19.25	4.02	19.41	8.62
Hungarian (HU)	135,254	20,457	16.46	15.01	5.07	21.12	9.29
Romanian (RO)	169,328	9,581	20.61	18.19	4.14	23.11	10.16
Russian (RU)	129,973	16,758	15.82	14.50	4.44	22.90	9.70

Table 3.5: Statistics of the MASS corpus.

English (EN)	`It is a fearful thing to fall into the hands of the living God`
Spanish (ES)	`Es terrible caer en manos del Dios vivo`
Basque (EU)	`Izugarria da Jainko biziaren eskuetan erortzea`
Finnish (FI)	`Hirmuista on langeta elävän Jumalan käsiin`
French (FR)	`C est une chose terrible que de tomber entre les mains du Dieu vivant`
Hungarian (HU)	`Félelmetes dolog az élő Isten kezébe esni`
Romanian (RO)	`Grozav lucru este să cazi în mâinile Dumnezeului celui viu`
Russian (RU)	Страшно впасть в руки Бога живаго

Figure 3.4: A tokenized multilingual parallel verse from our dataset (Hebrews 10, verse 31).

In order to insure the quality of the distributed corpus, a human evaluation was performed on a corpus subset (8 language pairs, 100 verses) by bilingual native speakers. This evaluation was performed through an online platform, and it was focused on the quality of the audios.[18] Results attested the quality of the alignments.

Lastly, we highlight that the presented pipeline can be applied to any translation of the Bible, and thus the current corpus can be easily extended to cover new languages. For ensuring reproducibility, we share all scripts and information needed for this extension together with our corpus[19] (Boito et al., 2020a), named **MaSS** for **M**ultilingu**a**l corpus of **S**entence-aligned **S**poken utterances.

[18]Transcriptions were provided as a form of supporting the audio evaluation. Full description and discussion about the human evaluation can be found in Boito et al. (2020a).

[19]Available at: `https://github.com/getalp/mass-dataset`

4 Contributions Overview

The goal of all the projects presented in this chapter was to propose more realistic low-resource datasets for conducting investigations. The three datasets presented correspond to the publications listed below. The remainder of this session briefly review some of the work performed using these resources.

- **Mboshi-French Parallel Corpus**: Godard, P., Adda, G., Adda-Decker, M., Benjumea, J., Besacier, L., Cooper-Leavitt, J., Kouarata, G.-N., Lamel, L., Maynard, H., Mueller, M., Rialland,A., Stueker, S., Yvon, F., and Boito, M. Z. (2018). **A very low-resource language speech corpus for computational language documentation experiments.** *International Conference on Language Resources and Evaluation (LREC 2018).*
  38 citations as in 06/04/2021.

- **Griko-Itallian Parallel Corpus**: Boito, M. Z., Anastasopoulos, Lekakou, M., A., Villavicencio, and A., Besacier, L. (2018). **A Small Griko-Italian Speech Translation Corpus.** *International Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU 2018).*
  8 citations as in 06/04/2021.

- **MaSS dataset**: Boito, M. Z., Havard, W. N., Garnerin, M., Le Ferrand, É., and Besacier, L. (2020). **MaSS: A large and clean multilingual corpus of sentence-aligned spoken utterances extracted from the bible.** *International Conference on Language Resources and Evaluation (LREC 2020).*
  10 citations as in 06/04/2021.

The **Griko-Itallian parallel corpus** was one of the endangered languages used in Wada et al. (2020) for the learning of contextualized cross-lingual word embeddings in zero resource settings. The **Mboshi-French parallel corpus** has been adopted as a test set for evaluating low-resource speech approaches by many: Anastasopoulos and Chiang (2018b) used it for testing their multitask model for transcription, translation and word discovery.

The Mboshi-French dataset was again used for speech-to-text translation approaches in Bansal et al. (2019), Sung et al. (2019) and Inaguma et al. (2019). In Scharenborg et al. (2020), it was used for representation learning and speech translation. It was also used for AUD in low-resource languages (Scharenborg et al., 2018; Ondel et al., 2018; Ondel et al., 2019; Yusuf et al., 2020; Feng et al., 2021), and for ASR (Scharenborg et al., 2018).

Still about the Mboshi-French parallel corpus, and focusing on the transcriptions, Anastasopoulos and Chiang (2018a) and Matsuura et al. (2020) investigated models for obtaining transcriptions from speech. In the context

of language documentation, Le Ferrand et al. (2020) proposes a new workflow for interactive transcription. Focusing on Unsupervised Word Segmentation (UWS), this dataset was used by us for a significant part of the thesis investigation (Boito et al., 2017; Godard et al., 2018c; Boito et al., 2019a,b, 2021). It was also used in Godard et al. (2018b, 2019).

Chapter IV

A Bilingual Attention-based Unsupervised Word Segmentation Model

We now present the model motivated in Chapter II for bilingual Unsupervised Word Segmentation (UWS) from speech. This model works in two steps: (1) Speech Discretization (SD), and (2) bilingual alignment and segmentation. The pipeline is illustrated in Figure 6.1.

The first step is responsible for producing discrete speech units (pseudo-phones) from the speech utterances. The second step works on the symbolic domain, aligning the discovered units with the translation words, and producing from this segmentation. Since the speech labels contain timestamp information, the output segmentation can be carried to the speech domain, producing segmentation over the speech input itself. This process is *bilingual*, as the segmentation is performed supported by the bilingual alignment discovered. In other words, translation words are used to *ground* the generated segmentation.

The pipeline nature of our model allows us to segment small datasets, a task that would be difficult to accomplish if we were to directly train speech-to-translation models. Moreover, we are supported by the studies that show that neural networks are able to learn linguistic features working with units smaller than words, such as sub-word units and characters (Kreutzer and Sokolov, 2018; Hahn and Baroni, 2019; Ataman et al., 2019), and therefore suitable for working with phonemes or discrete speech units.

In this chapter we focus our investigation in the second step of our speech UWS pipeline, which works on the symbolic domain. We start by validating this model in the ideal scenario of perfect speech discretization, replacing discrete speech units by the true phones (phonemes) from the language.[1] This

[1]We refer to it as the *perfect* speech discretization because there is no noise added (manually produced). We highlight that this is not necessarily the representation SD models *need* to reach, as there might exist different forms of meaningfully representing the same utterance. In this setting, it might even be possible for an SD model to produce a better representation (in terms of exploitability) to the speech signal than the phonetization itself.

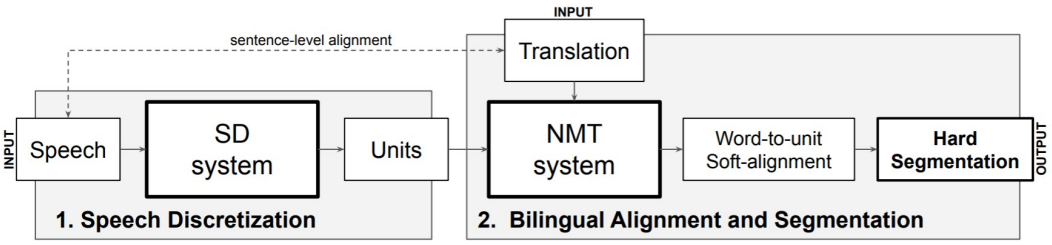


Figure 4.1: The general bilingual speech UWS pipeline. It requires as input a parallel dataset made of speech and sentence-level aligned translations. The system outputs word-level segmentation over the speech utterances. *Units* at the end of the first step correspond to the discrete speech units.

allow us to assess the *topline* performance that our models working from speech can accomplish.

Section 1 presents the core idea of our segmentation procedure: the use of source-to-target soft-alignment probability matrices for producing segmentation. Section 2 presents our two methods of evaluating performance. First, we assess the quality of the soft-alignment probability matrices produced by NMT training by using a task-agnostic metric we introduced in Boito et al. (2019a). Second, we evaluate the final product of our pipeline directly on the speech domain using UWS boundary metrics.

Section 3 presents the work from Boito et al. (2019a) and Boito et al. (2021). We compare soft-alignment probability matrices produced by three different attention-based NMT models for our UWS pipeline. We investigate how the different approaches for the attention mechanism impact our segmentation performance, and their robustness to low-resource settings. We also present an analysis of how the syntactic divergence between source and target sequences might impact the quality of the segmentation, and we illustrate how the task-agnostic metric we present in Section 2 can be used for increasing type discovery scores in low-resource settings.

Section 4 presents the work from Boito et al. (2020b). There, we investigate the impact that the bilingual supervision has over the discovered segmentation. For achieving that, we train several models by varying only the bilingual supervision, and we observe the difference in segmentation performance. We also use different segmentation targets, for illustrating that segmentation performance is language-dependent. Section 5 concludes the chapter by summarizing our findings.

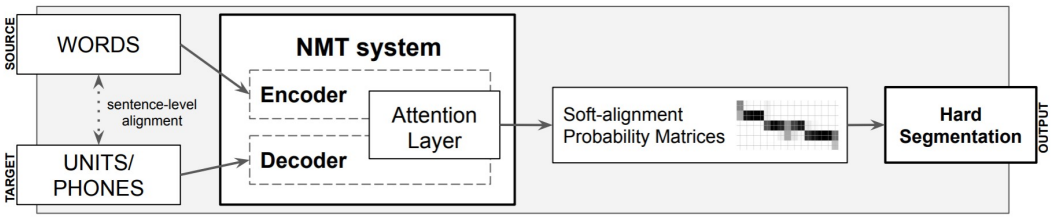


Figure 4.2: Detailed pipeline for the second step of our bilingual UWS pipeline from Figure 6.1. The discrete speech units, or phones, are fed to the decoder network. Soft-alignment probability matrices are extracted for each sentence after training, and transformed in speech segmentation.

1 Attention-based UWS Model

Our bilingual model for segmentation has as core component the attention-based NMT model described in Chapter II, Section 2.1. There are two features that make this architecture interesting for us. First, it incorporates sentence-level aligned text naturally, and second, its block-oriented nature allows us to have different representation levels in each coder. For instance, we can have text in one, and phones or speech in the other.

Figure 4.2 presents the detailed scheme for our bilingual alignment and segmentation step. We train a NMT model using sentence-level aligned examples (word-level for translations, unit/phone-level for the target language). Any attention-based NMT model able to produce soft-alignment between source and target sequences can be used for this task, and in Section 3 we compare segmentation performance by using three different NMT architectures.

Posterior to training, we retrieve the word-to-phone soft-alignment probability matrices, such as the ones presented in Figure 4.3. The soft-alignment is used to cluster together (segment) the target language phones into word-like units. We now describe how we perform segmentation from soft-alignment probability matrices.

Consider a given source and target sentence pair ($\boldsymbol{s}$, $\boldsymbol{t}$) of length $|\boldsymbol{s}|$ and $|\boldsymbol{t}|$ respectively. For every token in the target sequence, the attention layer outputs a probability $P(t_i, s_j)$, with $i \in [0, |\boldsymbol{t}|]$ and $j \in [0, |\boldsymbol{s}|]$, which quantifies the importance of the source token s_j for the generation of the target token t_i. We then have, for every target token t_i, a probability distribution over all tokens in the source sequence ($\sum_{j=0}^{|\boldsymbol{s}|} P(t_i, s_j) = 1$), which gives us the *soft-alignment* for that token.

For generating bilingual segmentation, we first transform this distribution into hard alignment by aligning each target token t_i to the source token with

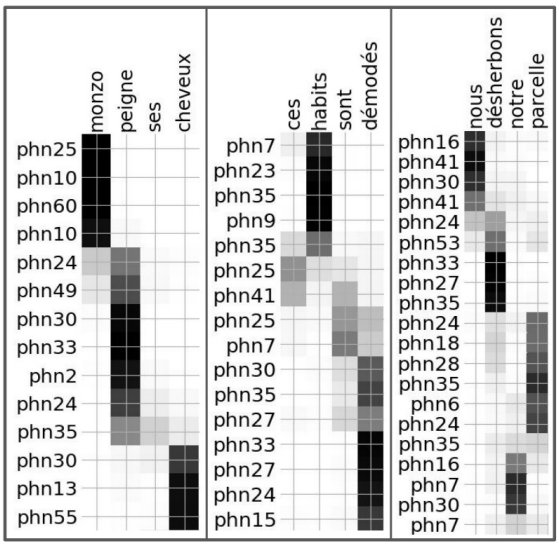


Figure 4.3: Soft-alignment probability matrix heatmaps between speech labels (rows) and french words (columns) for three sentences from the Mboshi-French Parallel Corpus (Chapter III). The higher the soft-alignment probability for a given pair, the darker is the square color.

maximum probability. That is: $\max_{\{0 \leq j \leq |\boldsymbol{s}|\}} (P(t_i, s_j))$. We then use the generated alignment to define which target tokens are segmented together. Consecutive target tokens aligned to the same source token are considered as part of the same word-unit. When an alignment shift occurs, and the next target token is aligned to a different source token, a boundary is inserted into the sequence.

For instance, in the example in the left in Figure 4.3, we obtain the following segmentation: `phn25-phn10-phn60-phn10` (aligned to `monzo`), `phn24-phn49-phn30-phn33-phn2-phn24-phn35` (aligned to `peigne`), `phn30-phn13-phn55` (aligned to `cheveux`). We observe that the word `ses` was not aligned to anything, being ignored. This kind of flexibility is necessary in order to account for the natural differences in morphology that languages have.

Previous to this thesis we validated this attention-based segmentation pipeline for the symbolic domain (characters instead of speech discretization).[2] During that investigation, we observed that one key aspect for successfully exploiting the attention mechanism is the translation *direction* for the NMT training.

This is because, as described above, the attention layer outputs probability distributions for every target token. When training word-to-phone NMT

[2]That pipeline, extensively described in Boito (2017), used as NMT architecture the RNN model from Bahdanau et al. (2015), applied to the Mboshi-French parallel corpus.

models, we retrieve probability distributions for every phone in the sequence, which allows us to choose between the source words, potentially ignoring some of them. In contrast, training phone-to-word NMT models means generating probability distribution over words, which will ignore a portion of the phones we aim to segment. Because of this, translation direction is important for our pipeline, and the target language input should always be fed to the decoder network.[3]

Finally, since the units we process and segment in this step of the pipeline come from the direct discretization of the speech, we also need to account for silence labels. These are automatically inserted into the sequences by SD models to represent periods of silence during an utterance. In preliminary experiments we investigated training the NMT models with this information, finding more benefit in removing it before training. We then reintroduce them after segmentation, since they provide a natural segmentation of the signal.

2 Model Evaluation

As mentioned in the last section, our segmentation pipeline uses soft-alignment probability matrices in order to segment. Therefore, we find important to evaluate the quality of these matrices as well, and not only the final segmentation performance. For this, we introduce a task-agnostic metric for assessing alignment quality (Section 2.1). In Section 2.2 we briefly go over the metrics used for evaluating segmentation performance in the speech domain.

2.1 Average Normalized Entropy

The ideal method for evaluating the quality of the soft-alignment probability matrices is to directly score their similarity to real word alignment. However, this is not realistic, especially in low-resource scenarios, as word-level alignment is not very often available for datasets. Thus, for assessing the overall quality of the soft-alignment probability matrices without having gold alignment information, in Boito et al. (2019a) we introduce Average Normalized Entropy (ANE).

Given the source and target sentence pair $(\boldsymbol{s}, \boldsymbol{t})$ of length $|\boldsymbol{s}|$ and $|\boldsymbol{t}|$ respectively, for every phone t_i, the Normalized Entropy (NE) is computed considering all possible words in $\boldsymbol{s}$ as in Equation 4.1, where $P(t_i, s_j)$ is the alignment probability between the phone t_i and the word s_j (a cell in the matrix). The

[3]Duong et al. (2016) differs from our approach by forcing the attention layer to output probability distributions on both directions: $\sum_{j=0}^{|\boldsymbol{s}|} P(t_i, s_j) = 1$ and $\sum_{i=0}^{|\boldsymbol{t}|} P(t_i, s_j) = 1$. This ensures that all source and target tokens are used.

ANE for a sentence is then defined by the arithmetic mean over the resulting NE for every phone from the sequence $\boldsymbol{t}$, as in Equation 4.2.

$$NE(t_i, \boldsymbol{s}) = -\sum_{j=1}^{|\boldsymbol{s}|} P(t_i, s_j) \cdot \log_{|\boldsymbol{s}|}(P(t_i, s_j)) \tag{4.1}$$

$$ANE(t, \boldsymbol{s}) = \frac{\sum_{i=1}^{|\boldsymbol{t}|} NE(t_i, s)}{|\boldsymbol{t}|} \tag{4.2}$$

From this definition, we can derive ANE for different granularities (sub or supra-sentential) by accumulating its value for the full corpus, for a single type or for a single token. *Corpus ANE* will be used to summarize the overall performance of the matrices produced by a NMT model on a specific corpus.

Token ANE extends ANE to tokens by averaging NE for all phones from a single (discovered) token. *Type ANE* results from averaging the ANE for every token instance of a discovered type.[4] Finally, *Alignment ANE* is the result of averaging the ANE for every discovered `(type, translation word)` alignment pair.

The motivation for using entropy comes from the fact that this metric summarizes the degree of *confusion* of our distributions. In other words, it assesses how concentrated (sharp) is a probability distribution between a given unit and the word candidates. The intuition that lower ANEs correspond to better alignments is exemplified in Figure 4.4.

2.2 Segmentation Evaluation on the Speech Domain

For directly evaluating the quality of our segmentation in the speech domain, we use the Zero Resource Challenge (ZRC) 2017 evaluation script (track 2) fully described in Dunbar et al. (2017).[5] This reference provides a standard for comparing performance of speech segmentation systems.[6]

For the task of UWS, it provides three metrics: boundary precision, recall and F-score. Boundary precision is the probability that the discovered boundaries are in the gold set of boundaries, and boundary recall is the probability that the gold boundaries are discovered. F-score is the arithmetic mean between boundary precision and recall. From this definition, metrics for Type

[4]Throughout this document we refer to a *token* as the collection of phones segmented into a word-like unit. *Types* are defined as the set of unique tokens (i.e. the lexicon).

[5]The ZRC evaluation scripts are available at: `https://github.com/bootphon/zerospeech2017`.

[6]The ZRC references for the datasets used during this thesis are available at: `https://github.com/mzboito/ZRC_corpora`.

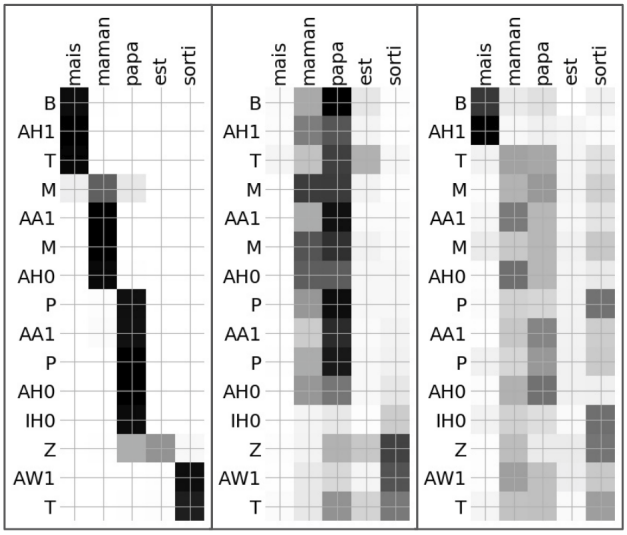


Figure 4.4: Soft-alignment probability matrices from the UWS task between English true phones (rows) and French words (columns). ANE values (from left to right) are 0.11, 0.64 and 0.83. The gold segmentation is "BAH1T MAA1MAH0 PAA1PAH0 IH0Z AW1T", which corresponds to the English sentence "But mama, papa is out".

and Token discovery scores can also be defined. In all cases, UWS segmentation results are computed over the totality of the corpora (training and validation sets).

3 Empirical Evaluation of NMT Models for UWS in Low-resource Settings

Recently the attention mechanism became an investigation target in sequence-to-sequence (seq2seq) models that process language, which resulted in the emergence of many different attention-based architectures for NMT. Here we investigate some of these different attention-based NMT approaches for generating the source-to-target soft-alignment probability matrices we use in our segmentation pipeline. We compare them with regards to their level of exploitability for the UWS task in low-resource settings.

We concentrate on the three NMT models presented in Section 2, Chapter II. These are: the Recurrent Neural Network (RNN) from Bahdanau et al. (2015), the original Transformer presented in Vaswani et al. (2017), and the 2D Convolutional Neural Network (CNN) from Elbayad et al. (2018). We apply them to the pipeline presented in Figure 4.2 using true phones as

input (topline performance), evaluating their UWS performance, as well as assessing the alignment quality by using the ANE metric. Throughout this investigation, we show that this ANE metric is correlated to the UWS performance of these attention-based NMT models. We also illustrate that it can be used for filtering the generated vocabulary, increasing type discovery scores.

Section 3.1 explains our experimental setup, detailing corpora and training regime. Section 3.2 presents the UWS results, and Section 3.3 discusses data size impact. We then focus on ANE, showing its correlation to boundary F-score in Section 3.4, discussing the impact of syntactic divergence in Section 3.5, and studying its use as a confidence metric for vocabulary filtering in Section 3.6. Section 3.7 summarizes our results.

3.1 Experimental Setup

For this investigation, we compare three NMT models, **RNN**, **2D-CNN** and **Transformer**, in low-resource settings. Section 3.1.1 presents the data we use for training, and Section 3.1.2 our training regime.

3.1.1 Datasets

We train the NMT models using the 5,130 parallel sentences from the Mboshi-French (MB-FR) Parallel Corpus (Chapter III). This corpus corresponds to a realistic setting of language documentation. Moreover, for assessing the sensitivity to low-resource data processing, we use a second dataset: the English-French (EN-FR) Parallel Corpus. This corpus is an extension from the Librispeech dataset[7] which includes automatically aligned French text sentences (Kocabiyikoglu et al., 2018).

We post-process the EN-FR corpus, retrieving only the high-quality alignments, and reaching a corpus which is made of 33,192 parallel sentences.[8] For providing a fair comparison, as well as to study the impact of corpus size, we down-sample it to 5,130 parallel sentences: to the exact same size as the MB-FR corpus.[9]

Lastly, we highlight that these datasets are quite distinct in nature. The EN-FR corpus presents larger vocabulary and longer sentences (literary text source). The MB-FR presents a more tailored environment, with short sentences and simpler vocabulary. Table 4.1 presents the statistics for the EN-FR corpus (both 33k and 5k), and for the MB-FR corpus.

[7]English audio books collected in Panayotov et al. (2015).

[8]Available at: `http://gitlab.com/mzboito/english-french-parallel-corpus`.

[9]Down-sampling was conducted preserving the original average number of tokens per sentence.

	#types		#tokens		avg token length		avg #tokens per sentence	
	source	**target**	**source**	**target**	**source**	**target**	**source**	**target**
EN-FR (33k)	21,083	33,135	381,044	467,475	4.37	4.57	11.48	14.08
EN-FR (5k)	8,740	12,226	59,090	72,670	4.38	4.57	11.52	14.17
MB-FR (5k)	6,633	5,178	30,556	42,715	4.18	4.41	5.96	8.33

Table 4.1: Statistics for the three source-target datasets.

3.1.2 Training Regime

For each NMT architecture, and for each one of the three corpora above mentioned, we train five models (runs) with different initialization seeds, reporting the standard deviation. The RNN, 2D-CNN and Transformer implementations come respectively from Bérard et al. (2016), Elbayad et al. (2018), and Ott et al. (2019). Before performing the segmentation explained in Section 1, we average all the generated soft-alignment probability matrices from the five different runs for each model. This can be seen as reaching an *agreement* between the alignments discovered by different runs. In preliminary experiments we saw that this increases our UWS boundary results.

Regarding optimization, our networks are optimized for the *monolingual task*, in which a phone sequence is segmented with regards to the corresponding word sequence (transcription) in the same language, hence monolingual.[10] The best parameters found for this monolingual setup are then used for training the bilingual models, and evaluation is performed in the *bilingual* segmentation condition, which corresponds to the real UWS task.

Regarding hyper-parameters, in Boito (2017) we performed an extensive study of dropout, batch size, embedding size, and number of layers for the RNN model in low-resource settings. We use our findings from there as the starting point for the optimization we perform in this study. Across all architectures, we use embeddings size of 64 and batch size of 32 (5k datasets), or embeddings size of 128 and batch size of 64 (33k dataset). Dropout of 0.5 and 550,000 steps for training are applied in all cases.

RNN models have only one layer, a bidirectional encoder, and cell size equal to the embedding size. 2D-CNN models use the hyper-parameters from Elbayad et al. (2018) with only 3 layers (5k dataset), or 6 (33k dataset), and kernel size of 3. Transformer models were optimized starting from the original parameters of Vaswani et al. (2017). Best results (among 50 setups) were achieved using 2 heads, 3 layers (encoder and decoder), warm-up of 5k steps, and using cross-entropy loss without label-smoothing. For selecting

[10]This task can be seen as an automatic extraction of a pronunciation lexicon from parallel words and phone sequences.

		Bilingual			Monolingual		
		P	**R**	**F**	**P**	**R**	**F**
EN 33k	**RNN**	**70.0**	**85.9**	**77.1**	**99.7**	**99.9**	**99.8**
	2D-CNN	63.9	80.5	71.3	97.8	99.3	98.6
	Transformer	48.1	58.2	52.7	92.0	98.1	94.9
EN 5k	**RNN**	**66.2**	75.2	**70.4**	**99.0**	99.5	**99.3**
	2D-CNN	44.5	75.2	55.9	98.1	**99.6**	98.8
	Transformer	37.4	**88.0**	52.5	70.7	94.5	80.9
MB 5k	**RNN**	**72.3**	75.9	**74.0**	**92.9**	**92.1**	**92.5**
	2D-CNN	65.9	70.6	68.2	89.6	90.1	89.8
	Transformer	56.6	**80.2**	66.4	79.8	87.7	83.5

Table 4.2: Precision (P), Recall (R), and F-score (F) UWS boundary results for the NMT models (RNN, 2D-CNN, Transformer) trained on the three corpora (EN 33k and 5k, MB 5k) in bilingual (real) and monolingual (topline) settings. Best results for each setting presented in bold.

which head to use for UWS, we experimented using the last layer's averaged heads, or by selecting the head with minimum corpus ANE. While the results were not significantly different, we kept the ANE selection.

Finally, we also present results for the baseline `dpseg`, presented in Chapter II. We use its unigram model, which yielded better results compared to the bigram model.[11] The hyper-parameters are replicated from the study for low-resource monolingual UWS performed in Godard et al. (2016).

3.2 UWS Results

The UWS boundary results from phone sequences (in Mboshi or English) are presented in Table 4.2, with monolingual results shown for information only, since they are a topline. For 2D-CNN and RNN, average standard deviation for the bilingual task computed over 5 runs and the 3 corpora is of less than 0.8%. For Transformer, it is almost 4%.

Looking at the monolingual results in the table, we verify that the soft-

[11] This was equally observed in Godard (2019).

alignment probability matrices produced by all three models trained in mid to low-resource settings are exploitable for solving this *easier* task.[12] Then, as expected, when word transcriptions are replaced by their translations (bilingual setup), the overall performance of the UWS models drops.

Looking at these bilingual results, we see that, surprisingly, RNN models outperform the more recent approaches (2D-CNN and Transformer). One possible explanation is the lower number of parameters (for a 5k setup, in average 700k parameters are trained, while 2D-CNN needs an additional 30.79% and Transformer 5.31%). However, for 33k setups, 2D-CNNs actually need 30% less parameters than RNNs, but still perform worse.[13] Thus, even if model size impacts performance (having more trainable parameters meaning needing more data to converge), it is still not the only factor for assessing the exploitability of the soft-alignment probability matrices in low-resource settings.

Transformer's low performance could be due to the use of several heads, which could be "distributing" alignment information across different matrices. Nonetheless, we evaluated averaged heads and single-head models, and these resulted in significant decrease in performance. This suggests that this architecture may not need to learn explicit alignment to translate, but instead it could be capturing different kinds of linguistic information. This was discussed in the original paper, and illustrated in the provided examples (Vaswani et al., 2017).

Also, on the decoder side, the behavior of the self-attention mechanism on phones is unclear and under-studied so far. For the encoder, Voita et al. (2019) performed after-training encoder head removal based on *head confidence*, showing that after initial training, most heads were not necessary for maintaining translation performance. Michel et al. (2019) reached a similar conclusion: removing all heads but one, they found a negligible loss in performance during decoding stage. Hence, we find the multi-head attention mechanism interpretation challenging, and maybe not suitable for a direct UWS application, especially in low-resource settings.

As in Godard (2019), our best UWS method (RNN) for the bilingual task does not reach the performance level of the strong NP Bayesian baseline `dpseg`, with F-scores of 89.80 (EN 33k), 87.93 (EN 5k), and 77.00 (MB 5k). However, our UWS approach has the benefit of providing bilingual annotation to the words discovered. These can be used, for instance, for increasing type discovery scores (Section 3.6). Moreover, Chapter VI will discuss how this baseline

[12]The monolingual task is considered easier because the discovered alignments will be very *diagonal*, with no order inversions or words needing to be ignored.

[13]The number of trainable parameters for each architecture are presented in the Appendix A, Table 1.1.

is less robust working from discrete speech units.

3.3 Robustness to Low-resource Settings

Looking at the EN 33k and EN 5k results of Table 4.2, we can observe the impact of data size on the NMT models. For the bilingual task, RNN performance drops by 7% on average, whereas the performance drop is bigger for 2D-CNN (14-15%). Transformer performs poorly in both cases, and increasing data size from 5k to 33k seems to help only for the monolingual setup.

The EN 5k and MB 5k results illustrate the impact of the language pair in our bilingual UWS pipeline. We know from Fourtassi et al. (2013) and Rialland et al. (2015) that English should be easier to segment than Mboshi, and this was confirmed by both *dpseg* and monolingual results. However, this trend is not confirmed in the bilingual scenario, where the quality of the (sentence-aligned) parallel corpus seems to have a greater impact (higher boundary F-scores for MB 5k than for EN 5k for all models).

As shown in Table 4.1, MB-FR corpus has shorter sentences and smaller lexicon diversity, while EN-FR is made of automatically aligned books (noisy alignments), which may explain our experimental results. In Section 4 we perform an in-depth investigation of the impact of the bilingual supervision in the quality of the segmentation.

3.4 Correlation Between ANE and Boundary Scores

We established that our UWS pipeline works in low-resource settings in the ideal scenario where the speech discretization is perfect (training models using the true phones), reaching our best UWS results by using the soft-alignment probability matrices from the RNN model. We now investigate the use of the ANE metric for assessing the quality of the soft-alignment probability matrices produced by the NMT models, starting by verifying its correlation to the Boundary F-scores.

Applying the methodology from Section 2.1, we reach the ANE scores presented in Table 4.3.[14] We then compute the Pearson's ρ correlation coefficients between them and boundary F-scores for all mono and bilingual runs of all corpora ($N = 30$). We find the following values: -0.98 (RNN), -0.97 (2D CNN), and $-0,66$ (Transformer), with p-values smaller than 10^{-5}. These

[14] A note about the Transformer's overall Corpus ANE performance: we highlight that due to the head selection based on Corpus ANE, the final Corpus ANE values for the runs from this model are expected to be lower than for the RNN and 2D-CNN, where no selection is performed. Moreover, the very low results for the bilingual setup on the EN 33k dataset seem to highlight an apparent lack of robustness for low-resource training for this model.

		EN 33k	EN 5k	MB 5k
Transformer	Monolingual	0.06	0.13	0.20
	Bilingual	0.18	0.68	0.59
2D-CNN	Monolingual	0.17	0.17	0.17
	Bilingual	0.56	0.73	0.58
RNN	Monolingual	0.02	0.03	0.14
	Bilingual	0.38	0.41	0.42

Table 4.3: Average Corpus ANE scores over the 5 runs for the different models we trained. Scores $\in [0, 1]$, smaller values being better (lower entropy).

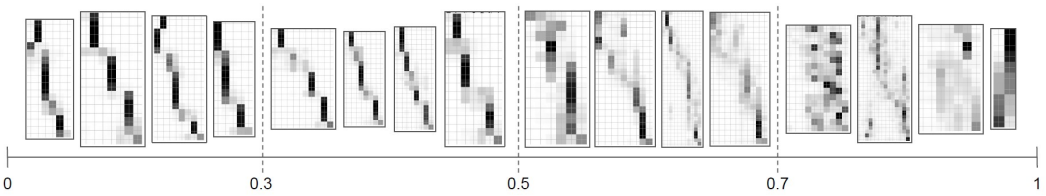


Figure 4.5: An illustration of the apparent correlation between Sentence ANE and soft-alignment quality. The heatmaps displayed correspond to random sentences sampled from the RNN model trained on the MB-FR language pair. This tendency is observed for all NMT models.

strong negative correlations confirm our hypothesis that lower ANEs correspond to sharper and better alignments. Figure 4.5 illustrates the degradation in the soft-alignment probability matrices' apparent quality as the ANE score increases.[15]

3.5 ANE and Syntactic Divergence

We commented that the monolingual setup is an easier task because of the direct equivalence between source and target sequences. Between phones and their word transcriptions, there is no word inversion (syntactic divergence), which makes the alignment an easier task. We now illustrate the relationship between alignment complexity and the quality of the discovered segmentation for the bilingual setup.

[15]More examples available at: `https://gitlab.com/mzboito/attention_study/-/tree/master/examples`.

For this experiment we use the tool `FastAlign` (Dyer et al., 2013) to obtain *alignment scores* for all sentences in the MB-FR corpus, using the reference segmentation in Mboshi and the French text. The resulting scores can be seen as the degree of *syntactic divergence* between source and target sentences: they measure how direct the bilingual word alignment is accordingly to `FastAlign`.[16]

We then create four *alignment complexity buckets* of equal size in number of matrices, for separating the corpus in four subsets with different degrees of complexity for our UWS task. For this analysis, we use the soft-alignment probability matrices produced by the RNN model. Figure 4.6 presents an example per bucket for sentences of equal source length: buckets one to four have increasing alignment complexity scores accordingly to `FastAlign`.[17]

For verifying the intuition that alignment quality will deteriorate as alignment complexity rises, we extract Alignment ANE scores for the matrices sets in every bucket. The alignment ANE score for a given `(discovered type, translation word)` pair gives us information about how confident the network is about that discovered pairing. The result of this is a collection of alignments and ANE scores for each bucket.[18]

Then, for each complexity bucket we evaluate its *precision* for the type discovery task. This allows us to verify if our model is more often correct on its segmentation (better overall precision) when working on sentences with *straightforward* alignment. Moreover, within each bucket, we sort the alignment pairs by their alignment ANE scores, computing precision for different Alignment ANE thresholds. This informs us about the quality of the *best* examples, in terms of alignment confidence, for each bucket.

Table 4.4 presents the type discovery precision scores for UWS using different Alignment ANE thresholds within each bucket. We notice that buckets with easier examples in terms of alignment probabilities (from `FastAlign`) have higher overall precision (see last row). This confirms that the quality of the alignments obtained is related to the syntactic divergence of the sentences. However, it is interesting to notice that even for the most challenging case (bucket 4), there are still a fair amount of alignments being retrieved.[19] We believe this highlights the robustness of the RNNs, that even in low-resource settings, are able to learn non-trivial equivalences between source and target sentences.

Focusing on the Alignment ANE threshold, we observe that it can be

[16]Since we use an automatic method, the resulting scores are an approximation.

[17]Alignment scores thresholds of, respectively, -10.61, -46.87, -60.18, and -78.15.

[18]For instance, in Figure 4.6 the first alignment pair for the matrix in the left is (`phn25-phn10-phn60-phn10`, `monzo`).

[19]We see a drop of 14.1 in precision from bucket 1 to 4.

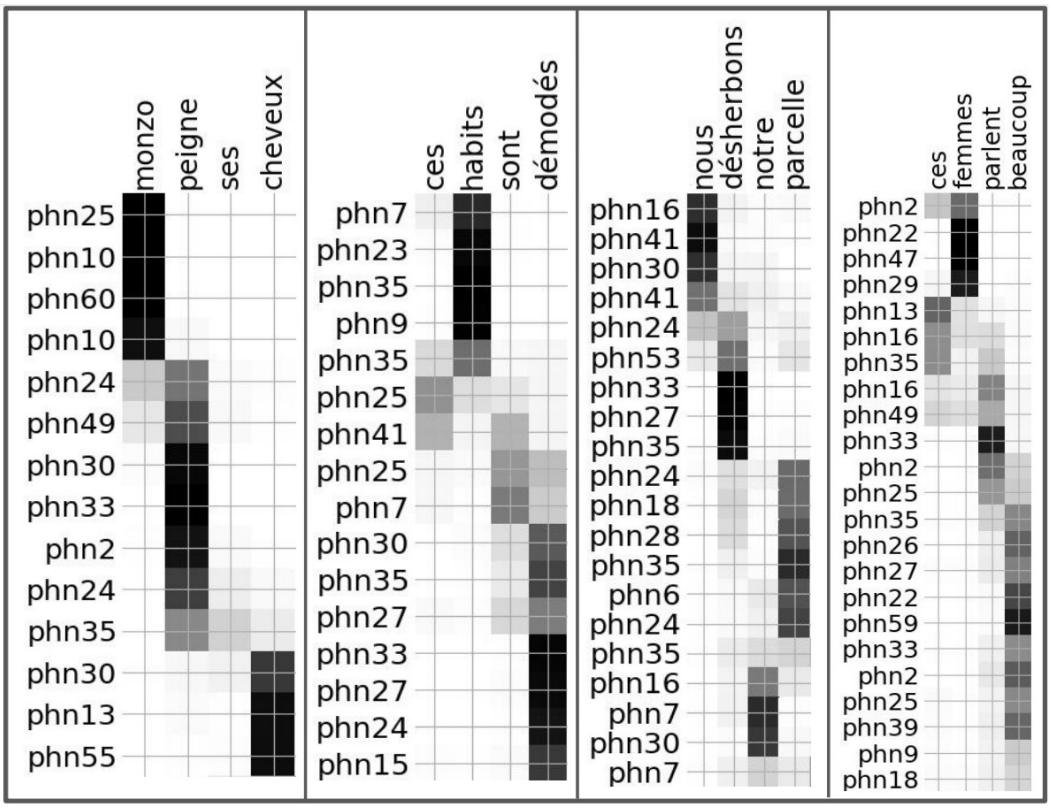


Figure 4.6: Soft-alignment probability matrices from the alignment complexity buckets 1 to 4 (left to right) for examples with same source length. Darker squares correspond to higher probabilities. The sentence ANE scores are, from left to right, 0.26, 0.40, 0.47 and 0.53. The language pair is French (words) and Mboshi (phonemes).

used for filtering the alignments, resulting in higher type discovery precision. For instance, looking at the bucket 1, we see that by limiting our search for the alignments which scored less or equal to 0.2 (first row), 68.8% of them correspond to real words (types) in the language. This illustrates the potential of ANE for vocabulary filtering, which we will continue to explore in the next section.

3.6 ANE for Vocabulary Filtering

Supported by the results in Sections 3.4 and 3.5, we now investigate the use of Alignment ANE as a confidence measure for vocabulary filtering. From the RNN models, we extract and rank the discovered alignment pairs by their Alignment ANE, and we examine if this metric can be used to separate true

ANE (≤)	Bucket 1	Bucket 2	Bucket 3	Bucket 4	All Buckets
0.2	68.8	59.2	56.4	47.8	49.0
0.4	44.8	41.4	38.0	31.8	32.6
0.6	38.3	34.5	30.6	25.3	24.7
0.8	36.8	32.4	28.8	22.8	22.2
1	36.7	32.4	28.8	22.6	22.1

Table 4.4: Type discovery precision scores for the alignment complexity buckets, and for the totality of the corpus (All buckets). Results in each of the rows are cumulative and use the Alignment ANE thresholds indicated in the first column.

words in the discovered vocabulary from the rest. For achieving this, we again evaluate our results for the type discovery task, but this time reporting results for all three metrics (precision, recall and F-score).

The results for low-resource scenarios (5k setups only) in Table 4.5 suggest that low ANE scores correspond to the portion of the discovered vocabulary the network is *confident* about, and these are, in most of the cases, true discovered lexical items (first row, $P \geq 70\%$).[20] As we add higher Alignment ANE values, we increase recall but lose precision. Still, for both languages and at a given ANE threshold, we are able to reach a higher type discovery F-score than by using the totality of the discovered vocabulary.

This suggests that, in a documentation setting, ANE could be used as a confidence measure by a linguist to extract a list of generated types with higher precision, without having to pass through all the discovered vocabulary. In Table 4.6 we exemplify this by presenting top low and high ANE results for our ranking using the EN 5k corpus.

Finally, in this work our focus lies on filtering the discovered types. However, as mentioned, our approach also retrieves the aligned information for the generated lexicon (translation candidates), and we observe in Table 4.6 that at least half of the alignment pairs in the top 10 low ANE entries present correct translations. We thus hypothesize that this aligned information could be explored for other documentation tasks, such as semantic retrieval.

[20]Type ANE, instead of Alignment ANE, was also investigated for this task, and results were positive, but slightly worse than the ones presented.

	EN 5K			MB 5K		
ANE (≤)	P	R	F	P	R	F
0.1	**70.97**	0.50	1.00	**72.13**	0.57	1.12
0.2	55.43	3.85	7.20	49.02	2.89	5.46
0.3	44.99	12.51	19.58	38.18	8.14	13.41
0.4	32.81	21.76	**26.17**	32.63	16.61	22.01
0.5	23.37	28.17	25.54	27.93	23.44	25.49
0.6	18.54	32.41	23.59	24.73	27.61	**26.09**
0.7	16.23	34.34	22.04	23.00	30.12	26.08
0.8	15.21	35.16	21.23	22.17	**30.95**	25.84
0.9	15.01	35.31	21.06	22.06	31.05	25.80
All	15.01	**35.34**	21.07	22.06	31.05	25.80

Table 4.5: Type discovery recall scores using Alignment ANE for keeping the most confident `(type, translation)` pairs. Results in each row are cumulative and use the Alignment ANE thresholds indicated in the first column.

	Top Low ANE	Top High ANE
1	SER1 (sir, *EOS_token*)	AH0 (a, convenablement)
2	HHAH1SH (hush, chut)	IH1 (INV, ah)
3	FIH1SHER0 (fisher, fisher)	D (INV, riant)
4	KLER1K (clerc, clerc)	N (INV, obéit)
5	KIH1S (kiss, embrasse)	YUW1 (you, diable)
6	GRIH1LD (grilled, grilled)	IH1 (INV, quen)
7	WUH1D (would, mennuierais)	AE1T (at, laquelle)
8	HHEH1LP (help, aidez)	Z (INV, bas)
9	DOW1DOW0 (dodo, dodo)	YUW1P (INV, *EOS_token*)
10	KRAE1BZ (crabs, crabes)	L (INV, parfaitement)

Table 4.6: Top 10 low and high ANE ranking for the discovered types (EN 5k), with gold transcription and aligned information between parentheses (respectively). "INV" means incorrect type.

3.7 Discussion

In this section we investigated the use of three different seq2seq attention-based NMT models for extracting the soft-alignment probability matrices we use in our bilingual attention-based UWS pipeline. We perform experiments across two different languages, and two different dataset sizes, using the true phones in English (EN) and Mboshi (MB) as a topline for the speech discretization step.

Our UWS results show that the RNN model produces the most exploitable soft-alignment probability matrices for our task in this low-resource setting, and across different languages (EN and MB) and dataset sizes (33k and 5k). Investigating the reasons for its quality, we scored syntactic divergence using the `FastAlign` tool, finding that this model is able to produce exploitable soft-alignment probability matrices even when sentence pairs are considered to be very distinct.

However, compared to the strong baseline `dpseg`, the F-score for this RNN model was inferior by 3 points. We still find these results encouraging, because our model also retrieves bilingual alignments to root its segmentation, and these can be used for many tasks.

Regarding the ANE metric, the UWS results were shown to be strongly negatively correlated to Corpus ANE scores for all NMT models. This informs us that this task-agnostic metric can be used to assess the quality of the soft-alignment probability matrices we use in our pipeline. Moreover, we illustrated that Alignment ANE can be used for filtering the generated vocabulary, increasing type discovery scores.

Lastly, using two different target languages (EN and MB), we verified that the supervision played a role in the quality of the segmentation obtained in bilingual settings. In the next section we will investigate in depth the effect of the bilingual supervision in our UWS pipeline.

4 Investigating Language Impact in the Bilingual UWS model

After verifying that we are able to exploit soft-alignment probability matrices from NMT models trained in low-resource settings, we now look at the impact caused by having different languages as source (annotation, translation words) and target (unsegmented phones) in our bilingual attention-based UWS pipeline. For this experiment, we use the eight languages from the multilingual speech-to-speech MaSS dataset, presented in Chapter III.

We then 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 our approach.[21]

The experiment is organized as follows. In Section 4.1 we detail our experimental setup, and in Section 4.2 we present segmentation and translation results for all the language pairs. Based on these results, we investigate the language impact by studying the language ranking obtained (Section 4.3), analysing the discovered vocabulary (Section 4.4), and assessing alignment confidence (Section 4.5). We present a final discussion in Section 4.6.

4.1 Experimental Setup

For this investigation, we train the RNN models as in the last section, but using the languages from the MaSS dataset (Section 4.1.1). The training regime is summarizing in Section 4.1.2, and evaluation is detailed in Section 4.1.3.

4.1.1 Dataset

We use the MaSS dataset (Boito et al., 2020a) described in Chapter III. This 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 than 100 in number of tokens. The resulting corpus contains 5,324 sentences, a size which is compatible with the experiments performed in the last section (5,130 sentences).

For the phonetic transcription of the speech (target side of the pipeline), we use the automatic phonetization from *Maus forced aligner* (Kisler et al., 2017). The transformation from word graphemes (original) to phonemes results in an average vocabulary reduction of 835 types, the smallest being for RO (396), and the most expressive being for FR (1,708).[22] The phonetization for the languages presents an average number of unique phonemes of 42.5. Table 4.7 presents statistics for the text expressed as graphemes (original) and as phonemes (from the phonetization performed).

[21] We highlight that we use a dataset of high-resource languages due to the lack of multilingual resources in documentation languages that could be used to investigate this hypothesis.

[22] This difference depends on the distance between phonetic and graphemic forms for each language.

	TEXT IN GRAPHEMES				TEXT IN PHONEMES			
	#types	#tokens	avg token length	avg #tokens per sentence	#types	#tokens	avg token length	avg #tokens per sentence
EN	5,232	90,716	3.98	17.04	4,730	90,657	3.86	17.03
ES	8,766	85,724	4.37	16.10	7,980	85,724	4.30	16.10
EU	11,048	67,012	5.91	12.59	9,880	67,012	6.94	12.59
FI	12,605	70,226	5.94	13.19	12,088	70,226	5.97	13.19
FR	7,226	94,527	4.12	17.75	5,518	93,038	3.21	17.48
HU	13,770	69,755	5.37	13.10	12,993	69,755	5.86	13.10
RO	7,191	88,512	4.06	16.63	6,795	84,613	4.50	15.89
RU	11,448	67,233	4.66	12.63	10,624	67,176	6.19	12.62

Table 4.7: Statistics for the subset of 5,324 sentences of the MaSS corpus.

4.1.2 Training Regime

We replicate the training regime from Section 3.1 for the RNN model, and for the baseline `dpseg`. Due to the considerable number of networks we need to train for this experiment, and supported by the low standard deviation we found for RNN models in the past, we reduce the number of runs per model, training only two. This results in the training of 112 NMT models. Regarding the data, 10% of the multilingual ids were randomly selected for validation, and the remaining were used for training.[23] This ensures all networks are trained with the same parallel information, and are therefore comparable.

4.1.3 Evaluation Protocol

For this experiment, we evaluate results on the symbolic domain, instead of using the ZRC evaluation protocol introduced in Section 2. We do so because at the time of this experiment we encountered difficulties to produce the reference necessary for speech-level evaluation using the MaSS dataset. As the utterances for this dataset do not exactly correspond to sentences, the average length is higher, and the computational memory cost of the ZRC scripts becomes very elevated. As we wanted to avoid filtering the corpus further for removing longer utterances, we opted for this form of evaluation.

Also, for these experiments, we evaluate the translation quality of the NMT models by using the BLEU score (Papineni et al., 2002). This allows us to investigate the correlation between UWS boundary results and the *translation quality* of these models. This investigation extends our previous study (Boito, 2017), in which we noticed that the best soft-alignment probability matrices were not necessarily produced by the best translation models.

[23] This is the same protocol applied for Section 3.

Boundary F-score

	EN	ES	EU	FI	FR	HU	RO	RU
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	-

BLEU Scores

	EN	ES	EU	FI	FR	HU	RO	RU
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	-

Table 4.8: UWS Boundary F-score (left) and BLEU score (right) results for all language pairs using the RNN model. The columns represent the target of the segmentation, while the rows represented the translation language used. Darker squares represent higher **column** scores. Best scores presented in bold. Better visualized in color.

4.2 UWS Results

UWS boundary F-score and BLEU score results are presented in Table 4.8. The `dpseg` UWS F-score results for these languages are considerably higher: 82.4 (EN), 79.2 (ES), 81.0 (EU), 80.0 (FI), 78.1 (FR), 75.5 (HU), 82.0 (RO), and 78.3 (RU).

We observe that segmentation and translation scores are strongly correlated for all eight languages, with an average ρ-value of 0.94 (significant to $p < 0.05$). Only one language (EU) presented correlation results (0.94) not significant to $p < 0.01$, and we believe the general lack of segmentation performance in this case could explain this result. Therefore, we conclude that higher BLEU scores will correspond to better, directly exploitable, soft-alignment attention matrices.

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 (69.1), while EU is the most challenging to segment using our bilingual attention-based approach (38.4). We also notice that, following intuition, some languages are more difficult to segment than others. In the following sections we investigate the impact of the supervision language (source), and the vocabulary and alignments obtained.

4.3 Source Language Impact

For assessing the impact of the language chosen for the supervision (source) in our bilingual UWS pipeline, we investigate the obtained language ranking in terms of the best translation languages for a given target segmentation language. Complementary to this, in the Appendix A we provide a study comparing the obtained results to a bilingual segmentation baseline.

There, we use a bilingual *proportional* model for studying the relationship between the UWS results and the ratio between the number of tokens per sentence in source and target sequences. We observe a large gap of almost 20 F-score points between applying this simple proportional method and ours. We conclude 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.

Looking into the quality of the segmentation results (Table 4.8) 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,[24] 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 statistical features. Table 4.7 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 models, results seemed to reproduce the trend from the proportional baseline (Appendix A), in which these pairs were also found to be the best. This could be the result of a low syntactic divergence between the languages of these pairs.

Finally, the language isolate EU is not a good choice for segmenting any language (worst result for all languages). Moreover, results for EU segmentation are both low (F-score and BLEU) and very close to the proportional baseline (average difference of 4.23), which suggests that these models were not able to learn meaningful bilingual alignment.

4.4 Analysis of the Discovered Vocabulary

Next we study the characteristics of the vocabulary produced by the bilingual models, focusing on the impact caused by the aligned translation. Table 4.9

[24] 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
EN	-	29.9	17.5	23.8	47.9	20.7	41.9	27.3
ES	49.3	-	22.0	25.4	**50.1**	21.6	40.9	27.5
EU	36.6	22.6	-	24.8	34.0	19.5	32.1	22.6
FI	48.3	29.2	**24.0**	-	40.5	27.9	39.9	**35.8**
FR	**58.5**	**38.4**	19.7	29.5	-	20.5	**50.6**	30.0
HU	42.2	29.1	23.8	**37.4**	36.4	-	37.4	31.8
RO	48.4	31.3	21.4	28.8	49.8	21.6	-	30.4
RU	47.5	28.2	21.2	37.4	41.5	**28.0**	41.4	-

Table 4.9: Type discovery recall scores for the bilingual-rooted UWS models. The columns represent the target of the segmentation, while the rows represented the translation language used. Darker squares represent higher **row** scores. Best (column) scores presented in bold. Better visualized in color.

presents the type discovery *recall* scores of our bilingual models.[25] This metric gives us information about the percentage of the true vocabulary the bilingual models were able to retrieve.

Looking at the rows, we see that FR>EN (58.5), FR>ES (50.6), ES>FR (50.1), RO>FR (49.8) and EN>FR (47.9) are the setups which retrieve most of the vocabulary, presenting the highest scores. The source for these models (FR, ES, RO and EN) are all fusional languages.[26]

We also notice that models for segmenting FI and HU (columns FI and HU in Table 1.2) present very low type discovery recall scores overall. 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 targets. In these conditions, the network must learn to align many translation words to the same structure in order to achieve the expected segmentation.[27] 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 segmenta-

[25]The boundary and F-scores results are presented in the Appendix A, Table 1.2.

[26]Fusional, or inflected, is the opposite of agglutinative, referring to languages in which one morpheme form can simultaneously denote multiple grammatical, syntactic, or semantic features.

[27]This is highlighted by the high average token length of the phonetic representation of these languages in Table 4.7.

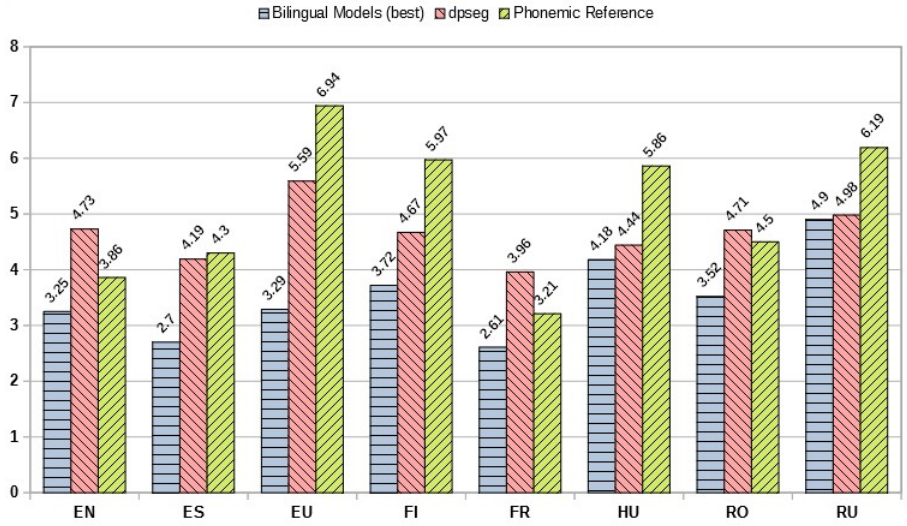


Figure 4.7: Average token length of the best bilingual UWS models from Table 4.8, `dpseg`, and reference.

tions produced in Figure 4.7, and supported by the overall low precision for type discovery (Appendix A, Table 1.2), we verify that our bilingual models tend to over-segment the output independent of the target language. This is probably due to the challenge of clustering the very long phoneme sequences into the many available source words (see statistics for words and phonemes per sentence in Table 4.7).

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 4.8 we see the EN soft-alignment generated by the FR and ES bilingual 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 produced alignment might be sub-optimal, and that a more in-depth evaluation of source-to-target correspondences would be ideal.

4.5 Alignment Confidence

The approach we use for bilingual UWS 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

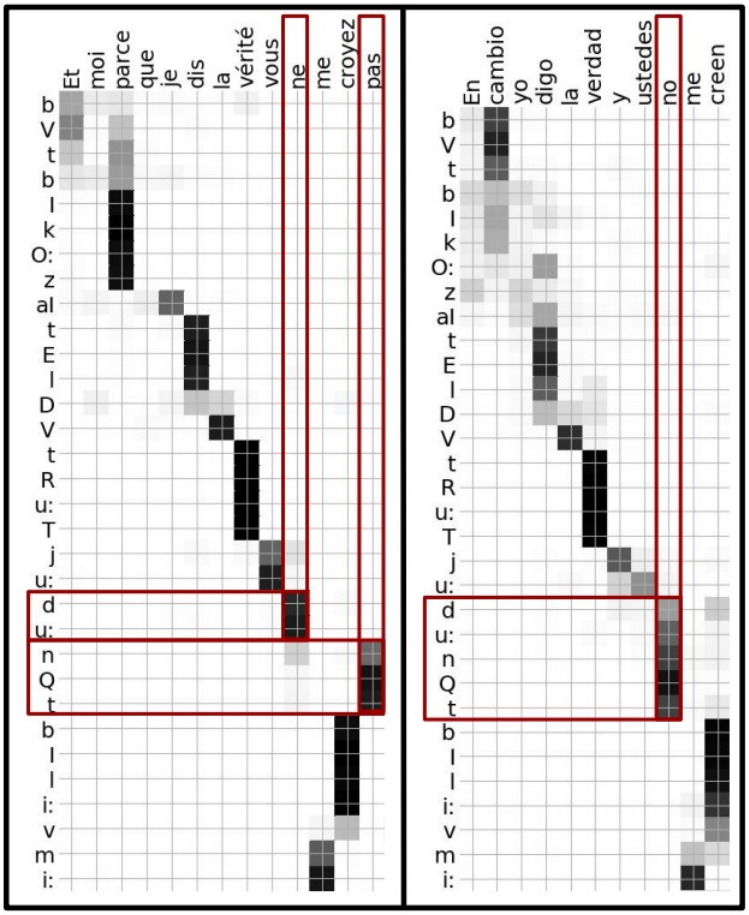


Figure 4.8: EN soft-alignment probability matrices generated by FR (left) and ES (right) bilingual models. The squares represent alignment probabilities (the darker the square, the higher the probability). The EN phonemization (rows) correspond to the following sentence: "But because I tell the truth, you do not believe me".

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

Table 4.10: Top 10 low Alignment ANE ranking for FR models trained with EN, ES and RU supervision. Each column brings the discovered types with gold transcription and aligned information between parentheses (respectively). "INV" means incorrect type.

extends the results from the previous section, that showed that models trained on different languages retrieve a different percentage of the vocabulary. We now aim to show that this difference in segmentation behavior comes from the different source-to-target correspondences discovered by the models with access to different languages.

We use the approach based on Alignment ANE from Section 3.6 for extracting the alignments the bilingual models are the most *confident about*. Table 4.10 presents the top 10 low ANE (high-confidence) pairs for FR models trained using 3 different translation languages (from Table 4.8, 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 purposes.

We observe a different set of discovered types depending on the language used, but it's noticeable that all languages learn a fair amount of biblical names and numbers, very frequent due to the nature of the dataset.[28] This highlights that very frequent types might be captured independently of the language used, but that 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 (`INV` in

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

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 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 being 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 the most, 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.

Moreover, this highlights how collecting data in *multilingual settings* (that is, using more than one translation language) could enrich low-resource approaches. In the next chapter we present our attempts to integrate multilingual information into our UWS pipeline.

4.6 Discussion

In language documentation scenarios, transcriptions are difficult to obtain. In order to ensure the interpretability of the recordings, a popular solution is to replace them by translations in high-resource 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 used for these translations in posterior automatic methods.

In this section we investigated the existence of a language-dependent behavior in our bilingual UWS pipeline. We simulated such a scenario by using the MaSS dataset for training 56 bilingual models, the combination of all the available languages in that 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 UWS results for every target language. We also verified that the languages are not all equally difficult to segment, but that this segmentation performance seems to be correlated to the translation capacity of the corresponding NMT model.

Moreover, while some of our *language ranking*, 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 statistical features (e.g. token length ratio between source and target sentences, and vocabulary size), related to the corpus, and not to the language features.

Finally, looking into the vocabulary produced by different bilingual models, we verified that those trained with the same parallel information, but using different languages to express that information, learned to focus on different bilingual structures. We believe this highlights the importance of carefully considering statistical and linguistic features for bilingual (and multilingual) language processing pipelines.

5 Conclusions

In this chapter we introduced our bilingual attention-based UWS pipeline for speech, which consists in two steps. The first step produces discrete speech units using SD models. The second uses these units, together with sentence-level word translations, to retrieve soft-alignment probability matrices from NMT models. The soft-alignment information is then used for segmentation.

The experiments we presented in this chapter focused on the second step: the task of bilingual alignment and segmentation. We investigated two important aspects that might impact its performance: (1) the attention-based NMT model used for generating the soft-alignment probability matrices, and (2) the language chosen for grounding the segmentation. In both cases, for reducing noise, we experimented with the true phones (phonemes) of the languages. This corresponds to the topline performance for models using an unsupervised discretization of the speech signal (full pipeline).

In our first experimental section, which corresponds to our work published in Boito et al. (2019a) and Boito et al. (2021), we investigated the use of different attention-based NMT models (RNN, 2D-CNN, Transformer) for producing the source-to-target soft-alignment probability matrices we use for segmentation. We found the RNN model to be the most exploitable in low-resource settings, reaching the best segmentation performance compared to the other two novel attention-based NMT approaches.

We also introduced a task-agnostic metric to assess the degree of exploitability of the soft-alignment probability matrices produced by NMT models. This metric, Average Normalized Entropy (ANE), can be accumulated across different representation levels (i.e. token, sentence, alignment, corpus). We showed that Corpus ANE is strongly correlated to the segmentation performance, and that Alignment ANE allows us to filter the generated vocabulary,

increasing type discovery scores.

Our second experimental section, which corresponds to our work published in Boito et al. (2020b), focused on the impact of language in our bilingual segmentation. We used a multilingual corpus for segmenting a given language supported by the same information in seven different languages. By varying the target language, we produced 56 bilingual models, allowing us to clearly verify the impact of the supervision in the generated segmentations.

Our results highlighted the existence of a relationship between language features and the segmentation performance for our approach. We verified that languages close in phonology and linguistic family scored better, while less similar languages yielded lower scores. While we find that our results are affected by linguistic features, we also believe that there is a non-negligible influence from corpus statistic features which can greatly impact neural approaches in low-resource settings.

The next chapter of this thesis studies extensions for the second step of the attention-based UWS pipeline presented here. Chapter VI then presents the SD step, and results for the complete pipeline, which works from speech.

Chapter V

Model Extensions for Attention-based UWS

In the previous chapter, we presented our pipeline for bilingual attention-based UWS from speech, and in low-resource settings. It is made of two different parts: a Speech Discretization (SD) component, and a bilingual alignment and segmentation component. Focusing on the latter, we investigated the impact of using different attention mechanisms for producing bilingual alignment, and we assessed the impact of the supervision's language. Before presenting in detail the SD step in Chapter VI, we focus on possible extensions for this bilingual alignment and segmentation component, with the goal of increasing UWS scores.

Inspired by documentation initiatives approaches, in Section 1 we investigate the leveraging of partial transcriptions from the bilingual corpus (i.e. monolingual data), and in Section 2 we study the leveraging of *boundaries suggestions* into the pipeline. Focusing on the training regime, in Section 3 we experiment with the extension proposed in Godard et al. (2019), in which some word-length bias is introduced into the produced soft-alignment probability matrices during training. Finally, Section 4 presents some less successful experiments regarding multilingual supervision for UWS, and Section 5 concludes this chapter, summarizing our findings.

1 Monolingual Data Leveraging

One of the motivations for our work lies in the impossibility of expecting extensive transcriptions for speech in low-resource settings, especially for oral-languages. However, it is not uncommon for a small portion of the produced documentation corpora to be manually transcribed and segmented. In these cases, it might then be interesting to use this annotation as a way of *informing* the UWS pipeline trained in bilingual settings.

In the past, we proposed to explicitly inject known segmentation into the bilingual models (Boito et al., 2017). Hypothesizing that a given number of types was known prior to training,[1] we segmented all occurrences of these

[1]This information could correspond to the lexicon a linguist is able to acquire after a

	#sentences	#types	#tokens	avg token length	avg tokens per sentence
Monolingual Set	1,000	2,159	5,934	4.20	5.93
Bilingual Set	4,130	5,812	24,622	4.18	5.96
All	5,130	6,633	30,556	4.18	5.96

Table 5.1: Statistics for the MB transcriptions for both sets, as well as for the totality of the corpus (All). The monolingual and bilingual sets have a type intersection of 1,338.

types, training networks in a mixed representation setting (characters and words in the decoder side). While we achieved a marginal performance increase in that setting, we find this option to be sub-optimal, as sub-word information is potentially lost.

Different from that, we now propose pretraining the NMT models from our pipeline on monolingual data, made of phones and their word-level transcriptions. This is a form of *warming up* the network. We hypothesize that a decoder trained in monolingual settings for a subset of the data is potentially better informed, and that it might perform better on the remainder of the corpus (bilingual setting).

We highlight that this experiment does not correspond to simply training a bilingual model using as starting point the monolingual ones presented in Chapter IV. This is because the monolingual models presented so far have access to transcriptions for the totality of the corpora, not corresponding to the real scenario of UWS. Instead, for this experiment we train monolingual models using only a fraction of the total dataset, in order to leverage a limited amount of monolingual supervision into the bilingual pipeline.

1.1 Experimental Setup

Dataset. As in the last chapter, we use the Mboshi-French (MB-FR) parallel corpus, randomly selecting 1,000 sentences for which we consider we have access to the transcription. We call this the *monolingual set*, while the other 4,130 sentences correspond to the *bilingual set*. For training, we maintain the data protocol from the last chapter, keeping 10% of the sentences for validation, and the rest for training. Table 5.1 presents some statistics for both sets.

few days of exchange with the local community.

Units (target)	<mono> phn16 phn35 phn26 phn16 phn27 phn16 phn49 phn31 phn47 phn30 phn35 phn8 phn35 phn6 phn55
Transcription (source)	bána bo báatúsá ambángé
Units (target)	<bi> phn16 phn35 phn26 phn16 phn27 phn16 phn49 phn31 phn47 phn30 phn35 phn8 phn35 phn6 phn55
Translation (source)	les enfants sont en train de cueillir les mangues

Figure 5.1: An example of the same target sequence, with its monolingual (<mono>) and bilingual (<bi>) aligned source information. Note that the tags are inserted at the decoder (target).

Training regime. We use our best NMT model (RNN), training each step for one third of the total number of epochs. We train[2] 5 runs using the MB-FR corpus, as in last chapter, and average the soft-alignment probability matrices obtained.

In preliminary experiments, we tried to apply a regular pretraining regime, training the bilingual network on top of the one trained with the monolingual subset. This however did not result in any benefit in the final model. We hypothesize that this happens because the monolingual subset is considerably smaller than the bilingual one.

We thus propose to train our models in three steps. First, we train the model using only the monolingual set (1st step), made of word transcriptions aligned to the unsegmented phones. Following this, the model is trained with a mixed input (2nd step), made of 1,000 sentences from the monolingual set, and the 4,130 remaining sentences with bilingual alignment only (bilingual set). Finally, in the 3rd step we remove all transcriptions, and the network is trained fully in bilingual settings. This includes training on the 1,000 sentences from the monolingual set, but replacing their transcriptions by their translations

Target side tags. We adapt our representation to include language tags in the target side (units), as in Johnson et al. (2017). This is necessary because the encoder annotations will vary by encoding transcriptions or translations. The tags in the target side are thus a way of better informing the decoder network of the type of source annotation it will attend to.

We use two language tags, <mono> and <bi>, for denoting unsegmented phones aligned to transcriptions and translations, respectively. These tags are added to the beginning of every sentence. In preliminary experiments, we noticed that including them increased our UWS scores. Figure 5.1 presents an example of the different supervision forms one sentence in the dataset can have.

[2]Other model settings, such as the training loss and hyperparameters, remain the same for all three steps.

	Boundary			Type		
	P	**R**	**F**	**P**	**R**	**F**
Base Model	72.3	75.9	74.0	21.6	28.8	24.7
1st step (monolingual)	-	-	-	29.4	17.7	22.1
2nd step (mixed)	**77.0**	**77.8**	**77.4**	**29.8**	**38.7**	**33.7**
3rd step (bilingual)	74.1	75.4	74.8	23.1	30.2	26.2

Table 5.2: UWS Precision (P), Recall (R), and F-score (F) for the discovered boundaries and types. The Base model corresponds to the RNN result obtained in Chapter IV. All segmentations are scored over the totality of the corpus (5,130 sentences), including type results for the 1st step.

1.2 Results

Table 5.2 presents UWS boundary and type discovery results computed over the totality of the bilingual corpus using the ZRC reference. Boundary results for the 1st step (1,000 sentences) are not reported, as they are not comparable.[3] Regarding type discovery scores for the 1st step, these correspond to 51.7 (P), 56.2 (R) and 53.8 (F) when scoring over the monolingual set only.

Looking at the results, we notice that the 2nd step achieves the highest boundary and type discovery scores, compared to other models trained on the full corpus (base and 3rd step). Then, by removing the monolingual information from this *warmer* model, the boundary and type scores go down. Even so, type scores for the 3rd model are superior compared to the base model. This hints that some of the pretraining information is still helping the model at this stage.

Regarding this decrease in boundary scores at the 3rd step, we experimented replacing it by a different model, which combined the totality of the bilingual corpus (5,130) and the monolingual set (1,000). Results for this network trained with 6,130 parallel sentences were not significantly different from the ones obtained using only the bilingual information (base).[4]

We believe this happens because once the network learns the alignments for the monolingual subset, adding their translations might lead to alignment

[3]While the vocabulary discovered by a subset can be compared against a larger one, directly comparing boundary scores generates an anomaly in precision scores.

[4]Evaluation settings were kept constant, scoring over the 5,130 sentences.

confusion. In this setting the same target sequence is aligned to different source information at different training steps. The same could also explain the performance drop between the 2nd and 3rd steps. We had hypothesized that the language tags would serve as enough guidance for the decoder network, but we might still be limited by the number of available examples.

About the training regime adopted, we experimented giving each step the totality of training steps from the base model, noticing that having too many epochs for the 1st and 2nd steps resulted in inferior UWS performance.[5] We also experimented removing the 1st step, and directly training the network with the mixed representation (2nd step), again noticing a significant decrease in performance.

In summary, our results suggest that it is beneficial to replace translations by their transcriptions, when these are available, and that it is possible to train our pipeline with this mixed representation. The vocabulary produced in this case (2nd step) seems to benefit from the monolingual supervision, while not being limited by it: it reached higher type recall scores than both monolingual (1st step) and bilingual models (3rd step).

2 Hybrid Bayesian-Neural Model

In the last chapter we showed that our best model for bilingual attention-based UWS did not surpass `dpseg`'s performance working with the true phones. However, although inferior in segmentation performance, our bilingual model has the advantage of incorporating annotations to the segmentation it produces. In this section we present a simple way of combining both approaches by creating a *hybrid* model which takes advantage of this Nonparametric (NP) Bayesian model's ability to correctly segment from small data while jointly producing translation alignments.

This investigation is inspired by the fact that several intermediate segmentations might be manually produced by linguists during language documentation. We then question if segmentation hypotheses, in this case represented by `dpseg`'s segmentation, could be included into our pipeline. In this scenario, a linguist could use the output of our model for validating their word hypotheses.

[5] We find that the NMT models get forgetful about the initial information (monolingual supervision) as we increase the training time.

2.1 Experimental Setup

Hybrid model. We inject `dpseg`'s segmentations into the unsegmented phone sequences, input of our NMT model. In this augmented input representation, illustrated in Figure 5.2, 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 phone sequence can be ignored by the NMT model, and new boundaries can be inserted as well. For instance, in Figure 5.2 `aintrat` becomes `a intrat` (boundary insertion), and `urat debine` becomes `uratdebine` (soft-boundary removal).

Training regime. The experimental protocol is the same from Section 1: we train 5 runs using the MB-FR corpus, and we average the produced soft-alignment probability matrices prior to segmentation. The soft-boundary tokens (`#`) are removed before UWS evaluation.

Syntactic Divergence. For understanding the impact of the soft-bounda-ries on the discovered soft-alignment probability matrices, we once again assess the relationship between type precision and the syntactic divergence of the sentences. We know from the last chapter that our model is more precise when segmenting sentences with low syntactic divergence. Now we want to investigate if including the soft-boundaries impacts this behavior. For this investigation we use the alignment complexity buckets produced in Chapter IV. An example of sentence pair for each bucket is presented in Figure 5.3.

2.2 Results

Table 5.3 presents UWS results for the RNN model (base) and `dpseg` from the last chapter, and for the proposed hybrid model. We notice that the hybrid model has a performance comparable to `dpseg` for boundary scores, and that it produces a better vocabulary (higher type discovery scores). This shows that the NMT model is learning to leverage the soft-boundaries from `dpseg` into the discovered alignment, instead of simply forcing a pre-established segmentation.

This information leveraging can be observed in the example of soft-alignment probability matrices produced by the hybrid model in Figure 5.3. There, some of the soft-boundaries (`#`) are ignored, with the phones next to them being aligned to the same translation word. We believe that the flexibility of not forcing a segmentation, and yet informing the model about possible boundaries, might be the reason why this setup successfully increased boundary and type discovery scores over the base model.

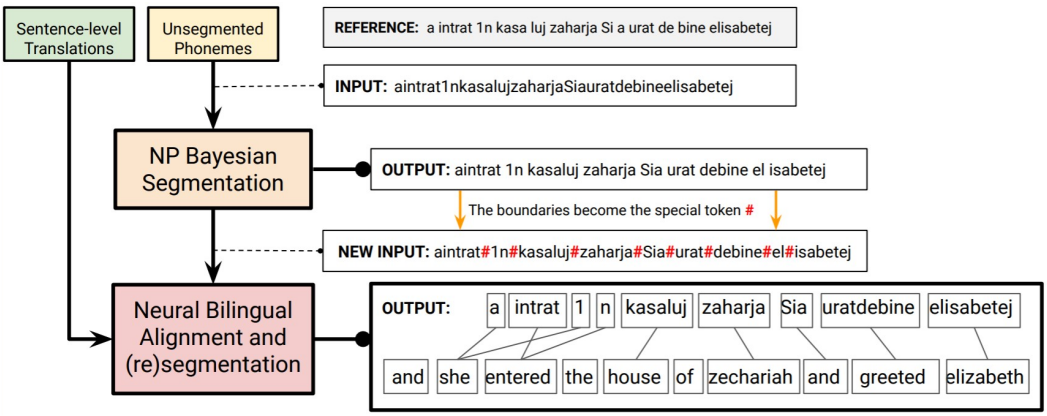


Figure 5.2: An illustration of the hybrid model using a sentence from the EN-RO language pair from the MaSS Corpus. The NP Bayesian model (`dpseg`) receives the unsegmented phonemes, producing segmentation. The discovered boundaries are then replaced by a special token (`#`), and bilingual alignment and re-segmentation are jointly performed.

	Boundary			Type		
	P	**R**	**F**	**P**	**R**	**F**
Base Model	72.3	75.9	74.0	22.1	31.0	25.8
`dpseg` **(unigram)**	71.9	**82.8**	**77.0**	21.1	30.0	24.8
Hybrid Model	**72.5**	81.1	76.5	**28.0**	**33.9**	**30.7**

Table 5.3: UWS boundary and type discovery scores for the RNN (base) and `dpseg` models from Chapter IV, and for the hybrid model.

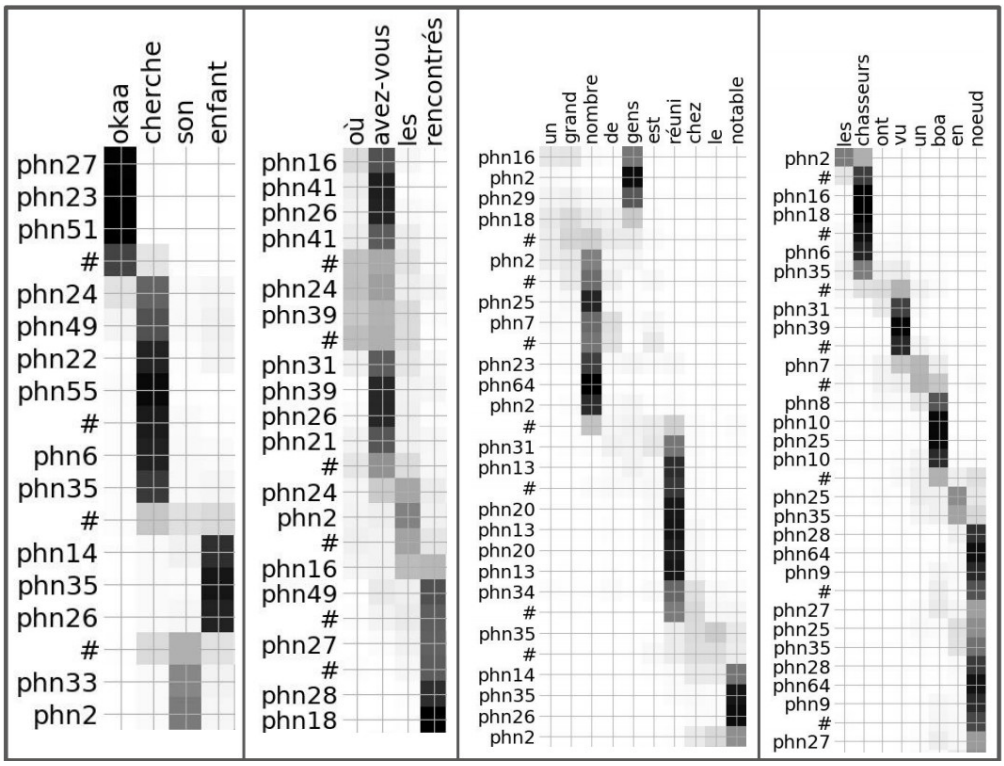


Figure 5.3: Soft-alignment probability matrix heatmaps for hybrid models trained on sentences from the MB-FR corpus. Darker squares correspond to higher pair alignment probability. The examples are ordered, from left to right, by alignment complexity buckets. The # is the soft-boundary symbol.

	#types	#tokens	avg token length	avg #tokens per sentence
Reference	6,633	30,556	4.2	6.0
`dpseg` **(unigram)**	2,343	37,458	2.5	7.3
Base Model	10,951	32,067	3.0	6.3
Hybrid Model	9,412	35,693	2.7	6.9

Table 5.4: General statistics for the produced segmentations.

Table 5.4 presents some statistics for the produced segmentations. We observe that all models, especially `dpseg`, over-segments the input, compared to the reference (higher number of tokens, and smaller average token length). Regarding vocabulary (number of types), we see that including the soft-boundaries helped our model reduce its size. There is a difference of 1.539 types between base and hybrid models.

Moreover, the neural approaches (base and hybrid) have a higher Type-to-Token Ratio (TTR), compared to `dpseg`. This means that in these models, types are not as often reused as it occurs in the NP Bayesian model. In fact, the `dpseg`'s implementation explicitly constrains the produced vocabulary, stimulating the reuse of the discovered units. In contrast to that, the neural models do not have any form of *global vision* over the produced alignments, and instead segmentation is produced at the sentence-level. This can result in excessively large lexicons being produced.

We now present type discovery precision for the syntactic buckets in Table 5.5, providing the difference between these and the scores obtained for the base model in Table 4.4, Chapter IV. We notice an expressive difference in type precision, compared to the base model. The augmented input representation seems to help this model especially for the intermediate buckets (2 and 3). This is interesting because it shows that the model gained capacity aligning more challenging sentence pairs.

In summary, we observed that we were able to successfully incorporate soft-boundaries into our attention-based UWS model, and that these resulted in a better vocabulary (Table 5.3 and 5.4), and capacity dealing with divergent sentence pairs (Table 5.5). However, this did not result in better boundary scores.

We also still need to investigate if this setup is feasible in the speech setting, working from discrete speech units from SD models. This is because,

ANE (<)	Bucket 1	Bucket 2	Bucket 3	Bucket 4	All buckets
0.2	82.0 (+13.3)	66.7 (+7.4)	69.0 (+12.5)	73.3 (+25.5)	64.5 (+15.5)
0.4	59.3 (+14.5)	55.4 (+14.1)	51.3 (+13.2)	45.6 (+13.8)	45.9 (+13.2)
0.6	47.9 (+9.6)	44.6 (+10.2)	40.8 (+10.2)	33.7 (+8.4)	32.7 (+8.0)
0.8	43.9 (+7.1)	40.7 (+8.2)	37.1 (+8.3)	29.2 (+6.4)	28.2 (+6.0)
1	43.7 (+7.0)	40.2 (+7.8)	36.9 (+8.1)	28.9 (+6.3)	28.0 (+6.0)

Table 5.5: Precision type discovery scores for the alignment complexity buckets, and for the totality of the corpus (All buckets), by using the matrices produced by the hybrid model. Results are cumulative and use the Alignment ANE thresholds indicated in the first column. The difference between the obtained scores and the ones from the base model (Table 4.4, Chapter IV) is displayed between parentheses. The buckets from 1 to 4 correspond to increasing alignment complexity scenarios (4 is the hardest).

the unsupervised discretization tends to be considerably longer than a manual phonetization, and then including soft-boundaries in that scenario might be too challenging in low-resource settings.

Lastly, we also applied this hybrid approach to the language pairs from the MaSS corpus, complementing the investigation of language impact presented in the last chapter. Our results showed that the target language affects the degree of *acceptance* of the soft-boundaries, with different languages having different degrees of overlap between the lexicon discovered by both `dpseg` and the hybrid model. Results are presented in Boito et al. (2020b) and in the Appendix B, Section 1.

3 Word-length biased NMT Training

In the last section we showed that the vocabulary produced by our model tends to be large. The lack of bias towards token length and reusability results in overly short or long tokens, fruit of, respectively, a very dispersed or clustered soft-alignment. Inspired by that, Godard et al. (2019) extended our attention-based UWS approach, proposing the constraining of the attention mechanism with a word-length bias during training. In this section we provide a comparison between their model and ours.

3.1 Model Definition

There are two differences from the neural model from Godard et al. (2019) to the one we use in this thesis. The first is the attention mechanism implementation (Equation 2.3, Chapter II), which they modify for including a bias towards longer words. They define attention over the source words as in Equation 5.1, where γ is a monotonically increasing function of the source word's length given by $|w_j|$. This modification is similar to the idea of *proportional segmentation* we proposed as a baseline in the last chapter: longer translation words should be aligned to more phone units than shorter words.

$$c_t = Att(H, s_{t-1}) = \sum_{j=1}^{|s|} \gamma(|w_j|)\alpha_{i,j}h_j \tag{5.1}$$

The second distinct feature from this model is the introduction of an auxiliary loss. Its goal is to control the number of words an alignment produces on the target side, encouraging it to become closer to the number of words in the source sentence. This is illustrated in Equation 5.2, where $|\boldsymbol{s}|$ and $|\boldsymbol{t}|$ are respectively the length of source (word-level) and target (phone-level) sentences. The last term sums over all target phones, resulting in a high value if there are few alignment shifts. This is because, if two consecutive phones i and $i+1$ are most strongly aligned to the same source word, then multiplying their alignment distributions $\alpha_{i,*}$ and $\alpha_{i+i,*}$ will result in a value close to one.

$$\mathcal{L}_{Aux}(\Omega|w) = ||\boldsymbol{t}| - |\boldsymbol{s}| - \sum_{i=1}^{|t|-1} \alpha_{i,*}^T \alpha_{i+1,*}| \tag{5.2}$$

3.2 Experimental Setup

We train the model from Godard et al. (2019), using their implementation and the MB-FR corpus. We follow the same experimental protocol from the last sections, training 5 runs per model, and averaging the soft-alignment probability matrices before scoring.

3.3 Results

Results for the base and for the word-length biased model from Godard et al. (2019) are presented in Table 5.6. We notice a slight performance gain using the proposed modification. Similar results were reported in Godard et al. (2019).

We believe that one possible drawback of the Godard et al. (2019) model is the over-constraining of the produced alignment: it forces the amount of words

	Boundary			Type		
	P	**R**	**F**	**P**	**R**	**F**
Base Model	72.3	**75.9**	74.0	22.1	**31.0**	25.8
Godard et al. (2019)	**78.2**	72.4	**75.2**	**24.0**	29.9	**26.6**

Table 5.6: UWS Boundary and Type scores for the RNN (Base) and word-length biased model.

	#types	**#tokens**	**avg token length**	**avg #tokens per sentence**
Reference	6,633	30,556	4.2	6.0
Base Model	10,951	32,067	3.0	6.3
Godard et al. (2019)	11,406	26,001	3.7	5.0

Table 5.7: General statistics for the produced segmentations.

produced to be close to the number of source translation words available, which potentially reduces the flexibility of the attention mechanism. For instance, in the third example in Figure 5.3, we see that some source words are almost completely ignored. As mentioned before, this might need to happen when source and target languages differ syntactically.

Regarding vocabulary, the statistics for the produced segmentations are presented in Table 5.7. We notice that the vocabulary for this word-length biased model is actually larger than than ours, and that the number of generated tokens is considerably smaller. This shows that, while forcing this source-to-target equivalence can help with over-segmentation (producing less tokens), it does not necessarily help reduce vocabulary size. This is because both models still suffer from a lack of constraining regarding the reuse of the discovered structures.

4 Multilingual Supervision for UWS

In the last chapter we trained bilingual models for the 8 languages from the MaSS dataset (56 bilingual pairs). During those experiments, we noticed that by choosing a different language for the supervision, the bilingual models

had the tendency to focus on different information. Thus, a natural extension would be to incorporate multilingual supervision during the training, for capturing these different structures using a single model. In this section we explore a multilingual model for our attention-based UWS approach.

There are multiple forms of creating a multilingual structure for NMT (Dabre et al., 2020). Here we focus on the *many-to-one* multilingual scenario, which corresponds to using one anchor language as target, training a single multilingual encoder structure made of different source languages (Johnson et al., 2017; Arivazhagan et al., 2019). The challenge of this type of approach lies in the network size, as the final model needs to have higher capacity than its bilingual equivalent.

4.1 Experimental Setup

Dataset. We use the MaSS dataset, training 8 multilingual models, which one with 7 source languages and one target language. This results in 36,638 parallel sentences.

Training Regime. We train the RNN NMT models, as in the previous sections, but with the multilingual setup from Johnson et al. (2017). That is, we add language tags in the decoder side, and we share the same encoder for all source languages.[6] For accommodating the larger vocabulary, we use the hyper-parameters from the 33k setup from the last chapter.[7]

Evaluation. We extract the soft-alignment probability matrices for every bilingual pair inside the multilingual model, scoring UWS in the symbolic domain, as in Section 4.1.3, Chapter IV. Therefore, we still produce bilingual segmentation, but using a NMT model trained with multilingual supervision.

4.2 Results

Table 5.8 presents our UWS results after bilingual (yellow, left) and multilingual (blue, right) NMT training. The former is included in order to facilitate comparison: results are the same from Chapter IV.

We notice that all our 8 multilingual models are worse than their bilingual counterparts. This trend is the same for BLEU scores, with all multilingual

[6]In preliminary experiments, we trained multi-encoder NMT models as well. We found worse results that we attribute to the larger number of parameters.

[7]We experimented with Byte-Pair Encoding (Sennrich et al., 2016) for source vocabulary reduction, but this resulted in worse UWS results. We hypothesize that increasing the source sequence length causes over-segmentation.

Bilingual									
Boundary F-score		EN	ES	EU	FI	FR	HU	RO	RU
	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	-

Multilingual									
Boundary F-score		EN	ES	EU	FI	FR	HU	RO	RU
	EN	-	48.9	35.2	48.5	**58.2**	47.0	51.4	48.4
	ES	55.3	-	35.8	46.0	56.3	46.5	50.8	46.8
	EU	50.9	45.3	-	44.5	50.2	44.8	47.7	44.2
	FI	56.8	47.4	**36.2**	-	54.1	**49.2**	50.4	**49.5**
	FR	**59.1**	**51.0**	35.4	47.7	-	47.2	**53.7**	48.5
	HU	53.5	47.2	35.9	48.5	51.9	-	49.5	48.4
	RO	57.0	49.0	35.4	47.3	58.0	47.2	-	48.5
	RU	56.4	47.9	35.8	**49.6**	54.6	48.8	51.4	-

Table 5.8: UWS Boundary F-score results for bilingual (left) and multilingual (right) UWS models. The bilingual results are the same from Table 4.8. Darker squares represent higher **column** scores. Best scores presented in bold. Better visualized in color.

models displaying lower translation capacity. This could be an indication that the number of languages and the size of the dataset used are incompatible with a multilingual setting, even for this *lighter* scenario in which all languages share the same encoder.

Another hypothesis would be that some of the languages are too dissimilar to share an encoder (e.g. RU and FR). Nonetheless, we also experimented with multilingual models with fewer source languages (from 3 to 6), and with models using only languages from the same language family (i.e. ES, FR and RO). Results in all cases were lower than the ones presented.

Moreover, we also investigated methods for combining multilingual supervision after training, merging the information learned by different bilingual models for the same target language. Our results, presented in Appendix B, Section 2, did not represent a clear improvement over the bilingual baseline.

Thus, we conclude that the use of multilingual information, especially in low-resource settings, is a difficult task, requiring a high degree of optimization and model expertise. Due to the negative results we obtained with the models presented in this section, we did not invest further in this direction.

5 Discussion

In this chapter we presented some extensions for our bilingual attention-based UWS pipeline. The first two models (monolingual pretraining and hybrid) focused on the case of extra supervision which could be leveraged during training. We also compared the word-length biased NMT model (Godard et al., 2019) to our base approach. Results for these three models were presented

		Boundary	Type
1	**Base Model (RNN)**	74.0	24.7
2	`dpseg` **(unigram)**	77.0	24.8
3	**Pre-trained Model (2nd step)***	***77.4***	*29.8*
4	**Hybrid Model**	76.5	**30.7**
5	**Word-length Biased Model**	75.2	26.6

∗ The model uses monolingual supervision.

Table 5.9: Boundary and Type UWS F-scores for base model (1), `dpseg` segmentation baseline (2), and the proposed model extensions (3-5).

in Boito et al. (2021). Lastly, we also presented some attempts to leverage multilingual information into our pipeline.

Focusing on the first three models presented, Table 5.9 presents a summary of their performance using the MB-FR corpus.[8] Looking at the assembled results, we notice that all modifications improved upon the base model, and some upon `dpseg`.

However, we highlight that although results for the pretrained model are the best ones in terms of boundary scores, this model uses monolingual supervision, whereas all other extensions depend on bilingual supervision only. Because of that, we find the hybrid model to be the most promising from the proposed extensions.

About this model, our general impression is that the gain in performance is due to the soft-boundaries helping the model to avoid under-segmentation. However, in this case, it is still unclear how dependent on the quality of the soft-boundaries (in terms of precision) the final model is. That is: *if the* ***dpseg*** *performance is not as good as the one presented, can its soft-boundaries still help the neural model?* In the next chapter we will address this research question.

Lastly, inspired by the notion that multiple translations could be a form of capturing deeper layers of meaning (Evans and Sasse, 2004), we also investigated the incorporation of multilingual supervision to our pipeline. Our

[8]We highlight that throughout this chapter we do not present Corpus ANE scores for the different models, as these are not comparable due to differences in: vocabulary (tag insertion for pretrained model and soft-boundaries for the hybrid model) and training procedure (attention mechanism for word-length biased model).

results, however, were not very encouraging, and we leave further exploration of this research branch as future work.

Chapter VI

Attention-based UWS for Speech

In this thesis we propose a bilingual attention-based pipeline for Unsupervised Word Segmentation (UWS) from speech, presented again in Figure 6.1. This pipeline has two steps: Speech Discretization (SD), and bilingual alignment and segmentation. So far, we focused on the second part of our pipeline, evaluating it using the true phones in the target languages (Chapters IV and V). This setting corresponds to a topline compared to using the discrete speech units generated by the unsupervised SD task.

Thus, after validating our approach working from the true phones, and in low-resource settings, we now focus on incorporating the SD step into the pipeline. Working from speech, we expect the input sequences for NMT training, which will come from the SD models, to have some noise. We are then interested in assessing how much our UWS performance deteriorates, in special against the robust `dpseg`.

This chapter is organized as follows. In Section 1 we study and compare the discrete speech units (also called *pseudo phones*) generated by five different SD models in low-resource settings. In Section 2 we then present results for our pipeline on its intended setting: bilingual attention-based UWS starting from a parallel corpus made of speech utterances and their textual translations. There, we compare our UWS results against the `dpseg` baseline and by using five different language pairs, investigating the quality and generalization capacity of the proposed approach. Finally, in Section 3 we study the use of the soft-boundaries from the `dpseg` model for increasing UWS results (hybrid model), and in Section 4 we discuss our results, concluding the chapter.

Lastly, our pipeline for bilingual attention-based UWS from speech was first presented in Godard et al. (2018c). There, we presented results using the RNN model from Chapter IV and the HMM SD model from Ondel et al. (2016). In this chapter, we revise and update that work, by including other four SD approaches, and five different languages.[1]

[1]The work presented in this chapter was submitted to Interspeech 2021, with the collaboration of Bolaji Yusuf and Lucas Ondel, from the Brno University of Technology. We thank them for all their help and expertise in SD models.

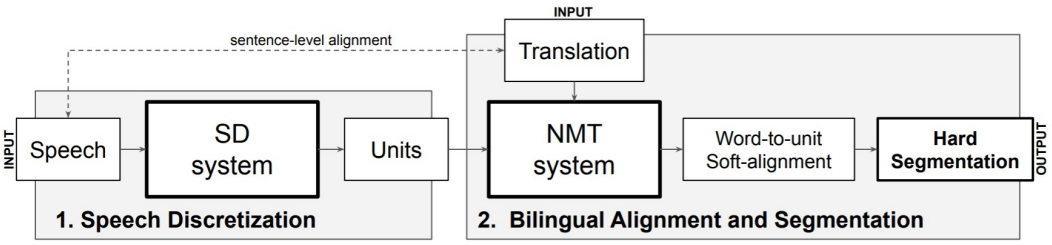


Figure 6.1: The general bilingual speech UWS pipeline. It requires as input a parallel dataset made of speech and sentence-level aligned translations. The system outputs word-level segmentation over the speech utterances. *Units* at the end of the first step correspond to the discrete speech units.

1 Comparing SD Approaches in Low-resource Settings

In Chapter II we described five models for SD. From these, three are Bayesian HMM-based approaches: **HMM** (Ondel et al., 2016), **SHMM** (Ondel et al., 2019) and **H-SHMM** (Yusuf et al., 2020). The other two are neural architectures inspired by Vector Quantization (VQ): **VQ-VAE** (van den Oord et al., 2017) and **VQ-WAV2VEC** (Baevski et al., 2020a).

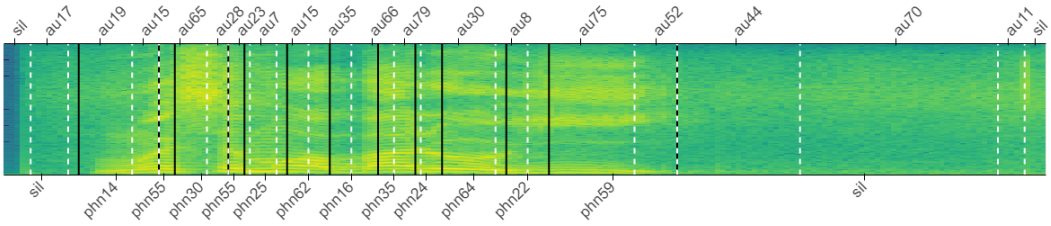
In this section we study the discrete representation produced by them using the Mboshi language. Section 1.1 explains the optimization and training regime for these different models, and Section 1.2 presents the generated discretization, discussing what makes them exploitable as a direct input for text-based UWS models.

1.1 Experimental Protocol

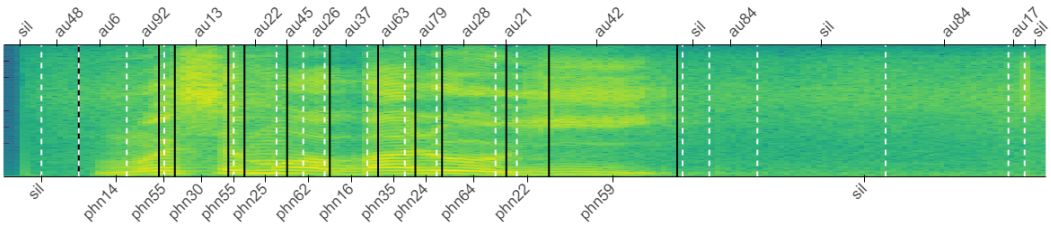
For the models presented in this chapter, the optimization is focused on the Mboshi-French parallel corpus, which is made of 5,130 utterances, corresponding to 4.28 hours of speech. Sections 1.1.1 and 1.1.2 present the training settings for Bayesian HMM-based and VQ-based models respectively. Section 1.1.3 explains the post-processing using silence labels.

1.1.1 Bayesian HMM-based Models

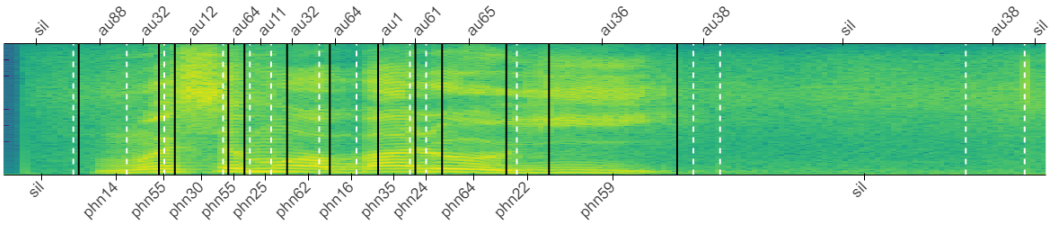
The Bayesian HMM-based models are trained with 4 Gaussians per HMM state, and using 100 for the Dirichlet process' truncation parameter. SHMM and H-SHMM use an embedding size of 100. For the H-SHMM models, this embedding is 7-dimensional (one per language).



(a) HMM: Discrete speech units (top), and reference (bottom).



(b) SHMM: Discrete speech units (top), and reference (bottom).



(c) H-SHMM: Discrete speech units (top), and reference (bottom).

Figure 6.2: Discrete speech unit segmentation for the same Mboshi utterance by each HMM-based SD system. The black lines denote the true boundaries, and the dashed white lines denote the discrete speech units' boundaries discovered by each system.

The subspace estimation for SHMM and H-SHMM uses the following languages: French, German, Spanish, Polish from the Globalphone corpus (Schultz et al., 2013), as well as Amharic (Abate et al., 2005), Swahili (Gelas et al., 2012) and Wolof (Gauthier et al., 2016) from the ALFFA project (Besacier et al., 2015). For each language, a subset of 2-3 hours is used, resulting in approximately 19 hours.

Further details for these three architectures are presented in the original papers (Ondel et al., 2016, 2019; Yusuf et al., 2020). The authors provided us with the trained models.[2] Figure 6.2 presents the representation produced by the HMM-based models for a given utterance, compared to the phonetic reference (true phones).

[2]Implementation available at: `https://github.com/beer-asr/beer/tree/master/recipes/hshmm`

1.1.2 VQ-based Models

For these models, we find that the direct application of their output to text-based UWS is challenging. This is because these self-supervised models tend to be quite inconsistent between consecutive predictions for the default 10 milliseconds window, and therefore the speech discretization produced for the utterances tends to be quite long in number of units.

These long unit sequences are then challenging to process and segment by both our attention-based approach and `dpseg`'s. For ours, it is because longer sequences are harder to cluster during bilingual alignment. For `dpseg`, it is due to an implementation hard limit for sequence size, which we were unable to circumvent. Because of that, our optimization focused in producing smaller sequences, sometimes in detriment to the size of the units vocabulary.[3]

VQ-VAE: The optimization of this model[4] for the Mboshi dataset was performed in Yusuf et al. (2020). The encoder is composed of 4 bidirectional LSTM layers, each with output dimension 128 followed by a 16-dimensional feed-forward decoder with one hidden layer. The number of discovered units (quantization centroids) is set to 50. This setting is unusually low, corresponding to less than a half of the standard value of 128, but this helps reduce the length of the generated sequences. Training is performed with Adam with an initial learning rate of 2×10^{-3}, which is halved whenever the loss stagnates for two training epochs. Finally, for the ℓ_2 losses, $k_1 = 2$ and $k_2 = 4$ are used.

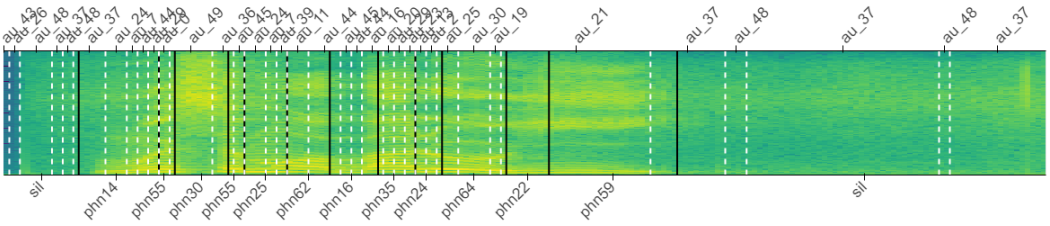
VQ-WAV2VEC: This model[5] was optimized starting from the settings provided for the *small model* in Baevski et al. (2020a), and using the EN 33k corpus from Chapter IV. The final model is trained on Mboshi, and it keeps the kernel sizes and strides from the original implementation, but uses only 64 channels, residual scale of 0.2, and warm-up of 10k. For vocabulary, we experimented having both 4 variables, resulting in 16 total units (V16), and 6, resulting in 36 units (V36). Larger vocabularies resulted in sequences that we were unable to apply to text-based UWS.

We also experimented reducing the representation by using Byte Pair Encoding (BPE) (Sennrich et al., 2016), hypothesizing that phones were being modeled by a combination of different units. In this setting, BPE serves as a method for identifying and clustering these patterns. Surprisingly, we found

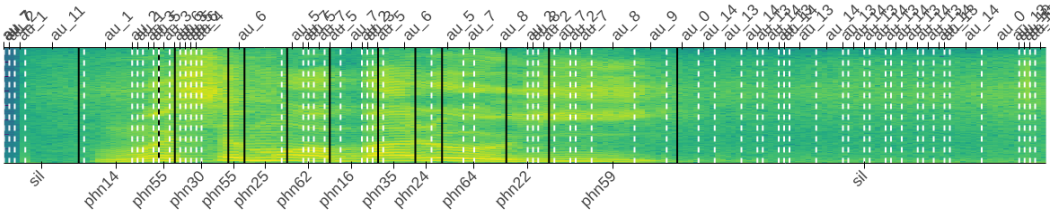
[3]Reducing the vocabulary is a way of forcing the model to be more consistent during prediction, as there are less options to choose from.

[4]Implementation available at: `https://github.com/BUTSpeechFIT/vq-aud`

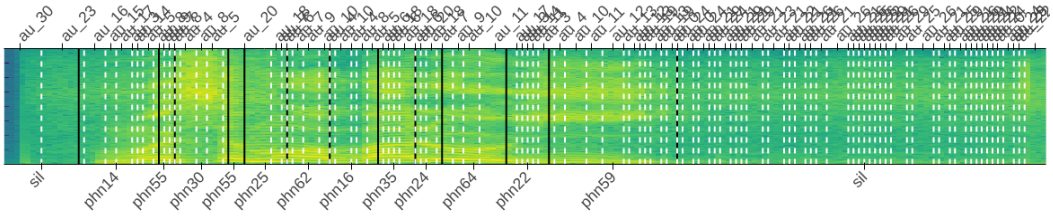
[5]Implementation available at: `https://github.com/pytorch/fairseq/tree/master/examples/wav2vec`



(a) VQ-VAE: Discrete speech units (top), and reference (bottom).



(b) VQ-WAV2VEC V=16: Discrete speech units (top), and reference (bottom).



(c) VQ-WAV2VEC V=36: Discrete speech units (top), and reference (bottom).

Figure 6.3: Discrete speech unit segmentation for the same Mboshi utterance by each VQ-based SD system. The black lines denote the true boundaries, and the dashed white lines denote the discrete speech units' boundaries discovered by each system.

that using BPE resulted in a decrease in UWS performance, which shows that the VQ-WAV2VEC model is not very consistent across utterances during labeling process.

Lastly, Figure 6.3 presents the representation produced by VQ-based models to a given utterance, compared to the phonetic reference. This figure is directly comparable with the example for HMM-based models (Figure 6.2).

1.1.3 Silence Post-processing

We experiment with reducing the representation by removing units predicted in silence windows according to the reference. This kind of annotation is inexpensive to obtain, and can be extracted from popular speech visualization tools such as *Praat* (Boersma, 2001). Moreover, this is an effective method

for reducing the length of the sequences from unsupervised models, letting us focus only on the units predicted at relevant segments, which correspond to true speech. Before UWS evaluation, the silence windows are reintroduced to ensure that their segmentation boundaries are taken into account.

1.1.4 Evaluation

We compare the discrete representation generated by the five SD models by focusing on two aspects: (1) their boundary recall over the words, and (2) their general statistics. Regarding (1), we decided to use the recall metric over the target words because, during our discretization process, we do not force our discretization to mimic the phonemes in the reference. For instance, one SD model can choose to use a sequence of units to represent one single phoneme, or describe the realization of two consecutive ones by one single unit.

In this setting, the most important thing is to reduce the cases where a word boundary is collapsed by the proposed discretization. This would result in noise for the UWS task, since some boundaries would be impossible to retrieve (loss of information).

Finally, regarding (2), the general statistics over the output representations will give us information about their degree of expressiveness (number of different units used to describe the utterances) and conciseness (average length of the sequences generated). We highlight that all evaluation is performed **after** merging consecutive 10ms windows that share the same unit prediction.

1.2 Resulting Representation

Table 6.1 presents the word boundary recall of the different representations. Table 6.2 summarizes the statistics for the obtained sequences.

For VQ-based models, we find that their very high boundary recall in Table 6.1 can be explained by the very long sequences that these models generate. This is because, by producing less clustered units, the probability of missing a boundary is smaller. Indeed, their average number of units per sequence are 3.4 (VQ-VAE) and 4.3 (VQ-WAV2VEC V16) times higher than the reference (Table 6.2). In this setting, adding the silence seems to reduce considerably the length of the sequences. Even so, comparing them against the HMM-based models, we see that the VQ-based sequences are not very concise, which might represent an issue in posterior UWS.

Regarding the HMM-based models, we observe that they are very concise, reaching a representation close to the reference even before silence post-processing (Table 6.2). Moreover, looking at Table 6.1 we see that this conciseness does not come at the cost of the word boundary recall, as they reach

	UNITS	UNITS + SIL
HMM	75.4	84.9
SHMM	82.5	90.5
H-SHMM	81.8	89.4
VQ-VAE	87.1	95.0
VQ-WAV2VEC V16	82.4	89.0
VQ-WAV2VEC V36	93.9	97.2

Table 6.1: Word Boundary Recall for the sequences generated by the 5 SD models before (left) and after (right) the silence post-processing. Results use the Mboshi utterances as input. For VQ-WAV2VEC, **VX** corresponds to the version of the model with a **V**ocabulary of **X** units.

	UNITS			UNITS + SIL		
	#units	avg # units per Sequence	max length	#units	avg #units per sequence	max length
HMM	77	27.5	83	75	20.9	69
SHMM	76	24.5	69	75	19.9	62
H-SHMM	49	21.7	63	47	19.4	60
VQ-VAE	50	65.2	217	50	43.4	143
VQ-WAV2VEC V16	16	81.7	289	16	52.6	229
VQ-WAV2VEC V36	36	111.0	361	36	76.2	271
REFERENCE				**68**	**18.8**	**51**

Table 6.2: Statistics for the produced discretization (unsegmented) using Mboshi utterances, and before (left) and after (right) the silence post-processing. For VQ-WAV2VEC, **VX** corresponds to the version of the model with a **V**ocabulary of **X** units.

high results (UNITS + SIL column). Finally, we observe a reduction in the number of units after post-processing (Table 6.2). This means that some units were modelling silence windows, even though these models already produce an independent token for silence.

In summary, we notice that all models produce an acceptable representation in terms of word boundary recall. This means that these models are not adding much noise into the posterior step of the pipeline. Regarding the statistics of the produced sequences, we notice that HMM-based models are more successful producing a representation close to the reference, while VQ-based models tend to produce longer sequences. In all cases, the silence post-processing positively affects the sequences by reducing their length.

2 Bilingual Attention-based UWS from Speech

In the last section we presented the application of five SD models to the Mboshi-French parallel corpus. That corresponds to the first step of the pipeline presented in Figure 6.1. We now use that generated representation for bilingual attention-based UWS (second step), using the settings for our best NMT model (RNN) from Chapter IV.[6]

Boundary F-score results for UWS models (ours and `dpseg`) trained using different discrete speech units, extracted from the Mboshi data, are presented in Table 6.3. We include results for both the direct output (RAW) and the post-processed version (+SIL). The RAW VQ-WAV2VEC V36 is not included as its average sequence length was excessively large for training our UWS models (Table 6.2).[7]

Looking at the results, we observe that in all cases post-processing the units with the silence information (+SIL) is beneficial for UWS, as it creates *easier* representations to learn from (higher scores for +SIL models). We believe this is due to the considerable reduction in the average length of the sequences (Table 6.2), as well as to the overall better phone boundary recall of these *filtered* representations (Table 6.1).

Focusing on the UWS models trained using the output of VQ-based SD models (rows 4-6), we see that the best result is achieved using the SD model with the smallest average sequence length (VQ-VAE). In general, we believe

[6] As in previous experiments, we train five of each model, averaging the soft-alignment probability matrices before UWS. Evaluation for all languages is performed on the speech domain, using the ZRC reference.

[7] An example of the output of the SD models using the Mboshi corpus is presented in the Appendix C, Figures 3.1 (HMM-based) and 3.2 (VQ-based).

		dpseg		Attention-based	
		RAW	**+SIL**	**RAW**	**+SIL**
1	**HMM**	32.4	59.9	**35.1**	**61.2**
2	**SHMM**	**43.7**	61.4	41.4	**64.7**
3	**H-SHMM**	**45.3**	61.4	44.8	**63.9**
4	**VQ-VAE**	**39.0**	52.7	32.1	**60.1**
5	**VQ-W2V-V16**	**37.4**	**52.2**	32.0	50.6
6	**VQ-W2V-V36**	-	48.0	-	**49.8**
7	**True Phones**	-	**77.1**	-	74.5

Table 6.3: Boundary F-scores results for the UWS models (`dpseg` and attention-based) using the SD models (1-6) and true phones (7, from Chapter IV), and applied to the Mboshi-French parallel corpus. Best results presented in bold.

that all VQ-based models under-perform due to the excessively long sequences produced, and that they are not a good choice for low-resource SD with the goal of direct application to text-based UWS. Regarding VQ-VAE, Chorowski et al. (2019) and Kamper and van Niekerk (2020) constrained its discretization mechanism, in order to produce a more concise representation. In Kamper and van Niekerk (2020), the constrained model was shown to be a better input for text-based UWS, compared to the standard VQ-VAE.

Overall, we find that UWS models trained using the discrete speech units from HMM-based models (rows 1-3) yield better results, in particular the SHMM and H-SHMM models. A noticeable difference between these two is the compression level: H-SHMM uses 27 less units than SHMM (Table 6.2).

Investigating the vocabulary discovered by these two approaches (type discovery recall results), we find that they scored 12.1% (SHMM) and 10.7% (H-SHMM), compared to the 31% reached by the topline model from Chapter IV. This illustrates that, even by using the best SD models, we still have a clear gap in comparison to using manual transcriptions. Moreover, we find that the SHMM models produced more types and less tokens, reaching a higher TTR (0.63) compared to H-SHMM (0.55).[8] This could be due to H-SHMM models having a smaller unit inventory.

Finally, focusing on the generalization of the presented SD models, we

[8]Statistics for the vocabulary generated by the SD models is presented in the Appendix C, Table 3.1.

		#types	#tokens	avg token length	avg #tokens per sentence	avg audio duration (s)
MB-FR	**MB**	6,633	30,556	4.2	6.0	4.28
	FR	5,162	42,715	4.4	8.3	-
MaSS	**FI**	12,088	70,226	6.0	13.2	8.19
	HU	12,993	69,755	5.9	13.1	7.57
	RO	6,795	84,613	4.5	15.9	8.08
	RU	10,624	67,176	6.2	12.6	8.06
	FR	7,226	94,527	4.1	17.8	-

Table 6.4: Statistics for the Mboshi-French (MB-FR) and MaSS datasets computed over the text (FR), or over the audio and phonetic representation (MB, FI, HU, RO and RU).

trained them using the languages from the MaSS dataset (Chapter III), with the same down-sampling from Chapter IV (5,324 utterances). We exclude English, French and Spanish, as these languages are present in the subspace prior from SHMM and H-SHMM models (Section 1.1). We also exclude Basque as the produced sequences were unfortunately too long for UWS training. Thus, the final language set is: Finnish (FI), Hungarian (HU), Romanian (RO) and Russian (RU). In all cases, the French translations are used as supervision for the attention-based UWS approach. Table 6.4 presents again the statistics for these languages, and for the Mboshi-French parallel corpus.

After training the MaSS models, we observed that due to the longer average duration of the utterances (Table 6.4), the VQ-based models produced sequences we were unable to directly apply to UWS training. This again highlights that these models need some constraining, or post-processing, in order to be directly exploitable for our task.

Focusing on the HMM-based models, which generated sequences directly exploitable for UWS, Table 6.5 presents UWS boundary results. We omit results for RAW, as we observe the same trend from the Mboshi results (Table 6.3). For the four languages, we again verify competitive results for SHMM and H-SHMM models, illustrating that these approaches generalize well to different languages.[9]

We also observe lower UWS results for the languages from MaSS dataset

[9]An example of the output of the SD models using the different languages from the MaSS dataset is presented in the Appendix C, Figures 3.3 (HMM), 3.4 (SHMM) and 3.4 (H-SHMM).

	dpseg				Attention-based			
	FI	**HU**	**RO**	**RU**	**FI**	**HU**	**RO**	**RU**
HMM	45.6	49.9	53.5	47.1	53.4	51.2	56.6	54.9
SHMM	49.0	52.3	53.5	50.5	56.0	**53.9**	57.7	**57.7**
H-SHMM	50.5	52.9	58.0	52.9	**56.1**	53.3	**59.6**	56.0
True Phones	**87.1**	**83.3**	**88.0**	**85.9**	68.4	63.4	75.7	68.4

Table 6.5: UWS Boundary F-scores for the MaSS dataset using HMM-based models (+SIL only) and true phones (Chapter IV). Best results for each language and SD model presented in bold.

(best result 59.6), compared to Mboshi (best result 64.7). We highlight that the data for the former comes from read text, and that the utterances correspond to verses, which can be considerably longer than sentences (see Table 6.4). Due to that, we consider it to be a more challenging setting for segmentation.

Lastly, focusing on the two UWS approaches (`dpseg` and ours), the UWS results over five languages show that our model produces better segmentation working from discrete speech units than `dpseg`, which in turn performs the best with the true phones (topline). The bilingual attention-based UWS models we proposed in this thesis have the advantage of their word-level aligned translations for grounding the segmentation process. We believe this might be attenuating the challenge of this task in this noisier scenario (longer sequences and larger phone vocabulary).

3 Hybrid Bayesian-Neural Model for Speech

In Chapter V we investigated extensions for increasing UWS scores by changing the NMT training, or by incrementing the input representation it receives. From the methods investigated, we obtained the best results by merging the output of the `dpseg` model into the input representation of our attention-based UWS approach. We called this the *hybrid model*.

We now investigate if this approach is also successful when working from discrete speech units, instead of the true phones. In this scenario, not only the sequences we have as input are longer, but the quality of the `dpseg` boundaries is also lower (see Table 6.3).

For this investigation, we focus on the Mboshi Language. Table 6.6 present

	dpseg			neural			hybrid		
	P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
HMM	52.8	69.1	59.9	**62.2**	60.3	**61.2**	43.5	**75.9**	55.3
SHMM	53.9	71.2	61.3	**68.5**	61.3	**64.7**	46.0	**77.5**	57.7
H-SHMM	55.5	68.8	61.4	**67.1**	61.1	**63.9**	47.7	**78.9**	59.4

Table 6.6: Precision (P), Recall (R), and F-score (F) boundary UWS results for the Mboshi-French parallel corpus using the HMM-based models (+SIL only). Best results presented in bold.

UWS results for the HMM-based SD models (+SIL only). We do not include results for VQ-based approaches, as the average sequence length in these cases is already elevated before including soft-boundaries (Table 6.2).

Looking at the results, we see that the hybrid approach under-performs in this noisier (true) UWS setting, reaching inferior performance due to over-segmentation (high recall, and low precision). We believe this happens due to two issues with this hybrid approach.

Firstly, for `dpseg` there is a considerable performance drop of 16.7 (F-score, Table 6.3) changing the representation from true to automatically generated phones. Thus, if `dpseg` was to serve as a proxy for assessing the insertion of high-quality segmentation information into the NMT training, this baseline is not a good fit anymore. In the last section, we showed that our attention-based model is competitive in this setting.

Secondly, throughout this chapter we discussed the challenge of treating long sequences. We attributed the success of HMM-based models in producing exploitable representations to the conciseness of their representations. Even though, the sequences are still longer than reference, and the resulting segmentation performance is inferior to the topline. In this setting, the addition of *more* information into the sequences in the form of soft-boundaries might be challenging to treat in low-resource settings.

4 Discussion

In this chapter, we investigated the first part of our UWS pipeline: SD models for producing discrete speech units from the speech signal. We compared five of these approaches: three Bayesian HMM-based models (Ondel et al., 2016, 2019; Yusuf et al., 2020), and two neural VQ-based models (van den Oord et al., 2017; Baevski et al., 2020a). In this comparison, our main goal was to

identify which model would produce the most exploitable representation. For us, an *exploitable* sequence from SD training needs to be concise, in order to be directly applied to text-based UWS models.

Comparing the SD models, we noticed that the VQ-based methods are not a good fit for our pipeline, as they output very long and inconsistent sequences, which are difficult to treat. This was also recently observed in Kamper and van Niekerk (2020).

Different from that, the HMM-based models output a good, yet concise, discrete representation, which we are able to successfully exploit for UWS. We believe this difference in performance is due to HMM-based models explicitly performing Acoustic Unit Discovery (AUD). This means the discretization produced by them aims not only to summarize the speech signal, but to correspond closely to the language's phonology.

Moreover, the subspace estimation performed by both SHMM and H-SHMM, might also play a significant role. This is because, these models are able to learn from an additional 19 hours of data in different languages. The other models (HMM and VQ-based models) do not have access to any form of pretraining or prior.

Regarding the UWS results obtained by applying the output of the SD models to the UWS task, we reached our best boundary results for Mboshi by using the SHMM and H-SHMM models. This same trend was also observed in four different languages from the MaSS dataset (FI, HU, RO, RU), verifying the generalization of the proposed pipeline.

Comparing our attention-based UWS approach against `dpseg`, we notice that we are very competitive in this setting, reaching better UWS boundary scores. This baseline is however better at segmenting true phones (topline scenario). About our approach, we also have the advantage of producing bilingual alignment as a form of grounding for the generated segmentation. In Chapter IV we showed that this grounding can be used for increasing type discovery.

Finally, in this chapter we also investigated applying the hybrid model from Chapter V to the true setting of UWS from speech. This model enriches the input representation for NMT training by using the `dpseg` output as *soft-boundaries*. We find that this model largely under-performs due to the degradation of `dpseg`'s performance in this noisier setting.

However, the motivation for this approach is to use `dpseg` as a proxy for assessing existing segmentations produced by a linguist. Therefore, we still believe that this method could potentially increase our UWS results if `dpseg` was to be replaced by annotations from a linguist, or a better UWS approach. Such an investigation is a suggestion for future work.

Conclusion

Chapter VII
Conclusion

Natural Language Processing is a very popular research domain, but language technology tends to be developed mostly in and for a very small portion of the existing languages in the world. These languages, the so-called high-resource, are used for proposing and testing approaches. In this naive approach, all languages are considered to be equal (to model and to learn from) as long as there is enough data to train data intensive machine learning approaches.

However, for many languages there is not, and there will probably never be, *enough* data. This is especially the case of minority dialects, which are not considered economically interesting for justifying the investment necessary for data gathering. Moreover, the evergrowing globalization indirectly pushes humanity towards a standardization of spoken languages. The result of this is the estimation that many (if not most) of existing languages will vanish within this century (Austin and Sallabank, 2011).

Meanwhile, *zero resource approaches* became popular in recent years, as they propose to reach the long tail of existing low-resource languages by proposing approaches adapted to settings with less data. In this context we highlight the need for not only developing with less, but the importance of using diverse data. Only by doing that can we truly test and understand the applicability of the methods we propose.

Moreover, there is a recent criticism about the meaning of this *zero* in zero resource approaches (Bird, 2020). Languages rarely exist in complete isolation, and rare are the ones with no existing lexicon or any initial or rudimentary documentation. The absence of interest in leveraging this information when proposing approaches can result in the products having marginal to no impact for the community of speakers.

Therefore, although the technological challenge of extracting knowledge with close to no information is attractive to scientists, if they aim to propose approaches for *computational language documentation*, they should collaborate with language experts and with the community. This way, they are sure to produce *for* the community, and not simply *from* their data.

In this thesis we investigated the task of Unsupervised Word Segmentation in the context of language documentation. Our main goal was to avoid the need for transcriptions, as these are known to be generally not available (Adda et al., 2016; Brinckmann, 2009).

Instead, we focused on segmenting audio into word segments using only a few hours of speech and by grounding this process in aligned annotations (translations). Our final segmentation is applied to the speech signal, accompanied with annotations in the form of potential translations. Our hope would be for this annotation to be useful for reviewing word candidates, and potentially even for building a bilingual lexicon of speech segments.

We now discuss in detail the contributions of this work (Section 1), as well as some limitations of the proposed approach (Section 2). Section 3 covers perspectives and directions for future work.

1 Contributions

This thesis proposed a pipeline approach for UWS in the speech domain. This approach grounded segmentation in translation words, and solved segmentation by using the soft-alignment produced by NMT models. Before alignment and segmentation, SD is performed in order to accommodate the challenge of low-resource speech processing. We now recapitulate the contributions listed in the introduction, elaborating on each topic.

C1: ***A thorough comparison of recent SD approaches for low-resource speech processing, focusing on their direct applicability to text-based UWS.***

The goal of SD models is to produce a sequence of discrete speech units from input utterances, without the use of any transcription. In Chapter VI we compared five of these models. Three of them were from the Bayesian HMM family: HMM (Ondel et al., 2016), SHMM (Ondel et al., 2019), H-SHMM (Yusuf et al., 2020), and the other two models were recent neural approaches based in Vector Quantizing: VQ-VAE (van den Oord et al., 2017) and VQ-WAV2VEC (Baevski et al., 2020a).

We optimized and trained these models in low-resource settings using five languages, assessing the quality of the produced discrete speech units. Our focus was the direct application to the text-based UWS approach.

Our results showed that the HMM-based models produced a concise output, close to the reference. For VQ-based models, we observed a very inconsistent speech labeling process, resulting in sequences which were challenging to apply to our task. The most exploitable SD models for UWS were the SHMM and HSHMM models. This work was submitted to Interspeech 2021.

C2: ***A study of the direct interpretability of the attention mechanism in NMT models, and in low-resource settings.***

In Chapter IV we investigated the use of the soft-alignment probability matrices obtained through NMT training for aligning translation words to an unsegmented sequence of phones. This soft-alignment, which is produced by the attention mechanism, is then used for clustering neighbor phones which share word alignment. We refer to this as *attention-based* UWS.

In order to assess the feasibility of this approach in low-resource settings, we compared three different attention-based NMT models: RNN (Bahdanau et al., 2015), 2D-CNN (Elbayad et al., 2018), and Transformer (Vaswani et al., 2017). We found the following ranking for the exploitability of these models, from best to worst: RNN, 2D-CNN, Transformer. Our results also showed that the soft-alignment discovered by the attention mechanism is still exploitable when the NMT is trained with only 5k sentences. We obtained our best segmentation results by using the *simple* RNN model, and our worst results by using the Transformer architecture. This work was presented in Boito et al. (2019a), and extended to a journal format in Boito et al. (2021).

C3: ***A comparison between UWS approaches: our attention-based model and two baselines.***

In this work we compared our attention-based UWS approach to two baselines in realistic settings (Godard et al., 2018c; Boito et al., 2019a, 2020b). We use only 5k sentences in the Mboshi language, and we include results in eight more languages: English, Spanish, Basque, Finnish, French, Hungarian, Romanian and Russian.

The first baseline is the proportional bilingual model. It allows us to assess the challenge of our alignment task. It is a naive approach which produces diagonal alignment, clustering units considering the length of translation words. As expected, our results with all the 56 language pairs from the MaSS dataset showed that this naive approach under-performs (Chapter IV). This illustrates that the bilingual segmentation task we target in this work is not trivial.

The second baseline is the model from Goldwater et al. (2009), which we refer as the `dpseg`. We find it to be very competitive: when working from the true phones of the language, this baseline was the one which produced the best segmentation results (Chapter IV). However, as we move to a more challenging scenario, where the input is noisier (in terms of consistency, length and vocabulary size), this baseline performed below or on par with our attention-based approach (Chapter VI). We believe this highlights how the grounding in bilingual information can help the discovery process in challenging settings.

C4: ***The investigation of language-related impact in our pipeline.***

Throughout this work we used diverse languages in order to assess the generality of the proposed pipeline. For assessing language-related impact using these different languages, there are two aspects to consider. The first one is the natural discrepancy that happens in unsupervised methods when segmenting different languages, as languages are not all equally hard to segment (Fourtassi et al., 2013). The second aspect is the impact of the bilingual grounding that exists in our pipeline, which guides segmentation through the words in the translation.

Focusing on the first aspect, in Chapter IV, we used the MaSS dataset for generating 56 language pairs from its eight languages, which we used for training our bilingual UWS models. Our results, published in Boito et al. (2020b), showed a clear gap in performance between models with different languages as target of segmentation.

Regarding the second aspect, we ranked the languages used for grounding with regards to the segmentation performance obtained for each of the eight target languages. We found that, although the final language ranking obtained seemed to be rooted in linguistic features, the impact of statistical features was non-negligible. This is because statistics such as vocabulary size and type-token ratio can impact the ability of the neural model to encode the input information. Thus, by having more favorable statistics (easier to learn in low-resource settings), some languages were superior as supervision for segmenting even unrelated languages.

C5: ***The proposal of pipeline extensions for incorporating extra information into the segmentation model.***

In Chapter V we proposed two methods for including extra knowledge in our models. The first one was the pre-training of the NMT models with a small portion of the transcriptions. This was motivated by the possible existence of these, produced by linguists during data collection. In this setting, after pre-training in this small portion of manually transcribed data, the NMT model is trained on the full bilingual dataset. Our results showed that this pre-training is helpful, increasing both boundary and type discovery.

We also proposed a hybrid model, which used the segmentation produced by `dpseg` to enrich the input sequences we have in our NMT model. These *soft-boundaries* seemed to inform our models, increasing segmentation results. The goal of this model was to assess the incorporation of word-hypotheses by linguists into the model. In this setting, the linguist could study the output of our model for validating existing hypotheses. Unfortunately, this model did not work in noisy settings. Both models were presented in Boito et al. (2021).

C6: ***The gathering and publishing of three datasets useful for low-resource and computational language documentation approaches.***

In Chapter III we presented the following datasets: Mboshi-French Parallel Corpus (Godard et al., 2018a), Griko-Italian Parallel Corpus (Boito et al., 2018), and MaSS Multilingual Dataset (Boito et al., 2020a). The **Mboshi-French parallel corpus** has been widely exploited for evaluating approaches in low-resource speech processing and language documentation (Anastasopoulos and Chiang, 2018a,b; Bansal et al., 2019; Sung et al., 2019; Inaguma et al., 2019; Scharenborg et al., 2018, 2020; Ondel et al., 2019; Yusuf et al., 2020; Godard et al., 2018b, 2019). The **Griko-Italian parallel corpus** is an interesting example of *extreme* low-resource scenario, being interesting for zero-shot learning approaches (Wada et al., 2020). Finally, the **MaSS multilingual dataset** has been mentioned by the community as an example of a dataset for studying diverse language pairs, helping attenuate the *English-centered* nature of current speech approaches.

In this thesis we used the Mboshi-French parallel corpus as our main target of study (Chapters IV to VI). We also used a down-sampled version of the MaSS dataset in order to investigate language impact (Chapters IV and VI). Results for the Griko-Italian parallel corpus were not presented here, as we found that this corpus was too small for NMT training.

2 Limitations

In this work we proposed a pipeline for solving UWS from speech in low-resource settings. The first limitation with such an approach is its pipeline structure. The lack of interaction between the process of speech discretization and segmentation means that errors in the former are propagated to the latter. Moreover, one can imagine that by grounding the discovery of units in their posterior usefulness for creating word-segments, a more robust model could be built.

A second limitation of our model is the data dependency in neural approaches. We were able to successfully train models using only 5,130 sentences however, in preliminary studies we failed to do the same for the small Griko-Italian dataset, made of only 330 sentences. In such limited settings, our model was largely inferior to the monolingual baseline `dpseg` working with the true phones. From discrete speech units, both models (ours and `dpseg`) failed to produce anything exploitable (Boito et al., 2018). This hints to the existence of a data threshold for the applicability of UWS models.

Moreover, it is notable that our model is constrained on the existence of aligned word translations. This means we cannot apply our pipeline for seg-

menting monolingual data from documentation initiatives. The motivation of our approach was exactly to propose something grounded on bilingual information, not covering the case of its absence. A recent monolingual neural model for UWS was proposed in Kawakami et al. (2019), but this model was deeply rooted in the characters representation, and it would need modifications for working from the output of SD models.

Finally, another limitation of our approach rests on the alignment procedure: the use of soft-alignment probability matrices from NMT training. These matrices are a by-product of translation, which means NMT models do not explicitly consider alignment in their optimization. Models such as Alkhouli et al. (2018) and Garg et al. (2019) focus on the joint optimization of translation and word-to-word alignment. However, their task is expected to be easier than our many-to-one units-to-words alignment. Godard et al. (2019) performed explicit optimization for attention-based UWS, finding marginal performance gain. This hints that a more sophisticated optimization for discovering word segments might be needed.

3 Future Work

We now summarize some research directions for the work presented in this thesis.

Clustering of Alignment Pairs. In Chapter IV we showed that we can use the Alignment ANE metric for filtering the structures discovered by the NMT model. This was based on the idea that low ANE scores correspond to alignments the network is *confident* about. By doing this, we were able to increase type retrieval scores for Mboshi and English.

One topic we wanted to investigate is the clustering of the discovered alignment pairs. That would mean putting together all the discovered types aligned to the same translation. From there, we wanted to have an agreement over the chosen type for that translation. This would make the model output less types, and it would remedy the *local* vision that our models have.[1]

In Table 7.1 we illustrate two examples from the ensemble of translation clusters we found for the EN-FR 5k model from Chapter IV. We can see that in both cases the correct segmentation is present, but in the middle of other type candidates which have extra or missing phones.

[1]We say that our models are *local* because the segmentation is based on sentence-level alignment, and there is no posterior analysis for assessing the generality of the alignment discovered for the remainder of the corpus.

Interpretation	Sequence	Translation	ANE	Interpretation	Sequence	Translation	ANE
Extra "W"	**MER1DER0ER0***W*	assassin	0.28	Missing "F"	**UW1D**	nourriture	0.37
Correct	**MER1DER0ER0**	assassin	0.31	**Correct**	**FUW1D**	nourriture	0.38
Missing "M" Extra "F"	**ER1DER0ER0***F*	assassin	0.32	Extra "IH0"	**FUW1D***IH0*	nourriture	0.47

Table 7.1: Examples of two translation word clusters, and the discovered types within. Extracted from the EN-FR 5k model from Chapter IV. The cluster on the left is *assassin* (murderer), the one on the right is *nourriture* (food). For ANE scores, lower is better.

Our idea would be then to propose something which would consider the ANE scores of the segments, and the phones present in each candidate. For instance, in the example in the left, we see that 100% of the candidates of that cluster found `ER1DER0ER0`, and 66% of them found the initial `M`. Moreover, as the latter were also the candidates with the lowest ANE scores (more confident alignments), we would then like to produce `MER1DER0ER0` (the correct type).

Finally, in this direction we would also like to explore clustering these groups in order to create *word sense clusters*. These would present discovered types which are expected to have a similar sense, considering the discovered alignment. This information could be useful for disambiguating the generated segmentation, or even for building bilingual lexicons.

Leveraging Knowledge into the Models. In Chapter V we showed some extensions for including extra information into our model. Moreover, during a previous work we also investigated the inclusion of segmented types directly into the NMT model (Boito et al., 2017). An interesting direction would be the introduction of this knowledge directly into the attention mechanism, such as in Alkhouli et al. (2018), in which the authors included an extra head into Transformer models with dictionary information.

End-to-end UWS from Speech. As mentioned in Section 2, a natural extension of this work would be the development of end-to-end UWS models from speech. Since speech processing is challenging in low-resource settings, we imagine that the combination of the HMM-based models for SD (Ondel et al., 2019; Yusuf et al., 2020) with `dpseg` could be an option for monolingual UWS from speech.

A different direction would be to ignore the process of discretization altogether, directly aligning speech with translation words. In this research direction, attention-based speech translation models (Besacier et al., 2006; Bérard

et al., 2016; Tjandra et al., 2019; Zhang et al., 2020) would be a possible solution, if these were to be successfully trained with such restricted amounts of data. Even then, it is unknown how exploitable the attention mechanism would be.

Visually Grounded Models for UWS. Going beyond language documentation, the core idea of our proposal is the grounding of word discovery in aligned information (translations). If we were to replace these translations by images or videos, we would then reach *visually grounded* models for UWS from speech. These models are an interesting source of investigation, as this visual grounding is something that naturally happens in children during language acquisition (Chrupała et al., 2017).

Part I

Appendix

Appendix A

Experiments from Chapter IV

		EN 33k	EN 5k	MB 5k
Transformer	Monolingual	3,497,728	739,712	607,616
	Bilingual	5,030,656	948,096	526,208
2D CNN	Monolingual	1,780,553	917,449	786,568
	Bilingual	3,060,553	1,126,089	704,904
RNN	Monolingual	3,370,000	700,000	570,000
	Bilingual	4,530,000	910,000	490,000

Table 1.1: Number of trainable parameters inside the models trained on different datasets (English (EN) 33k and 5k, Mboshi (MB) 5k) for both monolingual and bilingual settings from Section 3, Chapter IV. The amount of trainable parameters depend on the vocabulary size, due to the embedding layer and the softmax projection inside the decoder network.

TYPES PRECISION								
	EN	ES	EU	FI	FR	HU	RO	RU
EN	-	13.5	10.6	15.5	**17.7**	13.0	15.6	14.9
ES	12.7	-	10.1	13.5	15.5	12.0	13.4	12.7
EU	7.7	8.2	-	12.2	8.0	9.7	8.8	10.0
FI	11.2	9.8	9.7	-	9.9	13.9	11.1	**15.8**
FR	**18.1**	**17.4**	10.4	17.9	-	12.3	**19.3**	15.5
HU	8.7	9.3	9.2	17.1	8.4	-	10.1	13.4
RO	13.5	13.1	**10.6**	16.7	16.8	12.7	-	15.2
RU	11.0	9.9	9.3	**19.1**	10.5	**14.5**	12.5	-

TYPES RECALL								
	EN	ES	EU	FI	FR	HU	RO	RU
EN	-	29.9	17.5	23.8	47.9	20.7	41.9	27.3
ES	49.3	-	22.0	25.4	**50.1**	21.6	40.9	27.5
EU	36.6	22.6	-	24.8	34.0	19.5	32.1	22.6
FI	48.3	29.2	**24.0**	-	40.5	27.9	39.9	**35.8**
FR	**58.5**	**38.4**	19.7	29.5	-	20.5	**50.6**	30.0
HU	42.2	29.1	23.8	**37.4**	36.4	-	37.4	31.8
RO	48.4	31.3	21.4	28.8	49.8	21.6	-	30.4
RU	47.5	28.2	21.2	37.4	41.5	**28.0**	41.4	-

TYPES F-SCORE								
	EN	ES	EU	FI	FR	HU	RO	RU
EN	-	18.6	13.2	18.8	**25.8**	16.0	22.8	19.3
ES	20.2	-	13.9	17.6	23.7	15.4	20.2	17.4
EU	12.8	12.0	-	16.4	13.0	12.9	13.9	13.9
FI	18.2	14.7	13.8	-	15.9	18.5	17.4	**21.9**
FR	**27.7**	**23.9**	13.6	22.3	-	15.4	**28.0**	20.4
HU	14.5	14.0	13.3	23.5	13.6	-	15.8	18.9
RO	21.1	18.4	**14.2**	21.1	25.1	16.0	-	20.2
RU	17.9	14.6	13.0	**25.3**	16.8	**19.1**	19.2	-

Table 1.2: Type discovery precision, recall and F-score results for the bilingual models from Section 4, Chapter IV. The columns represent the target of the segmentation, while the rows represented the translation language used. Darker squares represent higher **column** scores. Best scores presented in bold.

	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 1.3: UWS Boundary F-score results for the proportional baseline. The columns represent the target of the segmentation, while the rows represented the translation language used. Darker squares represent higher **column** scores. Best scores presented in bold. Better visualized in color.

1 Investigating Language Impact: Bilingual Baseline Comparison

The results in Table 4.8 of Chapter IV confirm that there is an impact related to using different source languages for generating the segmentations. We identify interesting language pairs emerging as the most efficient, such as FI>HU (Uralic Family), FR>RO and FR>ES (Romance family).

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 phone sequences into words, having more or less translation words could be a determining aspect in the bilingual segmentation performed.

For this investigation, we use a naive bilingual baseline called proportional, introduced by us in Godard et al. (2018c). It performs segmentation by distributing phones equally between the words of the aligned translation, ensuring that words that have more letters, receive more phones (hence *proportional*). Results for the proportional baseline are presented in Table 1.3. The average difference between the best UWS segmentation (Table 4.8) and proportional (Table 1.3) results for the languages is 19.4 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 our bilingual-rooted segmentation and the 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 (0.71) is observed.

Our interpretation is that the token ratio between the number of tokens in source and target sentences have a significant impact on bilingual UWS. However, this ratio 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 suboptimal models.

Appendix B

Experiments from Chapter V

1 Hybrid Model for the MaSS Corpus

		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 2.1: UWS 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 **column** scores. Best results in bold. Better visualized in color.

Table 2.1 presents results for the base (neural) and hybrid models. Looking at the hybrid results, we verify that these models always 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 2.2 presents the intersection between the correct types discovered by both `dpseg` and hybrid models. Results show that while the monolingual

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

baseline `dpseg` *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, hinting that some languages might *accept* more easily the *soft-boundaries* proposed by `dpseg`.

2 Multilingual Selection

Multilingual training is not the only form of including multilingual supervision for generating segmentation. Since we generate soft-alignment probability matrices for all bilingual models, we also investigated combining the information present in these different matrices. This considers that, if the information from different languages aligned to the same speech utterance captures different optimal correspondences between source and target, their combination could lead to improved UWS. We studied two approaches for accomplishing this, which we detail below.

Multilingual Voting: This approach generates agreement over the boundaries inserted by different bilingual models, by selecting the number of models (languages), and an agreement threshold T. This threshold balances between accepting all the generated boundaries (zero agreement) and accepting only boundaries discovered by all systems (100% agreement). Values between these two extremes shed light on the different information learned and the utility of using more than one bilingual model for generating segmentation.[1]

[1] We experiment with $T \in [0, 0.25, 0.5, 0.75, 1]$.

Number of Languages	Threshold (best)	Voting	ANE selection
2	1.0	57.26	**64.26**
3	0.5	**59.65**	63.91
4	0.5	57.89	63.90
5	0.5	59.16	63.65
6	0.5	58.35	63.33
7	0.5	56.58	63.33

Table 2.3: Boundary UWS F-scores for the multilingual selection approaches applied to the RO language, from Table 5.8. The bilingual baseline scored 62.8.

Multilingual ANE Selection: In the last chapter we introduced ANE as a metric for assessing the quality of the produced soft-alignment discovered by bilingual models. Since Sentence ANE gives us a score for the quality of the bilingual alignment for a given sentence in the dataset, we can use this as a criteria for selecting matrices generated by different bilingual models, generating thus a multilingual set of alignments, from which we derive segmentation. In simple terms, using the set of bilingual models available for a target language, we perform selection by minimizing the Sentence ANE.

Results: For these experiments, we use the bilingual models trained for segmenting the RO language. Table 2.3 presents results for multilingual voting and ANE selection. In both cases, we start from the bilingual model using the best supervision language from Table 5.8 (RO column), which scored 62.8 (FR>RO). We then add languages accordingly to the obtained performance ranking in bilingual settings (respectively EN, RU, ES, FI, HU, EU).

Looking at the results in Table 2.3, we see that ANE Selection improves upon the best bilingual model (first row). However, adding more languages seems to be detrimental to the performance. We also experimented normalizing the ANE scores for every bilingual model, which resulted in even lower scores. We believe this happens because this normalization favors low-scoring models that otherwise would not be chosen very often.

In summary, the results for multilingual ANE selection suggests that *some* multilingual supervision is beneficial, but that simply selecting from the models without an explicit *weighting* system[2] might be detrimental to the perfor-

[2]We suppose, for instance, that similar languages should be favored when generating

mance when increasing the number of languages.

Focusing on the multilingual voting approach, we see that the results are worse than the ones observed for ANE selection. Moreover, the agreement over the languages for achieving the best results for each multilingual setting is low, being only 50% in most of the cases. Also, the results seem unstable, as we yield the best UWS results using three languages, and the second best using five. This also supports our conclusion that some form of explicit weighting would be necessary in order to filter the information being injected into these multilingual segmentation models.

segmentation.

Appendix C

Experiments from Chapter VI

	#types	#tokens	avg token length	max token length	avg # tokens per sentence	TTR
VQ-VAE	21,307	37,230	6.0	33	7.3	0.57
VQ-WAV2VEC V16	13,790	71,508	3.8	42	13.9	0.19
VQ-WAV2VEC V36	26,053	82,812	4.7	72	16.1	0.31
HMM	15,162	28,468	3.8	21	5.5	0.53
SHMM	16,017	25,534	4.0	14	5.0	0.63
H-SHMM	14,606	26,418	3.8	15	5.1	0.55
Reference	6,633	30,556	4.2	19	6.0	0.22

Table 3.1: Statistics for the segmentation produced by our UWS models for the Mboshi corpus, and by using the different SD approaches. TTR corresponds to Type-Token Ratio.

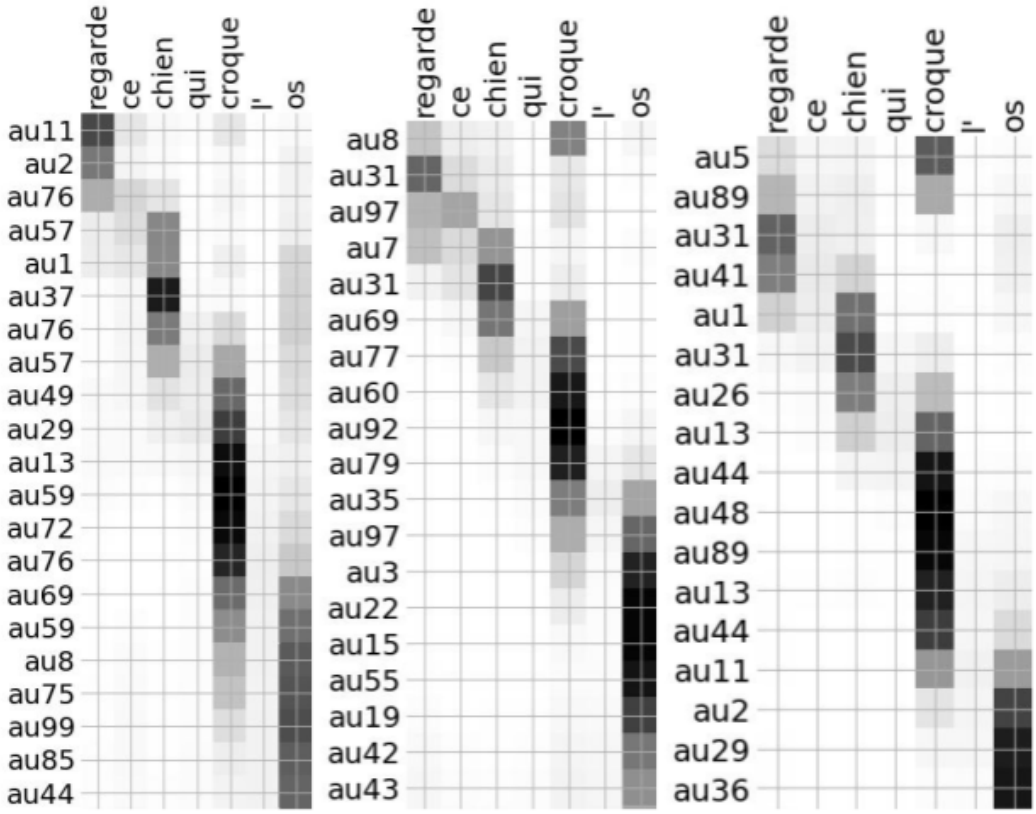


Figure 3.1: The soft-alignment probability matrices produced for the same sentence pair, but using different HMM-based SD approaches: HMM (left), SHMM (middle) and H-SHMM (right). Darker squares correspond to higher soft-alignment probabilities.

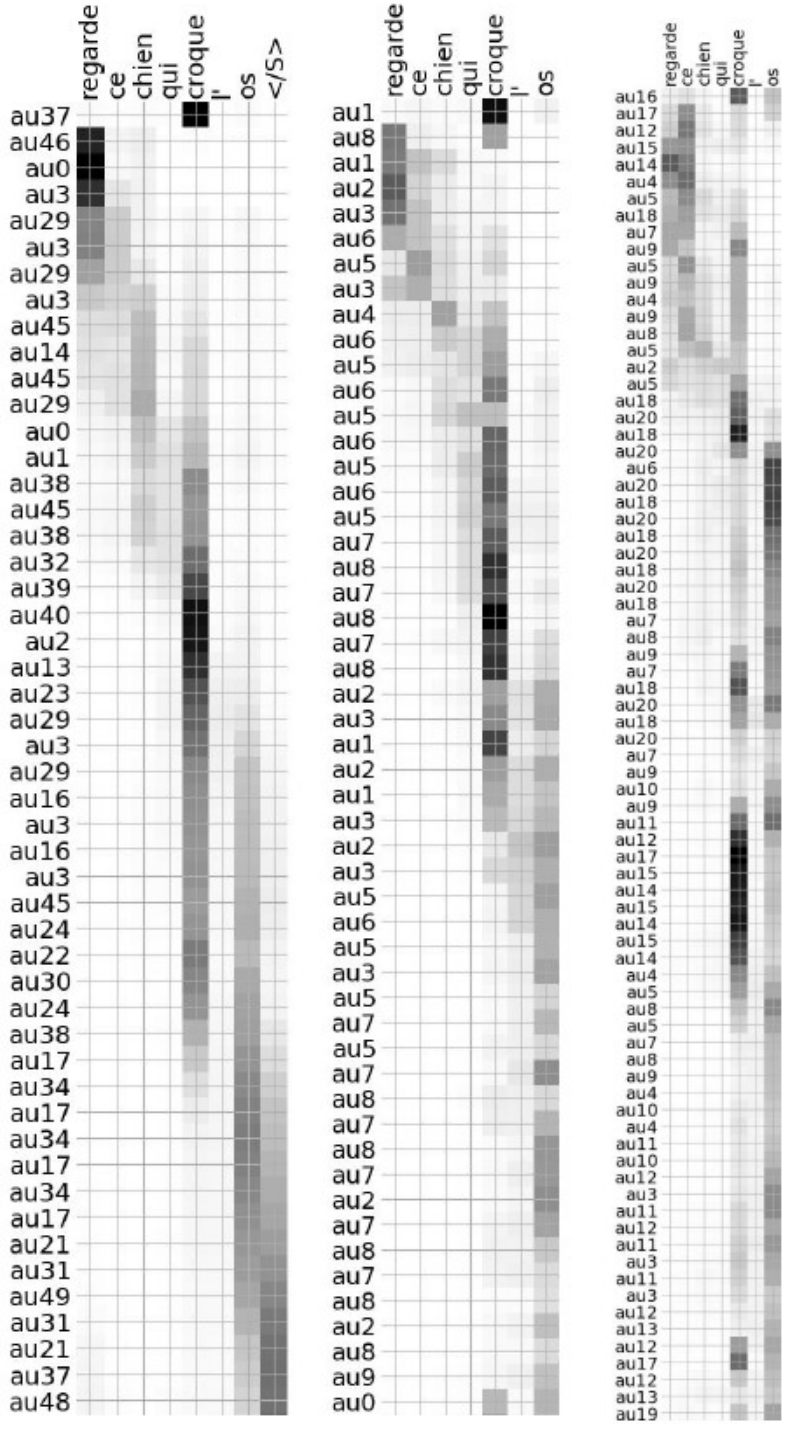


Figure 3.2: The soft-alignment probability matrices produced for the same sentence pair, but using different VQ-based SD approaches: VQ-VAE (left), VQ-WAV2VEC-V16 (middle) and VQ-WAV2VEC-V36 (right). Darker squares correspond to higher soft-alignment probabilities.

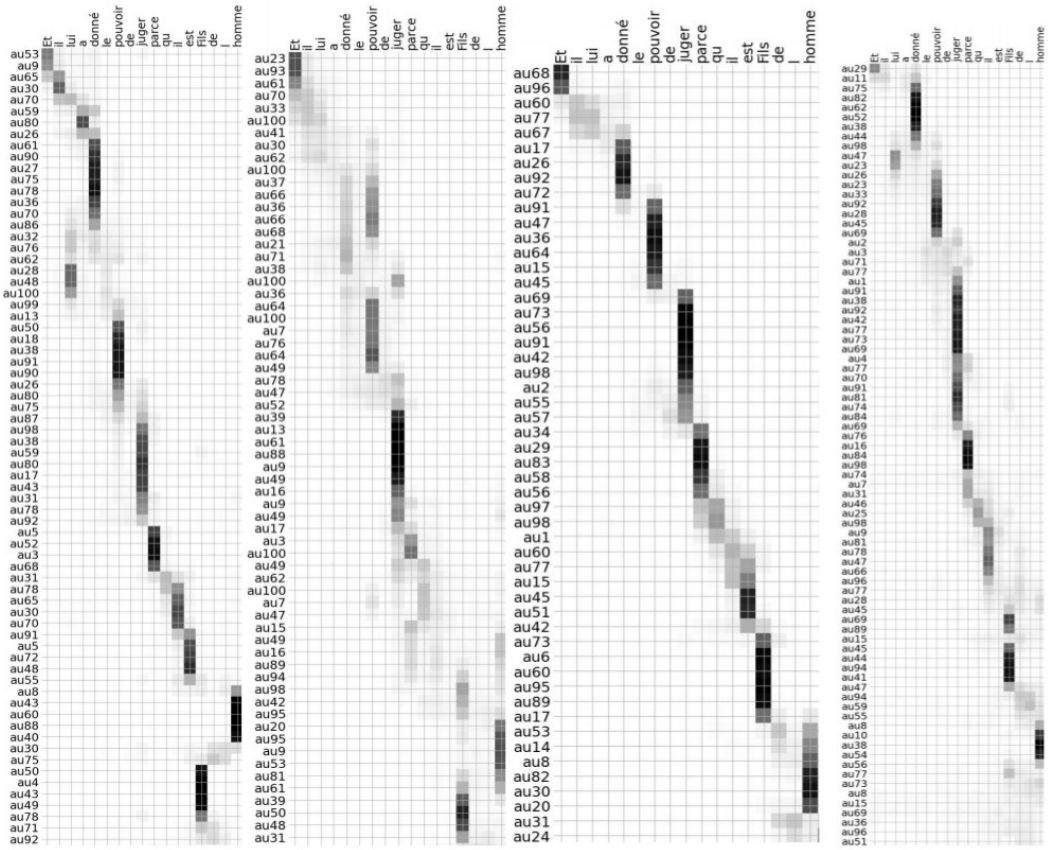


Figure 3.3: The soft-alignment probability matrices produced for the same sentence pair, but using different languages to train the **HMM SD model**: Finnish (left), Hungarian (left-to-center), Romanian (center-to-right), and Russian (right). Darker squares correspond to higher soft-alignment probabilities.

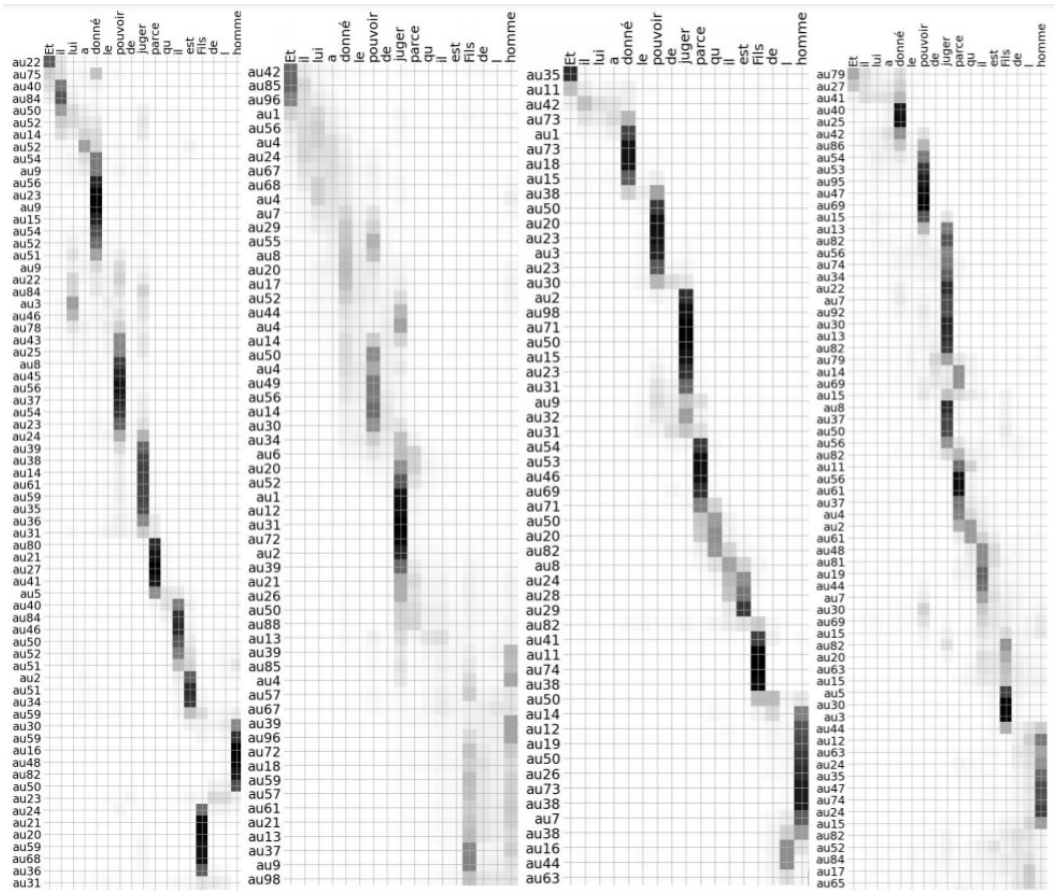


Figure 3.4: The soft-alignment probability matrices produced for the same sentence pair, but using different languages to train the **SHMM SD model**: Finnish (left), Hungarian (left-to-center), Romanian (center-to-right), and Russian (right). Darker squares correspond to higher soft-alignment probabilities.

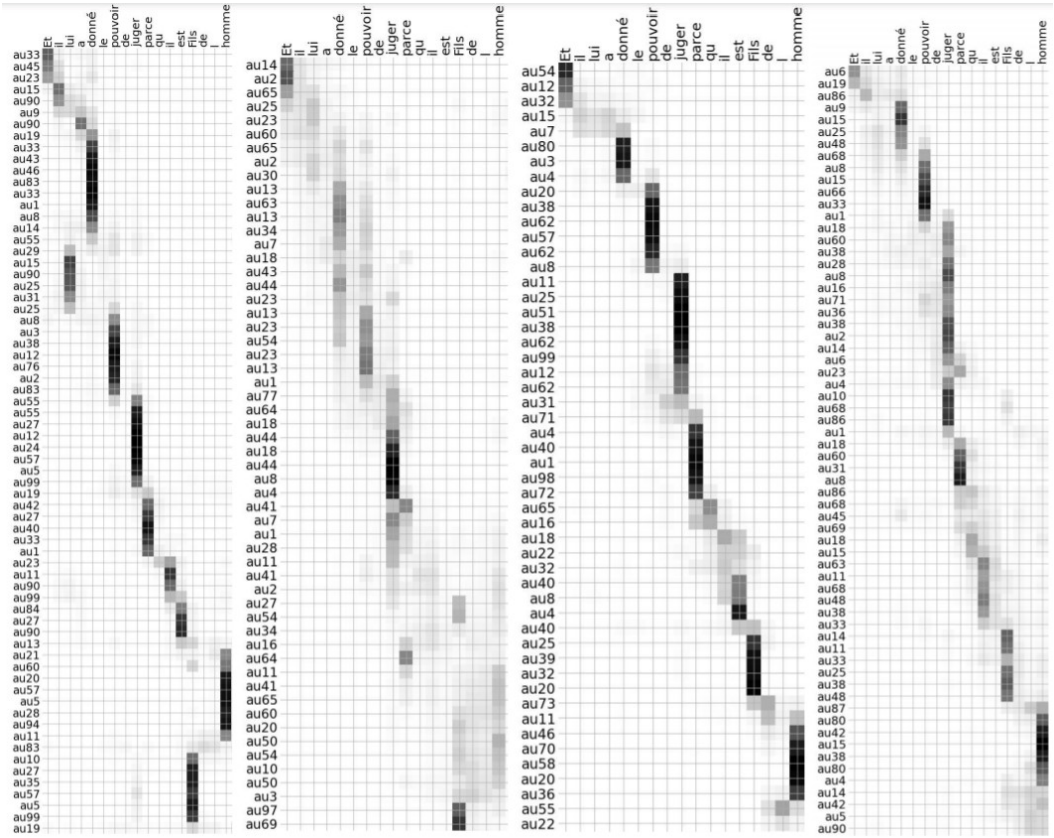


Figure 3.5: The soft-alignment probability matrices produced for the same sentence pair, but using different languages to train the **H-SHMM SD model**: Finnish (left), Hungarian (left-to-center), Romanian (center-to-right), and Russian (right). Darker squares correspond to higher soft-alignment probabilities.

Part II

French Translation

Annexe A

Introduction

La documentation des langues, telle que définie par Austin (2012), est le sous-domaine de la linguistique qui traite de la création d'enregistrements polyvalents des langues par des enregistrements audio et vidéo de locuteurs. Elle comprend l'annotation, la traduction, la préservation et la distribution du matériel résultant (par exemple, des grammaires, des dictionnaires, des collections de textes).

Le but de ce processus est de *documenter* les langues étudiées, de les préserver par la création de corpus et de ressources bien organisés et durables. Celles-ci peuvent être exploitées a posteriori pour des recherches ultérieures dans la langue cible, ou être utilisées pour des applications technologiques pratiques telles que la traduction automatique et la reconnaissance vocale. Ces données peuvent également être le point de départ d'initiatives de revitalisation de la langue cible (Pine and Turin, 2017).

L'une des principales cibles de la documentation des langues sont les *langues en danger*. Elles sont définies comme un sous-ensemble de langues existantes dont le nombre de locuteurs a considérablement diminué, ce qui les expose au risque de tomber en désuétude à mesure que leurs locuteurs périssent ou se tournent vers d'autres langues. Dans le *The Cambridge Handbook of Endangered Languages*, Austin and Sallabank (2011) a estimé que, sur les quelque 7 000 langues actuellement parlées, au moins 50% d'entre elles s'éteindront d'ici 2100.

Parmi les nombreuses raisons qui provoquent ce changement linguistique et l'homogénéisation des langues parlées à travers le monde, il faut noter l'impact du néocolonialisme et de la mondialisation (Austin and Sallabank, 2011). Ces langues en danger sont parlées dans des communautés isolées à travers le monde. Lorsque ces communautés commencent à être intégrées dans les circuits économiques, la langue parlée dans les grands centres économiques est transportée dans ces endroits. L'exode rural a également un impact, car les jeunes générations migrent vers les grandes villes à la recherche de meilleures opportunités d'emploi, ce qui réduit considérablement leur contact avec leur langue maternelle.

Certains soutiennent que l'extinction d'une langue dans son essence est un phénomène naturel (Ladefoged, 1992). Malgré cela, l'impact qu'elle provoque sur les communautés est largement reconnu. Les langues incarnent des visions

du monde, des systèmes de valeurs, des philosophies et des caractéristiques culturelles uniques. Leur extinction entraîne la perte irrémédiable de connaissances culturelles, historiques, spirituelles et écologiques uniques, utiles non seulement à la communauté, mais à d'innombrables autres (Drude et al., 2003; Bird, 2018; UNESCO, 2020). De plus, la perte de langues représente également un problème scientifique, car les futurs linguistes n'auront accès qu'à une fraction de la diversité linguistique mondiale disponible pour l'étude (Austin and Sallabank, 2011; Grenoble and Whaley, 1996; Nettle et al., 2000).

Dans ce contexte, le fait que la plupart des langues du monde ne sont pas activement écrites, même celles qui ont un système d'écriture officiel, pose un défi (Bird, 2011). Pour documenter ces langues orales, des enregistrements audio sont généralement collectés, puis transcrits. Cependant, cette transcription prend beaucoup de temps : on estime qu'une minute d'audio nécessite en moyenne une heure et demie de travail d'un linguiste (Austin and Sallabank, 2013).

De plus, le processus de documentation est itératif, et les transcriptions sont censées être révisées plusieurs fois avant le produit final (Crowley, 2007). Pour cette raison, les linguistes de terrain passent une grande partie de leur temps à transcrire et à réviser les documents, ce qui rend la documentation très coûteuse sur le plan humain et très lente. Brinckmann (2009) définit cela comme le *problème du goulot d'étranglement de la transcription* des initiatives de documentation.

Pour atténuer ce goulot d'étranglement, des travaux récents ont suggéré de remplacer les transcriptions par des liens multilingues, ajoutés aux enregistrements audio. Ces liens peuvent prendre la forme de traductions au niveau des phrases ou des mots (Adda et al., 2016), ou d'étiquettes superposées sur les fenêtres de temps dans l'audio (Bird, 2021). Ces approches mettent en évidence le contenu présent dans les audios, au lieu de créer des transcriptions exhaustives. Ce faisant, elles traitent la transcription comme une *observation* (Cucchiarini, 1993), au lieu de la considérer comme le but ultime de la documentation.

Cependant, afin de traiter et d'extraire des informations de cette nouvelle forme de corpus, la technologie doit intervenir en fournissant des méthodes informatiques robustes capables de traiter ces données qui sont : à faibles ressources, multilingues et parfois multimodales (par exemple, des images, des vidéos). L'émergence récente du domaine de la **documentation computationnelle des langues** (CLD) tente de proposer des réponses à cela. Il rassemble des linguistes et des experts en technologie afin de fournir des méthodologies et des modèles pour le traitement automatique des données et pour assister les linguistes, en atténuant les ressources humaines et le temps nécessaires à la documentation des langues.

Cette thèse s'inscrit dans le cadre des nombreuses approches CLD visant à produire une technologie utile pour le traitement des données dans le contexte de la documentation des langues. En particulier, nous proposons une approche pour la segmentation non supervisée de mots à partir de la parole. Résoudre une telle tâche à partir du signal de la parole, au lieu de segmenter dans le domaine textuel, est une façon de traiter le goulot d'étranglement de la transcription.

De plus, considérant les traductions comme un processus peu coûteux pour l'étiquetage des données (Adda et al., 2016), nous avons choisi de les inclure comme supervision faible de nos énoncés pendant la segmentation. Ainsi, nous considérons que notre processus de segmentation est *bilingue*, et au cours de cette thèse nous discutons de l'impact de la langue sur la qualité des segments découverts.

Notre modèle est composé de deux composants : (1) discrétisation de la parole, et (2) alignement basé sur le texte et segmentation de la parole. Cette séparation est nécessaire afin d'atténuer le défi que représente le traitement de la parole dans des environnements à très faibles ressources. Le but de la première composante est de produire des séquences d'unités discrètes de parole, exploitables dans des environnements à faibles ressources, en utilisant seulement quelques heures de parole. Par conséquent, dans cette thèse, nous étudions la qualité et l'exploitabilité des modèles de discrétisation de la parole dans notre cadre documentaire.

Pour la deuxième composante, nous utilisons des réseaux de neurones pour créer des matrices de probabilité d'alignement entre la discrétisation de la parole et sa traduction au niveau de la phrase. Cette opération est effectuée par une couche spéciale à l'intérieur des modèles neuronaux de traduction automatique appelée *attention*, dont la sortie peut être considérée comme un alignement souple bilingue. Cet alignement souple est utilisé pour produire une segmentation sur les unités discrètes de la parole, qui est ensuite reportée sur le signal vocal original. Ainsi, dans ce travail, nous étudions de manière approfondie la qualité et l'exploitabilité du mécanisme d'attention dans notre contexte, et nous introduisons également une métrique agnostique pour évaluer la *confiance dans l'alignement* des matrices de probabilité d'alignement souple (Boito et al., 2019a).

Le pipeline de segmentation de mots non supervisée en deux étapes que nous proposons est comparé à un modèle de référence bien établie (**?**), et à travers différentes langues (Godard et al., 2018c; Boito et al., 2019a, 2020b). En se concentrant sur les scénarios de documentation, nous proposons une extension qui prend en compte la disponibilité de transcriptions partielles, et un modèle qui exploite la segmentation préexistante dans le modèle d'alignement bilingue (Boito et al., 2021).

Enfin, le modèle que nous proposons nécessite un corpus bilingue composé d'audio et de traductions de phrases alignées. Afin de tester de manière réaliste nos modèles et de permettre à la communauté des chercheurs de faire le même, nous avons rassemblé et publié trois jeux de données, que nous présentons dans ce travail (Godard et al., 2018a; Boito et al., 2018, 2020a).

Annexe B

Résumé des Chapitres

1 Chapitre 3 : Les Resources

Actuellement, nous constatons un manque de corpus réalistes pour tester la généralisation des modèles proposés. Ainsi, de nombreux travaux s'appuient sur l'échantillonnage de langues à hautes ressources pour *émuler* le comportement attendu en utilisant des langues à faibles ressources. Cette méthodologie suppose que les différentes langues sont toutes aussi difficiles à *apprendre* et surtout, qu'elles sont apprises de la même manière. Le résultat de ce type d'hypothèse est la proposition de modèles qui pourraient être involontairement biaisés par rapport à une langue particulière à haute ressource, et qui pourraient ne pas fonctionner correctement lorsqu'ils sont appliqués à la cible réelle (Kawakami et al., 2019).

La solution à ce problème est donc de tester de manière approfondie les approches proposées dans des contextes réalistes et en utilisant de nombreuses langues, ce qui n'est généralement pas fait en raison d'un manque de données. Dans le but d'aider à combler ce manque de ressources disponibles dans les langues à faibles ressources, nous avons participé, au cours de cette thèse, à trois projets visant à mettre à la disposition de la communauté des corpus de parole réalistes à faibles ressources, que nous décrivons dans ce chapitre.

Nous avons publié deux ensembles de données provenant de langues en danger de disparition (Les corpus parallèles Mboshi-Français (Godard et al., 2018a) et Griko-Italien (Boito et al., 2018)) ; et un nouveau jeu de données multilingue au niveau de la parole (MaSS dataset (Boito et al., 2020a)) couvrant des langues avec des caractéristiques linguistiques intéressantes. Tous ces jeux de données mentionnés, ainsi que les références et scripts d'évaluation, sont disponibles gratuitement en ligne.

2 Chapitre 4 : Un modèle bilingue de segmentation de mots non supervisé basé sur l'attention

Dans ce chapitre, nous présentons notre modèle bilingue de segmentation non supervisée des mots (UWS) à partir de la parole. Ce modèle fonctionne en

deux étapes : (1) Discrétisation de la parole (SD), et (2) alignement et segmentation bilingue.

La première étape est responsable de la production d'unités discrètes de parole (ou pseudo-phones) à partir de l'audio. La deuxième étape travaille sur le domaine symbolique, en alignant les unités découvertes avec les mots de traduction, et en produisant à partir de cela une segmentation. Comme nos unités discrètes de parole contiennent des informations temporelles, la segmentation produite peut être transférée à l'audio, produisant une segmentation sur l'entrée de parole elle-même. Ce processus est *bilingue*, car la segmentation est exécutée en s'appuyant sur l'alignement bilingue découvert. En d'autres termes, les mots de traduction sont utilisés pour *guider* la segmentation générée.

La nature type pipeline de notre modèle nous permet de segmenter de petits ensembles de données, une tâche qui serait difficile à accomplir si nous devions entraîner directement des modèles de traduction de la parole. De plus, nous sommes soutenus par les études qui montrent que les réseaux neuronaux sont capables d'apprendre des caractéristiques linguistiques en travaillant avec des unités plus petites que les mots, comme les unités de sous-mots et les caractères (Kreutzer and Sokolov, 2018; Hahn and Baroni, 2019; Ataman et al., 2019), et donc adaptés pour travailler avec des phonèmes ou des unités discrètes de parole.

Dans ce chapitre, nous concentrons nos recherches sur la deuxième étape de notre pipeline UWS pour la parole, qui travaille sur le domaine symbolique. Nous commençons par valider ce modèle dans le scénario idéal d'une discrétisation parfaite de la parole, en remplaçant les unités discrètes de la parole par les vrais phones (phonèmes) de la langue. Cela nous permet d'évaluer la performance maximale que nos modèles travaillant à partir de la parole peuvent accomplir.

L'idée centrale de notre procédure de segmentation est l'utilisation de matrices de probabilité d'alignement souple entre source et cible pour produire la segmentation. Nous utilisons des modèles de traduction automatique neuronale (NMT) basés sur l'attention afin de récupérer l'alignement souple entre les unités de parole discrètes et les mots de traduction, en utilisant cette information pour inférer la segmentation bilingue des mots.

Nous disposons de deux méthodes pour évaluer la performance de nos modèles. Premièrement, nous évaluons la qualité des matrices de probabilité d'alignement souple produites par l'entrainement des modèles NMT à l'aide d'une métrique que nous avons introduite dans Boito et al. (2019a). Cette métrique est appelée Entropie Normalisée Moyenne (ANE), et elle nous donne le degré de confiance des alignements souples découverts par un modèle NMT. Deuxièmement, nous évaluons le produit final de notre pipeline directement sur le domaine de la parole en utilisant les métriques pour les frontières (*bound-*

ary metrics en anglais).

Les expériences que nous présentons dans ce chapitre se concentrent sur la deuxième étape de notre pipeline : la tâche d'alignement et de segmentation bilingue. Nous avons étudié deux aspects importants qui pourraient avoir un impact sur la performance : (1) le modèle NMT basé sur l'attention utilisé pour générer les matrices de probabilité d'alignement souple, et (2) la langue choisie pour guider la segmentation.

Dans notre première section expérimentale, qui correspond à nos travaux publiés dans Boito et al. (2019a) et Boito et al. (2021), nous avons étudié l'utilisation de différents modèles de NMT basés sur l'attention (RNN, 2D-CNN, Transformer) pour produire les matrices de probabilité d'alignement souple source-cible que nous utilisons pour la segmentation. Nous avons constaté que le modèle RNN est le plus exploitable dans des environnements à faibles ressources, atteignant la meilleure performance de segmentation par rapport aux deux autres approches plus modernes de NMT.

Nous avons également introduit une métrique pour évaluer le degré d'exploitabilité des matrices de probabilité d'alignement souple produites par les modèles NMT. Cette métrique, l'Entropie Normalisée Moyenne (ANE), peut être accumulée à travers différents niveaux de représentation (i.e. token, phrase, alignement, corpus). Nous avons montré que l'ANE du corpus est fortement corrélée à la performance de segmentation, et que l'ANE de l'alignement nous permet de filtrer le vocabulaire généré, augmentant ainsi les scores de découverte de type.

Notre deuxième section expérimentale, qui correspond à notre travail publié dans Boito et al. (2020b), s'est concentrée sur l'impact de la langue dans notre segmentation bilingue. Nous avons utilisé un corpus multilingue pour la segmentation d'une langue donnée soutenue par les mêmes informations dans sept langues différentes. En faisant varier la langue cible, nous avons produit 56 modèles bilingues, ce qui nous a permis de vérifier clairement l'impact de la supervision dans les segmentations générées.

Nos résultats ont mis en évidence l'existence d'une relation entre les caractéristiques de la langue et les performances de segmentation pour notre approche. Nous avons vérifié que les langues proches en termes de phonologie et de famille linguistique obtiennent de meilleurs résultats, tandis que les langues moins similaires donnent des résultats plus faibles. Bien que nos résultats soient affectés par les caractéristiques linguistiques, nous pensons également qu'il existe une influence non négligeable des caractéristiques statistiques du corpus, ce qui peut avoir un impact considérable sur les approches neuronales dans les environnements à faibles ressources.

3 Chapitre 5 : Extensions du modèle de UWS basé sur l'attention

Dans le chapitre précédent, nous avons présenté notre pipeline pour l'UWS bilingue basé sur l'attention à partir de la parole, et dans des environnements à faibles ressources. Il est composé de deux parties différentes : un composant de discrétisation de la parole (SD), et un composant d'alignement et de segmentation bilingue. En nous concentrant sur ce dernier, nous avons étudié l'impact de l'utilisation de différents mécanismes d'attention pour produire un alignement bilingue, et nous avons évalué l'impact de la langue de la supervision. Avant de présenter en détail l'étape de SD au chapitre 6, nous nous concentrons sur les extensions possibles de cette composante d'alignement et de segmentation bilingue, dans le but d'augmenter les scores de UWS.

Inspirés par les approches de documentation, nous étudions un modèle qui intègre l'exploitation des transcriptions partielles du corpus bilingue (c'est-à-dire des données monolingues), nous étudions également l'exploitation de *suggestions de frontières* dans le pipeline. En nous concentrant sur le régime d'entraînement, nous expérimentons dans ce chapitre l'extension proposée dans Godard et al. (2019), dans laquelle un biais de longueur de mot est introduit dans les matrices de probabilité d'alignement souple produites pendant l'entraînement. Enfin, nous présentons également quelques expériences moins réussies concernant la supervision multilingue pour UWS.

En comparant les résultats obtenus à travers les différentes extensions de modèle mentionnées dans ce chapitre, nous observons qu'elles ont toutes permis une amélioration dans les scores d'UWS du modèle de base, et certaines du modèle `dpseg`. Les meilleurs résultats ont été obtenus par le modèle pré-entraîné, qui avait accès à des informations monolingues. Parmi les modèles entièrement bilingues, l'extension la plus prometteuse est le modèle hybride qui incorpore les frontières intermédiaires de `dpseg` dans l'apprentissage NMT.

Pour ce dernier modèle, notre impression générale est que le gain de performance est dû au fait que les frontières souples aident le modèle à éviter la sous-segmentation. Cependant, dans ce cas, il n'est pas encore clair dans quelle mesure le modèle final dépend de la qualité des frontières souples (en termes de précision). C'est-à-dire : *si la performance du `dpseg` n'est pas aussi bonne que celle présentée, ses frontières douces peuvent-elles encore aider le modèle neuronal ?* Dans le prochain chapitre, nous aborderons cette question de recherche.

Enfin, inspirés par l'idée que les traductions multiples pourraient être une forme de capture de couches de sens plus profondes (Evans and Sasse, 2004), nous avons également étudié l'incorporation d'une supervision multilingue à

notre pipeline. Nos résultats, cependant, n'étaient pas très encourageants, et nous laissons l'exploration de cette branche de recherche comme travail futur.

4 Chapitre 6 : UWS basé sur l'attention au niveau de la parole

Dans ce chapitre, nous étudions la première partie de notre pipeline UWS : Les modèles SD pour produire des unités de parole discrètes à partir du signal de parole. Nous comparons cinq de ces approches : trois modèles bayésiens basés sur les HMM (Ondel et al., 2016, 2019; Yusuf et al., 2020), et deux modèles neuronaux de quantification vectorielle (van den Oord et al., 2017; Baevski et al., 2020a). Dans cette comparaison, notre objectif principal est d'identifier le modèle qui produirait la représentation la plus exploitable. Pour nous, une séquence *exploitable* issue de l'entraînement SD doit être concise, afin d'être directement appliquée aux modèles UWS basés sur le texte.

En comparant les modèles SD, nous avons remarqué que les méthodes basées sur VQ ne sont pas adaptées à notre pipeline, car elles produisent des séquences très longues et inconsistantes, qui sont difficiles à traiter. Ceci a également été récemment observé dans Kamper and van Niekerk (2020).

En revanche, les modèles basés sur les HMM produisent une bonne représentation discrète et concise, que nous sommes en mesure d'exploiter avec succès pour l'UWS. Nous pensons que cette différence de performance est due au fait que les modèles basés sur les HMM effectuent explicitement la découverte des unités acoustiques (AUD). Cela signifie que la discrétisation qu'ils produisent vise non seulement à résumer le signal vocal, mais aussi à correspondre étroitement à la phonologie de la langue.

En plus, l'estimation du sous-espace effectuée par les SHMM et les H-SHMM pourrait également jouer un rôle important. En effet, ces modèles sont capables d'apprendre à partir de 19 heures supplémentaires de données dans différentes langues. Les autres modèles (HMM et modèles basés sur VQ) n'ont accès à aucune forme de pré-entraînement ou d'antériorité.

En ce qui concerne les résultats UWS obtenus en appliquant la sortie des modèles SD à la tâche UWS, nous avons atteint nos meilleurs résultats de frontière pour le Mboshi en utilisant les modèles SHMM et H-SHMM. Cette même tendance a également été observée dans quatre langues différentes du jeu de données MaSS (FI, HU, RO, RU), vérifiant la généralisation du pipeline proposé.

En comparant notre approche UWS basée sur l'attention à `dpseg`, nous remarquons que nous sommes très compétitifs dans ce cadre, atteignant de meilleurs scores de limite UWS. Cette baseline est cependant meilleure pour

la segmentation des vrais phones (scénario de base du chapitre 4). Dans notre approche, nous avons également l'avantage de produire un alignement bilingue comme base de la segmentation générée.Dans le chapitre 4, nous avons montré que cette information peut être utilisée pour augmenter les scores de découverte de types.

Enfin, dans ce chapitre, nous avons également étudié l'application du modèle hybride du chapitre 5 à la situation réelle de l'UWS à partir de la parole. Ce modèle enrichit la représentation d'entrée pour l'entraînement NMT en utilisant la sortie `dpseg` comme des *frontières souples*. Nous constatons que ce modèle est largement sous-performant en raison de la dégradation des performances de `dpseg` dans ce contexte plus bruyant.

Cependant, la motivation de cette approche est d'utiliser `dpseg` comme un proxy pour évaluer les segmentations existantes produites par un linguiste. Par conséquent, nous pensons toujours que cette méthode pourrait potentiellement améliorer nos résultats UWS si `dpseg` était remplacé par des annotations d'un linguiste, ou une meilleure approche UWS. Une telle investigation est une suggestion pour un travail futur.

Annexe C

Conclusion

Le traitement du langage naturel est un domaine de recherche très populaire, mais la technologie pour les langues tend à être développée principalement dans et pour une très petite partie des langues existantes dans le monde. Ces langues, dites à hautes ressources, sont utilisées pour proposer et tester des approches. Dans cette approche naïve, toutes les langues sont considérées comme égales (à modéliser et à apprendre) tant qu'il y a suffisamment de données pour entraîner des approches d'apprentissage automatique à forte intensité de données.

Cependant, pour de nombreuses langues, il n'y a pas, et il n'y aura probablement jamais, de données suffisantes. C'est notamment le cas des dialectes minoritaires, qui ne sont pas considérés comme économiquement intéressants pour justifier l'investissement nécessaire à la collecte de données. De plus, la mondialisation croissante pousse indirectement l'humanité vers une standardisation des langues parlées. Il en résulte que l'on estime que de nombreuses langues existantes (si ce n'est la plupart) disparaîtront au cours de ce siècle (Austin and Sallabank, 2011).

Parallèlement, les *approches à zéro ressources* sont devenues populaires ces dernières années, car elles proposent d'atteindre la longue traîne des langues à faibles ressources existantes en proposant des approches adaptées à des contextes avec moins de données. Dans ce contexte, nous soulignons la nécessité non seulement de développer avec moins de ressources, mais aussi l'importance d'utiliser des données diverses. Ce n'est qu'en procédant ainsi que nous pourrons véritablement tester et comprendre l'applicabilité des méthodes que nous proposons.

De plus, il existe une critique récente sur la signification de ce *zéro* dans les approches à zéro ressources (Bird, 2020). Les langues existent rarement de manière totalement isolée, et rares sont celles qui ne disposent d'aucun lexique existant ou d'une documentation initiale ou rudimentaire. L'absence d'intérêt pour l'exploitation de ces informations lors de la proposition d'approches peut faire en sorte que les produits n'aient qu'un impact marginal ou nul pour la communauté des locuteurs.

Par conséquent, bien que le défi technologique consistant à extraire des connaissances à partir d'informations quasi inexistantes soit attrayant pour les

scientifiques, s'ils veulent proposer des approches pour la *documentation computationnelle des langues*, ils devraient collaborer avec des experts en langues et avec la communauté. De cette façon, ils sont sûrs de produire *pour* la communauté, et pas simplement *à partir* de leurs données.

Dans cette thèse, nous avons étudié la tâche de segmentation non supervisée de mots dans le contexte de la documentation des langues. Notre objectif principal était d'éviter le besoin de transcriptions, car celles-ci sont connues pour être généralement non disponibles (Adda et al., 2016; Brinckmann, 2009).

Au lieu de cela, nous nous sommes concentrés sur la segmentation de l'audio en segments de mots en utilisant seulement quelques heures de parole et en fondant ce processus sur des annotations alignées (traductions). Notre segmentation finale est appliquée au signal vocal, accompagnée d'annotations sous forme de traductions potentielles. Nous espérons que ces annotations seront utiles pour examiner les mots candidats, voire pour construire un lexique bilingue de segments de parole.

Nous discutons maintenant en détail les contributions de ce travail, ainsi que certaines limitations de l'approche proposée.

1 Contributions

Cette thèse a proposé une approche pipeline pour l'UWS dans le domaine de la parole. Cette approche base la segmentation dans les mots de traduction, et résout la segmentation en utilisant l'alignement doux produit par les modèles NMT. Avant l'alignement et la segmentation, le SD est effectué afin de relever le défi du traitement de la parole à faibles ressources. Nous récapitulons maintenant les contributions, en développant chaque sujet.

C1: ***Une comparaison approfondie des approches SD récentes pour le traitement de la parole à faibles ressources, en se concentrant sur leur applicabilité directe aux modèles UWS à base de texte.***

L'objectif des modèles SD est de produire une séquence d'unités vocales discrètes à partir d'énoncés d'entrée, sans avoir recours à une transcription. Dans le chapitre 6, nous avons comparé cinq de ces modèles. Trois d'entre eux appartiennent à la famille des HMM bayésiens : HMM (Ondel et al., 2016), SHMM (Ondel et al., 2019), H-SHMM (Yusuf et al., 2020), et les deux autres modèles sont des approches neuronales récentes basées sur la quantification vectorielle : VQ-VAE (van den Oord et al., 2017) and VQ-WAV2VEC (Baevski et al., 2020a).

Nous avons optimisé et entraîné ces modèles dans des environnements à faibles ressources en utilisant cinq langues, en évaluant la qualité des unités

vocales discrètes produites. Nous nous sommes concentrés sur l'application directe de l'approche UWS basée sur le texte.

Nos résultats ont montré que les modèles basés sur les HMM ont produit une sortie concise, proche de la référence. Pour les modèles basés sur VQ, nous avons observé un processus d'étiquetage de la parole très inconsistant, résultant en des séquences difficiles à appliquer à notre tâche. Les modèles SD les plus exploitables pour UWS étaient les modèles SHMM et HSHMM.

C2: *Une étude de l'interprétabilité directe du mécanisme d'attention dans les modèles NMT, et dans des contextes à faibles ressources.*

Dans le chapitre 4, nous avons étudié l'utilisation des matrices de probabilité d'alignement souple obtenues par l'entraînement NMT pour aligner les mots de la traduction sur une séquence non segmentée de phones. Cet alignement souple, qui est produit par le mécanisme d'attention, est ensuite utilisé pour regrouper les phones voisins qui partagent l'alignement des mots. Nous appelons cette méthode *attention-based* UWS (UWS basé sur l'attention).

Afin d'évaluer la faisabilité de cette approche dans des contextes à faibles ressources, nous avons comparé trois différents modèles de RNN basés sur l'attention : RNN (Bahdanau et al., 2015), 2D-CNN (Elbayad et al., 2018), et Transformer (Vaswani et al., 2017). Nous avons trouvé le classement suivant pour l'exploitabilité de ces modèles, du meilleur au pire : RNN, 2D-CNN, Transformer. Nos résultats ont également montré que l'alignement souple découvert par le mécanisme d'attention est toujours exploitable lorsque le RNN est entraîné avec seulement 5k phrases. Nous avons obtenu nos meilleurs résultats de segmentation en utilisant le modèle RNN, le plus *simple*, et nos pires résultats en utilisant l'architecture Transformer. Ce travail a été présenté dans Boito et al. (2019a), et étendu au format journal dans Boito et al. (2021).

C3: *Une comparaison entre les approches UWS : notre modèle basé sur l'attention et deux baselines.*

Dans ce travail, nous avons comparé notre approche UWS basée sur l'attention à deux baselines dans des contextes réalistes (Godard et al., 2018c; Boito et al., 2019a, 2020b). Nous utilisons seulement 5k phrases dans la langue Mboshi, et nous incluons des résultats dans huit autres langues : Anglais, Espagnol, Basque, Finnois, Français, Hongrois, Roumain et Russe.

La première baseline est le modèle bilingue proportionnel. Il nous permet d'évaluer le défi de notre tâche d'alignement. Il s'agit d'une approche naïve qui produit un alignement diagonal, regroupant les unités en tenant compte de la longueur des mots traduits. Comme prévu, nos résultats avec les 56 paires de langues du jeu de données MaSS ont montré que cette approche naïve est sous-performante. Ceci illustre que la tâche de segmentation bilingue que nous

ciblons dans ce travail n'est pas triviale.

La deuxième baseline est le modèle de Goldwater et al. (2009), que nous appelons le `dpseg`. Nous constatons qu'il est très compétitif : en travaillant à partir des vrais phones de la langue, cette baseline était celle qui produisait les meilleurs résultats de segmentation. Cependant, lorsque nous passons à un scénario plus difficile, où l'entrée est plus bruyante (en termes de consistance, de longueur et de taille du vocabulaire), cette baseline a obtenu des résultats inférieurs ou égaux à ceux de notre approche basée sur l'attention. Nous pensons que cela met en évidence la façon dont la supervision bilingue peut aider le processus de découverte dans des environnements difficiles.

C4: ***L'étude de l'impact de la langue dans notre pipeline.***

Tout au long de ce travail, nous avons utilisé diverses langues afin d'évaluer la généralité du pipeline proposé. Pour évaluer l'impact lié à la langue en utilisant ces différentes langues, il y a deux aspects à prendre en compte. Le premier est l'écart naturel qui se produit dans les méthodes non supervisées lors de la segmentation de différentes langues, car les langues ne sont pas toutes aussi difficiles à segmenter (Fourtassi et al., 2013). Le second aspect est l'impact de l'information bilingue qui existe dans notre pipeline, qui guide la segmentation à travers les mots de la traduction.

En ce qui concerne le premier aspect, dans le chapitre 4, nous avons utilisé le jeu de données MaSS pour générer 56 paires de langues à partir de ses huit langues, que nous avons utilisées pour entraîner nos modèles UWS bilingues. Nos résultats, publiés dans Boito et al. (2020b), ont montré un net écart de performance entre les modèles ayant différentes langues comme cible de segmentation.

En ce qui concerne le second aspect, nous avons classé les langues utilisées pour guider la segmentation en fonction des performances de segmentation obtenues pour chacune des huit langues cibles. Nous avons constaté que, bien que le classement final des langues obtenu semblait être ancré dans les caractéristiques linguistiques, l'impact des caractéristiques statistiques était non négligeable. En effet, des statistiques telles que la taille du vocabulaire et le ratio type-token peuvent avoir un impact sur la capacité du modèle neuronal à encoder les informations d'entrée. Ainsi, en ayant des statistiques plus favorables (plus faciles à apprendre dans des environnements à faibles ressources), certaines langues étaient supérieures comme supervision pour la segmentation de langues même non liées.

C5: ***La proposition d'extensions de pipeline pour incorporer des informations supplémentaires dans le modèle de segmentation.***

Au chapitre 5, nous avons proposé deux méthodes pour inclure des con-

naissances supplémentaires dans nos modèles. La première consistait à pré-entraîner les modèles NMT avec une petite partie des transcriptions. Ceci a été motivé par l'existence possible de celles-ci, produites par des linguistes pendant la collecte des données. Dans ce contexte, après le pré-entraînement sur cette petite portion de données transcrites manuellement, le modèle NMT est entraîné sur l'ensemble complet de données bilingues. Nos résultats ont montré que ce pré-entraînement est utile, augmentant à la fois la découverte des frontières et des types.

Nous avons également proposé un modèle hybride, qui utilise la segmentation produite par `dpseg` pour enrichir les séquences d'entrée que nous avons dans notre modèle NMT. Ces *soft-boundaries* semblaient informer nos modèles, augmentant ainsi les résultats de la segmentation. L'objectif de ce modèle était d'évaluer l'incorporation des mots-hypothèses par les linguistes dans le modèle. Dans ce contexte, le linguiste pouvait étudier la sortie de notre modèle pour valider les hypothèses existantes. Malheureusement, ce modèle n'a pas fonctionné dans des environnements bruyants. Les deux modèles ont été présentés en Boito et al. (2021).

C6: ***La collecte et la publication de trois ensembles de données utiles pour les approches de documentation des langues computationnelles à faibles ressources.***

Dans le chapitre 3, nous avons présenté les jeux de données suivants : Corpus parallèle Mboshi-Français (Godard et al., 2018a), Corpus parallèle Griko-Italien (Boito et al., 2018), et Jeu de données multilingues MaSS (Boito et al., 2020a). Le **Corpus parallèle Mboshi-français** a été largement exploité pour l'évaluation d'approches dans le traitement de la parole à faibles ressources et la documentation des langues (Anastasopoulos and Chiang, 2018a,b; Bansal et al., 2019; Sung et al., 2019; Inaguma et al., 2019; Scharenborg et al., 2018, 2020; Ondel et al., 2019; Yusuf et al., 2020; Godard et al., 2018b, 2019). Le **Corpus parallèle griko-italien** est un exemple intéressant d'*extrême* scénario à faibles ressources, étant intéressant pour les approches d'apprentissage du type *zero shot* (Wada et al., 2020). Enfin, le **MaSS multilingual dataset** a été mentionné par la communauté comme un exemple de jeu de données permettant d'étudier paires de langues diverses, contribuant ainsi à atténuer la nature *anglais-centré* des approches actuelles pour la parole.

Dans cette thèse, nous avons utilisé le corpus parallèle Mboshi-Français comme cible principale de notre étude (Chapitres 4 à 6). Nous avons également utilisé une version sous-échantillonnée de l'ensemble de données MaSS afin d'étudier l'impact de la langue. Les résultats pour le corpus parallèle Griko-Italien n'ont pas été présentés ici, car nous avons trouvé que ce corpus était trop petit pour l'entraînement NMT.

2 Limitations

Dans ce travail, nous avons proposé un pipeline pour résoudre l'UWS à partir de la parole dans des environnements à faibles ressources. La première limite d'une telle approche est sa structure en pipeline. L'absence d'interaction entre le processus de discrétisation et de segmentation de la parole signifie que les erreurs dans le premier processus sont propagées dans le second. De plus, on peut imaginer qu'en fondant la découverte d'unités sur leur utilité postérieure pour la création de segments de mots, un modèle plus robuste pourrait être construit.

Une deuxième limitation de notre modèle est la dépendance des données dans les approches neuronales. Nous avons pu entraîner avec succès des modèles en utilisant seulement 5 130 phrases ; cependant, dans des études préliminaires, nous n'avons pas réussi à faire de même pour le petit ensemble de données Griko-Italien, composé de seulement 330 phrases. Dans un cadre aussi limité, notre modèle s'est révélé largement inférieur à la baseline monolingue `dpseg` fonctionnant avec les vrais phones. À partir d'unités de parole discrètes, les deux modèles (le nôtre et le `dpseg`) n'ont rien produit d'exploitable (Boito et al., 2018). Cela suggère l'existence d'un quantité minimale de données pour l'applicabilité des modèles UWS.

De plus, il est à noter que notre modèle est contraint par l'existence de traductions de mots alignés. Cela signifie que nous ne pouvons pas appliquer notre pipeline pour segmenter les données monolingues des initiatives de documentation. La motivation de notre approche était précisément de proposer quelque chose de fondé sur l'information bilingue, et non de couvrir le cas de son absence. Un modèle neuronal monolingue récent pour UWS a été proposé dans Kawakami et al. (2019), mais ce modèle était profondément ancré dans la représentation des caractères, et il aurait besoin de modifications pour travailler à partir de la sortie des modèles SD.

Enfin, une autre limitation de notre approche repose sur la procédure d'alignement : l'utilisation de matrices de probabilité d'alignement souple issues de la formation NMT. Ces matrices sont un sous-produit de la traduction, ce qui signifie que les modèles NMT ne prennent pas explicitement en compte l'alignement dans leur optimisation. Des modèles tels que Alkhouli et al. (2018) et Garg et al. (2019) se concentrent sur l'optimisation conjointe de la traduction et de l'alignement mot à mot. Cependant, on s'attend à ce que leur tâche soit plus facile que notre alignement de plusieurs unités à un mot. Godard et al. (2019) ont effectué une optimisation explicite pour l'UWS basée sur l'attention, et ont trouvé un gain de performance marginal. Cela suggère qu'une optimisation plus sophistiquée pour découvrir les segments de mots pourrait être nécessaire.

Bibliography

Abate, S. T., Menzel, W., and Tafila, B. (2005). An Amharic Speech Corpus for Large Vocabulary Continuous Speech Recognition. In *INTERSPEECH-2005*. 111

Adams, O., Neubig, G., Cohn, T., and Bird, S. (2015). Inducing bilingual lexicons from small quantities of sentence-aligned phonemic transcriptions. In *12th International Workshop on Spoken Language Translation (IWSLT)*. 27, 49

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. iii, v, 14, 16, 17, 54, 89, 125, 152, 153, 162

Alishahi, A., Chrupała, G., and Linzen, T. (2019). Analyzing and interpreting neural networks for nlp: A report on the first blackboxnlp workshop. *Natural Language Engineering*, 25(4):543–557. 38

Alkhouli, T., Bretschner, G., and Ney, H. (2018). On the alignment problem in multi-head attention-based neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 177–185, Brussels, Belgium. Association for Computational Linguistics. 36, 130, 131, 166

Anastasopoulos, A. and Chiang, D. (2018a). Leveraging translations for speech transcription in low-resource settings. In *Proc. Interspeech 2018*, pages 1279–1283. 15, 61, 129, 165

Anastasopoulos, A. and Chiang, D. (2018b). 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)*, pages 82–91, New Orleans, Louisiana. Association for Computational Linguistics. 61, 129, 165

Arivazhagan, N., Bapna, A., Firat, O., Lepikhin, D., Johnson, M., Krikun, M., Chen, M. X., Cao, Y., Foster, G., Cherry, C., et al. (2019). Massively multilingual neural machine translation in the wild: Findings and challenges. *arXiv preprint arXiv:1907.05019*. 105

Ataman, D., Firat, O., Di Gangi, M. A., Federico, M., and Birch, A. (2019). On the importance of word boundaries in character-level neural machine translation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 187–193, Hong Kong. Association for Computational Linguistics. 33, 63, 156

Austin, P. (2012). Language documentation. 13, 151

Austin, P. K. and Sallabank, J. (2011). *The Cambridge handbook of endangered languages*. Cambridge University Press. 13, 14, 125, 151, 152, 161

Austin, P. K. and Sallabank, J. (2013). *Endangered languages*. Taylor & Francis. iii, v, 14, 152

Baevski, A., Schneider, S., and Auli, M. (2020a). vq-wav2vec: Self-supervised learning of discrete speech representations. In *International Conference on Learning Representations (ICLR)*. iii, v, 44, 45, 110, 112, 120, 126, 159, 162

Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. (2020b). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33. 43

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural Machine Translation by Jointly Learning to Align and Translate. In *ICLR 2015*, pages 3104–3112, San Diego, California, USA. iv, vi, 3, 30, 31, 32, 33, 39, 66, 69, 127, 163

Bansal, S., Kamper, H., Livescu, K., Lopez, A., and Goldwater, S. (2019). Pre-training on high-resource speech recognition improves low-resource speech-to-text translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 58–68, Minneapolis, Minnesota. Association for Computational Linguistics. 61, 129, 165

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. 15

Beapami, R. P., Chatfield, R., Kouarata, G., and Waldschmidt, A. (2000). Dictionnaire mbochi-français. *Brazzaville: SIL-Congo*. 54

Belinkov, Y., Durrani, N., Dalvi, F., Sajjad, H., and Glass, J. (2020). On the linguistic representational power of neural machine translation models. *Computational Linguistics*, 46(1):1–52. 33

Bengio, Y., Léonard, N., and Courville, A. (2013). Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*. 45

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. In *NIPS Workshop on End-to-End Learning for Speech and Audio Processing*. 49, 71, 131

Berg-Kirkpatrick, T., Bouchard-Côté, A., DeNero, J., and Klein, D. (2010). Painless unsupervised learning with features. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 582–590. 27, 28

Besacier, L. et al. (2015). Speech technologies for african languages: Example of a multilingual calculator for education. In *Interspeech*. 111

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. 131

Bird, S. (2011). Bootstrapping the language archive: New prospects for natural language processing in preserving linguistic heritage. *Linguistic Issues in Language Technology*, 6(4). 14, 152

Bird, S. (2018). Creating a world that sustains its languages. In Seyalioglu, H. and Hymes, K., editors, *Dialect - A Game about Language and How it Dies*. Thorny Games. 14, 152

Bird, S. (2020). Decolonising speech and language technology. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3504–3519. 15, 125, 161

Bird, S. (2021). Sparse transcription. *Computational Linguistics*. 14, 16, 152

Bisazza, A. and Tump, C. (2018). The lazy encoder: A fine-grained analysis of the role of morphology in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2871–2876, Brussels, Belgium. Association for Computational Linguistics. 34

Blachon, D., Gauthier, E., Besacier, L., Kouarata, G.-N., Adda-Decker, M., and Rialland, A. (2016). Parallel speech collection for under-resourced language studies using the lig-aikuma mobile device app. In *Proceedings of SLTU (Spoken Language Technologies for Under-Resourced Languages)*. 54

Black, A. W. (2019). Cmu wilderness multilingual speech dataset. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5971–5975. 57, 59

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot. Int.*, 5(9):341–345. 113

Boito, M. Z. (2017). Unsupervised word discovery using attentional encoder-decoder models. Master's thesis, University Grenoble Alpes (UGA), Grenoble, France. 66, 71, 82

Boito, M. Z., Anastasopoulos, A., Villavicencio, A., Besacier, L., and Lekakou, M. (2018). A Small Griko-Italian Speech Translation Corpus. In *Proc. The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages*, pages 36–41. iv, vi, 18, 20, 56, 129, 154, 155, 165, 166

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. 15, 62, 93, 131

Boito, M. Z., Havard, W. N., Garnerin, M., Ferrand, É. L., and Besacier, L. (2020a). Mass: A large and clean multilingual corpus of sentence-aligned spoken utterances extracted from the bible. *Language Resources and Evaluation Conference (LREC)*. iv, vi, 18, 20, 60, 81, 129, 154, 155, 165

Boito, M. Z., Villavicencio, A., and Besacier, L. (2019a). Empirical evaluation of sequence-to-sequence models for word discovery in low-resource settings. In *Proc. Interspeech 2019*, pages 2688–2692. iv, vi, 18, 20, 62, 64, 67, 90, 127, 153, 156, 157, 163

Boito, M. Z., Villavicencio, A., and Besacier, L. (2019b). How does language influence documentation workflow? unsupervised word discovery using translations in multiple languages. In *Scientific Meeting of the "Computational, Formal and Field Linguistics" Research Group. Orléans, France.* 54, 62

Boito, M. Z., Villavicencio, A., and Besacier, L. (2020b). Investigating language impact in bilingual approaches for computational language documentation. In *Proceedings of the 1st Joint SLTU and CCURL Workshop (SLTU-CCURL 2020)*. iv, vi, 18, 20, 64, 91, 102, 127, 128, 153, 157, 163, 164

Boito, M. Z., Villavicencio, A., and Besacier, L. (2021). Investigating alignment interpretability for low-resource nmt. *Machine Translation Journal: Special Issue on Machine Translation for Low-Resource Languages*. iv, vi, 18, 20, 62, 64, 90, 107, 127, 128, 153, 157, 163, 165

Bouquiaux, L. and Thomas, J. (1976). Enquete et methode de la description des langues ä tradition orale,(bd. i, ii, iii). 54

Brent, M. R. (1999). An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning*, 34(1-3):71–105. 25, 26

Brinckmann, C. (2009). Transcription bottleneck of speech corpus exploitation. In *Proceedings of the Second Colloquium on Lesser Used Languages and Computer Linguistics (LULCL II). Combining efforts to foster computational support of minority languages.* 14, 16, 17, 125, 152, 162

Brunner, G., Liu, Y., Pascual, D., Richter, O., and Wattenhofer, R. (2019). On the validity of self-attention as explanation in transformer models. *arXiv preprint arXiv:1908.04211*. 40

Chatzikyriakidis, S. (2010). *Clitics in four dialects of Modern Greek: A dynamic account*. PhD thesis, University of London. 55

Chen, H., Leung, C.-C., Xie, L., Ma, B., and Li, H. (2017). Multilingual bottleneck feature learning from untranscribed speech. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 727–733. IEEE. 43, 44

Chen, Y., Liu, Y., Chen, G., Jiang, X., and Liu, Q. (2020). Accurate word alignment induction from neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 566–576, Online. Association for Computational Linguistics. 36

Chorowski, J., Weiss, R. J., Bengio, S., and van den Oord, A. (2019). Unsupervised speech representation learning using wavenet autoencoders. *IEEE/ACM transactions on audio, speech, and language processing*, 27(12):2041–2053. 43, 44, 117

Chorowski, J. K., Bahdanau, D., Serdyuk, D., Cho, K., and Bengio, Y. (2015). Attention-based models for speech recognition. In *Advances in neural information processing systems*, pages 577–585. 30

Chrupała, G., Gelderloos, L., and Alishahi, A. (2017). Representations of language in a model of visually grounded speech signal. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 613–622, Vancouver, Canada. Association for Computational Linguistics. 132

Cotterell, R., Mielke, S. J., Eisner, J., and Roark, B. (2018). Are all languages equally hard to language-model? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 536–541, New Orleans, Louisiana. Association for Computational Linguistics. 53

Crowley, T. (2007). *Field linguistics: A beginner's guide*. OUP Oxford. 14, 152

Cucchiarini, C. (1993). *Phonetic transcription: a methodological and empirical study*. PhD thesis, Radboud Universiteit Nijmegen. 14, 152

Dabre, R., Chu, C., and Kunchukuttan, A. (2020). A comprehensive survey of multilingual neural machine translation. *arXiv preprint arXiv:2001.01115*. 105

Di Gangi, M. A., Cattoni, R., Bentivogli, L., Negri, M., and Turchi, M. (2019). Must-c: a multilingual speech translation corpus. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 2 (Short Papers)*, pages 2012–2017. 53

Ding, S., Xu, H., and Koehn, P. (2019). Saliency-driven word alignment interpretation for neural machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 1–12, Florence, Italy. Association for Computational Linguistics. 39

Douri, A. and De Santis, D. (2015). Griko and modern Greek in Grecìa Salentina: an overview. *L'Idomeneo*, 2015(19):187–198. 55

Doyle, G. and Levy, R. (2013). Combining multiple information types in bayesian word segmentation. In *Proceedings of the 2013 Conference of the*

North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 117–126. 28

Drude, S. et al. (2003). Language vitality and endangerment. 14, 152

Dunbar, E., Algayres, R., Karadayi, J., Bernard, M., Benjumea, J., Cao, X.-N., Miskic, L., Dugrain, C., Ondel, L., Black, A. W., Besacier, L., Sakti, S., and Dupoux, E. (2019). The Zero Resource Speech Challenge 2019: TTS Without T. In *Proc. Interspeech 2019*, pages 1088–1092. 44

Dunbar, E., Cao, X. N., Benjumea, J., Karadayi, J., Bernard, M., Besacier, L., Anguera, X., and Dupoux, E. (2017). The zero resource speech challenge 2017. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 323–330. IEEE. 28, 44, 54, 68

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. 15, 49, 67

Dyer, C., Chahuneau, V., and Smith, N. A. (2013). A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648. 76

Elbayad, M., Besacier, L., and Verbeek, J. (2018). Pervasive attention: 2D convolutional neural networks for sequence-to-sequence prediction. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 97–107, Brussels, Belgium. Association for Computational Linguistics. iv, vi, 3, 30, 36, 37, 69, 71, 127, 163

Elsner, M., Goldwater, S., Feldman, N., and Wood, F. (2013). A joint learning model of word segmentation, lexical acquisition, and phonetic variability. In *Proc. EMNLP*. 28

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. 89, 107, 158

Feng, S., Żelasko, P., Moro-Velázquez, L., and Scharenborg, O. (2021). Unsupervised acoustic unit discovery by leveraging a language-independent subword discriminative feature representation. 61

Flokstra, J. (1987). The phonology of parent child speech. 27

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. 15

Fourtassi, A., Börschinger, B., Johnson, M., and Dupoux, E. (2013). Why is english so easy to segment? In *Proceedings of the Fourth Annual Workshop on Cognitive Modeling and Computational Linguistics (CMCL)*, pages 1–10. 74, 128, 164

Galassi, A., Lippi, M., and Torroni, P. (2019). Attention, please! a critical review of neural attention models in natural language processing. *IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS*. 29

Garg, S., Peitz, S., Nallasamy, U., and Paulik, M. (2019). Jointly learning to align and translate with transformer models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4453–4462, Hong Kong, China. Association for Computational Linguistics. 36, 130, 166

Gauthier, E., Besacier, L., Voisin, S., Melese, M., and Elingui, U. P. (2016). Collecting Resources in Sub-Saharan African Languages for Automatic Speech Recognition: a Case Study of Wolof. *LREC*. 111

Gehring, J., Auli, M., Grangier, D., Yarats, D., and Dauphin, Y. N. (2017). Convolutional sequence to sequence learning. *Proceedings of the 34th International Conference on Machine Learning*, pages 1243–1252. 30, 38, 39

Gelas, H., Besacier, L., and Pellegrino, F. (2012). Developments of Swahili resources for an automatic speech recognition system. In *SLTU - Workshop on Spoken Language Technologies for Under-Resourced Languages*, Cape-Town, Afrique Du Sud. 111

Ghader, H. and Monz, C. (2017). What does attention in neural machine translation pay attention to? In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 30–39, Taipei, Taiwan. Asian Federation of Natural Language Processing. 38

Gibadullin, I., Valeev, A., Khusainova, A., and Khan, A. (2019). A survey of methods to leverage monolingual data in low-resource neural machine translation. *arXiv preprint arXiv:1910.00373*. 42

Godard, P. (2019). *Unsupervised Word Discovery for Computational Language Documentation*. PhD thesis, Université Paris-Saclay. 15, 72, 73

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*. 27, 72

Godard, P., Adda, G., Adda-Decker, M., Benjumea, J., Besacier, L., Cooper-Leavitt, J., Kouarata, G.-N., Lamel, L., Maynard, H., Mueller, M., Rialland, A., Stueker, S., Yvon, F., and Boito, M. Z. (2018a). A very low resource language speech corpus for computational language documentation experiments. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA). iv, vi, 18, 20, 54, 129, 154, 155, 165

Godard, P., Besacier, L., and Yvon, F. (2019). Controlling utterance length in nmt-based word segmentation with attention. *Proceedings of the 16th International Workshop on Spoken Language Translation (IWSLT)*. 62, 93, 102, 103, 104, 106, 129, 130, 158, 165, 166

Godard, P., Besacier, L., Yvon, F., Adda-Decker, M., Adda, G., Maynard, H., and Rialland, A. (2018b). Adaptor grammars for the linguist: Word segmentation experiments for very low-resource languages. In *Proceedings of the Fifteenth Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 32–42, Brussels, Belgium. Association for Computational Linguistics. 27, 62, 129, 165

Godard, P., Boito, M. Z., Ondel, L., Bérard, A., Yvon, F., Villavicencio, A., and Besacier, L. (2018c). Unsupervised word segmentation from speech with attention. In *Proc. Interspeech 2018*, pages 2678–2682. iv, vi, 18, 20, 62, 109, 127, 136, 153, 163

Goldwater, S., Griffiths, T. L., and Johnson, M. (2009). A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, 112(1):21–54. iv, vi, 15, 18, 19, 26, 28, 127, 164

Goldwater, S. J. (2007). *Nonparametric Bayesian models of lexical acquisition*. PhD thesis, Citeseer. 16, 18, 25, 26

Grenoble, L. A. and Whaley, L. J. (1996). Endangered languages: Currentissues and future prospects. *International journal of the sociology of language*, 118(1):209–223. 14, 152

Gulcehre, C., Firat, O., Xu, K., Cho, K., Barrault, L., Lin, H.-C., Bougares, F., Schwenk, H., and Bengio, Y. (2015). On using monolingual corpora in neural machine translation. *arXiv preprint arXiv:1503.03535*. 42

Gumbel, E. J. (1948). *Statistical theory of extreme values and some practical applications: a series of lectures*, volume 33. US Government Printing Office. 46

Hahn, M. and Baroni, M. (2019). Tabula nearly rasa: Probing the linguistic knowledge of character-level neural language models trained on unsegmented text. *Transactions of the Association for Computational Linguistics*, 7:467–484. 33, 63, 156

Harwath, D., Chuang, G., and Glass, J. R. (2018). Vision as an interlingua: Learning multilingual semantic embeddings of untranscribed speech. In *ICASSP 2018 - 2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, AB, Canada, April 15-20, 2018*, pages 4969–4973. 59

Haspelmath, M. (2011). The indeterminacy of word segmentation and the nature of morphology and syntax. *Folia linguistica*, 45(1):31–80. 49, 86

He, S., Tu, Z., Wang, X., Wang, L., Lyu, M., and Shi, S. (2019). Towards understanding neural machine translation with word importance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 953–962, Hong Kong, China. Association for Computational Linguistics. 39

Horrocks, G. (2009). *Greek: A History of the Language and its Speakers*. Wiley-Blackwell. 55

Hu, D. (2019). An introductory survey on attention mechanisms in nlp problems. In *Proceedings of SAI Intelligent Systems Conference*, pages 432–448. Springer. 29

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708. 37

Imankulova, A., Dabre, R., Fujita, A., and Imamura, K. (2019). Exploiting out-of-domain parallel data through multilingual transfer learning for low-resource neural machine translation. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pages 128–139, Dublin, Ireland. European Association for Machine Translation. 43

Inaguma, H., Duh, K., Kawahara, T., and Watanabe, S. (2019). Multilingual end-to-end speech translation. *IEEE Automatic Speech Recognition and Understanding Workshop, ASRU*. 61, 129, 165

Jain, S. and Wallace, B. C. (2019). Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics. 39, 40

Johnson, M., Christophe, A., Dupoux, E., and Demuth, K. (2014). Modelling function words improves unsupervised word segmentation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 282–292. 25

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. 16, 25, 27

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., et al. (2017). Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351. 95, 105

Kamper, H., Jansen, A., and Goldwater, S. (2016). Unsupervised word segmentation and lexicon discovery using acoustic word embeddings. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(4):669–679. 29

Kamper, H., Livescu, K., and Goldwater, S. (2017). An embedded segmental k-means model for unsupervised segmentation and clustering of speech. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 719–726. IEEE. 29

Kamper, H. and van Niekerk, B. (2020). Towards unsupervised phone and word segmentation using self-supervised vector-quantized neural networks. *arXiv preprint arXiv:2012.07551*. 117, 121, 159

Kann, K., Cho, K., and Bowman, S. R. (2019). Towards realistic practices in low-resource natural language processing: The development set. *arXiv preprint arXiv:1909.01522*. 42

Kawakami, K., Dyer, C., and Blunsom, P. (2019). Learning to discover, ground and use words with segmental neural language models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6429–6441, Florence, Italy. Association for Computational Linguistics. 26, 28, 53, 130, 155, 166

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*. 44

Kisler, T., Reichel, U., and Schiel, F. (2017). Multilingual processing of speech via web services. *Computer Speech & Language*, 45:326 – 347. 58, 81

Kocabiyikoglu, A. C., Besacier, L., and Kraif, O. (2018). Augmenting librispeech with French translations: A multimodal corpus for direct speech translation evaluation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA). 70

Kreutzer, J. and Sokolov, A. (2018). Learning to segment inputs for nmt favors character-level processing. In *15th International Workshop on Spoken Language Translation (IWSLT)*. 33, 63, 156

Ladefoged, P. (1992). Another view of endangered languages. *Language*, 68(4):809–811. 14, 151

Larsen, E., Cristia, A., and Dupoux, E. (2017). Relating unsupervised word segmentation to reported vocabulary acquisition. In *INTERSPEECH*, pages 2198–2202. 25

Le Ferrand, E., Bird, S., and Besacier, L. (2020). Enabling interactive transcription in an indigenous community. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3422–3428, Barcelona, Spain (Online). International Committee on Computational Linguistics. 62

Lee, C.-y. and Glass, J. (2012). A nonparametric bayesian approach to acoustic model discovery. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 40–49. 44

Lee, C.-y., O'Donnell, T. J., and Glass, J. (2015a). Unsupervised lexicon discovery from acoustic input. *Transactions of the Association for Computational Linguistics*, 3:389–403. 29

Lee, J., Cho, K., and Hofmann, T. (2017). Fully character-level neural machine translation without explicit segmentation. *Transactions of the Association for Computational Linguistics*, 5:365–378. 33

Lee, L.-s., Glass, J., Lee, H.-y., and Chan, C.-a. (2015b). Spoken content retrieval beyond cascading speech recognition with text retrieval. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(9):1389–1420. 59

Lekakou, M., Baldiserra, V., and Anastasopoulos, A. (2013). Documentation and analysis of an endangered language: aspects of the grammar of Griko. *Griko Project*. `http://griko.project.uoi.gr`. 56

Liang, P. and Klein, D. (2009). Online em for unsupervised models. In *Proceedings of human language technologies: The 2009 annual conference of the North American chapter of the association for computational linguistics*, pages 611–619. 27

Lignos, C. (2011). Modeling infant word segmentation. In *Proceedings of the fifteenth conference on computational natural language learning*, pages 29–38. 27, 28

Lignos, C. (2012). Infant word segmentation: An incremental, integrated model. In *Proceedings of the West Coast Conference on Formal Linguistics*, volume 30, pages 13–15. 27

Lignos, C. and Yang, C. (2010). Recession segmentation: simpler online word segmentation using limited resources. In *Proceedings of the fourteenth conference on computational natural language learning*, pages 88–97. Association for Computational Linguistics. 15

Lignos, C. and Yang, C. (2016). Morphology and language acquisition. *The Cambridge handbook of morphology*, page 743764. 25, 28

Lin, Y.-H., Chen, C.-Y., Lee, J., Li, Z., Zhang, Y., Xia, M., Rijhwani, S., He, J., Zhang, Z., Ma, X., Anastasopoulos, A., Littell, P., and Neubig, G. (2019). Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy. Association for Computational Linguistics. 43

Lin, Z., Feng, M., dos Santos, C., Yu, M., Xiang, B., Zhou, B., and Bengio, Y. (2017). A structured self-attentive sentence embedding. In *iclr*. 38

Littell, P., Mortensen, D. R., Lin, K., Kairis, K., Turner, C., and Levin, L. (2017). URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14, Valencia, Spain. Association for Computational Linguistics. 43

Luong, T., Pham, H., and Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics. 33, 38, 40

Lyzinski, V., Sell, G., and Jansen, A. (2015). An evaluation of graph clustering methods for unsupervised term discovery. In *Sixteenth Annual Conference of the International Speech Communication Association*. 29

MacWhinney, B. (2004). English macwhinney corpus. 28

Matsuura, K., Mimura, M., Sakai, S., and Kawahara, T. (2020). Generative adversarial training data adaptation for very low-resource automatic speech recognition. *arXiv preprint arXiv:2005.09256*. 15, 61

Maxwell, M. and Hughes, B. (2006). Frontiers in linguistic annotation for lower-density languages. In *Proceedings of the Workshop on Frontiers in Linguistically Annotated Corpora 2006*, pages 29–37, Sydney, Australia. Association for Computational Linguistics. 41, 53

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. *University of Hawaii Press*. 15

Michel, P., Levy, O., and Neubig, G. (2019). Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems*, pages 14014–14024. 36, 73

Mikolov, T., Karafiát, M., Burget, L., Cernocký, J., and Khudanpur, S. (2010). Recurrent neural network based language model. In Kobayashi, T., Hirose, K., and Nakamura, S., editors, *INTERSPEECH*, pages 1045–1048. ISCA. 28

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26:3111–3119. 45

Mochihashi, D., Yamada, T., and Ueda, N. (2009). Bayesian unsupervised word segmentation with nested pitman-yor language modeling. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 100–108. 27

Moradi, P., Kambhatla, N., and Sarkar, A. (2019). Interrogating the explanatory power of attention in neural machine translation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 221–230, Hong Kong. Association for Computational Linguistics. 40

Nettle, D., Romaine, S., et al. (2000). *Vanishing voices: The extinction of the world's languages*. Oxford University Press on Demand. 14, 152

Neubig, G. (2014). Simple, correct parallelization for blocked gibbs sampling. *Nara Institute of Science and Technology, Tech. Rep.* 27

Nguyen, T., Vogel, S., and Smith, N. A. (2010). Nonparametric word segmentation for machine translation. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 815–823. 27

Ondel, L., Burget, L., and Černockỳ, J. (2016). Variational inference for acoustic unit discovery. *Procedia Computer Science*, 81:80–86. iii, v, 44, 47, 48, 56, 109, 110, 111, 120, 126, 159, 162

Ondel, L., Godard, P., Besacier, L., Larsen, E., Hasegawa-Johnson, M., Scharenborg, O., Dupoux, E., Burget, L., Yvon, F., and Khudanpur, S. (2018). Bayesian models for unit discovery on a very low resource language. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5939–5943. 61

Ondel, L., Vydana, H. K., Burget, L., and Černocký, J. (2019). Bayesian Subspace Hidden Markov Model for Acoustic Unit Discovery. In *Interspeech*, pages 261–265. iii, v, 47, 48, 61, 110, 111, 120, 126, 129, 131, 159, 162, 165

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., and Auli, M. (2019). fairseq: A fast, extensible toolkit for sequence modeling. *arXiv preprint arXiv:1904.01038*. 71

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (2015). Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2015, South Brisbane, Queensland, Australia, April 19-24, 2015*, pages 5206–5210. 43, 70

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. 82

Pine, A. and Turin, M. (2017). *Language revitalization*. Oxford University Press (OUP). 13, 151

Powers, D. M. (1998). Applications and explanations of zipf's law. In *Proceedings of the joint conferences on new methods in language processing and computational natural language learning*, pages 151–160. Association for Computational Linguistics. 26

Räsänen, O., Doyle, G., and Frank, M. C. (2015). Unsupervised word discovery from speech using automatic segmentation into syllable-like units. In *Sixteenth Annual Conference of the International Speech Communication Association*. 29

Rekabsaz, N., Pappas, N., Henderson, J., Khonglah, B. K., and Madikeri, S. (2019). Regularization advantages of multilingual neural language models for low resource domains. *arXiv preprint arXiv:1906.01496*. 42

Rialland, A., Aborobongui, M. E., Adda-Decker, M., and Lamel, L. (2015). Dropping of the class-prefix consonant, vowel elision and automatic phonological mining in embosi (bantu c 25). In *Selected Proceedings of the 44th Annual Conference on African Linguistics*, pages 7–10. 74

Saffran, J. R., Newport, E. L., and Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of memory and language*, 35(4):606–621. 25, 26

Scharenborg, O., Besacier, L., Black, A., Hasegawa-Johnson, M., Metze, F., Neubig, G., Stüker, S., Godard, P., Müller, M., Ondel, L., Palaskar, S., Arthur, P., Ciannella, F., Du, M., Larsen, E., Merkx, D., Riad, R., Wang, L., and Dupoux, E. (2018). Linguistic unit discovery from multi-modal inputs in unwritten languages: Summary of the "speaking rosetta" jsalt 2017 workshop. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4979–4983. 61, 129, 165

Scharenborg, O., Besacier, L., Black, A., Hasegawa-Johnson, M., Metze, F., Neubig, G., Stüker, S., Godard, P., Müller, M., Ondel, L., Palaskar, S., Arthur, P., Ciannella, F., Du, M., Larsen, E., Merkx, D., Riad, R., Wang, L., and Dupoux, E. (2020). Speech technology for unwritten languages. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:964–975. 61, 129, 165

Scharenborg, O., Ebel, P., Ciannella, F., Hasegawa-Johnson, M., and Dehak, N. (2018). Building an asr system for mboshi using a cross-language definition of acoustic units approach. In *Proceedings of the 6th Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU)*. ISCA. 61

Schneider, S., Baevski, A., Collobert, R., and Auli, M. (2019). wav2vec: Unsupervised Pre-Training for Speech Recognition. In *Proc. Interspeech 2019*, pages 3465–3469. 43, 45

Schultz, T., Vu, N. T., and Schlippe, T. (2013). Globalphone: A multilingual text & speech database in 20 languages. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8126–8130. IEEE. 111

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics. 33, 105, 112

Sennrich, R. and 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*, pages 211–221, Florence, Italy. Association for Computational Linguistics. 41

Serrano, S. and Smith, N. A. (2019). Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Lin-*

guistics, pages 2931–2951, Florence, Italy. Association for Computational Linguistics. 39

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, R., et al. (2018). Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4779–4783. IEEE. 30

Stahlberg, F., Cross, J., and Stoyanov, V. (2018). Simple fusion: Return of the language model. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 204–211, Brussels, Belgium. Association for Computational Linguistics. 42

Stahlberg, F., Schlippe, T., Vogel, S., and Schultz, T. (2013). Pronunciation extraction from phoneme sequences through cross-lingual word-to-phoneme alignment. In *International Conference on Statistical Language and Speech Processing*, pages 260–272. Springer. 49

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. 15

Stüker, S., Adda, G., Adda-Decker, M., Ambouroue, O., Besacier, L., Blachon, D., Bonneau-Maynard, H., Godard, P., Hamlaoui, F., Idiatov, D., et al. (2016). Innovative technologies for under-resourced language documentation: The bulb project. *Proceedings of CCURL (Collaboration and Computing for Under-Resourced Languages: toward an Alliance for Digital Language Diversity)*. 54

Sung, T.-W., Liu, J.-Y., Lee, H.-y., and Lee, L.-s. (2019). Towards end-to-end speech-to-text translation with two-pass decoding. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7175–7179. IEEE. 61, 129, 165

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc. 30, 31

Tjandra, A., Sakti, S., and Nakamura, S. (2019). Speech-to-speech translation between untranscribed unknown languages. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 593–600. 59, 132

UNESCO (2017 (accessed May 13, 2020)). Frequent asked questions: Endangered languages. 14, 152

van den Oord, A., Vinyals, O., and kavukcuoglu, k. (2017). Neural discrete representation learning. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30, pages 6306–6315. Curran Associates, Inc. iii, v, 44, 45, 110, 120, 126, 159, 162

Vashishth, S., Upadhyay, S., Tomar, G. S., and Faruqui, M. (2019). Attention interpretability across nlp tasks. *arXiv preprint arXiv:1909.11218*. 40

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008. iv, vi, 30, 34, 39, 41, 69, 71, 73, 127, 163

Versteegh, M., Anguera, X., Jansen, A., and Dupoux, E. (2016). The zero resource speech challenge 2015: Proposed approaches and results. *Procedia Computer Science*, 81:67–72. 28

Versteegh, M., Thiolliere, R., Schatz, T., Cao, X. N., Anguera, X., Jansen, A., and Dupoux, E. (2015). The zero resource speech challenge 2015. In *Sixteenth annual conference of the international speech communication association*. 28, 44

Voita, E., Talbot, D., Moiseev, F., Sennrich, R., and Titov, I. (2019). Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, Florence, Italy. Association for Computational Linguistics. 36, 73

Wada, T., Iwata, T., Matsumoto, Y., Baldwin, T., and Lau, J. H. (2020). Learning contextualised cross-lingual word embeddings for extremely low-resource languages using parallel corpora. *arXiv preprint arXiv:2010.14649*. 61, 129, 165

Wang, Y., Skerry-Ryan, R., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., et al. (2017). Tacotron: Towards end-to-end speech synthesis. *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017*. 30

Watanabe, S., Hori, T., Kim, S., Hershey, J. R., and Hayashi, T. (2017). Hybrid ctc/attention architecture for end-to-end speech recognition. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1240–1253. 30

Weiss, R. J., Chorowski, J., Jaitly, N., Wu, Y., and Chen, Z. (2017). Sequence-to-sequence models can directly translate foreign speech. In Lacerda, F., editor, *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017*, pages 2625–2629. ISCA. 49

Wiegreffe, S. and Pinter, Y. (2019). Attention is not not explanation. *arXiv preprint arXiv:1908.04626*. 39, 41

Yang, S., Wang, Y., and Chu, X. (2020). A survey of deep learning techniques for neural machine translation. *arXiv preprint arXiv:2002.07526*. 31

Yusuf, B., Ondel, L., Burget, L., Cernocky, J., and Saraclar, M. (2020). A hierarchical subspace model for language-attuned acoustic unit discovery. *arXiv preprint arXiv:2011.03115*. iii, v, 48, 61, 110, 111, 112, 120, 126, 129, 131, 159, 162, 165

Zhang, C., Tan, X., Ren, Y., Qin, T., Zhang, K., and Liu, T.-Y. (2020). Uwspeech: Speech to speech translation for unwritten languages. *arXiv preprint arXiv:2006.07926*. 59, 132

Zoph, B., Yuret, D., May, J., and Knight, K. (2016). Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics. 43