Rule-Based Pronunciation Modeling for Resource-Scarce Varieties

Simone Ashby, José Pedro Ferreira, Silvia Barbosa, Catarina Silva

1Instituto de Linguística Teórica e Computacional (ILTEC), Lisbon, Portugal
{simone, zpferreira, silvia, catarina}@iltec.pt

Abstract. Narrowing the speech technology gap within and between nations requires, among other things, a strategy for coping with scarce linguistic resources. In this paper, we describe the ongoing research activities of the LUPo project to model non-standard varieties of spoken Portuguese from Mozambique and other resource-scarce countries and regions. We demonstrate how these models have been combined to form a rule-based architecture for generating accent-specific pronunciations, adapted from the Unisyn Lexicon for English. We describe some of the challenges of extending the Unisyn model to Portuguese, and of establishing models for non-standard varieties that receive considerable substratal interference through contact with local languages. Finally, we discuss some preliminary results, quality control efforts, and the merits and pitfalls of applying the Unisyn methodology to generate pronunciation dictionaries for resource-scarce varieties.

Keywords: rule based pronunciation modeling, pronunciation dictionaries, Portuguese, regional accents, Mozambique, resource-scarce

1 Introduction

Portuguese is the world’s sixth largest language, with major concentrations of speakers in Angola, Brazil, Cape Verde, East Timor, Guinea Bissau, Mozambique, Portugal, and São Tomé and Príncipe. The LUPo (Portuguese Unisyn Lexicon) project is dedicated to: (1) modeling standard and non-standard varieties of spoken Portuguese from around the globe; (2) delivering a free, open-source tool for the automatic generation of accent-specific pronunciation lexica within the existing Portal da Língua Portuguesa (http://www.portaldalinguaportuguesa.org); and (3) providing the research community with a free, online, searchable databank dedicated to the description of regional varieties of spoken Portuguese.

While numerous resources exist for the dominant Brazilian (BP) and European (EP) Portuguese standards – including electronic dictionaries, phonetically aligned corpora, pronunciation lexica, and detailed published studies – less dominant varieties of Portuguese typically have far fewer linguistic resources at their disposal. This is especially true of Mozambique and other non-western Lusophone nations, despite the fact that L1 (first language) speakers are fast on the rise in such countries, where “Portuguese is viewed as the language of science, knowledge, and power”, and that
which holds the most promise for obtaining employment and enhancing one’s upward mobility [1].

With few available linguistic resources, the less dominant varieties see little in the way of language technology adaptation. Speech applications, which rely on the existence of good quality pronunciation dictionaries, continue to be designed with end-users from Lisbon/Coimbra, Rio de Janeiro, or São Paulo in mind. As a result, researchers active in promoting spoken language technologies for development in areas such as northeastern Brazil or the Mozambican capital of Maputo must find alternate means of establishing pronunciation dictionaries for these and other resource-scarce varieties of Portuguese.

The Unisyn model described in [2] offers a unique approach to modeling pronunciation variation through investment in an accent-independent lexicon and rule system for generating high quality accent-specific phonetic output. Within this model, the accent-independent or ‘master’ lexicon contains orthographic entries and their underspecified “phonemic” representations, encoded via computer-readable key symbols (see Figure 1). These can best be understood as ‘metaphonemes’, and take from the ideas in [3] as a means of “[a]bstracting away from phonetics [so] that a single lexicon can represent numerous different accents” [4]. Additional symbols are used for marking stress, syllable, and morpheme boundaries, as well as for linking phonemes to their graphical counterparts.

<table>
<thead>
<tr>
<th>Numeric ID</th>
<th>Word</th>
<th>Part of Speech</th>
<th>Representation</th>
</tr>
</thead>
<tbody>
<tr>
<td>937</td>
<td>gorra</td>
<td>substantivo</td>
<td>g #o . r rr == a</td>
</tr>
<tr>
<td>102789</td>
<td>gorro</td>
<td>substantivo</td>
<td>g #o . r rr == u . o</td>
</tr>
<tr>
<td>104152</td>
<td>gosma</td>
<td>substantivo</td>
<td>g #O . o z . s . m a</td>
</tr>
<tr>
<td>78696</td>
<td>gosmar</td>
<td>verbo</td>
<td>g o z . s . m #a &gt; r</td>
</tr>
<tr>
<td>56651</td>
<td>gospel</td>
<td>substantivo</td>
<td>g O . o s . p #E . e 5</td>
</tr>
<tr>
<td>6187</td>
<td>gostar</td>
<td>verbo</td>
<td>g o s . t #a &gt; r</td>
</tr>
<tr>
<td>58007</td>
<td>gostinho</td>
<td>substantivo</td>
<td>g o s . t == #i . j . nh u . o</td>
</tr>
<tr>
<td>32523</td>
<td>gosto</td>
<td>substantivo</td>
<td>g #o s . t == u . o</td>
</tr>
</tbody>
</table>

Fig. 1. Sample master lexicon entries encoded in X-SAMPA, as adapted by LUPo. The numeric field signifies corresponding entries in the Portal da Língua Portuguesa and their record IDs. Syllable boundaries are marked with a dot '.', primary stress is denoted with a hash symbol '#', and orthographic linking is denoted with an underscore symbol '_'. The remaining non-alphabetic symbols denote morpheme boundaries.

Phonetic conversions involve the application of hand-written rules to the underspecified master lexicon representations, a system that is governed by a geographically organized regional accent hierarchy. In simple terms, this consists of a system of nodes corresponding to the levels COUNTRY, REGION, TOWN, and PERSON, for identifying which rules correspond to one or more of the varieties modeled. What is interesting about this hierarchy is the inheritance by each node of features from the higher level node, provided the inheritance is not broken by the introduction of competing rule(s) at a lower level. As such, the Unisyn framework enables the sharing of rules across geopolitical varieties, allowing for the representation of non-systematic relationships between branch nodes (e.g. the
realization of word-final /l/ as [ɫ] by a small sub-population of São Paulo speakers and among a much larger population of speakers in Angola, Cape Verde, Mozambique, Portugal, etc.), while gaining its economy from the broader notion that phonetic similarity is, to some extent, tied to geographical proximity (e.g. the widespread use of word-final [w] for /l/ in Brazil). It is also possible to represent sociolectal and idiolectal varieties by introducing rules at the low-level nodes that combine with a set of more generally applied rules to capture the unique segmental characteristics of a neighborhood, or an individual speaker.

The Unisyn framework not only offers a robust method for generating large-scale pronunciation dictionaries for multiple varieties, and handling out-of-vocabulary items, it is also extremely well suited for representing varieties with limited linguistic resources. Once established, the sets of hand-written, accent-specific rules and their lexical exceptions can be used to generate high-quality pronunciation dictionaries for use in ASR and TTS applications, and other spoken language technologies.

In the sections that follow, we describe the rule creation processes employed by LUPo for modeling the variety of Portuguese spoken in Maputo, Mozambique. We then show how these rules have been integrated into LUPo’s pronunciation generator, whereupon we highlight some of the challenges of modeling non-standard varieties that receive considerable interference from local languages, and, more generally, of adapting the Unisyn model to Portuguese. A description of LUPo’s preliminary results and evaluation criteria follows, along with an assessment of the Unisyn framework’s suitability for creating pronunciation dictionaries for resource-scarce varieties.

2 Modeling the Maputo Variety

2.1 Data Collection

Given LUPo’s aims and the fact that Portuguese is widely spoken throughout Mozambique’s larger urban centers, inclusion of the Maputo variety (Pmap) was identified as one of the project’s top priorities. As with other resource-scarce varieties, our initial efforts were spent collecting a corpus of read and spontaneous speech from male and female informants, aged 18-35, from the target area. The majority of speech recordings were conducted at the Eduardo Mondlane University in Maputo, with additional recordings performed at ILTEC, in Lisbon, Portugal. In total, we collected more than 21 hours of spoken data from 14 informants, nine whom regard Portuguese as their first language, while the remaining five informants reported learning Portuguese during the early years of their primary education.

Materials for the elicitation of read speech are based on those in [5], along with an additional set of words and phrases deemed necessary for capturing other relevant

---

1 In the cities of Mozambique, it is increasingly common to raise children with Portuguese as the L1. According to [6], Portuguese is spoken as a lingua franca by 50.37% of the population; an additional 10.7% regard Portuguese as their native language.
contexts. For the read speech elicitation task, informants were asked to read the individual phrases and sentences projected in front of them on PowerPoint slides. The elicitation of spontaneous speech data was conducted in the form of an oral questionnaire for obtaining general speaker information and attitudinal data. Recordings were performed with a Marantz digital voice recorder, with a microphone positioned on the table in front of the informant.

2.2 Accent Modeling and Rule Creation

By accent, we mean those features characteristic of a speaker’s pronunciation. While future plans involve the modeling of suprasegmental characteristics, our current efforts are focused on describing the broad segmental features of $P_{Mapo}$. In terms of style, this model reflects a slower and more carefully articulated variety than that found in conversational speech.

Corpus-based accent models were developed at the segmental level by first establishing a list of contexts in which the metaphonemes described in Section 1 were likely to have a different surface realization for the variety being modeled. This was accomplished by listening to read-speech data for three of the L1 informants, creating a new list entry when alternate surface realizations were heard, and noting the observed realization(s) in the form of rules. The list of contexts was then expanded and modified using observations from data for the remaining informants, with a focus on those contexts which yielded variable output in the original three informant recordings analyzed. In cases where a metaphonemic context resulted in variable output, frequencies were calculated across informants, with an indication of the most stable variant per context.

For example, in words such as Evangelho ‘gospel’ and melhor ‘better’, the palatal lateral approximant /ʎ/ was realized by our Maputo informants as [j$\text{l}$] or [j$\text{l}$j] (with ‘$\text{S}$’ denoting a syllable boundary). For a very small number of speakers, /ʎ/ was additionally realized as the simplified [l], although not systematically. Being most common, [j$\text{l}$] was selected as a variant of [ʎ], and the following rule was formed (here and throughout this paper using X-SAMPA notation): do_L_jl_resyllabify. This rule instructs LUPo’s pronunciation generator to perform a conversion when a word in the master lexicon is encountered containing /ʎ/ (or ‘L’), thus producing the resyllabified output forms ‘Evang[ej$\text{l}$]$\text{o}$’ and ‘m[ej$\text{l}$]$\text{o}$’.

Together, these rules – along with the features encoded in the master lexicon that are resistant to change, and the particular lexical exceptions that apply to a given variety – form a segmental accent model. In many cases, a rule already exists for one or more of the varieties modeled, in which case the applicability ‘score’ for that rule must be modified to include the new variety. Naturally, it is also of extreme importance that the rules be ordered. These and other topics related to LUPo’s implementation of the Unisyn system are described in Section 3.

2 All variants are organized per context and stored in LUPo’s free, searchable, online databank (forthcoming) for potential use in linguistic studies and ASR applications.
Substratal Interference. The diglossia that defines Mozambique linguistically means that Mozambican Portuguese speakers are also fluent in one or more Southern Bantu languages. This is no less true of our PMap informant population, which reported varying degrees of fluency in Changana, Chope, Gitonga, Ronga, Tshwa, and Zula, as well as French and English. The effects of local Bantu substrates on informants’ spoken Portuguese varieties are often attested in the literature on Bantu linguistics, and can be modeled in the form of rules. E.g. do_d_dd_resyllabify: gemination of the word-final onset /d/ in penultimately stressed words, such as esperança[d$du] esperançado ‘hopeful’ and ansiedade[dSdi] ansiedade ‘anxiety’; do_monothong: the monothongization of diphthongs in words such as [r][u]nião ‘meeting’ and ap[o$]o apoio ‘support’; and the combination do_devoicedC and do_V_aspirant_resyllabify for the realization of aspirated, devoiced consonants and vowel apocope in unstressed, word-final contexts, such as be[p] bebo ‘I drink’, and objet[i]f objetivo ‘objective’.

Other phenomena are more complex. For example, like BP varieties, consonant sequences are predictably broken up by epenthetic vowels. The problem for PMap lies in adapting the Unisyn’s search and replace style regular expressions to properly identify these vowels, which assimilate with the vowel in a neighboring syllable. E.g. o[moS][i]ciente omnisciente ‘omniscient’, _interre[geS]n interregno ‘interruption’, and eru[p]SsS[ão] erupçao ‘eruption’. A separate long-distance phenomenon occurs, in which the second in a non-contiguous series of voiced stops undergoes voice or place dissimilation. E.g. obe[d]ência obediência ‘obedience’, and de[d]ozinho dedozinho ‘little finger’. Contexts requiring look-ahead and look-behind methods of computation are described in Section 3.

As with a large number of Portuguese varieties, the rule for palatalizing the coda /s/ (i.e. do_s_S) takes effect before another consonant. However, there exists in PMap a circumstance in which the deletion of a subsequent syllable causes the resultant [ʃ] to become realized as a syllable onset, or in which [ʃ] undergoes metathesis in the context of a neighboring sibilant. E.g. invel[S]igações investigações ‘investigations’ (instead of inve[ʃ]igações), and onmi[S]lente omnissciente ‘omniscient’ (instead of onmi[S]ciente). In these and other cases, LUPo’s rule system must be carefully ordered and specified so that the necessary context is in place for applying the rule do_s_S, and so that predictable contexts for elision – e.g. /u/ in the dispreferred sequence [ʃ] – undergo subsequent resyllabification.

Variability. Inter- and intra-speaker variability issues must be accounted for in the treatment of any regional variety. However, given the degree of multilingualism described in Section 2.2.1, and the fact that the Portuguese language is still in the process of becoming indigenized in Mozambique, phonetic variability can be described as a defining feature of PMap, thus requiring adequate representation in LUPo’s pronunciation output.

Some of the contexts in which PMap is particularly prone to yielding variable output include: realization of word-final /t/ as an alveolar trill [ɾ], a flap [r], a voiceless velar fricative [x], or as an elided segment [*]; non-systematic pronunciation of “silent” consonants /p/ and /k/ in words such as ado[p]ção adopção ‘adoption’, and da[k]ilografia dactilografia ‘typing’, which are no longer phonetically represented in BP and EP; realization of /t/ as a stop [t] or a homorganic fricative [θ]; realization...
of the stressed nasal vowel /ɛ/ as a diphthong [ɛj] or a raised segment [I]; and variable realizations of the open and close vowel pairs /e/, e/ and /ɔ/, o/.

**Implications for the Master Lexicon.** As indicated, the metaforms that make up the master lexicon are abstract so as to minimize the number of rules necessary to generate target output for multiple regional varieties. Given the strong correspondence between phonemes and graphemes in Portuguese, the specification of segments was accomplished in a rather straightforward manner. For example, the metaform \{ a . b #O_o . b o . r a \} maps to the word *abóbora* ‘pumpkin’.

However, master lexicon development also involves a level of decision making that is informed by the specific varieties being treated. Again, this can be handled in a relatively straightforward manner if one already has a sufficient understanding of the regional variety in question. But, in the case of resource-scarce varieties, such as *P_map*, not all of the important considerations for master lexicon development will be obvious until partial modeling is done for this and the other varieties concerned. For example, based on what we already know about the strong tendency for gliding in standard BP and EP varieties in words such as *ensju*mar enciumar ‘to make jealous’ and *continwu* continuo ‘continuous’, it is tempting to encode these and other common types of gliding in the master lexicon to minimize the number of rules needed. With some modeling of *P_map*, however, the need to preserve the original vowel sequences became clear so as to generate the appropriate output forms *ensisu*mar and *continu*. Orthographic linking also emerged as particularly important in preserving the context for fully specified, word-final, nasal consonants, which occurs in *P_map* among non-high-frequency words. E.g. the specification of \{ m #o . == e ~_m \} (instead of \{ m #o . == e ~ \}) for *moem* ‘they grind’ for yielding the output nasal vowel [ɛ] + [mʃ] by *P_map* speakers, as opposed to just [ɛ#]. This type of encoding is essential for identifying the contexts of other orthographically motivated pronunciations, such as occurs with the “silent” graphemes <c, p> among *P_map* speakers (see also Section 2.2.2).

### 3 Integrating *P_map* Rules into LUPo

As indicated above, LUPo’s rules take effect through application to the underspecified metaforms stored in the master lexicon, each of which corresponds to a lemma in the Portal da Língua Portuguesa (hereafter referred to as the *Portal*). A component file system is used to: define LUPo’s regional accent hierarchy, assign rule scores for determining which rules apply to which varieties and the order in which they take place, define classes of sounds, govern phonotactic sequences, deal with lexical exceptions, and access word frequencies.

Integrating *P_map* rules into LUPo first required ensuring that the rule did not already exist. *P_map* shares a number of its rules with standard BP and EP, such as the aforementioned palatalization rule *do_s_S*, which characterizes the standard Rio de Janeiro and Lisbon/Coimbra varieties, and many other nonstandard varieties. The fact that LUPo already features rules for BP and EP meant that by simply switching these rules ‘on’ in the rule scores file, a significant part of the work was already done for
By the same token, introducing new rules for \( P_{\text{Map}} \) facilitates the work of adding more regional varieties from Mozambique and other Luso-African countries that share a common store of allophones. This is very meaningful for adapting rule systems to resource-scarce varieties and establishing tailor-made pronunciation dictionaries in a relatively low-cost manner.

Still, the introduction of novel rules is not trivial, in part given the need for careful attention to rule ordering and resyllabification phenomena. These matters being carefully attended to in the original Unisyn lexicon, the model holds up well to Portuguese adaptation. Specifically, each rule has a \texttt{ruleorder} assignment in the rule scores file that tells the system which rules apply to which varieties and in what order they should take effect. In terms of resyllabification, the Unisyn lexicon’s search and replace method and system for marking syllable and morpheme boundaries are well suited for generating resyllabified output. Returning to the example \texttt{investigações}, from Section 2.2.1, the \texttt{rulescore} for the newly created rule \texttt{simplify_cluster} was adjusted to take effect \textit{after} the existing rule \texttt{do_s_S}, and override the syllable boundary specified in the master lexicon with a new one, yielding the output \texttt{inve[dS]}\texttt{igações}. As such, the Unisyn framework has also proved highly suitable for dealing with sandhi phenomena.

What is more challenging in terms of computation, logistics, and quality control is: (1) the need to account for complex substratal interference (see Section 2.2.1); (2) the important role spoken varieties play in informing the design of the master lexicon; (3) the need to strike the right balance between the master lexicon and rule sets so as to minimize the number of rules required, while making it generalizable to the different varieties being treated; (4) the lack of standardization that currently exists among Luso-African and Luso-Asian varieties, and the magnitude of variability that comes with describing non-standardized varieties; and (5) the lack of resources for measuring output quality (see also Section 4).

With LUPo, we can address (1) by augmenting the standard form of regular expression strings used in the original Unisyn Lexicon with advanced Perl constructs. The look-ahead and look-behind assertions enable LUPo to identify the appropriate vowel to insert between disallowed clusters in words such as \texttt{omnisciente} and \texttt{interregno}, and to anticipate the conditions for long-distance dissimilation, such as occurs among voiced stops in the words \texttt{obediência} and \texttt{dedozinho}.

Challenges (2) and (3) above should offer the benefit of hindsight to anyone interested in applying the Unisyn framework to other languages. For, while it may seem logical to start with the development of the accent-independent master lexicon before continuing on to develop the accent-specific rules, large-scale expansion of the master lexicon should only occur \textit{after} sufficient efforts have been made to understand the basic characteristics of the varieties to be modeled. In the case of LUPo, which started with a small-scale master lexicon that was prematurely expanded, several iterations followed in which we aimed to properly address (3).

Challenges (4) and (5) above were known variables when setting about this work and remain LUPo’s foremost hurdles. In terms of (4), which poses far fewer problems for better sourced varieties, we have applied informant frequencies in our decision making concerning the modeling of \( P_{\text{Map}} \). Having collected detailed background information from our informants, we are also making use of the regional accent hierarchy (see also Section 1) to attribute certain allophonic alternations to a specific
speaker profile by introducing rules at the PERSON level that define the characteristics of a particular city zone. In general, while our methods for modeling regional varieties through direct contact with speakers introduces a level of variability that may be considered unwieldy, we regard variability as essential in creating rules for generating high-quality pronunciation output. By cataloging all observed variants with their context in the project’s free, online, searchable databank, we achieve the added goal of establishing a separate resource for use by other speech technologies, such as ASR, as well as serving linguists.

Our approach to challenge (5) above is described in the section that follows.

4. Preliminary Results and Conclusions

In this paper, we presented the work of the LUPo project to create a rule-based pronunciation model for the resource-scarce variety of Portuguese spoken in the capital city of Mozambique, based on the Unisyn framework developed by Fitt [2].

Currently, LUPo consists of a 31,850-word master lexicon, and a conversion architecture with rules for generating accent-specific output for standard varieties from Lisbon, Rio de Janeiro, and São Paulo, as well as non-standard varieties representing the greater Maputo area, and the island of Catembe.

In terms of quality control, golden test sets have been developed for the three standard varieties based on hand-transcriptions by native speakers. From our modeling efforts, we also have control sets for the PMap and Catembe varieties. For each of these – consisting of 177 hand-transcribed words, selected for adequate phonetic coverage – 100% accuracy was achieved in terms of generating output forms that matched the control set. However, beyond this preliminary quality control measure, the framework is limited in its capacity for verifying transcription quality over the entire set of pronunciations generated from the 31,850-word master lexicon. We aim to address this by: engaging language experts to validate output for the resource-scarce varieties, as part of the Common Orthographic Vocabulary project, or VOC, described in [7]; performing routine checks via LUPo’s administrative back-end; and building a user feedback mechanism into LUPo’s online Portal module.

In summary, although the Unisyn framework possesses no built-in methods for ensuring data integrity, supplementary measures can be taken to fine-tune the rules for varieties with few available resources. Overall, we hope to have shown that with careful attention to the staging of the master lexicon expansion and adequate handling of variability, the Unisyn Lexicon offers a potentially robust methodology for adaptation to other languages and resource-scarce varieties.

5. Acknowledgements

The authors gratefully acknowledge the support of the Fundação para a Ciência e a Tecnologia, and the cooperation of Dr. Susan Fitt, whose development of the original Unisyn Lexicon is the inspiration for this work. We are also extremely grateful to Inês Machungo and colleagues at the Eduardo Mondlane University.
6. References