Pronunciation Dictionary Development in Resource-Scarce Environments

Marelie Davel and Olga Martirosian

Human Language Technologies Research Group,
Meraka Institute, CSIR, South Africa.
mdavel@csir.co.za, olga.martirosian@gmail.com

Abstract

The deployment of speech technology systems in the developing world is often hampered by the lack of appropriate linguistic resources. A suitable pronunciation dictionary is one such resource that can be difficult to obtain for lesser-resourced languages. We design a process for the development of pronunciation dictionaries in resource-scarce environments, and apply this to the development of pronunciation dictionaries for ten of the official languages of South Africa. We define the semi-automated development and verification process in detail and discuss practicalities, outcomes and lessons learnt. We analyse the accuracy of the developed dictionaries and demonstrate how the distribution of rules generated from the dictionaries provides insight into the inherent predictability of the languages studied.

Index Terms: pronunciation dictionaries, dictionary verification, resource-scarce, bootstrapping, Southern Bantu languages.

1. Introduction

Spoken dialogue systems (SDSs) can assist with information dissemination in the developing world, where alternative infrastructure is typically limited, literacy low and language diversity high. When an SDS requires automatic speech recognition (ASR) or text-to-speech (TTS) components, a number of basic linguistic resources are required, including large annotated audio corpora. For the lesser-resourced languages, these linguistic resources typically do not exist, and even word lists, phone sets and pronunciation dictionaries can be difficult to come by. In addition, as linguistic expertise in these languages is often not readily accessible, the linguistic resource development process presents a significant challenge to researchers active in the field of spoken language technology for development (SLT4D).

In South Africa, one of the aims of the Lwazi project was to create basic linguistic resources in all eleven of South Africa's official languages, and to make these freely and easily available. (All resources can be downloaded from http://www.meraka.org.za/lwazi.) These linguistic resources were used to develop speech technology (ASR and TTS) systems supporting an SDS in the government service delivery domain. In this paper we describe one category of resources developed during this project, namely, a set of electronic pronunciation dictionaries suitable for integration in speech technology systems. We also define a general process for the development of such dictionaries in resource-scarce environments, which we hope will be useful to other SLT4D researchers encountering similar challenges.

2. Background

Many electronic pronunciation dictionaries were initially created as digital versions of similar printed dictionaries. Classical printed pronunciation dictionaries typically only list word base forms, and for each word base form its 'standard' pronunciation. Modern pronunciation dictionaries are often developed using the semi-automated approach of bootstrapping [1, 2, 3], can include multiple word forms and variants, and may be specialised for ASR, TTS or other purposes. Bootstrapping approaches all require a letter-to-sound formalism: when trained on the available dictionary, additional entries are predicted and can be verified efficiently by a human verifier. Prior research in this area includes the development of tools for automated error verification [4], the human factors to consider when bootstrapping pronunciation dictionaries [5] and a significant body of work related to the extraction of efficient letter-to-sound rules.

For the purposes of the Lwazi project, pronunciation dictionaries were required for all of South Africa's eleven official languages. Only for English were prior dictionaries available. The ten languages for which no electronic pronunciation dictionaries were available at the start of the project include the Southern Bantu languages: isiZulu (zul), isiXhosa (xho), isiNdebele (nbl) and siSwati (ssw), all from the Nguni family; Sepedi (nso), Setswana (tsn) and Sesotho (sot), from the Sotho-Tswana family; Xitsonga (tso) from the Tswa-Ronga family; and Tshivenda (ven) from the Venda family. One Germanic language, Afrikaans (afr) is also considered resource scarce. The largest language, isiZulu, is used as home language by approximately 10.7 million speakers, while the smallest language, isiNdebele, has a home language population of only 700,000 speakers [6].

3. Dictionary development process

The dictionary development process described here involves three groups of individuals: speech technologists familiar with speech technology but possibly unfamiliar with the target language, linguists with phonological and phonetic training in the target language but not necessarily a first language speaker of the language or dialect in question, and dictionary developers who are first language speakers of the target dialect but may have limited linguistic training and no experience in speech technology development. The aim of the process is to develop 5,000-word generic dictionaries that can later be extended (with additional words) and specialised to different applications, as required. The process relies heavily on the DictionaryMaker tool [7] and consists of a preparation phase (followed by an initial verification) and two main development phases (followed by intermediate and final verification). During the dictionary preparation phase, the following are addressed:

Selecting a dictionary developer. Dictionary developers are selected based on both their dialect and their ability to perform the task. Only first language speakers of standard dialects are considered: preferably speakers who have remained within a specific geographical region for their entire life. For this set of dictionaries, all developers were between 22 and 37, with an equal split across genders. Typically more than one developer would be selected to work on a single dictionary. Developers with linguistic training are preferred, but such individuals were not available for all the languages studied.

Phoneme set development. A consolidated phoneme set across all the languages was developed in a parallel process through interaction with language-specific linguists. The initial phoneme set was refined extensively throughout the dictionary development process. During dictionary preparation, each of the phonemes is recorded by the dictionary developer. Phonemes are recorded clearly articulated, with some of the inter-phone differences slightly exaggerated. Where necessary, consonants are followed by a short schwa, in order to increase their audibility. Unless any meta-aspect such as duration or tone is an important marker, phonemes are required to be more or less equal in duration and tone. Any silence is cut from the start and end of the phoneme, and all phonemes in the set are normalised with regard to amplitude. This ensures a set of consistent phonemes during training.

Word list development. Language specific textual data is obtained from a wide variety of sources (publishers, the Internet) and automatically cleaned, as far as possible. From this textual data, a set of the 7,000 most frequently occurring tokens is selected. (Even though only 5,000 words are required, it is expected that this list will contain a significant proportion of invalid words.)

DictionaryMaker set-up. The DictionaryMaker tool is initialised with the word list and phone set, and training is provided to the dictionary developer in the use of the tool. An important part of the training relates to ensuring that the dictionary developers understand both the process (as described next) and the phoneme set. Dictionary developers are required to describe the difference between closely related phonemes, and develop a set of personal reference words: clear examples where either the one phoneme or the other would be required.

Calibration. Where uncertainty exists with regard to the ability of a developer to perform the development task accurately (for example, if a linguistic specialist is not available to review work), it is recommended that an initial calibration phase is employed, whereby more than one developer develops a small set of words independently, according to the protocol described below. Results are compared, the developers are asked to discuss discrepancies, and the process repeated with another set of words until the results from one or both developers agree closely with the consensus dictionary. Only then is the formal development process initiated.

3.1. Dictionary development protocol

The actual dictionary development process is similar during each stage of development, and follows the following protocol:

Verifying the validity of the word itself. Before the pronunciation of the word is considered, it is first evaluated for correctness: foreign words, proper nouns, partial words, spelling errors or words missing diacritic symbols are all marked as invalid. When a word is marked as invalid, its pronunciation is not captured.

Verifying the pronunciation of the word. For a valid target-language word, the predicted pronunciation of the word is played ('sounded out') by the system, using the pre-recorded phonemes. (TTS systems are not used as these introduce additional artifacts, and accurate TTS systems cannot be developed prior to the first dictionary being built.) Developers are required to make use of audio assistance whether this is their preference or not, as this has previously been shown to improve quality [5]. Developers edit this pronunciation and either indicate that it is now 'correct' or, in exceptional cases, that they are 'uncertain' of the way a word should be pronounced.

Marking ambiguous words. For ambiguous words (words that have more than one variant) an option is provided to capture one of the variants, but also flag that additional variants are possible. As only one variant is used during letter-to-sound rule extraction, additional variants are not captured immediately, but may be added during post-processing.

Time limits. Dictionary developers are encouraged not to work for longer than 30 minutes in one development session, and to take at least a 10-minute break between sessions.

Quality verification. Once a development session is completed, the dictionary developer is asked to review the pronunciations provided during the previous session, specifically paying attention to exceptional pronunciations automatically flagged by the system as 'possible errors'.

3.2. Dictionary verification protocol

The dictionary verification protocol was designed to monitor the dictionary development process as well as analyse final results. The aim is to eliminate as much human error as possible.

3.2.1. Initial Verification

Initial verification is performed directly after dictionary setup, when the word list, graphemes, phonemes and recorded phoneme sounds have all been captured. These are analysed independently from the linguists overseeing the setup and the developers of the specific dictionary. The dictionary is verified for consistency until consensus is reached between the speech technologists, linguists and developers on all the above items.

The specific errors that were identified during initial verification of the Lwazi dictionaries mainly relate to misunderstandings with regard to the phoneme set, the inclusion of phonemes from similar languages and the recording of closely-related phonemes with indistinguishable sounds. Families of languages exist in South Africa and words are easily shared between similar languages, blurring the distinctions in the phoneme sets across languages. Also, when pronounced in isolation (outside of word context), it can be extremely difficult to produce the correct sound. These errors are important to correct immediately, as they can confuse the dictionary developer and cause core errors in the dictionary.

3.2.2. Intermediate Verification

During the intermediate development phase, the pronunciation dictionary is built by annotating 2,500 words from the word list. This is a crucial verification stage for identifying errors, as the pronunciation dictionary is now sufficiently large to analyse systematically. Three main items are verified: (1) whether all phonemes are being used (2) whether sufficient samples of all graphemes are included in the dictionary and (3) whether analysis indicates any systematic errors. During analysis, a letter-to-sound alignment of the dictionary is performed and misalignments flagged. An initial detection of "possible errors" is performed (see below) and severe or systematic errors identified by the speech technologists and corrected by the developers. Possible adaptation of the phoneme and grapheme sets in collaboration with the linguists may now also be required.

3.2.3. *Final Verification*

During the final development stage, the pronunciation dictionary is completed to 5,000 words. A very thorough verification process is followed prior to final acceptance of the dictionary. The full dictionary is examined, words requiring rework are identified, and the process repeated multiple times.

In addition to the verification mentioned above, a set of letter-to-sound rules is extracted and analysed. The dictionary is aligned using Viterbi alignment, and rules extracted using the Default&Refine algorithm[8]. The rules extracted by this algorithm are ordered according to the number of words to which the rule is applied. By extracting those words in which rare pronunciations occur, possible errors can be isolated. For each one of the anomalous rules, word behaviour is analysed and systematic patterns identified by the speech technologists. All outlier behaviour is reported for re-verification.

The developers analyse the list of exceptional words and differentiate between true linguistic exceptions and erroneous pronunciations, editing the erroneous pronunciations in the dictionary. This process is repeated until all the exceptional pronunciations generated by the verifiers can be accounted for as true linguistic exceptional pronunciations, and are so accepted by the relevant linguists. For illustration, the isiNdebele pronunciation dictionary underwent 11 verification iterations before being accepted. During the first iteration, 19 problematic patterns were identified, affecting over 100 words. During the second iteration, 8 additional patterns were identified, affecting 13 words. During each iteration words were corrected, which allowed new exceptional pronunciations to be identified. At the final verification, all exceptional pronunciations were accounted for as authentic. This was the general pattern during the final verification stage, with the number of iterations ranging from as few as 5 to as many as 15 for different languages.

4. Categorisation and analysis of errors observed

The verification of multiple pronunciation dictionaries reveals a number of systematic errors that are likely to occur.

4.1. Similar sound confusion

Fairly similar phonemes often exist across languages. For example, the languages isiNdebele, Sepedi, Setswana and Sesotho all contain the four sounds /tS_>/, /tS_h/, /ts_h/ and /ts_>/ (using extended SAMPA notation). These sounds describe the alveolar and post-alveolar affricates in their aspirated and ejective forms. As these sounds are phonetically very similar, it can be difficult to annotate each one correctly. Fortunately, the techniques used to find possible errors are very successful in identifying this type of error. Similar confusion was observed between the /o/ and /O/ and between the /e/ and /E/ for the Sotho languages with developers confusing vowel quality and vowel tone – these errors were more difficult to identify, as the correct pronunciation is not systematically predictable (without morphological analysis and semantic knowledge).

4.2. Compound sounds vs individual sounds

Many complex sounds can also be considered as a compound of separate sounds, for example affricates (a stop released as a fricative), diphthongs (two vowels with a smooth transition) and other double articulations. Whether a word should be annotated with two individual phonemes or a single compound sound is not always a point of agreement between linguists and developers. Linguists tend to lean toward the existence of compound sounds, while developers find that the distinction between the two options may not be required, and select either the one or the other arbitrarily.

Once the requirement for a compound sound has been confirmed, the dictionary developer is requested to take care where the distinction between individual and compound sounds needs to be made. Even though the verification process is typically very successful in identifying inconsistencies of this nature, it is suggested that the choice to include compound sounds is evaluated critically during phoneme set development (and again during further verifications) as unnecessary compound sounds complicate the development process.

4.3. Missing diacritics, misspelt words

A pronunciation dictionary for a specific language is built from a word list that may contain incorrect spellings and borrowed words from other languages. (Accurate spell checkers typically do not exist for resource-scarce languages.) These errors tend to confuse developers and result in inconsistent letter-to-sound rules, flagging large numbers of possible errors. Even though explicitly requested not to, annotators tend to attempt to correct a word via its pronunciation rather than marking it as invalid. This is especially the case where words include slight spelling errors or missing diacritics. Depending on the extent of difference between the word and the pronunciation and the frequency of invalid words being annotated as correct, the effect of these words on the dictionary can be severe. This point should be emphasised during both training and development.

4.4. Differences of opinion

Even linguists specialising in pronunciation systems do not always agree on conventions to use when developing phoneme sets or describing pronunciations. This problem is exacerbated where a single linguist does not have experience in all the target languages, and existing literature is limited or contradictory. It is recommended that a consensus opinion among a number of linguists is obtained, where possible. For the 11 official South African languages, a consolidated phoneme system was developed in order to allow linguists to compare conventions across similar languages. This phone set was refined during dictionary development, informed by developer experience and the phoneme counts obtained.

4.5. When automated tools are detrimental

Dictionary developers may become too dependent on the tools developed to assist them. The specific verification algorithms implemented in this study perform well when identifying exceptional pronunciations in a dictionary that has few errors, but are less efficient when many errors exist. When developers start to rely too heavily on automated suggestions – accepting any suggestion that seems fairly reasonable and assuming that any errors would be identified at a later stage – it may be required to either retrain or replace a developer. Such behaviour may not be easy to identify, although the development logs do provide some indications that the behaviour of a developer may need to be verified. Similarly, the communication of exceptional words between the speech technologies and the developers is especially important when a dictionary has a particularly high number of exceptional words, as developers may be tempted to simply make the dictionary consistent, instead of correct.

5. Analysis of the Lwazi dictionaries

In this section we provide an overview of the contents of the Lwazi-1.0 dictionaries, and highlight some interesting observations with regard to the predictability of the languages studied. The basic statistics (unique graphemes, unique phonemes, number of words, total phonemes, total graphemes) for the various dictionaries are listed in Table 1.

	#unique graphs	#unique phons	#words	#total graphs	#total phons
afr	31	37	4,998	35,647	32,075
nbl	25	47	5,152	40,548	37,523
nso	22	45	5,112	36,419	32,608
sot	25	43	5,015	35,888	31,316
ssw	25	39	5,069	42,324	38,606
tsn	23	36	5,012	40,008	32,545
tso	26	55	5,065	36,990	33,960
ven	28	39	5,598	35,815	36,775
xho	26	53	5,064	36,546	34,287
zul	25	45	5,020	38,425	36,072

Table 1: *General statistics for the Lwazi-1.0 dictionaries.*

In practice, the core dictionaries are used as training data for letter-to-sound rules that are then utilised in TTS or ASR systems. By cross-validating each dictionary, it is possible to obtain a clear indication of its predictive accuracy. During 10-fold cross-validation, 90% of the dictionary is used as training data to extract letter-to-sound rules, 10% as test data to verify the rules, and the average accuracy over 10 runs calculated. Results are displayed in the first two columns of Table 2. As can be seen from this table, the (fairly small) 5,000-word dictionaries have good predictive capability, and perform well.

	phone acc	word acc	#rules
ssw	99.96	99.66	101
zul	99.95	99.66	86
xho	99.95	99.66	115
ven	99.93	99.55	164
nbl	99.88	99.14	124
tso	99.83	98.93	165
sot	98.23	89.61	504
afr	97.84	88.03	906
nso	97.10	83.74	896
tsn	95.53	75.54	1,138

Table 2: *Cross-validated letter-to-sound accuracy.*

These results immediately raise the questions: How regular is each of these languages? How many rules are required to specify the training data with 100% precision and recall? As the standard Default&Refine rule set meets this objective, we extract Default&Refine rules, and list the number of rules extracted per language in the next column.

For all of the Southern Bantu languages, the majority of graphemes generate very simple rule sets, with only a few graphemes requiring more complex rule sets to be described fully. This distribution is depicted in Fig 1 for four of the Southern Bantu languages: isiZulu and siSwati (similar behaviour to isiNdebele, isiXhosa and Xitsonga), Setswana and Sepedi (similar behaviour to Sesotho); shown in comparison with a less regular language such as Afrikaans. It is clear that very regular letter-to-sound relationships are interspersed with highly irregular relationships for a limited number of graphemes only. The lower accuracy of the Setswana rule set is due to vowel confusion which is currently undergoing further analysis.

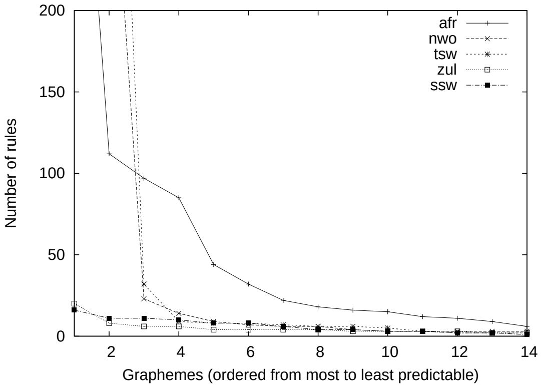

Figure 1: *Distribution of rules per grapheme for representative languages. Note the 3 distinct types of behaviour (relating to the number of graphemes requiring complex rule sets).*

These dictionaries have since been used to build ASR systems, achieving phone recognition accuracies of between 66% and 76%, which is sufficiently high to support the development of Spoken Dialogue Systems [9]. (These phone recognition accuracies result in word error rates of 2-12% when tested on general 10-concept grammars.)

6. Conclusion

Using a process such as the one described in this paper, and taking care with both the development and verification protocol employed, we feel it is possible to develop practically usable pronunciation dictionaries with minimal resources. This is especially important in the developing world, where obtaining a pronunciation dictionary is often one of the first hurdles to overcome when deploying SLT4D applications.

7. References

[1] S. Maskey, L. Tomokiyo, and A.Black, “Bootstrapping phonetic lexicons for new languages,” in *Proceedings of Interspeech*, Jeju, Korea, October 2004, pp. 69–72.

[2] M. Davel and E. Barnard, “Bootstrapping for language resource generation,” in *PRASA*, South Africa, Nov 2003, pp. 97–100.

[3] P. Mertens and F. Vercammen, “Fonilex manual,” Tech. Rep., K.U.Leuven CCL, 1998.

[4] O.M. Martirosian and M. Davel, “Error analysis of a public domain pronunciation dictionary,” in *PRASA*, South Africa, Nov 2007, pp. 13–18.

[5] M. Davel and E. Barnard, “The efficient creation of pronunication dictionaries: human factors in bootstrapping,” in *Interspeech*, Jeju, Korea, Oct. 2004, pp. 2797–2800.

[6] Pali Lehohla, *Census 2001: Census in brief*, Statistics South Africa, 2003.

[7] Meraka-Institute, “DictionaryMaker,” 2009, Online: http://dictionarymaker.sourceforge.net/.

[8] M. Davel and E. Barnard, “Pronunciation prediction with Default&Refine,” *Computer Speech and Language*, vol. 22, pp. 374–393, 2008.

[9] J. Badenhorst, C. van Heerden, M. Davel, and E. Barnard, “Collecting and evaluating speech recognition corpora for nine Southern Bantu languages,” in *Proceedings of EACL*, Athens, Greece, October 2009, pp. 1–8.