Adapting the Unisyn Lexicon to Portuguese: Preliminary issues in the development of LUPo

Simone Ashby, José Pedro Ferreira, Sílvia Barbosa

Instituto da Linguística Teórica e Computacional (ILTEC), Lisbon, Portugal

{simone,zpferreira,silvia}@iltec.pt

Abstract

This paper presents some preliminary issues and proposed solutions in the development of an accent-independent pronunciation lexicon for Portuguese, known as the Portuguese Unisyn Lexicon (LUPo). LUPo's objectives are presented within the context of the Portal da Língua Portuguesa knowledge base. Key considerations are addressed for encoding morphological boundaries, treating orthographic forms, and handling loan words. Here, it is argued that the knowledge-driven paradigm exemplified in the original English Unisyn Lexicon, along with the Portal da Língua Portuguesa's relational structure and rich lexicographic content present a good foundation for establishing a tightly integrated and well informed system.

Index Terms: lexicon, pronunciation, Portuguese accents, morphology, orthography, loan words, dictionary, relational database, speech synthesis

1. Introduction

Adapting speech technologies to accommodate a wider number of speakers, and represent regions and countries for whom such development concerns have largely been overlooked, carries significant economic and political weight in narrowing the global digital divide. 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 use of morphology and other types of lexicographic information. Moreover, these projects typically occur in isolation, and are governed by private sector interests that prohibit the sharing of data and tools.

Fitt's Unisyn Lexicon [1] presents a 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' keywords concept [2]. Key symbols, which can additionally 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 presents some of the preliminary issues and proposed solutions in the development of an accent-independent lexicon and rule system for generating accent-specific pronunciations in Portuguese, otherwise known as the Portuguese Unisyn Lexicon (LUPo). Our methodologies will be a reformulation of those originally employed by Fitt to adapt this successful paradigm to Portuguese, and take advantage of the relational structure and rich lexicographic content of the *Portal da Língua Portuguesa* [3] knowledge base (hereafter referred to as the 'Portal') to create a more integrated and well informed system. LUPo will capitalize on having direct access to a morphological parser, part of speech information, syllable boundaries, mappings of European and Brazilian Portuguese spelling variants, and etymological relationships.

Our approach to this work is explicitly knowledge-driven, as motivated by: (1) the need for greater linguistic input in statistically derived speech processing algorithms; (2) the success of the English Unisyn model in creating a highly scalable, extendible, and customizable lexicon for accommodating a large number of regional variants; and (3) complementary objectives for the establishment of a Portuguese cross-dialectal database and the first freely available online resource of its kind to provide phonetic transcriptions.

In subsequent sections of this paper, the design objectives and principal architectural components of LUPo are presented, followed by a brief description of the Portal. Three preliminary issues in the development of LUPo are then discussed concerning the encoding of morphological boundaries, treatment of orthographic forms, and the handling of foreign loan words.

2. LUPo

The LUPo project will produce an accent-independent pronunciation lexicon for Portuguese, along with tools for generating accent-specific output for lexical entries and multi-word texts. Users will have the option of accessing the open-source lexicon and tools as a standalone application or via the Portal (http://www.portaldalinguaportuguesa.org). The Portal module will be accessible as part of the page view for each lexical entry, wherein the user can select a desired accent to view the corresponding transcription for a given word. Online and offline users will also have access to a tool for entering a fixed amount of text, selecting a desired accent, and generating multi-word transcribed output for that accent, while also having the option to show the rules where they apply.

Our methodology will incorporate many of the strategies used by Fitt to create the English Unisyn Lexicon. Standard Brazilian Portuguese (BP) and European Portuguese (EP) lexicons will be merged to form the basis of the master lexicon. Lexical items will be represented as underspecified forms using key symbols. Regional hierarchical relationships will be encoded within the system to enable the inheritance of accent related features. Regional accents will be modeled one by one, based on the representativeness of the accent and the availability of data or informants. And the tools for generating surface output will be developed from Perl scripts.

Unlike Fitt's Unisyn Lexicon, LUPo will be stored in the multi-dimensional and lexicographically rich Portal database, thereby enabling the cross-referencing of semantically informed morphological parses, part of speech information, spelling variants, and foreign loan word attributes. 'Blackout', for example, which is stored in the Portal's dictionary of *estrangerismos* (loan words), will automatically be excluded from the application of post-

lexical rules, thereby eliminating the need to hard-code it (and other loan words) in an exception dictionary. Instead, the appropriate morpho-phonological rules will be re-routed to the *aportuguesamento* (Portuguese spelling adaptation) to which this word is mapped.

The Portal's morphological parser will enable LUPo to provide better lexical coverage while eliminating the need to encode redundant information. Thus, LUPo can be used to generate transcriptions for the words 'actividade' and 'practicamente' without the need to store these and other inflected nominal forms in the lexicon.

The redundancy of treating spelling variants as separate entries in the master lexicon will be avoided by making use of the Portal's existing system of cross referencing BP and EP forms. For example, the master pronunciation for the Portuguese superlative meaning 'great' will have a single entry in the master lexicon that maps to the respective BP and EP spellings *ótimo* and *óptimo*, along with corresponding forms from previous recent orthographic accords. These and other topics concerning utilization of the Portal's relational structure and lexicographic content are discussed in greater detail in section 4.

2.1. Objectives

In keeping with LUPo's development initiatives, our objectives for the project are as follows:

- Create an accent-independent master lexicon for Portuguese using an extended set of X-SAMPA-based typographical symbols that account for morphological boundaries and other phenomena for encoding lexical entries and processing conversions.
- Use a knowledge-driven approach to create a system of post-lexical morpho-phonological rules for processing conversions from the master lexicon to accent-specific targets.
- Develop tools for automatically generating accent-specific output for individual lexical entries and multi-word texts.
- Establish a regional accent hierarchy for specifying which rules apply to one or more accents, along with default inheritances for all the sub-nodes of a large geographic area, and a system for overriding these values.
- Create pronunciation models for: standard BP and EP; the Lisbon accent and at least one additional EP accent; the two major BP accents, as they are actually spoken in Rio de Janeiro and São Paulo, plus one or more other accents specific to Brazil; and three or more accents from the continents of Africa and Asia.
- Enhance the Portal by introducing richly detailed and varied phonetic content, and open it up to a wider audience.
- Provide the research community and general public with a freely available electronic data standard for: testing the results of different speech processing systems, conducting empirical analyses across multiple Portuguese accents, and facilitating L2 studies of Portuguese.
- Facilitate the entry of lesser or undocumented regional variants into the digital domain.
- Establish the basis for a subsequent project aimed at developing a TTS module for inclusion in the Portal and as a freely available, open-source standalone application.

2.2. Regional accent hierarchy

LUPo's architecture will be framed by a regional accent hierarchy that feeds accent-specific transformations by specifying whether and how the rules apply to geographically defined entities. As in Fitt's Unisyn Lexicon, it will consist of a system of files organized by COUNTRY, REGION, TOWN, and PERSON. Each node will inherit the features of the previous node, provided the inheritance is not broken by the introduction of new features at a lower level. As the lowest level, attributions to PERSON will override all other level specifications. Similarly, features attributed to local areas will override the settings of wider geographical areas.

Applying an example from Brazil, a general rule may be attributed to the COUNTRY node for expressing the coronal plosives /t/ and /d/ as [d] and [t] (Figure 1). However, given what we already know about accents specific to Salvador, Belo Horizonte, and Rio de Janeiro, it will be necessary to introduce a separate rule at the TOWN level for transforming these sounds into the affricates /ʧ/ and /ʤ/, thereby overriding the previous rule for these variants.

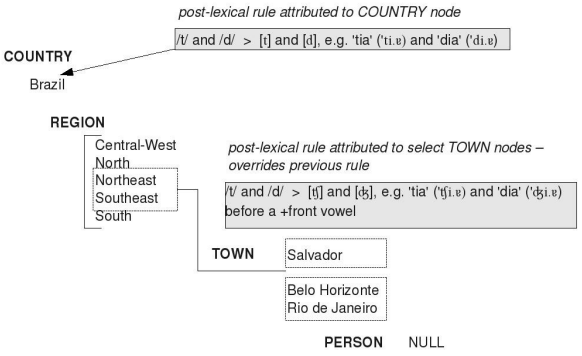

Figure 1: Sample regional accent hierarchy and post-lexical rule application for Brazil.

2.3. Master lexicon and key symbols

Accent-specific transformations will be derived from a master lexicon represented through the use of key symbols, based on an extended interpretation of the X-SAMPA alphabet. In addition to representing consonants and vowels, typographical markers will be used to encode stress, syllables, morphological constituents, and other phenomena, e.g. deletion of word-final rhotics for some BP varieties.

The specification of consonants and vowels requires an analysis of Portuguese phonology and the underspecification of segments. This work will be guided by the research team's phonologists, who will be involved in distinguishing types of systematic variation (for defining global symbols) from cases of allophonic variation (for defining post-rule symbols). Based on the work involved in developing the original Unisyn Lexicon [4], we predict that the more intensive aspect of this task will be the specification of morphological boundaries, as these will be instrumental in either triggering or blocking post-lexical rules.

The development of an accent-independent master lexicon will proceed as follows:

- Preprocess the 1961 edition of the *Dicionário da Língua Portuguesa da Academia Brasileira de Letras* (the only known large-scale lexical resource to contain phonetic transcriptions for BP) to adapt transcriptions to the criteria followed for the phonetic information in the Portal.
- Extract from the Portal the 2,217 high-frequency lemmas that correspond to [5]. Establish links to the relevant databases for getting spelling variants, syllable boundaries, part of speech information, foreign loan word attributes, frequency data, and morphological constituents.

- Merge the lexical entries in the Portal with those contained in the *Dicionário da Língua Portuguesa da Academia Brasileira de Letras*.
- Extract differences between the two resources and ensure that key symbols account for every confirmed instance of variation.
- Introduce frequency data and a system for weighting high-frequency words.
- Fully integrate LUPo into the Portal.

The end results of these actions will be the project's first instantiation of an accent-independent lexicon, which will provide the basis for applying post-lexical rules and generating accent-specific target output. The subsequent task of creating and evaluating morpho-phonological post-lexical rules will feed from this data, and result in an iterative set of improvements to the master lexicon through the modeling of new accents and application to a wider number of lemmas.

2.4. Morpho-phonological rule sets

The task of creating post-lexical morpho-phonological rule sets can be broken down into four successive types of activities. These include: (1) using semi-automatic methods to model the rules for generating standard BP and EP output (effectively recreating the transcriptions in the Portal and the processed version of the *Dicionário da Língua Portuguesa da Academia Brasileira de Letras*); (2) expanding the base lexicon and post-lexical rules for standard BP and EP to the 55K transcribed lemmas in the Portal, along with corresponding inflected forms; (3) and extending the rule sets to include as many actual spoken accents of Portuguese from Africa, Asia, Europe, and South America as the project's resource and time constraints allow.

For steps 1-3 above, we will also be performing a number of subroutines. For example, the modeling of each new accent will first be done for the 2,217 high-frequency lemmas in [5] before undergoing thorough evaluation by the project's dialect consultants and informants. Each time the data and rules are evaluated, it will be necessary to refine the key symbols, make changes to master lexicon, and add mappings. Once the initial set of pronunciations and rules have been thoroughly checked and modified, we will extend the rule set for that accent to the entire list of lemmas and inflected forms before, again, subjecting the resulting pronunciations and rules to a system of spot-checking and revision (Figure 2).

Throughout all of these processes, it will be necessary to define new entries in the exceptions dictionary, and create and adjust mappings to LUPo's regional accent hierarchy.

3. Portal da Língua Portuguesa

The Portal is an online knowledge base dedicated to providing the general public with a set of free Portuguese language resources, as well as serving as an open-source repository of lexicographic information for the research community. Its modular architecture enables it to extend far beyond the bounds of a traditional dictionary, while its relational structure offers the advantage of dealing with homographs, spelling variants, inflected forms, loan words, and etc.

The Portal was originally conceived as a lexical database focused on representing Portuguese inflectional morphology (i.e. *MorDebe*), and began in 2004. It continues to be maintained by the Instituto da Linguística Teórica e Computacional (ILTEC) in Lisbon, and contains more than 150,000 lemmas and close to 1.5 million word forms.

The Portal currently receives 4000-4500 hits by unique users per day and is increasingly regarded as a standard resource for inquiries about the Portuguese language. 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 free online resource to provide any manner of phonetic transcription data for Portuguese.

4. Preliminary development issues

In this section, three key issues in the initial development of LUPo are discussed in relation to the structural and lexicographic attributes of the Portal.

4.1. Morphology

As previously indicated, one of the advantages LUPo will have over the original Unisyn Lexicon is that it will reside within a set of relational databases containing detailed lexicographic information, e.g. syllable boundary, stress, and morphological encodings. Here, we intend to show how having these data will enhance the reliability of LUPo, while greatly reducing the amount of manual labor required to develop such a lexicon.

Vowel height presents one of the most challenging aspects of devising a suitable grapheme-to-phoneme system for Portuguese. Mid and low vowels are usually raised in an unstressed position, especially in standard EP. This height variation is predictable to an acceptable extent by knowing the underlying phonologic segment, syllable boundaries and stress position. For instance, /o/ will typically be pronounced as [o] in a stressed position and as [u] in an unstressed one (*sopa*, ['so.pɐ], 'soup'; *ensopado*, [ẽ.su.'pa.ðu] 'stew'), while /ɔ/ will always be pronounced [ɔ] in a stressed position and [u] or [ɔ] in an unstressed one (e.g. *roda*, ['ʀɔ.ðɐ]; 'wheel'; *rodinha,* [ʀɔ.'ði.ɲɐ]; 'small wheel' *rodagem*, [ʀu.'ða.ʒẽj], 'running in [e.g. an engine]'). As these examples show, the rules can be extended to words sharing the same root.

The problem is that orthography, stress position, and syllable structure are not enough to determine whether the 'o' in *roda* and *sopa* corresponds to a [-low] or [+low] segment. This also applies to the [-high] front vowels /e, ɛ/, both represented by the grapheme 'e'. Without a workaround for this problem, most of the entries containing a non-final 'o' or 'e', along with their morphologically related words, would need to be checked manually. This problem is ubiquitous in any phonetically transcribed lexicon of Portuguese.

The Portal will soon feature a database containing morphological information. The morphological database includes morphological boundary encodings, along with the identification of roots and affixes so that any given root or affix will bear a unique record ID. The completion of this database will enable us to: (1) improve the overall Portal architecture by tying words to their morphological constituents; and (2) predict the phonologic behavior of any word that contains a previously analyzed constituent (the latter of which should be particularly relevant in the development of LUPo). Through pursuit of this methodology, we expect to greatly reduce the need for performing manual checks.

LUPo will not only benefit from the lexicographic content contained in the Portal, but it will itself have an impact on the lemma list in the Portal's central database. Consider the following examples:

1a) *molho*, masc. noun, ['mɔ.ʎ + u], 'bundle'
1b) *molho*, masc. noun, ['mo.ʎ + u], 'sauce'
2a) *molhada*, fem. noun, [mɔ.'ʎ + að + ɐ], 'large bundle'
2b) *molhada*, fem. noun, past [mu.'ʎ + aðɐ], 'wet'[1]

Currently, the Portal is a meaningless dictionary, and homographs of the same class are treated as a single entry,

[1] The transcriptions provided in section 4.1 are for standard EP. The plus symbol '+' is used to mark morphological boundaries. The assumptions about the phonological system of Portuguese are based on the analyses of [6].

provided there is no formal feature (such as different inflectional paradigms) telling them apart. Once fully integrated in the Portal, the data contained in LUPo will enable us to identify homographs such as those in examples (2a) and (2b) above, and split them into different entries through the application of a set of formal features that, until recently, have been lacking in the Portal, i.e. phonetics.

4.2. Orthography

An example of a formal feature that currently sets homographs apart is an entry's inflectional paradigm. For example, *til* takes two plurals, *tis* and *tiles*, the former corresponding to a botanic species endemic of the Portuguese archipelagos and the latter to the graphical 'tilde' symbol. These words currently have two separate entries in the Portal.

There is an explicit link between entries that corresponds to alternative ways of spelling the same words, as in the case of *luzecu* and *luze-cu*, 'firefly'. When LUPo is fully integrated into the Portal, we will be able to profit from having this information by checking whether the transcription for each orthographic form is the same, thus guaranteeing correctness.

One problem associated with this setup is the fact that some alternative orthographic forms are only accepted in specific countries. This stems in part from the poor lexicographic tradition of many of the Portuguese speaking countries, the fact that only Brazil has a state-mandated language normalizing organization, and the lack of coordination between countries where Portuguese is recognized as an official language.

Fortunately, the Portal contains a database that explicitly links these entries, in each case identifying the country where such forms are acceptable. By using the information in this database, we will avoid generating pronunciations for spellings that are incorrect for a given regional variant. In other words, we will have an automatically generated exclusion list. Such is the case for the graphical pair *lambugem* (EP) and *lambujem* (BP). While in this particular example, the pronunciation will be unaffected (since *lambugem* and *lambujem* are pure orthographic variants), other country-specific word pairs exist that have distinct pronunciations.

Albeit the fact that Portuguese orthography is overtly phonological, and that a unified orthography has been in place since 2008, phonetic differences in each of the countries where Portuguese is spoken still surface at the orthographic level. An example is the generalized difference in the pronunciation of [+round, +front] vowels before a nasal consonant in proparoxytones (words with stress on the antepenultimate syllable). These are almost always produced as [-low] in Brazil and [+low] everywhere else. Given that diacritics are mandatory in Portuguese proparoxytones, this creates an orthographic divide between the politically defined regions where Portuguese is spoken, e.g. anônimo (BP) vs. anónimo (EP). By using the information already contained in the Portal, we will avoid generating double pronunciations for orthographic forms that, in fact, only exist in one system. We will also save a potentially huge amount of manual effort required to track them down, as this problem, alone, accounts for close to 1.2% of the lexicon.

4.3. Loan words

It has already been suggested that the Portal's dictionary of *estrangerismos* will be useful in excluding foreign loan words from the application of post-lexical rules. By mapping word borrowings in LUPo's master lexicon to their corresponding *aportuguesamento* in the *estrangerismos* dictionary, we avoid the need to treat these forms as exceptions. Thus, the Portal will be instrumental in both the identification of loan words and the rerouting of LUPo queries to the Portuguese spelling adaptation, if a reliable one exists.

The issue lies in the inconsistent handling of loan words in Portuguese. For, while some have been fully adapted, such as *holígane* for 'hooligan' and *surfe* for 'surf', others function more as graphical-phonetic hybrids. E.g. the representation of 'iceberg' as *icebergue*, which retains the [aɪ] sound from 'ice', thus barring the otherwise standard grapheme-to-phone mapping of 'i' > [i]. To further complicate matters, some *aportuguesamentos* lack authority in the spoken world. The Portal currently maps 'bluff' to the graphical *blefe*, which appears in dictionaries and is pronounced in standard BP as ['blɛ.fɨ], but is dispreferred for ['blɐf] in standard EP. Indeed, this last example helps to illustrate yet another problem in the treatment of loan words: the use of different spelling (and pronunciation) adaptations for regional variants of Portuguese.

Given the above problems and their implications for generating accurate pronunciations in LUPo, we will adopt a phased approach to dealing with loan words that makes use of the Portal's current ability to identify these forms, while steadily improving the manner in which *aportuguesamentos* are encoded in the *estrangerismos* dictionary, and formulating morpho-phonological rules that apply to the language of origin.

5. Conclusions

The Unisyn Lexicon presented in [1] offers a set of strategies and methodologies suitable for extending this model to Portuguese. By incorporating many of the same design constituents as Fitt's model and utilizing the existing architecture of the Portal, LUPo will function as a free, open-source accent-independent pronunciation lexicon for Portuguese. Issues concerning the representation of morphology, orthography, and loan words will no doubt present us with a variety of challenges. However, by constructing LUPo as a tightly integrated module that has full access to the lexicographically rich Portal data, we expect to overcome many of these problems, while enhancing the manner in which lexical entries are represented in the Portal.

6. Acknowledgements

The authors gratefully acknowledge the support of the Fundação para a Ciências e Tecnologia, and the cooperation of Susan Fitt, whose development of the original English Unisyn Lexicon is the inspiration for this work.

7. References

[1] Fitt, S., "Documentation and user guide to UNISYN lexicon and post-lexical rules," Technical Report, Centre for Speech Technology Research, University of Edinburgh, 2000. Online: http://www.cstr.ed.ac.uk/projects/unisyn/, accessed on 10 October 2008.

[2] Wells, J. C., *Accents of English*. Cambridge: Cambridge University Press, 1982.

[3] "http://www.portaldalinguaportuguesa.org."

[4] Fitt, S. "Morphological approaches for an English pronunciation lexicon," in *Proc. of Eurospeech*, Aalborg, Denmark, 2001.

[5] INIC and CLUL, *Português Fundamental: Vocabulário e Gramática*. Lisboa: Instituto Nacional de Investigação Científica, Centro de Linguística da Universidade de Lisboa, 1984.

[6] Mateus, M. H. and d'Andrade, E. *The Phonology of Portuguese*. Oxford: Oxford University Press, 2000.