A Comparative Description of GtoP modules for Portuguese and Mirandese using Finite State Transducers

D. Caseiro[*], **I. Trancoso**[*], **C. Viana**[†], and **M. Barros**[†]

[*] INESC-ID/IST
R. Alves Redol 9, 1000-029 Lisbon, Portugal

[†] CLUL
Av. Prof. Gama Pinto 2, 1749-016 Lisbon - Portugal

ABSTRACT

Mirandese is the second official language in Portugal. For ages it was only preserved as an oral language and it was the object of a recent orthographic convention. This paper describes our efforts in porting our grapheme-to-phone module from European Portuguese to Mirandese. We describe the main differences between the two languages that can affect this module and the set of new SAMPA symbols that had to be defined for the phonetic transcription of Mirandese. We then briefly cover our rule formalism and the composition of the various transducers involved in the grapheme-to-phone conversion, and describe the results obtained for the two languages. The use of finite state transducers allowed a very flexible and modular framework for deriving and testing new rule sets. Our experience led us to believe that grapheme-to-phone modules could be helpful tools for researchers involved in the establishment of orthographic conventions for other oral transmission languages.

1 INTRODUCTION

On the northeastern border of Portugal, there is a small set of villages where people use two languages in their daily life: Portuguese and Mirandese. The use of the latter is typically confined to conversations among relatives and neighbors. Altogether the population that speaks Mirandese does not exceed 12,000. Its recognition as official language is fairly recent (1999). Nowadays, it is also taught as an optional course in high school.

Mirandese (*Mirandés*) is a romance language, related to Asturian-Leonese, and for several centuries it was preserved only as an oral transmission language. Despite belonging to this geolinguistic domain, Mirandese is different from all the languages spoken in the contiguous territories. The first attempt at creating a spelling norm for Mirandese was done by a nineteenth century linguist José Leite de Vasconcelos [1], who based the norm on the pronunciation of the village *Duas Igrejas* and used diacritics in order to faithfully encode this pronunciation. These diacritics created so many doubts in whoever tried to follow him, that the end result was the appearance of many spelling modes. This provided a strong motivation for the recent efforts to create an orthographic convention [2] in order to establish unifying criteria for writing in this language [1]. In the design of the convention, it was not attempted to annotate all local pronunciation variants, but only those that either show up regularly or represent a phenomenon that was once regular in the language and nowadays is realized differently in the set of villages.

The motivation for deriving grapheme-to-phone rules for Mirandese was to build a tool that may help native speakers to learn how to read and write, as well as students interested in that language.

As a starting point, we used the rules that we had derived for European Portuguese (EP). The first grapheme-to-phone module that we developed for EP was in the context of a rule-based system (DIXI)[3]. In fact, none of the data-driven tools that we had developed since then (either based on neural networks [4] or *CARTs* - Classification and Regression Trees [5]) were suited for Mirandese, given the small amount of training material.

Our most recent efforts in terms of grapheme-to-phone conversion were based on Finite State Transducers (*FSTs*) [6], motivated by their flexibility in integrating multiple sources of information and other interesting properties such as inversion. The knowledge-based approach using *FSTs* is flexible enough to allow easy porting to similar languages or other varieties of Portuguese.

This paper describes our efforts in porting our *FST*-based grapheme-to-phone module from Portuguese to Mirandese, starting with a very brief discussion of the main affinities and differences between the two languages, and our proposal for extending our machine

[1] http://mirandes.no.sapo.pt

ISBN 1-876346-48-5

readable phonetic alphabet in order to cover Mirandese as well (Section 2). We then describe the rule formalism we adopted and the composition of the various transducers involved in the grapheme-to-phone conversion (Section 3). The following two sections describe some transducer details for each of the two languages and the corresponding grapheme-to-phone results.

2 CONTINUITIES AND DIFFERENCES BETWEEN THE TWO LANGUAGES

A very detailed description of the main continuities and differences between Portuguese and Mirandese can be found in [7]. Here, for the sake of space, we only try to emphasize the most relevant ones.

Mirandese exhibits several continuity features relative to either what is commonly designated as standard European Portuguese or the one spoken in the Northern part of the country: the existence of initial *f-*; unvoiced affricate *ch* derived from the Latin *cl-*,*pl-* and *fl-*; voiced palatal consonant derived from the Latin *ly* and *c ´l*; vocalization of the *-ct-* group as *-it-*; existence of oral diphthongs *ei* and *ou*; voiced and unvoiced postalveolars; four voiced and unvoiced sibilants; absence of *v* (*b* used instead); existence of nasal vowels and diphthongs; etc.

Some of the main differences are: the preservation of the Latin intervocalic *-n-* and *-l-* (originating differences in syllabic structure); palatalization of the Latin *-nn-*, *-mn-*, and *-ll-*; existence of initial *lh-* originating from the Latin *l-*; the tendency for the non-existence of high unstressed vowels in initial position; the reduction of the initial *des-* to a sibilant consonant; male form of the definite article reduce to *l*, with two phonetic values depending on the beginning of the next word and the ending of the previous one; existence of raising diphthongs *ie* and *uo* (although with several differences in realization from village to village); etc.

The SAMPA phonetic alphabet for EP[2] was defined in the framework of the SAM_A European project and includes 38 phonetic symbols. We tried to limit to the minimum the creation of new SAMPA symbols to cover Mirandese, which means that the same symbol may be used with slightly different realizations in EP and Mirandese (e.g., while the symbols [E] and [O] account for low vowels in EP, they correspond to mid-low ones in Mirandese). New symbols were only introduced to account for either phonological contrasts involving 3 pairs of coronal fricatives, or important differences in contextual variation (e.g. while in standard EP *e* is always realized as [e˜] when followed by a tautosyllabic nasal consonant, in Mirandese it corresponds to [E˜] and [@˜] in stressed and unstressed positions, respectively). Table 1 lists the additional symbols, together with some examples.

SAMPA	Orthography	Transcription
@˜	centelha	s@˜t"ejL6
E˜	benga	b"E˜g6
B	chuba	tS"uB6
D	roda	R"OD6
G	pega	p"EG6
s_	sol	s_"Ol˜
z_	rosa	R"Oz_6

Table 1: Additional SAMPA symbols for Mirandese.

3 TRANSDUCER COMPOSITION

The rule formalism used in our approach was described in detail in [6], and will therefore be only summarized here very briefly. The rules are specified using a rule specification language, whose syntax resembles the BNF (Backus Naur Form) notation, allowing the definition of non-terminal symbols (e.g. `$Vowel`). Regular expressions are also allowed in the definition of non-terminals. Transductions can be specified by using the *simple transduction* operator $a \rightarrow b$, where a and b are terminal symbols. The most general command is:

`OB_RULE` $n,\ \phi \rightarrow \psi / \lambda \_ \_ \rho$

where n is the rule name and $\phi, \psi, \lambda, \rho$ are regular expressions. `OB_RULE` specifies a context dependent *obligatory rule*, and is compiled using Mohri and Sproat's algorithm [8]. The meaning of the rule is the following: when ϕ is found in the context with λ on the left and ρ on the right, ψ will be applied, replacing it.

This means that we are preserving the semantic of the original DIXI's rules. Although some similarities may be found between DIXI's and a Two-Level Phonology approach ([9], DIXI's rules were not two-level rules: contexts were not fully specified as strings of two-level correspondences and within the set of rules for each grapheme, a specific order of application was required. Default rules needed to be the last and in some cases in which the contexts of different rules partially overlapped, the most specific rule needed to be applied first.

The rules of the grapheme-to-phone module are organized in various phases, each represented by transducers that can be composed to build the full module. Figure 1 shows how the various phases are composed.

The first and last phases deserve a particular explanation. First, the grapheme sequence $g_1, g_2, ..., g_n$, is transduced into $g_1, \_, g_2, \_, ..., \_, g_n$, where _ is an *empty* symbol, used as a placeholder for phones (`introduce-phones`). Each rule will replace _ with

[2] http://www.l2f.inesc-id.pt/˜imt/sampa.html

the phone corresponding to the previous grapheme, but keeping it. The context of the rules can now freely refer to the graphemes. In this way, we avoid rule dependencies that would be necessary if we had just replaced graphemes by phones: the first rule would only have graphemes in its context, while the last ones have mainly phones. The very last rule removes all graphemes, leaving a sequence of phones (`remove-graphemes`).

The second phase (`exception-lexicon`) contains the pronunciation of words not covered by the rules. This lexicon also includes all the short function words that are not stressed and would be so, if covered only by the rules of the following phase. In fact, this third phase (`stress`) consists of rules that mark the stressed vowel of the word. The fourth phase (`prefix-lexicon`) consists of pronunciation rules for compound words, namely with roots of Greek or Latin origin such as *tele* or *aero*. The fifth phase (`gr2ph`) forms the bulk of the system, and consists of a set of rules that convert the graphemes (differentiating between diacritics) to phones. The sixth phase (`sandhi`) implements word co-articulation rules across word boundaries. Due to the fact that our test set consists of isolated words, this phase was not tested in this paper.

```
introduce-phones   o
exception-lexicon  o
     stress        o
 prefix-lexicon    o
     gr2ph         o
     sandhi        o
remove-graphemes
```

Figure 1: Phases of the knowledge based system.

The following example (in EP) illustrates the specification of 2 `gr2ph` rules for deriving the pronunciation of grapheme *g*: either as /Z/ (e.g. *agenda, gisela*) when followed either by *e* or *i*, or as /g/ otherwise (SAMPA symbols used).

```
OB_RULE 0200, g EMPTY -> g _Z \
    / NULL ___ ($AllE | $AllI)

OB_RULE 0201, g EMPTY -> g _g \
    / NULL ___ NULL
```

4 RESULTS FOR EUROPEAN PORTUGUESE

The transducer approach for EP involved a large number of rules: 27 for the `stress` transducer, 92 for the `prefix-lexicon` transducer, and 340 for the `gr2ph` transducer. The most problematic one was the latter. We started by composing each of the other phases into a single *FST*. `gr2ph` was first converted to a *FST* for each grapheme. Some graphemes, such as *e*, lead to large transducers, while others lead to very small ones. Due to the way we specified the rules, the order of composition of these *FSTs* was irrelevant. Thus we had much flexibility in grouping them and managed to obtain 8 transducers with an average size of 410k. Finally, `introduce-phones` and `remove-graphemes` were composed with other *FSTs* and we obtained the final set of 10 *FSTs*.

In runtime, we can either compose the grapheme *FST* in sequence with each *FST*, removing dead-end paths at each step, or we can perform a lazy simultaneous composition of all *FSTs*. This last method is slightly faster than the DIXI system.

In order to assess the performance of the *FST*-based approach, we used a pronunciation lexicon built on the PF ("Português Fundamental") corpus. The lexicon contains around 26,000 forms. 25% of the corpus was randomly selected for evaluation. The remaining portion of the corpus was used for training or debugging. As a reference, we ran the same evaluation set through the DIXI system, obtaining an error rate of 3.25% at a word level and 0.50% at a segmental level.

The first test of the *FST*-based approach was done without the *exception lexicon*. The *FST* achieved almost the error rate of the DIXI system it is emulating, both at a word level (3.56%) and at a segmental level (0.54%). When we integrate the exception lexicon used in DIXI, the performance is exactly the same as for DIXI. We plan to replace some rules that apply to just a few words with lexicon entries, thus hopefully achieving a better balance between the size of the lexicon and the number of rules.

5 RESULTS FOR MIRANDESE

The porting of the *FST*-based approach from EP to Mirandese involved changing the `stress` and `gr2ph` transducers. The stress rules showed only small differences compared to the ones for EP (e.g. stress of the words ending in *ç*, *n*, and *ie*). The `gr2ph` transducer was significantly smaller than the one developed for EP (around 120 rules), reflecting the much closer grapheme-phone relationship.

The porting effort was done in a bottom up way i.e., we took only the most generic rules for EP and made the necessary modifications to adapt these rules to Mirandese, namely to the ones involving the new phonetic symbols. We then consecutively added more rules to this set of default ones. The rules for consonant conversion are very simple. The ones for the 5 vowels take up around 60% of the rules. As an example of differences between the two rule sets, let us take the ones

ISBN 1-876346-48-5

that involve the conversion of graphemes *p* and *x*. Both graphemes have a single rule in Mirandese whereas in EP we needed 8 rules to account for the cases where *p* was not pronounced, and 16 to account for the 4 possible pronunciations of *x*.

The hardest step in this porting effort involved the definition of development and test corpora for Mirandese. Whereas for EP the choice of the reference pronunciation (the one spoken in the Lisbon area and most often observed in the media), was fairly easy, for Mirandese it was a very hard task, given the differences between the pronunciations observed in the different villages of the region. This called for a thorough review of the lexicon, and checking with native speakers. For development, we used a small lexicon of about 300 words extracted from the examples in [2]. For testing, we used a manually transcribed lexicon of around 1,100 words, built from a corpus of oral interviews conducted by CLUL in the framework of the ALEPG project (*Atlas Linguístico-Etnográfico de Portugal e da Galiza*). As a starting point, we selected the interviews collected in the village of *Duas Igrejas*, which was also the object of the pioneering studies of Mirandese by Vasconcelos.

Our first tests were done without an exceptions lexicon. In our small development set, we obtained only 11 errors (3.68% error rate at a word level), all of which are exceptions (foreign words, function words, etc.). For the test set, a similar error rate was obtained (3.09%). Roughly half of the errors will have to be treated as exceptions, and half correspond to stress errors.

This reduced error rate was expected, given the coherence of the recent orthographic convention. In fact, the number of errors obtained when we first run the test set through the grapheme-to-phone module was very high. However, after checking with native speakers, it became clear that the majority of the mispronunciations were due to incorrectly spelled entries in this corpus. Most of them were observed for words that according to the convention should be written with *ie*, but were written instead with *e* or *i*, because, alongside with the pronunciation [je], we can also find [e] or [i], depending on the village.

6 CONCLUDING REMARKS

This paper described an *FST*-based approach to grapheme-to-phone conversion that was first developed for European Portuguese and later ported to the other official language in Portugal - Mirandese. The hardest part of this task turned out to be the establishment of a reference pronunciation lexicon that could be used as the development corpus, given the observed differences in pronunciation between the inhabitants of the small villages in that region.

The use of finite state transducers allows a very flexible and modular framework for deriving new rule sets, testing the consistency of the orthographic conventions and helping in the establishment of coherent orthographic forms for dictionary entries. Based on this experience, we think that grapheme-to-phone systems could be useful tools for researchers involved in the establishment of orthographic conventions. Moreover, such tools could be helpful in the design of such conventions for other oral transmission languages in the CPLP (Community of the Portuguese speaking countries).

7 ACKNOWLEDGMENTS

We gratefully acknowledge the help of António Alves, Matilde Miguel, and Domingos Raposo.

REFERENCES

[1] J. Vasconcellos. *Estudos de Philologia Mirandesa*. Imprensa Nacional, Lisboa, 1900.

[2] M. Barros-Ferreira and D. Raposo, editors. *Convenção Ortográfica da Língua Mirandesa*. Câmara Municipal de Miranda do Douro - Centro de Linguística da Universidade de Lisboa, 1999.

[3] L. Oliveira, M. Viana, and I. Trancoso. A rule-based text-to-speech system for portuguese. In *Proc. ICASSP '1992*, San Francisco, USA, March 1992.

[4] I. Trancoso, M. Viana, F. Silva, G. Marques, and L. Oliveira. Rule-based vs. neural network based approaches to letter-to-phone conversion for portuguese common and proper names. In *Proc. ICSLP '94*, Yokohama, Japan, September 1994.

[5] L. Oliveira, M. C. Viana, A. I. Mata, and I. Trancoso. Progress report of project dixi+: A portuguese text-to-speech synthesizer for alternative and augmentative communication. Technical report, FCT, January 2001.

[6] D. Caseiro, I. Trancoso, L. Oliveira, and C. Viana. Grapheme-to-phone using finite state transducers. In *Proc. 2002 IEEE Workshop on Speech Synthesis*, Santa Monica, CA, USA, September 2002.

[7] M. Barros. A situação actual da língua mirandesa e o problema da delimitação histórica dos dialectos asturo-leoneses em portugal. *Revista de Filología Románica*, 18:117–136, 2001.

[8] M. Mohri and R. Sproat. An efficient compiler for weighted rewrite rules. In *34th Annual Meeting of the Association for Computational Linguistics*, Santa Cruz, USA, 1996.

[9] K. Koskenniemi. *Two-Level morphology: A general Computational Model for Word-Form Recognition and Production.* PhD thesis, University of Helsinki, 1983.