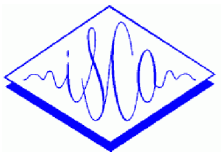
ISCA Archive
http://www.isca-speech.org/archive



INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

A Rule Based Pronunciation Generator and Regional Accent Databank for Portuguese

Simone Ashby [1], *Sílvia Barbosa*[1], *Silvia Brandão*[2], *José Pedro Ferreira*[1], *Maarten Janssen*[3], *Catarina Silva*[1], *Mário Eduardo Viaro*[4]

[1]Instituto de Linguística Teórica e Computacional (ILTEC), Lisbon, Portugal
[2]Universidade Federal de Rio de Janeiro (UFRJ), Rio de Janeiro, Brazil
[3]IULA, Universitat Pompeu Fabre, Barcelona, Spain
[4]Universidade de São Paulo, São Paulo Brazil

[1]{simone, silvia, jpferreira, catarina}@iltec.pt,
[2]silvia.brandao@terra.com.br, [3]maartenpt@gmail.com, [4]maeviaro@usp.br

Abstract

One of the major obstacles in deploying spoken language technologies (SLTs) in the developing world is a lack of key linguistic resources – e.g. electronic dictionaries, phonetically aligned corpora, pronunciation lexicons, etc. – that describe the non-dominant varieties spoken in such countries and regions. In this paper, we describe the work of the LUPo (Portuguese Unisyn Lexicon) project to model standard and non-standard varieties of spoken Portuguese from around the globe, and: (1) deliver a free, open-source tool for the automatic generation of accent-specific pronunciation lexica within the existing online lexical knowledge base, the *Portal da Língua Portuguesa*; and (2) provide the research and speech technology communities with a free, online, searchable database, the Portuguese RADbank, dedicated to the description of regional varieties of spoken Portuguese. Both resources are presented as bases for adapting SLTs to regional varieties spoken in the Luso-African and Luso-Asian world, as well as to non-standard varieties of Brazilian and European Portuguese.

Index Terms: rule based pronunciation modeling, pronunciation dictionaries, Portuguese regional accents, searchable database

1. Introduction

One of the essential motivations for this work is an effort to bridge the digital divide between the developed and the developing world. With respect to Portuguese SLTs, this means leveraging existing technologies, such as TTS and ASR applications – designed with speakers from Lisbon/Coimbra, Rio de Janeiro, and São Paulo in mind – for adaptation to additional Portuguese varieties around the globe.

The LUPo project is dedicated to the creation of a pan Lusophone speech corpus, and the description – via broad segmental models – of standard and non-standard varieties of spoken Portuguese as it is actually spoken in cities and towns of Angola, Brazil, Cape Verde, East Timor, Guinea Bissau, Mozambique, Portugal, and São Tomé and Príncipe. LUPo's accent models, which contain rules for converting orthographic input into accent-specific phonetic output, provide the basis for delivery of a free, open-source tool for establishing pronunciation dictionaries for regional variants of Portuguese. We are also in the process of publishing the corpus itself, along with accent models and sociolinguistic interviews, in the form a free, online, searchable database, known as the Portuguese RADbank, for use by linguists and speech technologists.

Instead of expending thousands of man hours to transcribe a complete dictionary for just one accent, our methodology involves a careful modeling of the accent's sound system. This information is interpreted as a set of hand-written rules, which are applied to an accent-independent lexicon (i.e. a list of words with their metaphonemic representations) for generating accent-specific phonetic transcriptions. In this way, LUPo, like its English language precursor [1], succeeds in dramatically reducing the investment spent per regional variety, while yielding high-quality pronunciation output.

2. Pronunciation generator[1]

As indicated above, LUPo's pronunciation generator is modeled after the Unisyn Lexicon for English (developed at the University of Edinburgh), which offers a unique approach to describing pronunciation variation through investment in an accent-independent lexicon and rule system for generating high quality accent-specific phonetic output.

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*). Additional components include: a regional accent hierarchy and rule score system for determining which rules apply to which varieties and the order in which they take place, and an exceptions database.

LUPo's accent-independent lexicon, or *master lexicon*, currently consists of 31,850 entries, each formed from an extended set of X-SAMPA-based key symbols that capture a rough approximation of a lemma's underlying phonological form. 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 key symbols are used for marking stress, syllable, and morpheme boundaries, as well as for linking phonemes to their graphical counterparts. For example, encoding morphological boundaries in the non-

[1] See [2] for a discussion on how LUPo fits within the context of the online *Portal da Língua Portuguesa*.

hyphenated word compounds *coigual* 'coequal' and *coutente* 'co-user' (shown below) enables LUPo to properly interpret the contiguous vowel sequences /oi/ and /ou/ as contexts for hiatus, instead of interpreting these sequences as diphthongs.

coigual < k_c o < . { i . g w_u #a 5_l }
coutente < k_c o < . { u . t #e ~_n . t i_e }

The master lexicon was generated semi-automatically, with manual checking of each entry to ensure strict conformity with LUPo's metaform standards. Entries were selected according to frequency, using word lists from sources such as [5] and [6].

The regional accent hierarchy – a system of files containing variant specifications and rule scores – enables the application of rules across geographic entities, i.e. towns, regions and countries, while allowing rules to be inherited from larger nodes to smaller, hierarchically linked nodes. E.g. the realization of word-final /l/ as [ɬ] by a small sub-population of São Paulo speakers and among a much larger population of speakers in Angola, Cape Verde, Mozambique, Portugal, etc., or the widespread use of word-final [w] for /l/ in Brazil. As with the English Unisyn system, 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 speakers' unique segmental characteristics. E.g. the retroflexization of /r/ and /l/ that is produced by some Mozambican speakers.

Through LUPo's pronunciation generator, rule sets have been established for standard varieties from Rio de Janeiro and São Paulo (Brazil), and Lisbon (Portugal), as well as for the non-standard varieties representing Maputo and neighboring Catembe (Mozambique). LUPo's rules exploit morphological boundaries and interact with the metaforms stored in the master lexicon, along with the system's regional accent hierarchy, to generate accent-specific transcriptions. Integrating topolectal rules into LUPo requires first ensuring that the rule does not already exist. The fact that LUPo already features rules for the dominant standard varieties means that by simply switching these rules 'on' in the rule scores file, a significant part of the work is already done for introducing additional non-standard varieties. Similarly, introducing rules for a non-standard topolect facilitates the work of adding other regional varieties from a nearby country or region that shares 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.

LUPo's online interface (available in Autumn 2012 at www.portaldalinguaportuguesa.org) will enable users to select from a list of available topolects and generate accent-specific pronunciations for lexical entries stored in the Portal. A separate function will allow users to generate a complete accent-specific pronunciation dictionary for all of the lemmas currently stored in LUPo's master lexicon. Specialty users will also be able to download the open-source pronunciation generator (in Perl), with detailed instructions for expanding LUPo to include additional dialectal, sociolectal, and idiolectal varieties.

3. The Portuguese RADbank

In 2013, the Portuguese Regional Accent Databank will become available to the research community in the form of a free, searchable, online databank for: accessing recordings, testing the results of different speech processing systems, conducting empirical analyses across multiple Portuguese accents, monitoring contact language effects, and facilitating the entry of lesser documented regional variants into the digital domain. As shown in Figure 1, the RADbank will contain all audio recordings and transcription files, sociolinguistic interviews, accent models, lexical exceptions, and phonetic output generated by the LUPo project.

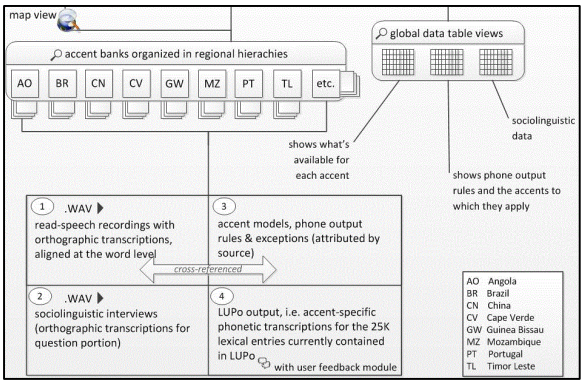


Figure 1: *The Portuguese RADbank architecture.*

4. Conclusions and future work

In this article, we introduced our work on the development of a free accent-independent lexicon and rule system for generating phonetic transcriptions for regional accents of Portuguese. We also described the architecture of the Portuguese RADbank, which is currently under development, and which presents linguists and speech technologists with a searchable database of all audio recordings, transcriptions, accent models, and phonetic output generated by the LUPo project.

Work is currently under way to model *non-standard* varieties from Rio de Janeiro, São Paulo (Brazil), Mindelo, Praia (Cape Verde), Dili (East Timor), Porto, and Lisbon (Portugal). These and other accents will gradually be integrated into LUPo, and their raw data made accessible via the Portuguese RADbank.

5. Acknowledgements

The authors gratefully acknowledge: the support of the FCT, Susan Fitt (whose original Unisyn Lexicon is the inspiration for this work), Inês Machungo, and Dora Pires.

6. References

[1] Fitt, S., "Documentation and user guide to UNISYN lexicon and post-lexical rules", Online Technical Report, Centre for Speech Technology Research, http://www.cstr.ed.ac.uk, 2000.

[2] Ashby, S. and Ferreira, J. P., "Reuse of lexicographic data for a multipurpose pronunciation database and phonetic transcription generator for regional variants of Portuguese", in 14th Euralex Proc., 241-244, 2010.

[3] Wells, J., The Accents of English. New York: Cambridge University Press, 1982.

[4] Fitt, S. and Isard, S., "Synthesis of regional English using a keyword lexicon", in Eurospeech Proc., 823-826, 1999.

[5] Rocha, P. and Santos, D., "CETEMPúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa", in PROPOR Proc., 131-140, 2000.

[6] Santos, D., "Linguateca's infrastructure for Portuguese and how it allows the detailed study of language varieties", in J. B. Johannessen [Ed], Language Variation Infrastructure, Oslo Studies in Language 3(2): 113-128, 2011.