Introducing the Portuguese Unisyn Lexicon (LUPo): An Accent-Independent Pronunciation Lexicon for Portuguese
Simone Ashby, José Pedro Ferreira, and Sílvia Barbosa
Instituto de Linguística Teórica e Computacional (ILTEC)
{simone.ashby, zpferreira, silvia}@iltec.pt

Successful speech technologies require the ability to account for variation in the speech signal. Most TTS systems are built using data from a single accent, usually what is considered to be the standard accent, or dialect, for a given language. While users of these technologies represent an ever widening speaker base, the prospect of developing separate lexicons to account for regional pronunciation variants is a very costly one. Semi-automatic approaches for exploiting regularities between graphemes and phones have yielded good results. However, such systems rarely extend to multiple accents, and make limited or no use of morphology. Moreover, these projects typically occur in isolation, and are governed by private sector interests that prohibit the sharing of data and tools.

Fitt (2000) presents an adaptable paradigm for minimizing the costs of representing multiple pronunciation variants by using knowledge-driven approaches to specify correspondences between a master lexicon and different accent-specific targets. Implicit in the notion of a master lexicon is the expression of phonological variation in the form of key symbols, a kind of metaphoneme based on Wells’ (1982) keywords concept. Key symbols, which can also be used to encode stress, syllables, and morphology, make up the lexical entries and set them apart as accent-independent. Instead of creating hundreds of thousands of phonetic transcriptions for each new accent, such data are generated automatically through the application of accent-specific post-lexical rules. By framing this information within the context of a regional accent hierarchy, a single rule can be used to describe a number of accents.

This paper describes our approach towards building an accent-independent lexicon and rule system for generating accent-specific pronunciations in Portuguese, also known as the Portuguese Unisyn Lexicon (LUPo). With the consultancy of Susan Fitt, author and developer of the Unisyn Lexicon for English, our methodologies will be a reformulation of those used by Fitt to adapt this largely successful paradigm to Portuguese, and take advantage of the MorDebe database´s relational structure and rich lexicographic content to minimize confusability and create a more integrated and well informed system. Our model will capitalize on having direct access to mappings of European and Brazilian Portuguese spelling variants, part of speech information, etymological relationships, and a morphological parser.

The end product will be a set of open-source tools for generating accent-specific output for lexical entries, along with the ability to produce transcriptions for multi-word texts. Pronunciation models will be included for European and Brazilian Portuguese standards, plus actual spoken accents representing the continents of Africa, Asia, Europe, and South America. Cross-dialectal data, phonetic transcriptions, the master lexicon, allophonic rules, and tools will be made freely available to the research community and general public via the Portal da Língua Portuguesa knowledge base. Inclusion of LUPo in the Portal will greatly enhance the Portal as a pan Lusophone resource and the only one of its kind to provide richly detailed and varied phonetic output for a large number of Portuguese accents. Indeed, it will be the first online resource to provide any manner of phonetic transcription data for Portuguese. LUPo also has the potential role of becoming used as a standard against which Portuguese speech applications may be evaluated, and phonological and diachronic theories tested.