Pronunciation Extraction from Phoneme Sequences Through Cross-Lingual Word-to-Phoneme Alignment

Felix Stahlberg[1], Tim Schlippe[1], Stephan Vogel[2], and Tanja Schultz[1]

[1] Karlsruhe Institute of Technology, Cognitive Systems Lab
Adenauerring 4, 76131 Karlsruhe, Germany
felix.stahlberg@student.kit.edu, {tim.schlippe,tanja.schultz}@kit.edu
[2] Qatar Foundation, Qatar Computing Research Institute
Al-Nasr Tower A, 21st Floor, Doha, Qatar
svogel@qf.org.qa

Abstract. With the help of written translations in a source language, we cross-lingually segment phoneme sequences in a target language into word units using our new alignment model *Model 3P* [17]. From this, we deduce phonetic transcriptions of target language words, introduce the vocabulary in terms of word IDs, and extract a pronunciation dictionary. Our approach is highly relevant to bootstrap dictionaries from audio data for Automatic Speech Recognition and bypass the written form in Speech-to-Speech Translation, particularly in the context of under-resourced languages, and those which are not written at all.
Analyzing 14 translations in 9 languages to build a dictionary for English shows that the quality of the resulting dictionary is better if the vocabularies in source and target language are about the same size, shorter sentences, more word repetitions, and formal equivalent translations.

Keywords: pronunciation dictionary, under-resourced languages, speech-to-speech translation, word segmentation

1 Introduction

There are over 7,000 living languages and dialects in the world [8]. Automatic Speech Recognition (ASR) and Machine Translation (MT) systems exist only for few of them. Porting rapidly and economically language technology to new unseen and under-resourced languages is in particular required in situations where languages with few linguistic resources suddenly appear in the focus of interest. Another challenge is the merely spoken nature of many languages and dialects, some of which are widespread despite the lack of a written script [16, 13]. However, language technology generally requires a written script nowadays.

In [17] and in this work, we take first steps towards gathering training data for ASR and MT systems for an unseen target language rapidly and at low cost: We segment phoneme sequences into word units using information from another language. We then deduce word pronunciations from these units, introduce the

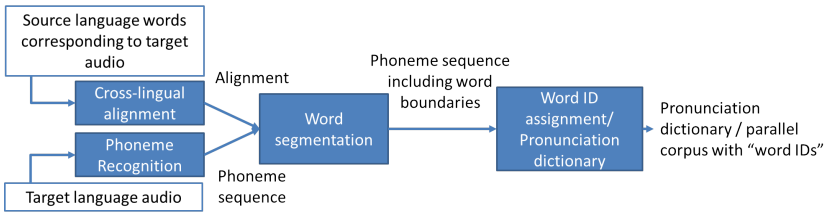


Fig. 1. Long-term scenario

vocabulary in terms of word IDs, and extract a pronunciation dictionary. Dictionaries are used to train speech processing systems by describing the pronunciation of words in manageable units such as phonemes [12]. As dictionaries are so fundamental, much care has to be taken to select a dictionary that is as free of errors as possible. Thus our approach is highly relevant for Speech-to-Speech Translation (S2S) of under-resourced languages, and those which are not written.

We explore 14 translations in 9 languages to build a dictionary for English. Our method benefits from the fact that written sentences are available in several economically viable languages such as Spanish. We assume that a speaker is available who understands Spanish and who speaks translations of the Spanish sentences in his or her mother tongue. This is a weak assumption, since human simultaneous translations happen frequently in the real world, e.g. in the context of humanitarian aid operations or in multilingual parliament sessions [7]. Our goal is to exploit the phonetic output of such human translators, so that the following scenario comes within reach (Fig. 1):

1) We recognize the spoken translations with a language independent phoneme recognizer. 2) We build an alignment between words in the written Spanish sentence and phonemes in the corresponding recognized phoneme sequence in the target language. 3) Using this cross-lingual alignment, we segment the phoneme sequence into word units. 4a) The word segmentation induces phonetic transcriptions of target language words, which are used in a pronunciation dictionary for ASR systems. 4b) The segmented phoneme sequence is replaced by a sequence of word IDs. This results in a parallel training corpus on the word level for a Statistical MT (SMT) system as described in [2]. Our final goal is to bootstrap an S2S system without any linguistic knowledge of the target language.

While we have focused on *step 2* and *3* in [17], we tackle *step 4a* in this paper – the pronunciation extraction. We test our algorithms on parallel data from the Christian Bible since it is available in many different languages in written form and in some languages also as audio recordings. A variety of linguistic approaches to Bible translation [21] enables us to compare different translations within the same source language. In our experiments, English takes the role of the under-resourced target language. English is by no means under-resourced and comprehensive pronunciation dictionaries are readily available [24]. However, for this exploratory work we feel that understanding the target language gives a deeper insight in the strengths and weaknesses of our algorithms.

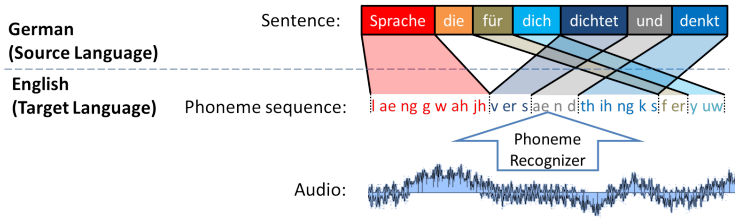


Fig. 2. Word segmentation through word-to-phoneme alignment

2 Word Segmentation

Cross-lingual word-to-phoneme alignments introduced in [2, 19, 20] and tackled by us with our new alignment model *Model 3P* [17] are the basis for our pronunciation extraction algorithm in Sec. 3. Therefore, this section summarizes the concepts of [17] in condensed form. The word segmentation problem describes the task of segmenting phoneme sequences into word units. We have shown in [17] that unsupervised learning of word segmentation is more accurate when information of another language is used. *Model 3P*[1] for cross-lingual word-to-phoneme alignment extends the generative process of IBM Model 3 by a word length step and additional dependencies for the lexical translation probabilities. Those alignments can be used for the segmentation task as illustrated in Fig. 2. Using *Model 3P* for the alignment between English words and correct Spanish phoneme sequences on the BTEC corpus [10] resulted in 76.5% F-Score (90.0% accuracy [22]) and thus outperformed a state-of-the-art monolingual word segmentation approach [9] by 42% absolute in F-Score (18.2% in accuracy).

3 Word Pronunciation Extraction

3.1 Formal Framework

Let V_{src} be the vocabulary of the source language and $PhonemeSet_{trgt}$ the phoneme set of the target language. The data source we explore in our scenario is a set $DB \subset {V_{src}}^{+} \times {PhonemeSet_{trgt}}^{+}$ of pairs containing a written sentence in the source language and its spoken translation in the target language. As described in Sec. 2, we use *Model 3P* to find word-to-phoneme alignments for each sentence-phoneme sequence pair in DB. An alignment $A_{s,t}$ consists of a mapping between the words in the source language sentence $s \in {V_{src}}^{+}$ and the phonemes in the target language phoneme sequence $t \in {PhonemeSet_{trgt}}^{+}$ segmented into word units. We formalize $A_{s,t}$ as a word over an alphabet containing pairs of source language words and target language phoneme sequences.

$$A_{s,t} \in (PhonemeSet_{trgt}{}^{+} \times V_{src})^{+}$$

[1] A multi-threaded implementation is available at `http://pisa.googlecode.com/`

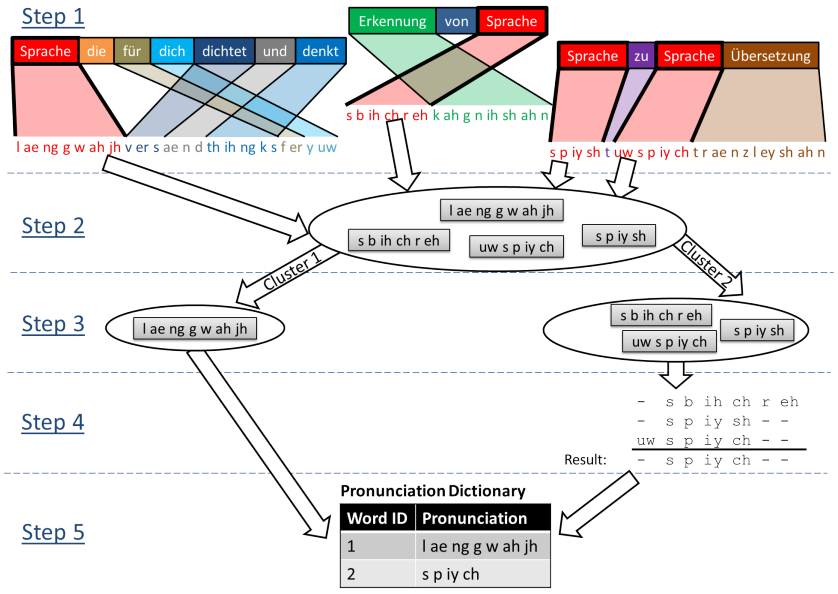


Fig. 3. Steps 1-5 on a German-English example ("Sprache zu Sprache Übersetzung" → "Speech to speech translation", "Sprache die für dich dichtet und denkt" → "Language verses and thinks for you", "Erkennung von Sprache" → "Speech recognition")

Each element in $A_{s,t}$ contains a hypothetical target language *word* represented by its phonemes and the source language word aligned to it. We postulate that the source language words are elements in s, and that concatenating all target language words results in the complete phoneme sequence t.

3.2 Pronunciation Extraction Algorithm

We extract pronunciations based on the assumption, that phoneme sequences, that are aligned to the same source language word, are likely to represent the same target language word. They only differ due to phoneme recognition and alignment errors. From the linguistic point of view, this is not always the case: in Fig. 3, the German word *Sprache* has two different English translations (*Speech* and *Language*). *Step 3* of our algorithm addresses this special case.

We build the pronunciation dictionary iteratively by repeating the following steps until all source language words are marked. The steps are visualized in Fig. 3 with German as source language and English as target language.

1. Select the most frequent unmarked source language word $v \in V_{src}$ and mark it.
2. Collect the set $P \subset PhonemeSet_{trgt}{}^{+}$ of all phoneme sequences, which are aligned to v (hypothetical target language words):[2]

[2] For technical reasons, we define the $\in$ sign for a symbol $x \in \Sigma$ and a word $w \in \Sigma^{+}$ as $x \in w :\Leftrightarrow \exists i \in \mathbb{N} : x = w_i$

$$P \leftarrow \{h | \exists (s,t) \in DB : (h,v) \in A_{s,t}\}$$

3. Group the phoneme sequences into clusters $C \subset 2^P$. We applied the clustering algorithm `DBSCAN` [6] ($\epsilon = 1$, $minPts = 3$) implemented in the `ELKI` [1] environment with the Levenshtein distance metric. This step aims to separate elements in P from each other, which do not represent the same target language word.
4. At this point, the clusters should contain phoneme sequences representing the same target language word, but differing due to alignment and phoneme recognition errors. Thus we try to reconstruct the correct phoneme sequence for each cluster by merging its elements with the `nbest-lattice` [18] program.We obtain a set $H \subset {PhonemeSet_{trgt}}^+$ of phoneme sequences, which are now assumed to correspond to real target language words.
5. For each pronunciation $h \in H$, we choose a new word ID $id_h \in \mathbb{N}$ and add both to the pronunciation dictionary $Dict$.

When we apply the general algorithm above to the example in Fig. 3, the variables have following values:

1. $v =$ `Sprache`
2. $P = \{$`s b ih ch r eh`, `l ae ng g w ah jh`, `uw s p iy ch`, `s p iy sh`$\}$
3. $C = \{\{$`l ae ng g w ah jh`$\}, \{$`s b ih ch r eh`, `uw s p iy ch`, `s p iy sh`$\}\}$
4. $H = \{$`l ae ng g w ah jh`, `s p iy ch`$\}$
5. $Dict = \{(1,$ `l ae ng g w ah jh`$), (2,$ `s p iy ch`$)\}$

4 Experiments

4.1 Corpus

We tested our pronunciation extraction algorithm on parallel data from the Christian Bible. A variety of linguistic approaches to Bible translation (Dynamic equivalence, formal equivalence, and idiomatic translation [21]) enables us to compare different translations within the same source language. In our experiments, English takes the role of the under-resourced target language. For this exploratory work we feel that understanding the target language gives a deeper insight in the strengths and weaknesses of our algorithms. The English Standard Version (ESV) [5] is a literal English translation of the Christian Bible [3]. Half of the words in the vocabulary occur three times or more in the text, 30.5% have only one occurrence. High word frequencies are suitable for our extraction algorithm since we merge more phoneme sequences in *Step 4* which leads to better error correction as shown in Sec. 4.4. Verses in the ESV Bible are identified by unique verse numbers (such as *Galatians 5:22*), which are consistent with verse numbers in other Bible translations. Based on these numbers, we extracted a parallel and verse-aligned corpus consisting of 30.6k English Bible verses (target language) and 14 written translations of them (Tab. 1).

To generate the target language phoneme sequences, we replaced the words in the ESV Bible with their canonical pronunciations and removed word boundary

markers. Thereby we simulate the output of a perfect phoneme recognizer (0% Phoneme Error Rate) and refrain from dealing with pronunciation variants and phoneme recognition errors. However, we design our algorithms to be robust against recognition errors. The pronunciations were taken from the CMUdict [24] or generated with a grapheme-to-phoneme model trained on it (39 phonemes).

4.2 Evaluation Measures

We measure the quality of the word segmentation (Sec. 2) in terms of **accuracy** [22]. Additionally, we suggest 3 different evaluation measures, which address different aspects of the quality of the extracted dictionary.

Let I be the set of all word IDs in the extracted dictionary $Dict : I \rightarrow {PhonemeSet_{trgt}}^{+}$. We measure the structural quality of $Dict$ by the **Out-Of-Vocabulary rate (OOV)** on a subset of the English ESV Bible. The OOV rate can not be calculated directly since $Dict$ contains word IDs instead of written words consisting of graphemes like in the ESV Bible. Therefore, a mapping between the word IDs and the written words is required. Let V_{trgt} be the vocabulary of the ESV Bible (written words) and $Dict_{ref} : V_{trgt} \rightarrow {PhonemeSet_{trgt}}^{+}$ the reference dictionary with the correct pronunciations. The mapping $m : I \rightarrow V_{trgt}$ assigns each word ID to the written word with the most similar pronunciation.

$$m(n) = \arg\min_{v \in V_{trgt}} d_{edit}(Dict(n), Dict_{ref}(v)) \quad (1)$$

where d_{edit} denotes the edit distance. The set $m(I)$ of matched vocabulary entries in $Dict_{ref}$ is then used to calculate the OOV rate.

While the OOV rate indicates the coverage of $Dict$ on a Bible text, the **Phoneme Error Rate (PER)** reflects the quality of the extracted pronunciations on the phoneme level. It is defined as the average edit distance between

Table 1. Overview of used Bible translations

ID	Language	Full Bible Version Name	Number of running words
bg	Bulgarian	Bulgarian Bible	643k
cs	Czech	Bible 21	547k
da	Danish	Dette er Biblen på dansk	653k
de1	German	Schlachter 2000	729k
de2	German	Luther Bibel	698k
es1	Spanish	Nueva Versión Internacional	704k
es2	Spanish	Reina-Valera 1960	706k
es3	Spanish	La Biblia de las Américas	723k
fr1	French	Segond 21	756k
fr2	French	Louis Segond	735k
it	Italian	Nuova Riveduta 2006	714k
pt1	Portugese	Nova Versão Internacional	683k
pt2	Portuguese	João Ferreira de Almeida Atualizada	702k
se	Swedish	Levande Bibeln	595k
en	*English*	*English Standard Version*	*758k*

the entries in $Dict$ and the closest entry in the reference dictionary $Dict_{ref}$:

$$PER = \frac{\sum_{n \in I} d_{edit}(Dict(n), Dict_{ref}(m(n)))}{|I|} \quad (2)$$

The **Hypo/Ref ratio** indicates how many hypothesis entries in $Dict$ are mapped by m to a single reference entry in $Dict_{ref}$ on average ($|I|$ divided by $|m(I)|$). The higher the Hypo/Ref ratio, the more pronunciations are extracted unnecessarily.

4.3 Which Source Translation Is Favorable?

Fig. 4 shows the distribution of the edit distances between the extracted pronunciations and the closest entries in the reference dictionary (pairs $(n, m(n))$) for each of the 14 translations in Tab. 1. For example, the length of the dark blue bar above the *es3* label shows, that using the Spanish *La Biblia de las Américas* translation, 4,464 of the 21,561 extracted pronunciations (20.7%) contain no or only minor phoneme errors (edit distance lower than 0.1). The translations are sorted by accuracy (descending from left to right). We can observe, that the red bar (interval $[0.1, 0.2)$) is small, because a word has to contain at least 6 phonemes (and 1 phoneme error) to fall into this class and English words are usually shorter. Apart from these side effects, the edit distance usually seems to be approximately uniformly distributed in $[0, 0.6)$, and only a few outliers have higher edit distances. Exceptions are *bg* and *cs*. The red line marks the actual size of the ESV Bible vocabulary. Fig. 5 breaks down the extracted pronunciations by the differently colored absolute number of insertions, deletions, and substitutions. 20% of all entries contain no phoneme error, 50% no more than one error. Only about 30% of all entries contain 3 or more phoneme errors.

We investigate the impact of four factors to our evaluation measures.

- **Δ Vocabulary size**. The difference between the vocabulary size of the source translation and the size of the ESV vocabulary.
- **Δ Average number of words per verse**. The difference between the average verse length in the source translation and in the ESV Bible.

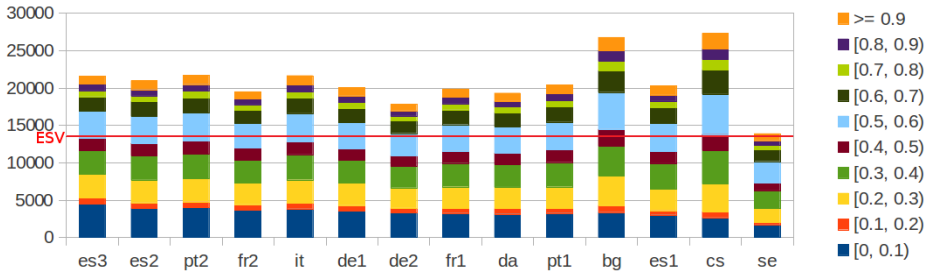


Fig. 4. Distribution of the edit distances between the extracted pronunciations and the nearest entry in the reference dictionary for all 14 source translations

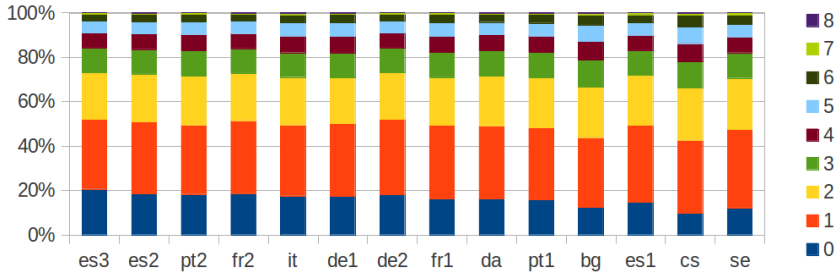


Fig. 5. Distribution of the Levenshtein distance (absolute) between the extracted pronunciations and the nearest entry in the reference dictionary

- **Δ Average word frequency**. The difference between the average number of word repetitions in the source language and in the ESV Bible.
- **IBM-4 PPL**. To measure the general correspondence of the translation to IBM-Model based alignment models, we run GIZA++ [14] with default configuration on the word level and use the final perplexity of IBM Model 4 [4].

Tab. 2 shows the Pearson's correlation coefficient $|r|$ [15] between those four factors and our evaluation measures from Sec. 4.2. Fig. 6 plots some of the point clouds with their regression line. We observe a rather weak linear correlation between the OOV rate and the word segmentation accuracy in Fig. 6 (a) ($r = 0.68$): The better the word segmentation, the closer the extracted and the reference dictionary structurally. The dominant factor for the OOV rate is the IBM-4 PPL (Fig. 6 (b), $r = 0.96$). This suggests, that a literal translation is more important than cross-lingual linguistic dissimilarities. This hypothesis is supported by the wide variance of the evaluation measures between different translations within the same source language: *es3* has 5.5% higher accuracy, 10.7% lower OOV rate, and 3.9% lower PER (absolute) than *es1* since *es3* is a very literal translation [11]. Similar results can be observed for French and Portuguese.There is only a weak linear correlation of the average word frequency with the accuracy (Fig. 6 (c)), but a stronger correlation with the PER. Consequently, frequent word repetitions improve the quality of the extracted pronunciations on the phoneme level since

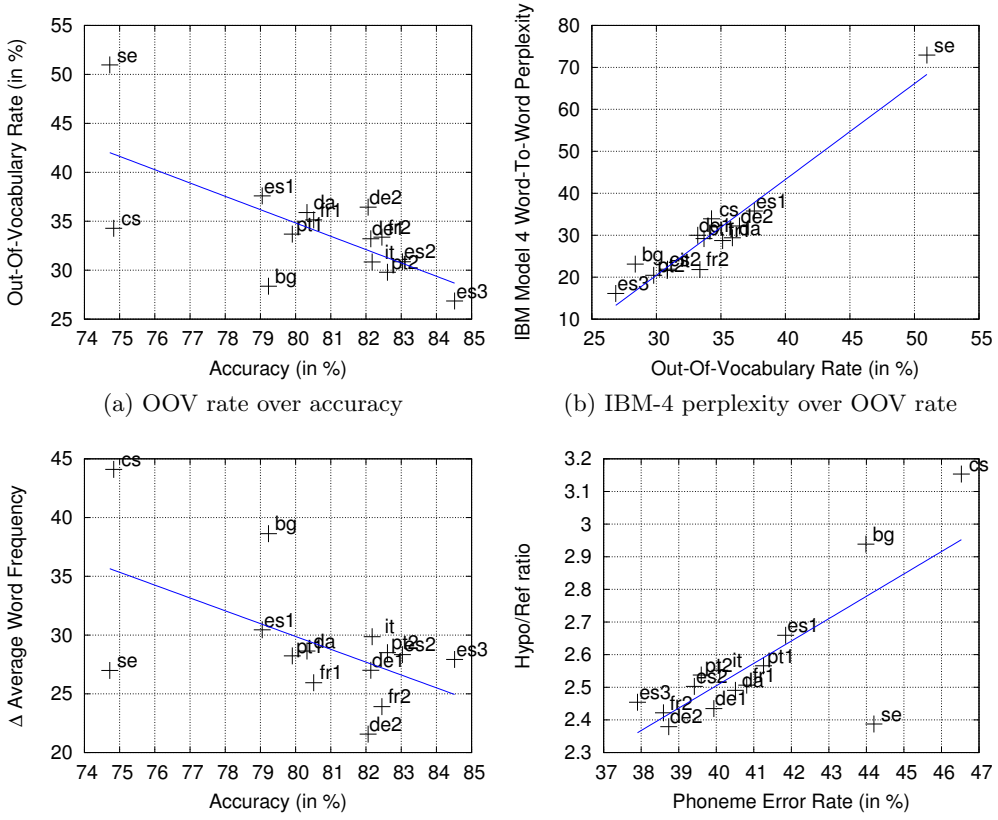
Table 2. Absolute correlation coefficients $|r| \in [0, 1]$ between our evaluation measures and different influencing factors (high $|r|$ - high linear correlation)

$\|r\|$	**Accuracy**	**PER**	**Hypo/Ref ratio**	**OOV rate**
Δ Vocabulary size	0.47	0.71	0.98	0.31
Δ Average number of words	0.59	0.72	0.85	0.06
Δ Average word frequency	0.55	0.79	0.97	0.21
IBM-4 PPL	0.77	0.54	0.10	0.96
PER	0.94	-	-	-
Hypo/Ref Ratio	0.53	0.77	-	-
OOV rate	0.68	0.40	0.24	-

Step 4 in our extraction algorithm in Sec. 3.2 merges many phoneme sequences and can correct errors more effectively. The Hypo/Ref ratio is highly correlated with both the vocabulary size and the average word frequency. This suggests, that *Step 3* in our extraction algorithm needs to be improved: Often one single cluster per source language word is generated, and *Step 4* merges words which are different in the target language. This high correlation also uncovers another point for improvement: Pronunciations extracted from different source language words can not be merged. For example, all three German definite articles are translated to *the*, so there are at least three dictionary entries for *the* alone.

4.4 Which Words Are Extracted Correctly?

This section describes the characteristics of words which are likely to be extracted correctly when the source translation *es3* is used. Experiments with other source translations show similar results. Fig. 7 indicates, that frequently repeated words



(a) OOV rate over accuracy (b) IBM-4 perplexity over OOV rate

(c) Divergence of vocabulary size over accuracy (d) Hypo/Ref ratio over Phoneme Error Rate

Fig. 6. Different influencing factors for the evaluation measures in Sec. 4.2

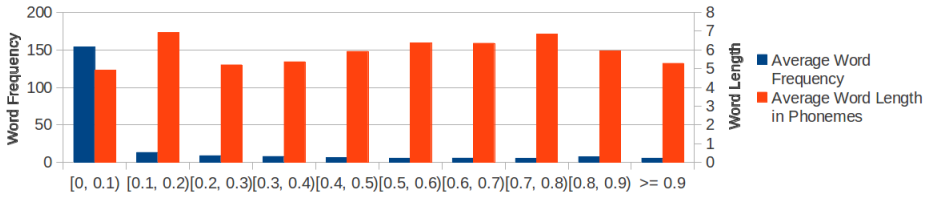


Fig. 7. Average word frequency and number of phonemes per word over the PER (*es3*)

tend to contain no or only minor errors on the phoneme level (blue bar) while there is no such clear correlation with the number of phonemes per word (red bar). A look at some extracted pronunciations reveals two major sources of errors for words with only 1-2 phoneme errors:

1. Single phonemes are added or dropped in the beginning or end of a word because of off-by-one alignment errors:
   - `z f ih s t s` instead of `f ih s t s` (fists)
   - `ih k s t` instead of `f ih k s t` (fixed)
2. Different words with the same stem are merged together:
   - `s ih d uw s ih t` instead of `s ih d uw s t` (seduced) or `s ih d uw s ih ng` (seducing)
   - `ih k n aa l ih jh m` instead of `ih k n aa l ih jh` (acknowledge) or `ih k n aa l ih jh m ah n t` (acknowledgement)

Entries with two phoneme errors or more often contain two words because of missing word boundaries between words often occurring in the same context:

- `w er ih n d ih g n ah n t` (were indignant)
- `f ih n ih sh t ih t` (finished it)

We assume that this kind of errors would not be very critical when using the dictionary in an S2S system since those words are likely to be stuck together as *phrase* later in the training process of the translation model anyway.

4.5 Combining Multiple Translations

In case of several written translations, we first extract the pronunciation dictionary with each source translation separately, and then combine all of them in a single dictionary. To combine the set of dictionaries, we first add the translation tags (i.e. *es3*, *de2*...) to the word IDs to obtain globally unique IDs. Second, we concatenate all dictionaries and remove homophones. Starting out from the *es3* dictionary, we successively combined more dictionaries of other translations ordered descending by the word segmentation accuracy. Fig. 8 suggests, that the OOV rate decreases slightly exponentially with the number of combined translations. At the same time, the Hypo/Ref ratio increases linearly. The PER only increases slightly. Combining all 14 translations results in a dictionary with only 7.9% OOV rate, but more than 9 of 10 dictionary entries are extracted unnecessarily (Hypo/Ref ratio 10.7:1). Such a dictionary is far too noisy for practical

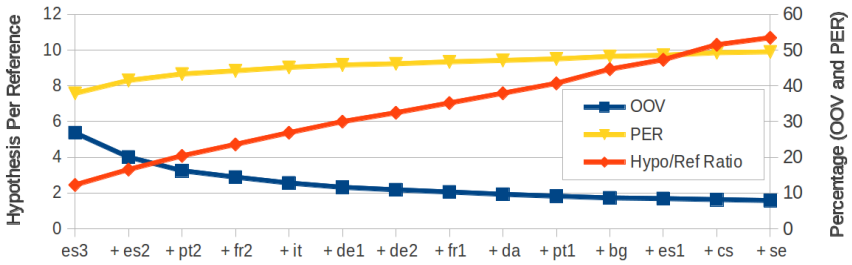


Fig. 8. Evaluation measures over the number of combined source translations

use, but it shows, that experiments with different source translations extract different English words. Therefore, our future work will also focus on how to remove this noise and explore the synergy of multiple translations.

5 Conclusion And Future Work

Using written translations in one or many source languages, we cross-lingually segmented phoneme sequences in a target language using our alignment model *Model 3P* [17]. We proposed a new algorithm for extracting a pronunciation dictionary with word IDs from these segmentations and alignments, which can be used in an S2S system bypassing the written form of a non-written or under-resourced target language. In our exploratory experiments, we extracted English pronunciations by using 14 different translations in 9 languages. With a Spanish translation (*es3*), we built a dictionary for the ESV Bible [5] with 26.9% OOV rate, in which most of the pronunciations contain not more than one wrong phoneme. Combining dictionaries from multiple translations drops the OOV rate to 7.9%, but increases the number of unnecessary entries. This shows, that depending on the used translation, different English words are extracted.

In the future, we plan to enhance our pronunciation extraction algorithm based on the results from Sec. 4.3: *Step 3* needs to be improved to separate pronunciation variants and different words with the same translation more reliably. The algorithm needs to be adjusted to allow merging of pronunciations generated by distinct source language words. Off-by-one pronunciation errors due to alignment errors may be reduced by reinforcing the alignments with the extracted pronunciations after each iteration of our algorithm. Monolingual word segmentation methods as in [9] may give additional hints. When combining multiple dictionaries, a mechanism is to be found to filter accurate entries and benefit from the lower OOV rate while keeping the Hypo/Ref ratio constant. In a next step, we will use a phoneme recognizer to obtain the phoneme sequences. Such a phoneme recognizer can be bootstrapped using recognizers from other languages and adaptation techniques as presented in [23]. Furthermore, we intend to use the extracted dictionaries in a speech recognizer for a truly under-resourced language. The final goal is to build an S2S system without any linguistic knowledge of the target language.

References

1. Achtert, E., Goldhofer, S., Kriegel, H.P., Schubert, E., Zimek, A.: Evaluation of Clusterings–Metrics and Visual Support. In: ICDE (2012)
2. Besacier, L., Zhou, B., Gao, Y.: Towards Speech Translation of Non-Written Languages. In: SLT (2006)
3. Borland, J.A.: The English Standard Version-A Review Article. Faculty Publications and Presentations p. 162 (2003)
4. Brown, P.F., Pietra, V.J.D., Pietra, S.A.D., Mercer, R.L.: The Mathematics of Statistical Machine Translation: Parameter Estimation. Computational Linguistics 19(2), 263–311 (1993)
5. Crossway: The Holy Bible: English Standard Version (2001)
6. Ester, M., Kriegel, H.P., Sander, J., Xu, X.: A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases With Noise. In: KDD (1996)
7. Gollan, C., Bisani, M., Kanthak, S., Schlüter, R., Ney, H.: Cross Domain Automatic Transcription on the TC-STAR EPPS Corpus. In: ICASSP (2005)
8. Gordon, R.G., Grimes, B.F.: Ethnologue: Languages of the World. SIL International, 15th edn. (2005)
9. Johnson, M., Goldwater, S.: Improving Non-Parameteric Bayesian Inference: Experiments on Unsupervised Word Segmentation with Adaptor Grammars. In: HLT-NAACL (2009)
10. Kikui, G., Sumita, E., Takezawa, T., Yamamoto, S.: Creating Corpora for Speech-to-Speech Translation. In: Eurospeech (2003)
11. Lockman: La Biblia de las Américas. `http://www.lockman.org/lblainfo/` (1986), Accessed on 28th February 2013
12. Martirosian, O., Davel, M.: Error Analysis of a Public Domain Pronunciation Dictionary. In: PRASA (2007)
13. Nettle, D., Romaine, S.: Vanishing Voices: The Extinction of the World's Languages. Oxford University Press (2000)
14. Och, F.J., Ney, H.: A Systematic Comparison of Various Statistical Alignment Models. Computational Linguistics 29(1), 19–51 (2003)
15. Rodgers, J.L., Nicewander, W.A.: Thirteen Ways to Look at the Correlation Coefficient. The American Statistician 42(1), 59–66 (1988)
16. Schultz, T., Kirchhoff, K. (eds.): Multilingual Speech Processing. Academic Press, Amsterdam (2006)
17. Stahlberg, F., Schlippe, T., Vogel, S., Schultz, T.: Word Segmentation Through Cross-Lingual Word-to-Phoneme Alignment. In: SLT (2012)
18. Stolcke, A., Konig, Y., Weintraub, M.: Explicit Word Error Minimization in N-best List Rescoring. In: Eurospeech (1997)
19. Stüker, S., Waibel, A.: Towards Human Translations Guided Language Discovery for ASR Systems. In: SLTU (2008)
20. Stüker, S., Besacier, L., Waibel, A.: Human Translations Guided Language Discovery for ASR Systems. In: Interspeech (2009)
21. Thomas, R.L.: Bible Translations: The Link Between Exegesis and Expository Preaching. The Masters Seminary Journal 1, 53–74 (1990)
22. VIM: International Vocabulary of Basic and General Terms in Metrology. International Organization pp. 09–14 (2004)
23. Vu, N.T., Kraus, F., Schultz, T.: Rapid Building of an ASR System for Under-Resourced Languages Based on Multilingual Unsupervised Training. In: Interspeech (2011)
24. Weide, R.: The Carnegie Mellon Pronouncing Dictionary 0.6 (2005)