Reuse of Lexicographic Data for a Multipurpose Pronunciation Database and Phonetic Transcription Generator for Regional Variants of Portuguese

Simone Ashby and José Pedro Ferreira[1]
Instituto de Linguística Teórica e Computacional (ILTEC)

Among the benefits of a flexible and modular lexical database are: the facility of building new modules from existing ones, the reuse of lexicographic data to both enhance the user experience and achieve NLP aims, the time saved in accomplishing these objectives, and the economy that comes from minimizing redundancy (van der Eijk, Bloksma, and van der Kraan 1992). LUPo, or the Portuguese Unisyn Lexicon, is one of the first speech-dedicated applications to take full advantage of a collection of lexical resources as the basis for a text-to-speech system. Consisting of a pronunciation lexicon and rule system for generating accent-specific phonetic transcriptions for Portuguese, 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 theories.

1. Introduction

This paper presents a description of LUPo's functions for online use, and the architecture and administrative layer that support this application. Implications for practical lexicography are also presented in terms of the emerging role of the multi-dimensional and lexicographically rich Portal database as a pan Lusophone resource and basis for addressing natural language processing (NLP) problems. Our presentation will showcase the LUPo system, with a focus on its setup, results for the end user, and an easy-to-use lexicographic back end for maintaining and expanding the pronunciation database. More in-depth information about LUPo and the English Unisyn lexicon upon which it is based may be found in Ashby, Ferreira, and Barbosa (2009) and Fitt (2000), respectively.

2. Background

Unconstrained by the traditional expectations of a dictionary, lexical databases have the capability of being more dynamic in the types of functions they serve, audiences they target, and information they reuse. The Portal da Lingua Portuguesa (Janssen 2007), hereafter referred to as the *Portal*, is one such collection of lexicographic resources that is designed both for human consumption and computational exploitation (Janssen 2005). The Portal's modular architecture, and aims for extending this database to a global audience, including the general public and research communities alike, provide an ideal background for developing and supporting functions designed to enhance the user experience and serve as inputs to NLP systems.

One of the issues of reaching out to a pan Lusophone audience, or even say a Brazilian one, is the difficulty of selecting pronunciations that are not so abstract as to alienate users. Traditionally, this has not been of great concern to lexicographers, including authors of pronunciation dictionaries, who avoid presenting 'a variant of contestable status, be it regional, stylistic or social, to the extent that they wish to continue to describe the norm' (de Caluwe and van Santen 2003: 71-82). The problem for Brazilian Portuguese, pushing global

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

circumstances aside, is that there is no well accepted standard for describing how a word should sound. This difficulty extends to the problem of adapting speech technologies to multiple accents of Portuguese, where a paucity of such data exists.

2.1. LUPo

Users in search of accent-specific pronunciations for Portuguese will soon be able to access LUPo at http://www.portaldalinguaportuguesa.org using the Portal's existing word lookup interface. LUPo is currently capable of producing phonetic transcriptions for the Rio de Janeiro and São Paulo standards, and the actual spoken variants corresponding to these two cities (Brazil); along with standard European Portuguese, and the Braga and Lisbon accents (Portugal). Additional Portuguese language accents will become available during successive phases of the project, as we aim to include a multitude of accents spanning Africa, Asia, Europe, and South America.

LUPo currently generates accent-specific pronunciations for lemmas.[2] Pronunciations are generated via LUPo's hierarchically based rule system, 'on the fly', and do not occupy separate storage in the Portal. To access LUPo, the user types a lemma or word form in the search box, selects the phonetic transcription check box, and then chooses from a drop-down list of Portuguese accents in which to display the transcription. If the word exists in the Portal, the lemma, IPA transcription (with region label), and inflectional paradigm are displayed on the results page.

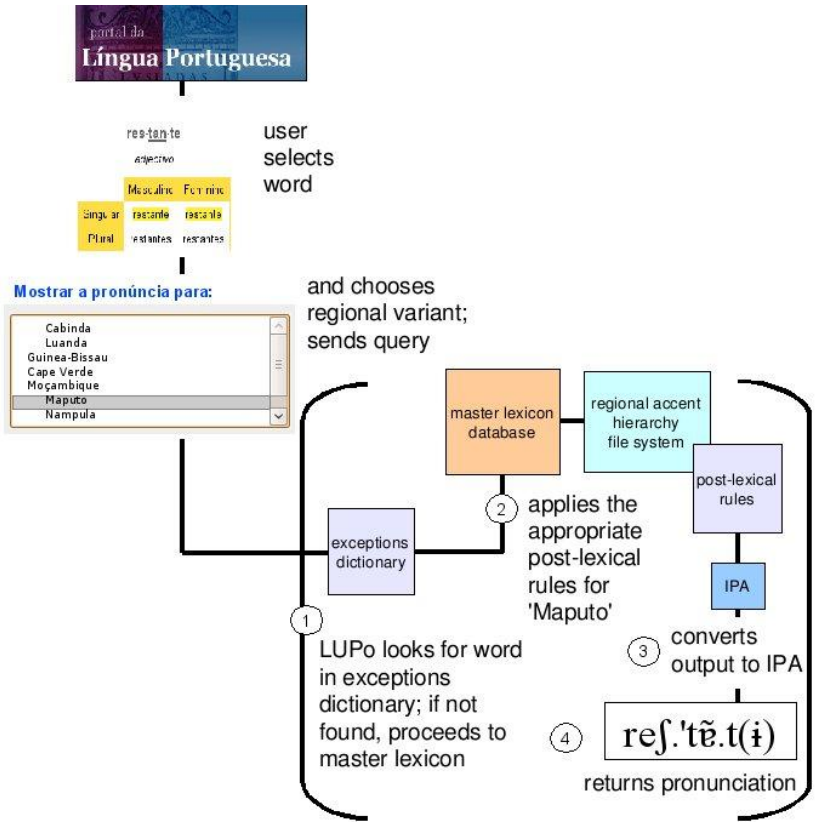

Figure 1. LUPo module made accessible via the *Portal da Língua Portuguesa*

2 A future version is in development for extending this capability to inflected forms, and for extending the module to multiword texts.

2.2. How it works

The Portal is a collection of lexical resources organized in MySQL tables, which are editable via a web-based PHP back office and accessible to the general public through the Portal website. A central table containing the list of lemmas (now over 200K entries) links all the tables. While each table, or resource, has the same general structure, the information it contains is unique and, therefore, non-redundant. New resources benefit from the existing infrastructure, and are controlled through a set of dedicated, unified, and user friendly back office modules for performing routine maintenance and expansion tasks.[3]

LUPo, being one such module, is unlike other grapheme-to-phone systems in that it resides within and has immediate access to a federated system of lexical relational databases. LUPo's core component, the master lexicon, was developed using this collected intelligence, and contains meta-phonemic transcriptions for every lemma in the Portal. This was accomplished in a relatively short period of time by executing scripts that made use of the lemmas' derivational links, spelling variants, part of speech information, foreign loan word or toponym attributes, and morphological boundaries. In short, we built LUPo's master lexicon using data that were, for the most part, already in place.

Perl scripts store the system of rules for transforming master lexicon entries into accent-specific phonetic transcriptions. When a user types a lemma, selects the desired accent, and clicks Enter, the system indexes the corresponding base pronunciation in the master lexicon and applies the appropriate post-lexical rules, as defined by LUPo's regional accent hierarchy.

Besides being built into a unified back office module for adding new entries, LUPo has its own editing module, thereby facilitating consistency checks and providing lexicographers with a broader view of the data it contains. By making use of free and open-source programming languages, LUPo may easily be deployed to any Unix-based server. The Portal's web-based back office makes it easy to have decentralized editing teams working from different locations. This is crucial, since the Portuguese language partners helping to expand LUPo are spread across several continents.

3. Conclusions

We sought to make the most of the Portal's modular design and flexible architecture by reusing lexicographic information to create a minimally specified pronunciation lexicon for accommodating a large number of regional variants of Portuguese. Applications include a Portal module for displaying regionally defined pronunciation variants, the basis of a subsequent text-to-speech application, and a Portuguese cross-dialectal database. In these and other important ways, we show how the reduced constraints on a lexical database (as distinct from dictionaries) enables the Portal to achieve the type of ‘heterogeneous client applications’ described in van der Eijk, Bloksma, and van der Kraan (1992), and meet the needs of a growing audience of online users.

3 See Barbosa, Ferreira, and Janssen (2008) for more information about the Portal's back end management system.

References

Ashby, S.; Ferreira, J. P.; Barbosa, S. (2009). 'Adapting the Unisyn Lexicon to Portuguese: Preliminary Issues in the Development of LUPo'. In *Proceedings of Iberian SLTech*. Porto Salvo.

Barbosa, S.; Ferreira, J. P.; Janssen, M. (2008). 'MorDebe Admin: A Lexicon Management System'. In *Proceedings of Euralex*. Barcelona.

De Caluwe, J.; van Santen, A. (2003). 'Phonological, Morphological and Syntactic Specifications in Mono-lingual Dictionaries'. In van Sterkenburg, P. (ed.). *A Practical Guide to Lexicography*. Amsterdam/Philadelphia: John Benjamins. 71-82.

Fitt, S. (2000). *Documentation and User Guide to UNISYN Lexicon and Post-lexical Rules*. *Technical Report*. Edinburgh: Centre for Speech Technology Research, University of Edinburgh.

Janssen, M. (2005). 'Lexical vs. Dictionary Databases: Design Choices of the MorDebe System'. In *Proceedings of COMPLEX 2005*. Budapest.

Janssen, M. (2007). *Portal da Língua Portuguesa* [on-line]. Lisbon: Instituto de Linguística Teórica e Computacional (ILTEC). http://www.portaldalinguaportuguesa.org.

Van der Eijk, P.; Bloksma, L.; van der Kraan, M. (1992). 'Towards Developing Reusable NLP Dictionaries'. In *Proceedings of COLING-92*. Nantes.