INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Language Modeling for Speech Analytics in Under-Resourced Languages

Simone Wills[1], *Pieter Uys*[1], *Charl van Heerden*[1], *Etienne Barnard*[1]

[1]Saigen (Pty) Ltd

{simone, pieter.uys, charl}@saigen.co.za, etienne.barnard@gmail.com

Abstract

Different language modeling approaches are evaluated on two under-resourced, agglutinative, South African languages; Sesotho and isiZulu. The two languages present different challenges to language modeling based on their respective orthographies; isiZulu is conjunctively written whereas Sotho is disjunctively written. Two subword modeling approaches are evaluated and shown to be useful to reduce the OOV rate for isiZulu, and for Sesotho, a multi-word approach is evaluated for improving ASR accuracy, with limited success. RNNs are also evaluated and shown to slightly improve ASR accuracy, despite relatively small text corpora.

Index Terms: speech recognition, kaldi, subword modeling, multi-word modeling, lattice rescoring, RNNLM

1. Introduction

Speech analytics is one of the most successful commercial applications of Automatic Speech Recognition (ASR). However, the uptake of speech analytics has largely been restricted to developed-world markets by the fact that available commercial systems cater for well-resourced languages, as opposed to under-resourced languages which dominate the linguistic landscape of less developed markets. This denies both companies, agents and customers in developing market call-centers the clear benefits of speech analytics. Where developing markets have attempted to adopt international speech analytics products in an effort to ensure, for example, legal compliance, it has become clear that this is a double-edged sword: on the one hand, agents are more consistently communicating important legal and regulatory information to customers than they would have been inclined to do so otherwise. On the other hand, they switch over to a well-resourced language such, as English, as they are monitored by a speech analytics system in the well-resourced language. This pattern forces customers to receive and interpret important information in a language for which they often lack the necessary level of comprehension.

The obvious answer to this dilemma is to develop speech analytics in the languages customers prefer, which in turn requires accurate ASR in that language. However, the financial cost of creating the necessary resources is often prohibitive [1]. Moreover, porting speech technology to an under-resourced language poses challenges beyond that of resource acquisition. It often requires innovative and adaptive language and acoustic modeling techniques to address challenging socio-linguistic characteristics of under-resourced languages [1].

One such adaptation is the use of alternative language modeling units, such as subword units, in place of traditional word-based language modeling. This offers a different language modeling approach to be taken in the case of morphologically complex languages. This paper investigates whether any gains can be made through the use of subword and multi-word units, both within an n-gram language model and a Recurrent Neural Network (RNN) model, when applied to isiZulu and Sesotho. These are two under-resourced languages spoken in South Africa, for which traditional language modeling techniques are less effective due to their agglutinative nature.

The paper is structured as follows. Section 2 provides background information on speech analytics and under-resourced languages. Sections 3, 4 and 5 present an overview of the language modeling approaches under evaluation. The experimental setup and results follow in Section 6. The paper concludes with a discussion and summary of findings in Sections 7 and 8.

2. Background

2.1. Speech Analytics in Contact Centres

Speech analytics is the extraction and analysis of information from speech data, with the purpose of providing insight, ranging from speech content to speaker sentiment [2]. It has emerged as a powerful tool for customer contact centres to leverage the vast unstructured speech data at their disposal. While other communication media, such as chat bots, are becoming increasingly popular, contact centres are still the main source of contact for many customers, and a primary channel for delivering services. To maintain a competitive edge in customer service, companies are looking to reduce costs while simultaneously improving the quality of customer interaction [2].

Speech analytics tools cater for these needs by improving operational efficiency, agent monitoring, evaluation, risk mitigation, amongst others. Moreover, it provides the qualitative insights needed to improve and ensure the quality of customer service. However, the performance of speech analytics tools is highly dependent on the underlying ASR technology accuracy, which remains a challenge for under-resourced languages.

2.2. isiZulu and Sesotho

isiZulu and Sesotho are two of South Africa's eleven official languages, and are representative of the two primary branches of Bantu languages in South Africa; the Nguni branch and the Sotho–Tswana branch. Languages within each group are, for the most part, mutually intelligible for native speakers. This makes isiZulu and Sesotho important languages for speech analytics within call-centres, as call-centre agents are driven to negotiate the communicative interaction between themselves and a broad base of linguistically diverse customers.

isiZulu is the most widely spoken first language in South Africa, with a speaker population size equivalent to that of Finish and Danish combined, while Sesotho is comparative to that of Norwegian. Despite speaker population size, they are both considered under-resourced languages. This is not unusual for Non-European languages, even widely spoken ones, which tend to be less well-studied, and thus resourced, than their European counterparts.

Both isiZulu and Sesotho are agglutinative languages, with a highly productive morphological strategy of augmenting word

http://dx.doi.org/10.21437/Interspeech.2020-1586

stems with affixes [3]. They are characterised by an expansive vocabulary [4], similar to the more well-known agglutinative languages Finnish, Turkish and Hungarian. The crucial distinction between isiZulu and Sesotho, from a language modeling perspective, is that orthographically, isiZulu is written conjunctively whereby affixal morphemes are joined to the word stem when written. In contrast, Sesotho is disjunctive, allowing for affixes to be written separately from word stems [5].

2.3. Word-Level N-Gram Modeling

Agglutinative languages are well-known for challenging traditional word-level n-gram language modeling, primarily due to a high number of out-of-vocabulary words (OOVs) and large lexicons. N-gram language models using word units are often chosen for being an effective, count-based, statistical approach which can achieve a high level of accuracy with limited parameter requirements. However, these models perform poorly at modeling previously unseen words [6]. This problem is exacerbated for under-resourced languages which lack access to the large training corpora generally required to produce accurate and statistically relevant models [6, 7].

For isiZulu, the characteristically large vocabulary is associated with data sparseness, resulting in poor n-gram parameter estimations and high OOV rates, directly influencing word error rate (WER) [8]. Sesotho unintentionally benefits from its disjunctive orthographic system which reduces the incidence of OOVs and produces a smaller vocabulary. However, it does increase the frequency of mono- and bimorphemic words which also leads to poor n-gram parameter estimation. This is a result of the limited contextual information available in each n-gram. Furthermore, in the case of isiZulu and Sesotho as under-resourced languages, the paucity of resources discourages data selectivity which lends itself to the inclusion of sub-optimal or noisy training data. Using alternative language modeling units potentially addresses these challenges.

3. Subword Language Modeling

Subword modeling is a common approach used to reduce OOVs and address the data sparsity associated with very large vocabularies. Subword units are obtained by splitting words into smaller parts. Depending on the segmentation method, subword units may be linguistic units such as phonemes, morphemes or syllables, or a morpheme-like unit if a data-driven approach is used [9]. The use of subword units reduces vocabulary size, while still allowing for the recognition of larger units through unit concatenation. This significantly reduces model complexity, improving the model's efficiency [10]. The concatenation of subword units also enables the model to create an extensive vocabulary, thus reducing the number of OOVs. [10] claim an absolute WER reduction of over 5% by using subword modeling, comparative to other publications on the same datasets.

3.1. Segmentation methods

Two data-driven segmentation methods are compared; Morfessor [11] and a consonant-vowel based syllable approach [12].

a. Morfessor is a tool used for morphological segmentation that uses an unsupervised algorithm based on Minimum Description Length principle. The goal is to find language units which resemble the surface forms of morphemes [10]. Morfessor has been a popular method for segmenting agglutinative languages such as Finnish and Estonian, for ASR [10].

b. Syllable based approach employs an algorithm that creates syllables by splitting words at consonant clusters. "Valid" consonant clusters are learned by observing consonant clusters at the beginning and end of words in the training vocabulary. Where a consonant cluster consists of more than one consonant C_1C_2, the cluster is split at that point where C_2 is as big as possible and valid, while C_1 is also valid. If no such split exists, the cluster is split to maximize the length of C_2 while still valid, otherwise the split is applied before the cluster.

3.2. Boundary markers

Recognizing OOV's using subword approaches requires the ability to reconstruct words from the subword units. One such approach entails adding word boundary tags to subwords to indicate where it was split from a word [10]. A marker is applied on either side of the subword unit if the subword occurred in-between other subwords. For example, "two slippers", split into the subwords [two, slipp, er, s] would be rewritten as "two slipp+ +er+ +s". To reconstruct words from these boundary marked subwords, all +[space]+ sequences in the one-best path are deleted, resulting in "words".

To ensure that only valid subword sequences are recognized, [10] creates a special subword L-FST, where position dependent phones are applied to subword units as they would have been in the original word; in other words, the subword "slipp+"'s pronunciation would be /s_B l_I @_I p_I/ instead of /s_B l_I @_I p_E/ (the final phone is still a word-internal phone, and not a word-end phone). The corresponding L-FST thus ensures that only valid sequences of subwords which can all join together to form words, are recognized during decoding.

In order to find the corresponding phonetic pronunciations for these subword units, the word pronunciation dictionary was G2P-aligned, and the corresponding pronunciations extracted.

4. Multiword Language Modeling

Due to the disjunctive orthography used by Sesotho, many of the lexical units are mono- or bimorphemic. As such, language modeling for Sesotho faces the same challenges as using morphemic language modeling units. In ASR, short lexical units are more frequently misrecognised than longer words [13]. Acoustically, they are more readily confused and in terms of n-gram language modeling, the span of the language model is shortened, reducing context. A solution to reduce the confusion caused by short morphemes is to merge these morphemes with neighbouring lexical units to create longer, compounded lexical units [13]. This approach has been shown to improve ASR performance in various tasks [14, 15, 16], although [13] found that a word-based model still outperformed their morpheme- based models with concatenation.

4.1. Lexical Sequence Selection

Different measures have been proposed in literature for selection of the lexical units to merge. Based on a comparison with the use of mutual information (MI) as a measurement, whereby pairs are chosen to maximise the MI, [9] suggest selecting pairs based on the product of their direct and reverse bigram probability. [9] achieved the best results by concatenating pairs with a probability and frequency count greater than a set threshold. A similar approach was used by [17] who based their selection on the geometric average of the direct and reverse bigrams. Others [13] have used the frequency of the lexical sequences in the training data, selecting sequences above an optimal frequency threshold. Lexical unit size (as determined by character, consonant or syllable count) under a set threshold has also been used as a selection criterion [18]. In this paper, bigram frequency and lexical unit size were used as the main selection criteria.

Multi-word tokens were created by concatenating the final selected bigrams and adding an identifying tag to the concatenated bigrams. The identifier aids the recovery of the original word sequence during post-processing. The new multi-word tokens are added to the vocabulary and used to replace their respective bigram sequences in the training text. The multi-word tokens are also added to the dictionary and assigned the combined pronunciation of the underlying words.

4.2. Phone Modeling Adaptations

The changes in word boundaries caused by word concatenation need to be adjusted for in the L-FST, because of the position dependent phone approach in Kaldi, described in Section 3.2. To ensure multi-word pronunciations are modeled correctly, the lexicon is re-generated with the multi-word phone-dependent sequences which reflect the phone positions as they are in the original words. Silence modeling is also affected by changes to the word boundaries. Kaldi permits optional silences on word boundaries, but not within words [10]. To preserve the phonetic modeling of the original words which make up each concatenation, optional silences are added for the multi-words where the original word boundaries lie in the L-FST.

5. RNN Language Model

RNNs, and more recently BERT [19], have improved the state-of-the-art in NLP significantly. In neural language models words are modeled in a continuous vector space, in which the spatial proximity of words automatically correlates with semantic similarity. This allows the model to infer word relationships beyond those present in training data [20]. This gives these LMs the ability to better generalise than traditional n-grams, leading to state-of-the-art results on many challenging NLP tasks [21].

RNNs, in particular, have been used for some time to train LMs, and are typically applied by rescoring lattices. While RNNs typically require large amounts of data, we were interested in evaluating the potential benefit of RNNs when using subword-like features [22] for morphologically complex under-resourced languages. This is achieved by representing each word w as a bag of character n-grams, where special boundary symbols $<$ and $>$ are added at the word boundaries to distinguish prefixes and suffixes from other character sequences. For example, for the word "hello" the character n-gram are: <he, hel, ell, llo, lo>, and the special sequence <hello> [22].

6. Experiments and Results

A common problem in under-resourced domains is having insufficient amounts of text for language modeling. For this reason, we report results for two scenarios: ideally having some in-domain text for training LMs, versus having only out-of-domain text such as government documents available.

6.1. Datasets

The two corpora used for training language models and evaluation of the subsequent ASR systems are summarized in Table 1. All text was text normalised by removing punctuation, lowercasing, and converting digits to spoken numbers. The OOV rates shown in Table 1 are significantly higher than one would see for non-agglutinative languages. An OOV rate of 1-2% has been reported for a 64K-word English vocab based on the Wall Street Journal corpus, and 3-4% for the same size Spanish vocab drawn from the Spanish Broadcast News corpus [23]. The difference in OOV rate between isiZulu and Sesotho reinforce the key role their written systems play in language modeling. The global increase in OOV rate for the out-of-domain test data is expected given the dissimilarity to the training data which represents a highly specific domain.

6.1.1. SASAL Corpus

The SASAL Corpus is an in-house speech and text corpus collected as part of the Speech Analytics for South African Languages (SASAL) project being run by Saigen (Pty) Ltd, with support from the South African Department of Sports, Arts and Culture (DSAC). The corpus mainly consists of narrow-band conversational speech in Afrikaans, Sesotho and isiZulu, sourced from call-centre audio. The Afrikaans and Sesotho data includes supplementary speech from broadcast news approximate to the speech style of the call-centre audio. Subsets of the Sesotho and isiZulu transcriptions were used as training text for the respective language models.

6.1.2. NCHLT CTexT Corpora

The second set of language modeling training text was the National Centre for Human Language Technology and Centre for Text Technology (NCHLT CTexT) isiZulu and Sesotho text corpora [24][1]. This is text sourced from South African government documents produced by various language units and crawled from gov.za websites.

Table 1: *Training text corpora vocabulary size, number of words and % of words in the evaluation text that are OOV.*

Corpus	Subset	Vocab	# Words	%OOV
SASAL	isiZulu	106 629	1 179 964	38,04
	Sesotho	16 655	472 239	19,38
NCHLT CTexT	isiZulu	234 567	2 256 091	50,48
	Sesotho	37 070	1 846 259	34,63

6.2. Speech Recognition System

The ASR systems were built using Kaldi [25]. For the acoustic models, speaker-dependent Gaussian Mixture model triphone models were trained in order to generate phone alignments. These alignments were then used to train Factored Time Delay Neural Network acoustic models [26]. The Sesotho acoustic models were trained on 55 hours of the SASAL speech corpus, while the isiZulu model was trained on the 55 hour SASAL as well as and additional 250 hour in-house corpus. A combination of proprietary pronunciation dictionaries and the NCHLT Sesotho and isiZulu dictionaries provided the pronunciations for the lexicon. For words not in the dictionaries, a joint-sequence model was used to performed language identification and the pronunciations were then generated by a grapheme-to-phoneme model. The ASR systems were evaluated on a 5-hour held-out set from the SASAL corpora, for each language.

6.3. Baseline N-gram Models.

Two 3-gram isiZulu and two 5-gram Sesotho word-level baseline language models (henceforth WB) were trained, one on each corpus. As expected, both the Sesotho and isiZulu ASR

[1] Additional language resources: https://repo.sadilar.org/

system using the NCHLT CTexT language model have a higher WER than the systems using SASAL training text. This is a result of the domain dissimilarity between the NCHLT training text and the SASAL evaluation set. The results of all the isiZulu ASR models are given in Table 3 and the results for the Sotho models are given in Table 4.

6.4. Subword N-gram Models

Four subword 3-gram models were trained in total; two using morfessor to create the subword units (henceforth SWM), and the other two using the novel syllable based algorithm (henceforth SWS) described in Section 3.1.1. For each segmentation approach, a model was trained on the SASAL corpus and a model was trained on the NCHLT corpus.

6.4.1. *Morfessor vs. Syllable Segmentation*

The syllable algorithm generated more subword units on both text corpora, than Morfessor. When trained on the SASAL corpus, the SWS model performed marginally (0.1-0.2%) better than both the word-level model and the SWM model at 54.7%. However, amongst the NHCLT CTexT trained models, the WB model achieved the best WER of 68.5%. Although, SWS model still performed better than the SWM model.

Table 2: *Number of subword units by segmentation method.*

Segmentation	Corpus	# Subword Units
Morfessor	NCHLT	2 093
	SASAL	2 166
Syllable	NCHLT	4 297
	SASAL	4 336

6.5. Multiword N-Gram Models

Two multi-word 5-gram models were trained; one on each corpus. Using bigram counts based on the training data, the 40 most frequent Sesotho bigrams, composed of words 5 characters or less, were selected for concatenation if the multi-word unit produced by the concatenation did not already exist as a unigram in the data. Similar to [13] the multi-word models did not improve upon the WB results, except for a small win (0.02%) on the SASAL corpus comparing the WB and MW RNNLMs.

Table 3: *WERs for isiZulu using different LM approaches. Word baseline (WB), RNNLM and RNNLM with character n-grams (RNN Char) results are shown, and subword modeling approaches using syllables (SWS) and Morfessor (SWM). % R-OOV refers to the percentage of OOVs recognized correctly.*

	SASAL		NCHLT	
Method	WER	% R-OOVs	WER	% R-OOVs
WB	54.9	-	68.5	-
WB RNN	54.2	-	67.6	-
WB RNN Char.	54.2	-	67.5	-
SWS	54.7	19.74	68.9	18.58
SWS RNN	53.6	19.10	68.3	14.69
SWM	54.8	22.98	69.1	17.41
SWM RNN	53.9	22.47	68.6	16.76

6.6. RNN Language Models

The RNNLMs were trained for 30 epochs with an embedding size of 200 and regularization parameter of 0.005. For isiZulu, adding subword units increases the RNNLM's gain from 0.7% to 1.1% on the SASAL corpus. On the SASAL data, the addition of n-gram character modeling to the RNNLM architecture improves performance for both languages. Conversely, on the Sesotho NCHLT corpus, the RNNLM performance sees a significant deterioration with an increase of 1.61%.

Table 4: *WERs for Sotho using different LM approaches. The word baseline results (WB) are shown in comparison to the results with RNNs, with and without character ngrams (RNN Char), and the Multiword (MW) modeling approach.*

Method	SASAL	NCHLT
WB	33.91	48.69
WB RNN	33.02	49.95
WB RNN Char.	33.01	51.56
MW	34.09	48.92
MW RNN	32.98	50.22

7. Discussion

Both isiZulu subword modeling approaches were useful in recognising OOVs. However, no statistically significant improvement in WER was observed, which is similar to what was observed for Latvian [8], but contrary to results obtained for Turkish [27]. While a significant part of the Latvian corpus also contained spontaneous speech from different sources, the Turkish corpus consisted of Broadcast News audio, which is typically an easier ASR task, as the speech is typically better enunciated. We have also seen some evidence of words with significant vowel reduction in isiZulu being recognized "as pronounced" by the subword system, ie, with missing syllables. This is currently considered an error by our scoring system, but could partly explain why we do not see WER improvements.

A minimal gain was observed in Table 3 and 4 on the SASAL data when rescoring word-lattices with RNNs, even though the LM corpora were relatively small compared to what is typically used to train these models. Further gains were observed on the in-domain LM experiments when rescoring the subword lattices with subword RNNLMs. For Sotho, gains where made when rescoring the word lattices with RNNLMs trained on the in-domain corpora, however performance deteriorated significantly when rescoring with RNNLMs trained on the out-of-domain corpora. It can be noted that the WERs remain high, particularly for isiZulu given the additional 250 hour in-house training corpus. Additionally, the RNN gains were smaller than anticipated. This could be attributed to noisy data which includes a large degree of code-switching - a key characteristic of South African call-centre speech. These matters should be investigated in future work.

8. Conclusion

In this paper, we investigated different language modeling approaches to improve the WER and OOV rate of two agglutinative South African under-resourced languages. The use of subword modeling was shown to significantly reduce the OOV rate in isiZulu, which is orthographically conjunctive, while multi-word modeling was not useful to significantly reduce the WER for the disjunctively written Sesotho, especially for out-of-domain LMs. RNNLMs were found to slightly reduce WER, even when trained on small amounts of in-domain data.

9. Acknowledgements

This work was sponsored by the South African Department of Sports, Arts and Culture (DSAC).

10. References

[1] Laurent Besacier et al. "Automatic speech recognition for under-resourced languages: A survey". In: *Speech Communication* 56 (Jan. 2014), pp. 85–100.

[2] Scott Scheidt and QB Chung. "Making a case for speech analytics to improve customer service quality: Vision, implementation, and evaluation". In: *International Journal of Information Management* 45 (Apr. 2019), pp. 223–232.

[3] Sonja E Bosch and Laurette Pretorius. "A computational approach to Zulu verb morphology within the context of lexical semantics". In: *Lexikos* 27.1 (2017), pp. 152–182.

[4] Peter Smit. "Modern subword-based models for automatic speech recognition". PhD thesis. Aalto, Finland: Department of Signal Processing and Acoustics, Aalto University, 2019. URL: https://aaltodoc.aalto.fi/handle/123456789/38073.

[5] Winston N Anderson and Petronella M Kotzé. "Finite state tokenisation of an orthographical disjunctive agglutinative language: The verbal segment of Northern Sotho." In: *Proc. LREC*. Genoa, Italy, May 2006, pp. 1906–1911.

[6] Tomas Mikolov et al. *Subword language modeling with neural networks*. Tech. rep. Brno, Czech Republic: Faculty of Information Technology, Brno University of Technology, 2012.

[7] Sopheap Seng et al. "Which units for acoustic and language modeling for Khmer automatic speech recognition?" In: *Proc. SLTU*. Hanoi, Vietnam, May 2008, pp. 33–38.

[8] Askars Salimbajevs and Jevgenijs Strigins. "Using sub-word n-gram models for dealing with OOV in large vocabulary speech recognition for Latvian". In: *Proc. NODALIDA*. Vilnius, Lithuania, May 2015, pp. 281–285.

[9] Kadri Hacioglu et al. "On lexicon creation for Turkish LVCSR". In: *Proc. Eurospeech*. Geneva, Switzerland, Sept. 2003, pp. 1165–1168.

[10] Peter Smit, Sami Virpioja, Mikko Kurimo, et al. "Improved Subword Modeling for WFST-Based Speech Recognition". In: *Proc. Interspeech*. Stockholm, Sweden, Aug. 2017, pp. 2551–2555.

[11] Mathias Creutz and Krista Lagus. "Unsupervised discovery of morphemes". In: *Proc. SIGPHON*. Vol. 6. Philadelphia, July 2002, pp. 21–30.

[12] Marelie Davel et al. "Exploring minimal pronunciation modeling for low resource languages". In: *Proc. Interspeech*. Dresden, Germany, Sept. 2015, pp. 538–542.

[13] George Saon and Mukund Padmanabhan. "Data-driven approach to designing compound words for continuous speech recognition". In: *IEEE Transactions on Speech and Audio Processing* 9.4 (May 2001), pp. 327–332.

[14] Christel Beaujard and Michéle Jardino. "Language modeling based on automatic word concatenations". In: *Proc. Eurospeech*. Budapest, Hungary, Sept. 1999, pp. 1563–1566.

[15] Michael Finke and Alex Waibel. "Speaking mode dependent pronunciation modeling in large vocabulary conversational speech recognition". In: *Proc. Eurospeech*. Rhodes, Greece, Sept. 1997, pp. 2379–2382.

[16] Egidio P Giachin. "Phrase bigrams for continuous speech recognition". In: *Proc. ICASSP*. Detroit, MI, USA, May 1995, pp. 225–228.

[17] Mijit Ablimit, Askar Hamdulla, and Tatsuya Kawahara. "Morpheme concatenation approach in language modeling for large-vocabulary Uyghur speech recognition". In: *Proc. Oriental COCOSDA*. Hsinchu, Taiwan, Oct. 2011, pp. 112–115.

[18] Oh-Wook Kwon. "Performance of LVCSR with morpheme-based and syllable-based recognition units". In: *Proc. ICASSP*. Vol. 3. Istanbul, Turkey, June 2000, pp. 1567–1570.

[19] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". Version 2. In: *ArXiv* abs/1810.04805 (May 2019).

[20] Tomas Mikolov et al. "Distributed representations of words and phrases and their compositionality". In: *Proc. NIPS*. Vol. 2. Harrahs and Harveys, Lake Tahoe, USA, Dec. 2013, pp. 3111–3119.

[21] Hlib Babii, Andrea Janes, and Romain Robbes. *Modeling Vocabulary for Big Code Machine Learning*. Apr. 2019. URL: arxiv.org/abs/1904.01873.

[22] Piotr Bojanowski et al. "Enriching word vectors with subword information". In: *Transactions of the Association for Computational Linguistics* 5 (2017), pp. 135–146.

[23] William Byrne et al. "Large vocabulary speech recognition for read and broadcast Czech". In: *Proc. TSD*. Heidelber, Berlin, Sept. 1999, pp. 235–240.

[24] Roald Eiselen and Martin J Puttkammer. "Developing Text Resources for Ten South African Languages." In: *Proc. LREC*. Reykjavik, Iceland, May 2014, pp. 3698–3703.

[25] Daniel Povey et al. "The Kaldi Speech Recognition Toolkit". In: *Proc. ASRU*. Hilton Waikoloa Village, Big Island, Hawaii, USA, Dec. 2011.

[26] Daniel Povey et al. "Semi-Orthogonal Low-Rank Matrix Factorization for Deep Neural Networks". In: *Proc. Interspeech*. Hyderabad, India, Sept. 2018, pp. 3743–3747.

[27] Haşim Sak, Murat Saraclar, and Tunga Güngör. "Morphology-based and sub-word language modeling for Turkish speech recognition". In: *Proc. ICASSP*. Istanbul, Turkey, Jan. 2010, pp. 5402–5405.