Introducing Non-Standard Luso-African Varieties into the Digital Domain

Simone Ashby, Sílvia Barbosa, and José Pedro Ferreira

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

Abstract

In this paper, we describe the work of the LUPo project to integrate non-standard Luso-African varieties from Cape Verde and Mozambique into a rule system for generating accent-specific phonetic transcriptions for these and other spoken variants of Portuguese. Here, we present a description of LUPo's functions for online use, LUPo's architecture, and the manner of linguistic data collection and analysis that support this application. Implications for a subsequent text-to-speech (TTS) module are also presented in terms of yielding high-quality pronunciation lexica for regional variants of Portuguese, and towards facilitating the entry of hitherto untreated speech varieties into the digital domain.

Index Terms: pronunciation lexicon, speech synthesis, Luso-African speech varieties, digitally endangered speech varieties

1. Introduction

This is a demonstration of the LUPo online prototype, as well as a presentation of some preliminary results for observing non-standard Luso-African pronunciation varieties from Cape Verde and Mozambique. This work marks the first phase of a three-year research project dedicated to the creation of an accent-independent lexicon and rule system for generating accent-specific pronunciations for regional variants of Portuguese. More in-depth information about LUPo and the English Unisyn Lexicon upon which it is based can be found in [1] and [2], respectively.

Here, we present a seminal effort towards developing systematized, multiple, complete phonetic models for non-standard varieties of Portuguese, as it is actually spoken in different parts of the world. We developed broad phonetic segmental models[1] based on six individual speakers, or *idiolects*: three from different locations on the island of Santiago (the largest in the Cape Verdean archipelago), and three from different locations in Mozambique.

The presentation provides a window into these segmental models, as contrasted with those for the São Paulo and standard Lisbon dialects. A selection of phonological rules is also presented, along with a description of how one of LUPo's key components, the regional accent hierarchy, enables the sharing of rules across pluridimensional dialectal and sociolectal varieties. Finally, we demonstrate the LUPo system as it currently exists, and provide some preliminary results for observing sub-national and national varieties of Cape Verdean Portuguese.

The motivation for this research is based around the development of high-quality pronunciation lexica for a pan Lusophone TTS system. As speech technologies become an increasing part of our everyday lives, the users of these technologies represent an ever widening speaker base. Thus, adapting such technologies to a wider number of speakers – and *topolects* – and representing countries and regions for whom such development concerns have been largely overlooked carries significant economic and political weight in narrowing the global digital divide, and promoting further research among lesser studied varieties.

Through the establishment of a linguistically derived rule system for the explicit treatment of allophones within and across regional varieties, LUPo circumvents the cost of producing high-quality phonetic transcriptions by hand, while attracting a wider pan Lusophone audience to the lexical database in which it resides, and providing the research community with a vast resource of Portuguese accent data for evaluating speech applications and testing diachronic, phonological and sociolinguistic theories.

2. Background

Portuguese is a pluricentric language spoken by one-fifth of the world's population, and with regional variants spanning Africa, Asia, Europe, and South America. In addition to Brazil and Portugal, Portuguese is a recognized official language in Angola, Cape Verde, East Timor, Equatorial Guinea, Guinea-Bissau, Macau, Mozambique, and São Tomé and Principe.

According to Alan Baxter, one of the leading scholars in the study of Portuguese-based creoles, there is a growing sense of identity among Luso-African speakers concerned with elevating the level of prestige associated with local varieties of Portuguese. Baxter also cites evidence to suggest that BP may have increasing role in affecting some of the Portuguese varieties spoken in Africa [3]. Unfortunately, and by Baxter's own admission, a dearth in the literature concerning varieties of Portuguese spoken in Africa makes these claims difficult to assess.

In the interest of promoting pluridimensional studies of Portuguese, this work aims to make a meaningful contribution to the understanding of the different sound systems featured among Luso-African varieties.

2.1. Cape Verdean Portuguese

Cape Verdean Portuguese is distinct from Mozambique in that its speakers have a Portuguese based creole, Kabuverdianu, as their mother tongue. The latter is used in everyday communication, while Portuguese is used in the media, education, and official communications. Roughly one-third of the population of Cape Verde is considered fluent in Portuguese; more than half possess a basic understanding of the language [4].

As an archipelago consisting of 10 islands[2], Cape Verde is host to a number of Kabuverdianu and Portuguese accents and dialects. Typically, these varieties correspond to a division of the Windward islands in the north and the Leeward islands in the south. However, a sociolinguistic shift dividing the western islands from the eastern half appears to be on the rise [5].

[1]We also intend to treat cross-word phenomena, such as external sandhi, as part of the LUPo project. Acoustic modeling and suprasegmental feature descriptions will be undertaken in the follow-up TTS project.

[2]All but one island are inhabited.

2.2. Mozambican Portuguese

While Portuguese is the official language of Mozambique, it generally spoken as a second language, after one of a variety of local languages and creoles. Portuguese is spoken as a lingua franca by more than one-third of the population. A very small minority of Mozambicans speak Portuguese at home.

Speakers commonly refer to European Portuguese (EP) as the norm. Nevertheless, the influence of indigenous languages, and the prestige many Mozambicans associate with Brazilian Portuguese (BP) varieties, along with a steady supply of Brazilian television shows, have resulted in the emergence of distinctively Mozambican ways of speaking Portuguese [6].

3. Multidialectal speech technologies

3.1. Portuguese systems

As Portuguese speech technologies continue to advance in the treatment of "standard" BP and EP varieties, efforts to develop computational models of *non-standard* varieties from these and other countries in the world where Portuguese is spoken are conspicuously absent.

One of the more relevant studies to date concerns an accent verification system for Portuguese, based on acoustic, phonotactic, and prosodic cues, for recognizing BP, EP, and what the authors refer to as African Portuguese (AP) varieties [7]. Sub-national differences are not evident due the selection of corpora used in this and related experiments. Two of the experiments, however, examine the ability of the language/accent verifier to distinguish Portuguese varieties from Angola, Cape Verde, Guinea-Bissau, Mozambique, and São Tomé and Principe[3].

The authors of [7] report a limited ability on the part of their system to cope with different accents of Portuguese, with some – the African varieties, in particular – resulting in the degraded performance of their speech recognizer. A separate experiment in which African varieties were bundled into one category revealed a correct identification rate of just 42% for this class. Misidentification of the AP class was also found among the study's human participants, representing Angola, Brazil, Cape Verde, Mozambique, and Portugal.

What these findings show, although not explicitly stated by the authors, is that the training of speech recognizers on BP or EP corpora is insufficient for handling non-standard input. Further, the results in [7] show that bundling Luso-African varieties together not only leads to an extremely low accent verification rate, it is a step in the wrong direction towards creating speech technologies that can cope with these and other regional variants.

3.2. Other models

For a handful other languages, such as Arabic [8], English [2], German [9], Irish [10], Mandarin [11], Romani [12], and Spanish [13], non-standard lectal variants are slowly attracting the attention of speech technologists. However, *slowly* is the operative word, considering the potential value posed by dialectal and sociolectal models for improving spoken dialog systems, enhancing the training of speech recognizers, and creating more "natural" sounding synthetic speech.

[3]It is important to note that the corpora used in these experiments have been drawn exclusively from RTP Africa television news media, for which it can be assumed that journalists and presenters use a marked EP variety, despite that EP speakers were eliminated from this portion of the variety verification corpus.

4. LUPo

By taking an integrative approach and focusing not only on the recognized centers of Brazil and Portugal, but also on the sub-national dialectal and sociolectal varieties from locations around the globe, LUPo aims to create pronunciation models for as many regional variants of Portuguese as the project's time and resources allow.

In section 4.3, we describe the future accessibility of LUPo via the existing free, online lexical knowledge base, the *Portal da Língua Portuguesa*, hereafter referred to as the 'Portal'. A separate project deliverable will be the subsequent release of a free, searchable, online database containing all of LUPo's data and rules.

4.1. Architecture

LUPo's core components include: an exceptions dictionary, an accent-independent master lexicon of underspecified pronunciations (including part of speech and frequency information), a regional accent hierarchy, and the application (through Perl scripts) of morpho-phonological rules that transform the master lexicon pronunciation into the the target output.

4.1.1. Regional accent hierarchy

The model for the regional accent hierarchy is based on that of the original English Unisyn Lexicon [2]. It is made up a system of files containing variant specifications and rule scores. Applying an example from Mozambique (see figure 1), the first set of lines is an entry in the file 'lupo_towns', with 'map' representing the capital city of Maputo, and the next set of abbreviations representing a system of levels that correspond to COUNTRY, REGION, TOWN, and PERSON.

The next set of lines in figure 1 is taken from a file called 'lupo_scores', wherein a general rule is attributed at the country 'CNY' level for the deletion of word-final /r/ across Brazilian and Mozambican varieties.

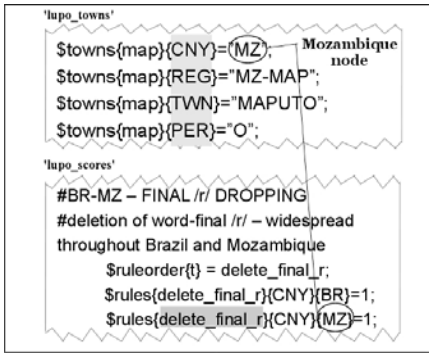

Figure 1: Sample of regional accent hierarchy.

What is interesting about this hierarchical system is the inheritance by each node of features from the previous node, provided the inheritance is not broken by the introduction of a competing feature (or features) at a lower level. As the lowest level in the hierarchy, rules attributed at the person 'PER' level override competing specifications from all the higher levels. Thus, by adding a relatively small number of rules to 'lupo_scores', we can model the unique segmental features that characterize a mesolectal variety of young urban speakers, or even that of an individual – say Mozambique's current president Armando Emílio Guebuza – while implicitly treating the remaining set of allophones as inherited features from the upper nodes TOWN, REGION, and COUNTRY.

In terms of the specific rules presented in figure 1, it is worth noting that word-final /r/ undergoes elision by the Mozambican informants, a phenomenon that is widespread throughout Brazil. Thus, we see possible evidence for the influence of BP on Mozambican varieties.

4.1.2. Rule system

The system stores allophonic rule sets that exploit morphological boundaries to express different accent-specific rules, most of which are post-lexical. Encoding morphology in the pronunciation rules enables the system to identify the correct pronunciation in opaque orthographic contexts, such as the assignment of EP vowel height in the lexically related word pairs 'm[o]lho' and 'm[u]lhada', and 'm[ɔ]lho' and 'm[ɔ]lhada'. Perl scripts are then used to apply rules to the master lexicon and generate accent-specific output. A closer look at the rules is presented in section 4.3.

4.2. Data collection and analysis

The collection and modeling of accent data involves using multiple means – from published studies, to the use of linguistically trained informants, to the collection and analysis of new speech data – to construct complete segmental models for spoken variants of Portuguese.

For each accent treated, a complete segmental model consists of: a list of morphophonological contexts (especially those which are most vulnerable to change) and their corresponding phonetic realizations, i.e. a set of morphophonological post-lexical rules; conditions for the ordering of rules; and a list of lexical exceptions.

For the Luso-African and Luso-Asian varieties in particular, we have initiated a long-term effort aimed at recording Portuguese speakers from capital cities and smaller towns alike. Materials for the elicitation of read speech are based on those developed as part of [14]. The elicitation of spontaneous data is conducted in the form of an oral questionnaire for obtaining general speaker information and attitudinal data[4].

For corpus-based accent models, the assessment of segmental data is performed by trained phoneticians, who use Praat [15] to identify and label target segments.

4.3. How it works

General users will soon be able to access LUPo via the Portal (http://www.portaldalinguaportuguesa.org) to select from a list of available topolects and generate accent-specific pronunciations[5]. With LUPo's online interface, users can select from one of the 11 accents we have modeled so far and search for a given word, as illustrated in figure 2.

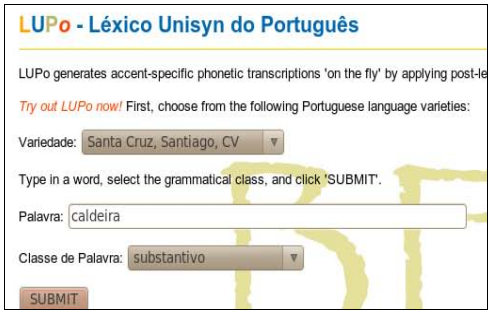

Figure 2: LUPo online prototype.

[4]General speaker information and attitudinal data will be made available through the searchable, online database.

[5]This capability is currently restricted to lemmas. Ultimately, LUPo will be extended to handle word forms and multi-word texts.

In figure 3, the result is displayed for the municipality of Santa Cruz, located in the eastern part of Santiago (Cape Verde) for the noun 'caldeira' (kettle). Here, we see the monothongization of /ej/ as /e/, along with dentalization of the lateral approximant /l/ in coda position.

Figure 3: Pronunciation of 'caldeira' in Santa Cruz.

A quick comparison with our speaker from Praia, Cape Verde's capital city, shows that this speaker maintains the diphthong /ej/, while producing a velarized lateral approximant, such as is common in EP (see figure 4).

Figure 4: Pronunciation of 'caldeira' in Praia.

In figures 5 and 6, respective results are shown for the same word as spoken in the cities of Lisbon (PT) and São Paulo (BR). Here, we see that for speakers of the standard Lisbon variety, the tonic vowel /ej/ is realized as the raised diphthong [ɐj]. Alternatively, for the same phonemic context, there is free variation in the production of both [ej] and [e] among São Paulo speakers, similar to that which was observed for the Santa Cruz (CV) informant.

Figure 5: Pronunciation of 'caldeira' in Lisbon.

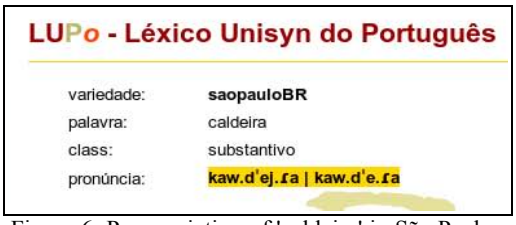

Figure 6: Pronunciation of 'caldeira' in São Paulo.

Figure 5 also reveals a velarized lateral approximant [ɬ] typical of EP, as was observed for the Praia (CV) informant, while figure 6 presents the realization of this sound as the labio-velar approximant [w], common throughout Brazil.

The specific rules applied to generate LUPo's accent-specific transcription output are printed in the lower half of the results page (see figure 7). These are not phonological rules in the strict sense, but rather the transformations the master lexicon entry had to undergo to become the sort of

output displayed in figures 3, 4, 5, and 6. At the bottom of the page, we include a description of the rules in plain language to make them easier to understand.

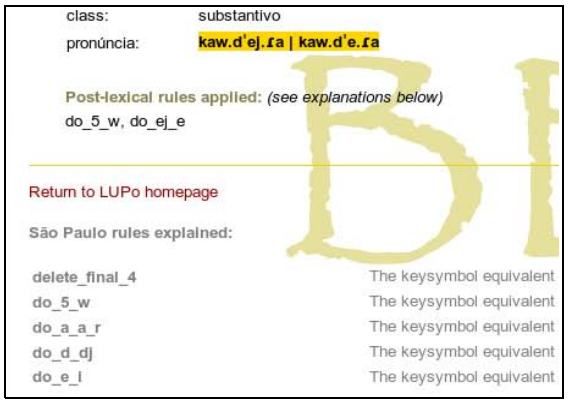

Figure 7: LUPo results page and rule descriptions.

Ultimately, LUPo's online interface (and searchable database) will allow users to observe more than one variant at a time. The map perspective in figure 8 offers one such glimpse at how LUPo is poised to provide linguists with a huge list of varying points and bundled phenomena – along with tangible data links – for testing notions of linguistic similarity and distance, and evaluating the pulling effect of different linguistic centers.

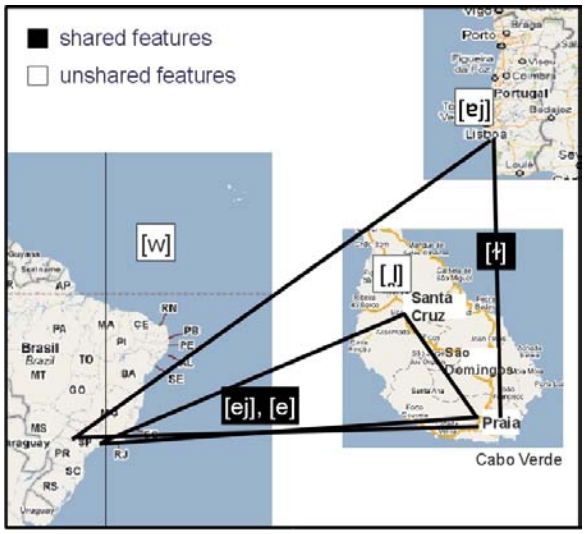

Figure 8: Tangible data links.

5. Conclusions

We have briefly introduced our work on the development of an accent-independent lexicon and rule system for generating phonetic transcriptions for regional accents of Portuguese. We have presented an initial prototype of the online LUPo system, along with a window into the phonetic segmental modeling of Luso-African idiolectal varieties from Cape Verde and Mozambique.

It has been shown that LUPo is designed to handle variability at the national and sub-national levels. This is achieved economically, through the sharing of rules across pluridimensional varieties (as demonstrated in the description of LUPo's regional accent hierarchy), while acknowledging those salient segmental features that are essential in distinguishing one variety from another, and resulting in more "natural" sounding synthetic speech.

In this vein, we seek to contribute to the improvement of Portuguese language speech technologies by providing high-quality pronunciation lexica, derived from linguistic rules, and covering as many topolectal variants as possible. We further anticipate that our work will have a positive impact on non-standard, "digitally endangered" [12] varieties of Portuguese, their enhanced prestige, and as varieties worthy of study in their own right.

6. Acknowledgements

The authors gratefully acknowledge the support of the Fundação para a Ciência e a Tecnologia (PTDC/CLE-LIN/100335/2008), and the cooperation of Dr. Susan Fitt, whose development of the original English Unisyn Lexicon is the inspiration for this work.

7. References

[1] Ashby, S., Ferreira, J. P., and Barbosa, S., "Adapting the Unisyn Lexicon to Portuguese: Preliminary Issues in the Development of LUPo", in Proc. Iberian SLTech, 2009.

[2] Fitt, S., "Documentation and User Guide to UNISYN Lexicon and Post-Lexical Rules", technical report, Centre for Speech Technology Research, University of Edinburgh, 2000.

[3] Baxter, A. N., "Portuguese as a Pluricentric Language", in Clyne, M. G. (ed.) Pluricentric Languages: Differing Norms in Different Languages, Mouton de Gruyter, Berlin, 11-44, 1992.

[4] Évora-Sagna, M., Gray, V., Minges, M., "A Internet num PMA Lusófono: Estudo de Casos de Cabo Verde", International Case Studies report, International Telecommunications Union, 2002.

[5] Neves, A. C., "Young Learners and Phonological Variation in the Portuguese Mesolect of Cape Verde", in Linguistica ONLINE 11, 2009.

[6] Albarrán Alves de Carvalho, M. J., "Aspectos Sintáctico-Semânticos dos Verbos Locativos no Português Oral de Maputo", Dissertação de Mestrado, Universidade de Lisboa, 1987.

[7] Rouas, J.-L., Trancosco, I., Viana, C., and Abreu, M., "Language and Variety Verification on Broadcast News for Portuguese", in Speech Communication,50(11-12):965-979, 2008.

[8] Kirchhoff, K. and Vergyri, D. "Cross-Dialectal Data Sharing for Acoustic Modeling in Arabic Speech Recognition", in Speech Communication 46(1):37-51, 2005.

[9] Pucher, M., Schabus, D., Yamagishi, J., Neubarth, F., and Strom, V., "Modeling and Interpolation of Austrian German and Viennese Dialect in HMM-Based Speech Synthesis", in Speech Communication 52(2):164-179, 2010.

[10] Mac Lochlainn, M., "Sintéiseoir 1.0: a Multidialectal TTS Application for Irish", in ReCALL 22(2):152-171, 2010.

[11] Zheng, Y., Sproat, R., Gu, L., Shafran, I., Zhou, H., Su, Y., Jurafsky, D., Starr, R., and Yoon, S.-Y., "Accent Detection and Speech Recognition for Shanghai-Accented Mandarin", in Proc. Interspeech, 2005.

[12] Rusko, M., Darjaa, S., Trnka, M., Zeman, V., and Glovna, J., "Making Speech Technologies Available in (Serviko) Romani Language", in Sojka, P. et al. (eds.) Lecture Notes in Artificial Intelligence 5246:501-508, Springer-Verlag Berlin Heidelberg, 2008.

[13] Huang, R. and Hansen, J. H., "Gaussian Mixture Selection and Data Selection for Unsupervised Spanish Dialect Classification", in Proc. Interspeech, 2006.

[14] Rodrigues, M. C. M., "Lisboa e Braga: Fonologia e Variação", Fundação Calouste Gulbenkian, Lisboa, 2003.

[15] Boersma, P. and Weenink, D., "Praat: Doing Phonetics by Computer", computer program, vs. 5.1.43, retrieved from http://www.praat.org/ 4 August 2010.