Casa de la Lhéngua: a set of language resources and natural language processing tools for Mirandese

José Pedro Ferreira[*] [1,2,3], Cristiano Chesi[1,4,5], Daan Baldewijns[1], Miguel Sales Dias[1,5], Daniela Braga[5,6], Fernando Miguel Pinto[1], Hyongsil Cho[1,5], Margarita Correia[3,7], Amadeu Ferreira[2]

[1]Microsoft Language Development Center, Portugal, [2]Associaçon de la Lhéngua Mirandesa, [3]Instituto de Linguística Teórica e Computacional - ILTEC, [4]Instituto Universitario di Studi Superiori di Pavia - IUSS, [5]ISCTE-IUL University Institute of Lisbon, [6]Voicebox, [7]Universidade de Lisboa

*contact author: jpferreira@gmx.com

Abstract

This paper describes the efforts for the construction of Language Resources and NLP tools for Mirandese, a minority language spoken in North-eastern Portugal, now available on a community-led portal, *Casa de la Lhéngua*. The resources were developed in the context of a collaborative citizenship project led by Microsoft, in the context of the creation of the first TTS system for Mirandese. Development efforts encompassed the compilation of a corpus with over 1M tokens, the construction of a GTP system, syllable-division, inflection and a Part-of-Speech (POS) tagger modules, leading to the creation of an inflected lexicon of about 200.000 entries with phonetic transcription, detailed POS tagging, syllable division, and stress mark-up. Alongside these tasks, which were made easier through the adaptation and reuse of existing tools for closely related languages, a casting for voice talents among the speaking community was conducted and the first speech database for speech synthesis was recorded for Mirandese. These resources were combined to fulfil the requirements of a well-tested statistical parameter synthesis model, leading to an intelligible voice font. These language resources are available freely at *Casa de la Lhéngua*, aiming at promoting the development of real-life applications and fostering linguistic research on Mirandese.

Keywords: language resources, minority language, Mirandese, speech synthesis, lexical database.

1. Introduction

While current discussion on NLP resources for the dominant European languages focuses on theoretical proposals to further develop them and on ways to bridge the gap with the most advanced systems of mainstream languages such as English (Branco et al. 2012), some, Mirandese being the prototypical example, still lack the most basic language resources. Prior to the 1990s, when the language was officially recognized and normalization efforts began (Barros Ferreira 2002), little attention had been given to it, and to this day the most complete studies are still those done at the turn of the 19th century (Leite Vasconcelos 1900). Lack of an established writing standard, little written production, and low perceived financial value were responsible to the near-inexistence of modern studies and resources. The inexistence of such resources contributes to low perceived sociolinguistic value of the language among its speakers and lack of access to modern technology in their native tongue.

This paper describes the development of basic NLP tools for Mirandese from scratch, having a speech synthesis system (Braga et al. 2010) as the final objective. It intends to describe the process, illustrating, through the adaptation and reuse of existing tools for phylogenetically-close languages, how the completion of the task was possible.

All the resources and tools were developed via collaboration between Microsoft, the Instituto de Linguística Teórica e Computacional (ILTEC) and the Mirandese Language Association (ALM), and are being made freely available through the open access web portal Casa de la Lhéngua (http://casadelalhengua.org) along with the Microsoft Language Development Centre (MLDC, www.microsoft.com/pt-pt/mldc/default.aspx).

2. On Mirandese

Mirandese is a minority language spoken in North-eastern Portugal. It belongs to the group of Astur-Leonese languages, being closely related to Asturian, spoken to this day in areas of the Asturias and Leon regions in Spain, with which Mirandese no longer retains a linguistic continuum, due mostly to linguistic policies during the Franco years in Spain. Unwritten for most of its history, Mirandese was first identified and studied in the 19th century.

Throughout the 20th century, strong demographic changes, namely the exodus of large numbers through emigration in the 1940s and in the 1970s, and an influx of non-Mirandese-speaking workers from various parts of the country in the 1950s and 1960s, along with the rise of Portuguese-spoken-only media, led to inter-generational transmission to be gradually abandoned, which left the language seriously threatened. Today, it is estimated that Mirandese is spoken by no more than 5000 people as a first language, and by at most 15.000, even when taking into account second language speakers, including those living outside of Terra de Miranda.

In the 1990s, strong efforts began to be made to make the survival of Mirandese possible: a spelling convention was developed by a group of linguists and native speakers, and the language was introduced into the formal education curricula locally. The Portuguese Government finally granted the language co-official regional status in 1999. These initiatives had a strong impact on the community: what used to be a reason for shame for the diglossic community that speaks Mirandese, this language increasingly started showing up on book shelves and in the local and national media, in social media and on the Web. Albeit seriously threatened as a mother tongue, it is currently used and learned by a large part of the population of Miranda in increasingly more formal contexts, currently enjoying a period of non-artificial revival.

3. Language Resources for Mirandese

Before the work described in this paper, there were few resources available for Mirandese (Trancoso 2003) (Caseiro 2003). The most detailed linguistic descriptions are to this day still those made by Leite Vasconcelos in 1900, more than 100 years ago, in part due to the lack of available data (Barros Ferreira, 2002), leaving academic researchers, school teachers and students with little up-to-date basic language resources for the formal study of the language. On the other hand, the fact that Mirandese is present in more and more contexts and digital supported formats, seems to be a decisive factor in the way its speakers perceive it, granting it a higher sociolinguistic profile (Barros Ferreira 2002). These facts, along with the joint will, from the academia and the industry, to develop a Text-to-Speech (TTS) system for Mirandese, was the spark behind the effort to create the resources presented in this paper, under the framework of a citizenship project.

To be in a position to achieve the end goal, an intelligible TTS system for Mirandese, a number of languages resources and NLP tools that are usually available for languages spoken by a larger communitywere created: a large text corpus in digital format, a phone-set and a large phonetic lexicon with part-of-speech (POS) tagging, a series of NLP tools, such as a tokenizer, an inflectioner, a hyphenator, a text normalization module (TN) and a grapheme-to-phoneme (GTP) converter.

For the complete language dependent frontend (including the text analysis that uses the mentioned language resources and tools and the voice font), and language agnostic HMM-based backend (run-time) pieces of the TTS system, as well as for the methodology of the selection and recording of the voice talent, standard Microsoft processes, tools and technologies were used. These ensure high quality of the resulting TTS voice, and were adapted from previously established and mature processes, designed for larger and better-resourced languages, such as European Portuguese.

The text corpus was compiled using data provided by the Mirandese Language Association, a non-for-profit local organization, most of it generously provided by publishers and newspapers sources. After being cleaned up, it contained over 1M tokens. The phonetic lexicon was built starting with a 25K-entry dictionary (Ferreira & Ferreira 2001), and later complemented with data retrieved from the corpus, after having it tokenized, inflected and POS-tagged using customizations of García & Gamallo (2007) and Janssen (2012). Simultaneously with these tasks, a voice talent was selected for high-quality recording sessions of several thousand prompts retrieved from the corpus. Over 7 hours of speech were collected. A rule-based TN module was developed, following a previously adopted approach (Ferreira et al. 2011a). The rule-set for European Portuguese was used as a starting point for Mirandese. The proximity between the two languages enabled re-using this pre-existing resource, thus speeding up the development process. These resources and tools are described in more detail below.

3.1. Corpus

The corpus was created from raw textual data collected by ALM, most of it generously provided by publishers, newspapers and the authors themselves. Those data were then complemented with data crawled from the web using the work developed by Scannell (2007), which allowed us to harvest authentic data from sources such as blog posts, comments and personal websites, increasing the total size of the corpus to over one million tokens.

After being tokenized, lemmatized and POS-tagged using an adapted version of NeoTag (Janssen, 2012) The data in the corpus was used to retrieve the text prompts used in the recordings, to extract lexical entries which were added up to the lexicon, and to test the text normalization (TN) module.

3.2. TN module

A TN module for Mirandese was developed taking advantage of the availability of a counterpart file for European Portuguese, using in-house developed software (Chesi et al. 2011, Cho et al. 2011). The proximity between the two languages and the great influence Portuguese has over written Mirandese made it possible to reuse the pre-existing resource to a great extent, changing only minimally several hundreds of the thousands of rules and terminals that compose the module, thus greatly speeding up the TN module development process.

The final TN module is composed of about 5.000 (contextual) rules dealing with the expansion of cardinal numbers (“12” ↔ “twelve”), date-time spell out (“12:00” ↔ “noon”), for instance. The following TN categories were developed for Mirandese are: cardinals, ordinals, percentage expressions, simple mathematical expressions, date and time expressions, currency, phone numbers, roman numerals, fractions, measurement expressions, titles, addresses, URLs and email, and file paths. Where needed, context-sensitive rules were developed, for instance to ensure that the gender agreement in noun phrases containing a cardinal number is correct.

3.3. Lexicon

The lexicon was based on a 25K lemma list of the ongoing work on a Mirandese-Portuguese dictionary (Scannell 2007), complemented with the most frequent lemmas in the compiled text corpus, which was previously tokenized, lemmatized and POS-tagged using customizations of Müller et al. (2000) and Ratnaparkhi et al. (1996). The resulting lemma list was then inflected using a version of Zen et al. (2007), syllabified, stress-marked and converted to IPA phonetic transcription using an in-house adaptation of an unpublished two-step Perl-based GTP tool originally developed for European Portuguese (Zheng et al. 2000). This simple regular expression string replacement set of scripts starts by marking stress and syllable divisions in the orthographic forms and, in a subsequent phase, applies an ordered set of grapheme-phone transformation rules based on syllable position and stress, taking advantage of the relatively shallow phonemic orthography of Mirandese.

In the end, we succeeded in compiling a fully annotated large lexicon, consisting of 124.360 word forms, for which standard orthography, pronunciation (syllabified, stress marked IPA transcription), POS (e.g. VER for verb, ADJ for adjective) as well as other morphological information (like mood, tense, person and number features) is provided, as exemplified in (1):

(1) *Word* | *Pronunciation* | *POS* | *morphological features*
abacelhe | ax - b ax - s eh 1 - lh aex | VER |
subjunctive, present, 3person, sing
melhor | m aex - lh oh r 1 | ADJ | qualifying, masc,
sing

Using algorithms like the one discussed in Zheng et al. (2000), we managed to easily generate syllabification rules that allow us to segment words out of lexicon and pair them with their correct pronunciation and the most likely stress.

3.4. Phone set

A crucial resource developed is a complete phone set for Mirandese, consisting of 43 distinct phones, determined partly by an accoustic study (Ferreira et al. 2011b). This resource specifies the full list of available phones paired to distinctive features and parameters that are used to train the voice model (v=voiced, -v=voiceless, a=alveolar, af= affricate, b =bilabial, c=central, d=dental, f=fricative, n=nasal, l=lateral, lb=labiodental, r=rounded, s=sonorant, u= uvular, ve=velar, w=vowel):

(2)

Phone	*features*
a	v, c, low s, w,
aex	v, c, high s, w,
an	v, c, low s, n, w,
ax	v, c, mid-high s, w,
b	v, b, p,
ch	-v, post-a, af
d	v, d, p,
eh	v, front mid-high s, w
ehn	v, front mid-low s, n, w
exn	v, c, high s, n, w
f	-v, ld, f
g	v, ve, p
i	v, front high s, w
in	v, front high s, n, w
j	v, p, semi w
je	rising front high s, diphthong
jen	rising front high s, n, diphthong
k	-v, ve, p
l	v, a, l, s
lg	v, l, ve, a, s
lh	v, p, l, s
m	v, b, n, s
n	v, a, s, n
ng	v, ve, n, s
nh	v, p, n, s
oh	v, back mid-high r, s, w
ohn	v, back mid-low r, s, n, w
p	-v, b, p
r	v, a, tap,
rr	v, a, trill,
s	-v, a, f
sh	-v, post-a, f
ss	-v, apico-a, f
t	-v, d, p
u	v, high r, s, w
un	v, high r, s, n, w,
w	v, l-ve, semi-w,
wo	rising back high s, diphthong
won	rising back high s, n, diphthong
x	-v, u, f
z	v, a, f
zh	v, post-a, f
zz	v, apico-a, f

3.5. Voice recordings

A voice talent was selected for high-quality recording sessions using 5.132 prompts retrieved from the corpus. The prompts consisted of full sentences whose selection was based on character length and phonological relevance (richness of phonological contexts), determined by an existing algorithm of the Microsoft TTS system software suite.

The voice talent was selected from a pool of candidates by a jury of 20 native speakers of varying ages, provenances and sociolinguistic profiles. Public advertisement in the local media and through speaking community networking helped greatly in getting a reasonable number of candidates with the correct profile: native speakers, having at least undergone undergraduate studies and no older than 40. The jury listened to recordings of each candidate reading an expressive text, and filled in a short questionnaire. The two highest ranked candidates in this first phase underwent one hour of pure speech studio recording under loose scrutiny and were again ranked by the jury, who this time had to fill in a more thorough questionnaire developed for subjective pleasantness assessment, using a methodology published before (Braga et al. 2008).

Finally, the selected voice talent was recorded over two weeks in a high-quality studio under the close supervision of a Language Expert, who monitored the clarity, accent, and completeness of the recording process, simultaneously checking the adequacy of each prompt and making textual corrections where needed. The recording process yielded over 7 hours of speech data.

Those data were semi-automatically trimmed and chopped into individual files using a standard acoustic marker inserted between prompts during the recording process, making it easier to map each individual recording file to a prompt. All the individual recordings were listened to by a Language Expert, and removed from the pool of available data when quality or conformity with the prompt was not met.

4. The Text-to-Speech System

The Linguistic Resources described in §3 have been used for training the TTS backend and building the Voice Font. Here we will briefly describe the full procedure we used to build the first fully intelligible font ever built for Mirandese.

4.1. Runtime Text Analysis of the Frontend

In runtime, the input user prompts are analyzed after being properly segmented (both sentence segmentation and word segmentation) and fully normalized (using the TN rules described in §3.2). This procedure is automatically carried out using rule-based sentence breakers, contextual Text-Normalization rules (as discussed before), and POS taggers (e.g. Ratnaparkhi et al. 1996). Once the single words are normalized and categorized, the correct pronunciation is retrieved from the lexicon and assigned to the current word. As a result, the prompts are enriched as shown in (3) ("w"= token, "v"=written form, "p"= pronunciation). The found pronunciations then drive the TTS backend system that generates the audio form.

(3) Por baixo de las saias de las rapazas ("in under of the skirts of the ladies")

```
<w v="Por" p="p . oh 1 . r" type="normal" length="3" />
<w v="baixo" p="b . a 1 . j - ch . u" type="normal"
     offset="4" length="5" />
<w v="de" p="d . aex" type="normal" offset="10"
     length="2" />
<w v="las" p="l . ax . ss" type="normal" offset="13"
     length="3" />
<w v="saias" p="ss . a 1 . j - ax . ss" type="normal"
     offset="17" length="5" />
<w v="de" p="d . aex" type="normal" offset="23"
     length="2" />
<w v="las" p="l . ax . ss" type="normal" offset="26"
     length="3" />
<w v="rapazas" p="rr . ax - p . a 1 - z . ax . ss"
     type="normal" br="4" offset="30" length="7" />
<w v="." type="punc" br="4" />
```

4.2 Training the TTS backend and building the Voice Font

With the fully annotated prompts, we run the voice font training and building procedure. The approach used to train the font model is referred to as Statistical Parameter Synthesis (SPS) and it is based on standard hidden-Markov-model approaches to TTS (HTS, Zen et al. 2007, Zen et al. 2009). The basic idea is that the waveform is stable during short time phrases and can be approximated by Gaussian models that represent the parameter distribution.

Given a sequence of observations (O_t), we expect an observation O_i at time i to belong to one state Q (e.g. 1, 2 or 3). In i+1, O_{i+1} might still be Q or a different state and this must be modeled depending on the transition probability built on the previous observation based on the aligned wave–prompts pairs that we have used for training.

We used a decision tree based on relevant distinctive feature associated to a given state (expressed by Gaussian models), to better estimate the state sequence (as well as its duration and excitation). The features used are mainly the ones used to described the phone set (cf. §3.4).

Notice that the SPS procedure not only allows us to use distinctive phonetic features (Linear Spectrum Pair, LSP model, Zheng et al. 2000) as parameters, but also prosodic cues, like pitch (F0). This allows us to both keep the advantages of having an HTS Voice Font (high flexibility, small font size) and to limit its disadvantages (muffle voice quality, flat prosody). The training process uses a gradient descent algorithm (Minimum Generation Error, MGE, Wu et al. 2006).

In the end, the (trained) decision tree is used in generation to select and concatenate the state models by maximizing the likelihood of the parameter sequence. The result is a fully intelligible TTS voice font.

5. Conclusions and Future Work

In this paper, we presented the results of a collaborative citizenship project between Microsoft, a speaking-community association, ALM, and an R&D institution, ILTEC, which created the world´s first comprehensive set of a language resources for an under-resourced language, the Mirandese. Additionally the team has jointly created the very first intelligible TTS system for this language. The Casa de la Lhéngua interface is a free access portal that makes these resources available to the speaking and scientific community. The same resources will be available in the Microsoft Language Development Center portal. We believe that granting Mirandese a stronger sociolinguistic profile within its speaking community, will be aided by the availability of these language resources and tools.

Future work should include the conversion of the resources to internationally standardized formats such as those proposed in the context of TEI and LMF and a community-centred MOS evaluation of the intelligible TTS system.

6. References

Barros Ferreira, M. (2002). O mirandês, língua minoritária. In Mira Mateus, M. H. (org.), *Uma política de língua para o português*. Lisboa: Colibri, pp. 137-145.

Beckman, M. E., & Hirschberg, J. (1994). *The ToBI* *annotation conventions*. Columbus, OH: Ohio State University.

Braga, D., Campillo, F., Dias, M.S., García-Mateo, C., Méndez, F., Mourín, A., Silva, P. (2010). Building high quality databases for minority languages such as Galician. In Calzolari, N. et al. (eds.). *Proceedings of LREC'10*. La Valletta: ELRA, pp. 113-116.

Braga, D., Coelho, L., Resende, F. G., Dias, M. (2008). "Subjective and Objective Evaluation of Brazilian Portuguese TTS Voice Font Quality". In Kacic, Z. and Markus, A. (eds.), Advances in Speech Technology – *Proceedings of the 14th International Workshop. Maribor: Faculty of Electrical Engineering and Computer Science*: 129-138.

Branco, A., Mendes, A., Pereira, S., Henriques, P., Pellegrini, T., Meinedo, H., Trancoso, I., Quaresma, P. S. de Lima, V. L. (2012). *The Portuguese Language in the Digital Age*. Berlin: Springer.

Caseiro, A., D., Trancoso, I., Guerreiro, M., C., Ribeiro, V., Barros, M. (2003), "A Comparative Description of GtoP modules for Portuguese and Mirandese using Finite State Transducers", In ICPhS'2003 - 15th International Congress of Phonetic Sciences, Barcelona, Spain, August 2003

Chesi, C., Baldewijns, D., Hyongsil, C., Braga, D., Sales Dias, M. (2011). A TN/ITN Framework for Western European languages, *Abstract Proc. of CLIN21*.

Cho, H, Braga, D., Chesi, C., Baldewijns, D., Ribeiro, M., Saarinen, K., Beck, J., Rustullet, S., Henriksson, P, Dias, M., Rahmel, H. (2010). "A Multi-lingual TN/ITN Framework for Speech Technology". In García Mateo, C., Campillo Díaz, F., Méndez Pazó, F. (eds.), *Proceedings of FALA 2010*. Vigo, Universidad de Vigo: 213-216.

Ferreira, A., Ferreira, J. P. (2001). Dicionário Mirandês-Português. Lisbon: authors. Retr. 15/05/2013.

Ferreira, J. P., Cho, H., Braga, D., Silva, P., Dias, M. (2011a). A guided quest for a spoken standard. Presentation at *Ethics and Methodology in Discourse Interaction Research*, 2011, FCSH-UNL, Lisbon (http://download.microsoft.com/download/A/0/B/A0B1A66A-5EBF-4CF3-9453-4B13BB027F1F/apres_etica_e_metodologia.pdf).

Ferreira, J. P., Cho, H., Braga, D., Silva, P., Dias, M. (2011b). Acoustic notes on stressed Mirandese vowels. Poster presentation at *Phonetics and Phonology in Iberia* 2011, URV, Tarragona, June 21-22 2011 (http://download.microsoft.com/download/A/0/B/A0B1A66A-5EBF-4CF3-9453-4B13BB027F1F/PaPI2011_MLDC_poster.pdf).

García, M., Gamallo, P. (2010). Análise Morfossintáctica para o Português Europeu e Galego: Problemas, Soluções e Avaliação. *Linguamática*, 2(2), pp. 59-67.

Janssen, M. (2012). NeoTag: a POS Tagger for Grammatical Neologism Detection. In Calzolari, N. et al. (eds.). *Proceedings of LREC'12*. Istanbul: ELRA, pp. 2118-2124.

Leite Vasconcelos, J. (1899-1900) [1990]. *Lições de Filologia Mirandesa*. Miranda do Douro: CMMD.

Ratnaparkhi, A. (1996). A maximum entropy model for part-of-speech tagging. *Proc. EMNLP*. New Brunswick, New Jersey: Association for Computational Linguistics

Scannell, K. (2007). The Crúbadán Project: Corpus building for under-resourced languages. *Cahiers du Cental* 4 (2007), pp. 5-15,

Trancoso, I., Ribeiro, V., Barros, M., Caseiro, A., D., Paulo, S., (2003). "From Portuguese to Mirandese: fast porting of a letter-to-sound module using FSTs". In Mamede, N. J. et al. (eds.), Proc. of PROPOR'2003. LNCS. Berlin / Heidelberg: Springer: 49-56.

Wu, Y.-J., Wang, R-H. (2006). Minimum generation error training for HMM-based speech synthesis, in *Proceedings of ICASSP*, pp. 89-92.

Yi-Jian Wu, and Ren-Hua Wang. (2006). "Minimum generation error training for HMM-based speech synthesis". In *Proc. of ICASSP*: 89-92.

Zen, H., T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. Black and K. Tokuda (2007). The HMM-based speech synthesis system version 2.0. *Speech Synthesis Workshop*, Bonn, Germany, 294-299.

Zen, H., Tokuda, K., & Black, A. W. (2009). "Statistical parametric speech synthesis". *Speech Communication*, 51(11):1039-1064.

Zheng F., Z. Song, W. Yu, F. Zheng, W. Wu, (2000) The distance measure for line spectrum pairs applied to speech recognition, *Journal of Computer Processing of Oriental Languages*, March 2000, vol. 11, pp. 221-225