13

Language Documentation & Conservation Special Publication No. 9 (January 2016):
Language Documentation and Conservation in Europe
ed. by Vera Ferreira and Peter Bouda, pp. 150–158
http://nflrc.hawaii.edu/ldc/
http://hdl.handle.net/10125/24676

The first Mirandese text-to-speech system

José Pedro Ferreira[a], Cristiano Chesi[bc], Daan Baldewijns[c], Daniela Braga[d], Miguel Dias[ce], and Margarita Correia[af]

[a]CELGA-ILTEC (Universidade de Coimbra), [b]Instituto Universitario di Studi Superiori (Pavia), [c]Microsoft Language Development Center (Lisbon), [d]DefinedCrowd, [e]ISCTE - University Institute of Lisbon, [f]Universidade de Lisboa

This paper describes the creation of base NLP resources and tools for an under-resourced minority language spoken in Portugal, Mirandese, in the context of the generation of a text-to-speech system, a collaborative citizenship project between Microsoft, ILTEC, and ALM – Associaçon de la Lhéngua Mirandesa. Development efforts encompassed the compilation of a large textual corpus, definition of a complete phone-set, development of a tokenizer, inflector, TN and GTP modules, and creation of a large phonetic lexicon with syllable segmentation, stress mark-up, and POS. The TTS system will provide an open access web interface freely available to the community, along with the other resources. We took advantage of mature tools, resources, and processes already available for phylogenetically-close languages, allowing us to cut development time and resources to a great extent, a solution that can be viable for other lesser-spoken languages which enjoy a similar situation.

1. INTRODUCTION. While the current discussion on NLP resources for some of the most widely spoken European languages focuses on theoretical proposals to further develop them and on how to bridge the gap with the most advanced systems available for English (Branco et al. 2012), some lesser-spoken ones, Mirandese being the prototypical example, still lack the most basic language resources, with severe consequences for research, teaching, and applications available to citizens. Prior to the 1990s, when this language was officially recognized and normalization efforts began (Barros Ferreira 2002), little attention had been given to this language – to this day, the most complete studies are still those done at the turn of the 19th century (Leite Vasconcelos 1899–1900). Lack of an established writing standard, little written production, and low perceived economic value were responsible for the near-inexistence of modern studies and language resources. This, in turn, contributed throughout the years to a low perceived sociolinguistic value of the language among its speakers and lack of access to modern technology in their native tongue.

This paper outlines the tasks that were carried out in a citizenship project, a collaboration between Microsoft, ILTEC, and the not-for-profit, speaking-community-led ALM – Associaçon de la Lhéngua Mirandesa (Mirandese Language Association), that aimed at developing basic NLP tools for Mirandese from scratch, having a high quality speech synthesis system generation process (Braga et al. 2010) as the final goal. This paper de-

Licensed under Creative Commons
Attribution Non-Commercial No Derivatives 4.0 International License

ISBN: 978-0-9856211-5-5

scribes such process, illustrating how the language resources development tasks evolved in relatively little time and with relatively few human resources and investment, through the adaptation and reuse of existing tools and resources available for phylogenetically-close languages. All the language resources that were developed will be made freely available for R&D purposes, through an open-access web interface, *Casa de la Lhéngua*, a portal which is to be managed by ALM.

2. ON MIRANDESE. Mirandese is a minority language spoken in Northeastern Portugal. It belongs to the Astur-Leonese group of West Iberian languages, being closely related to Asturian, spoken to this day in areas of the Asturias and Leon autonomous communities in Spain, with which Mirandese no longer retains a linguistic continuum. Unwritten for most of its history, Mirandese was first scientifically identified and studied in the late 19th century (Leite Vasconcelos 1899–1900).

Throughout the 20th century, strong demographic changes, namely the exodus of large numbers of people through emigration in the 1940s and in the 1970s, an influx of non-Mirandese speaking workers from various other 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, leaving the language seriously threatened. Today, it is estimated that Mirandese is spoken by no more than 5,000 people as a first language, and by at most 15,000 in total, counting heritage and second language speakers, including those living outside of Terra de Miranda (Barros Ferreira 2002).

In the 1990s, strong efforts began to be made to make the survival of Mirandese possible: A group of linguists and native speakers managed to reach an agreement for a spelling convention common to different varieties, and the language was introduced into the formal education curricula locally, although with limited scope and only as an option. The Portuguese State finally granted the language co-official regional status in 1999 (Barros Ferreira 2002).

These initiatives had a strong impact on its speakers: What used to be perceived as a reason for shame by many in the diglossic community increasingly started showing up in book shelves, in the local and national media, and on the Web. Albeit seriously threatened as a mother tongue, it is currently learned and used 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 AS A MEANS AND A GOAL. Before the efforts this paper describes were made, although some initial efforts towards developing speech technologies had been set about (Trancoso et al. 2003, Caseiro et al. 2003), there were little or no available modern language resources for Mirandese. The most detailed linguistic descriptions are to this day those made by Leite Vasconcelos (1899–1900), more than 100 years ago, in part due to the lack of available data (Barros Ferreira 2002), leaving researchers in need of conducting original fieldwork to get in touch with actual large-scale data, and school pupils with little up-to-date base tools for the formal study of the language. Additionally, the fact that Mirandese is present in more and more support formats and usage contexts seems to be a decisive factor in the way its speakers perceive the language, granting it a higher sociolinguistic profile (Barros Ferreira 2002). These facts, along with the will to develop a Text-to-Speech (TTS) system for Mirandese, were the spark behind the effort to create the resources presented in this paper.

To achieve the end goal of creating a TTS system, a number of language resources that are usually available for more widely spoken European languages had to be developed from scratch. For the voice font generation, an existing and proven process could be followed, using previously existing tools designed for larger and better-resourced languages (Braga et al. 2008b). Work for this project encompassed the creation of a large text corpus, the definition of a complete phone-set, the development of tokenizer, inflector, text normalization (TN) and grapheme-to-phoneme (GTP) modules, and the creation of a large phonetic lexicon with part-of-speech (POS) classification.

The corpus was compiled 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), increasing the total size of the corpus to over one million tokens.

The lexicon was built using the currently 25K-strong lemma list of the ongoing work on a Mirandese-Portuguese dictionary (Ferreira & Ferreira 2001–), complemented with the most frequent lemmas in the compiled text corpus, which was previously tokenized, lemmatized, and POS-tagged using customizations of García & Gamallo (2010) and Janssen (2012). The resulting lemma list was then inflected using a version of Janssen (2011), 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 (Janssen & Santos 2012). This simple regular expression string replacement set of scripts starts by marking up stress and syllable divisions over 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.

A TN module for Mirandese was also put in place for the first time for this language, its rule-set being developed with the aid of previously in-house developed software (Cho et al. 2010) and taking advantage of the availability of a counterpart file for European Portuguese. The proximity between the two languages made it possible to reuse the pre-existing resource 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.

4. CREATION OF THE TEXT-TO-SPEECH SYSTEM. The Voice Font Building procedure that allowed us to create the TTS system discussed here can be described at three levels:

1. TTS front-end (preparation of language resources described in Section 3, text analysis);

2. Voice font building (creation of text prompts, voice talent selection and recording of uttered prompts, voice font compilation);

3. TTS back-end (voice font training using SPS – Statistical Parameter Synthesis, prosodic model creation and tuning).

4.1. LANGUAGE RESOURCES PREPARATION AND TEXT ANALYSIS. As detailed in Section 3, the preparation of language resources consisted of creating, collecting, annotating, and validating the linguistic information that is required for the voice font building procedure. This included the compilation of 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) are provided, as exemplified in (1):

(1)

Word	Pronunciation	POS	morphological features
Abacelhe	*ax - b ax - s eh 1 - lh aex*	*VER*	*subjunc., present, 3p, s*
melhor	*m aex - lh oh r 1*	*ADJ*	*qualifying, m, s*

Another important resource that needed to be developed, was a complete phone set for Mirandese, consisting of 46 distinct phones, which takes us forward from prior work (Trancoso et al. 2003). This resource specifies the full list of available phones paired to distinctive features and parameters that are used to train the voice model (e.g. voiced, velar, dental, liquid, main stress, secondary stress, etc.) as exemplified schematically in (2).

(2)

Phone	features
b	*voiced bilabial plosive*
t	*voiced dental plosive*

The text normalization module is composed of about 5,000 (contextual) rules dealing with the expansion of cardinal numbers ("12" > "twelve") or date-time spell out ("12:00" > "noon"), for instance. The following TN categories were developed for Mirandese: 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 made for instance to ensure the correct gender agreement in noun phrases containing a cardinal number.

From the resources mentioned above, we managed to easily generate syllabification rules (using algorithms like the one discussed in Janssen & Santos (2012)) that allow us to segment words in the lexicon and pair them with their correct pronunciation and the most likely stress.

4.2. PROMPTS CREATION, VOICE SELECTION AND RECORDING. On par with these development tasks, a voice talent was selected for high-quality recording sessions of 5,132 prompts retrieved from the corpus. The prompts consisted of full sentences selected based on character length and phonological relevance (richness of phonological contexts), determined by an existing algorithm of the text 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 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 by Braga et al. (2008a).

Finally, the now elected 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 with 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.

4.3. THE TTS BACK-END. After being properly segmented (both sentence segmentation and word segmentation is needed) and fully normalized, the prompts used for guiding the recording procedure were analyzed. This procedure was automatically carried out using rule-based sentence breakers, contextual text normalization (TN) rules (as discussed in Section 4.1), and POS taggers (e.g. Ratnaparkhi 1996). Once single words are normalized and categorized, the correct pronunciation is retrieved from the lexicon and assigned to the current word.

In the end, each prompt was enriched with several types of information, as shown in Figure 1:

```
<text>Este era pa la missa de las seis de la manhana .</text>
<words>
  <w v="Este" p="e 1 s - t e" type="normal" pos="d"/>
  <w v="era" p="e 1 - r a" type="normal" pos="v"/>
  <w v="pa" p="p e 1 &amp;amp; a 1" type="normal" pos="p"/>
  <w v="la" p="l a 1" type="normal" pos="d" />
  <w v="missa" p="m i 1 s - s a" type="normal" pos="N"/>
  <w v="de" p="d e 1" type="normal" pos="p"/>
  <w v="las" p="l a 1 s" type="normal" pos="d"/>
  <w v="seis" p="s e 1 j s" type="normal" pos="u" />
  <w v="de" p="d e 1" type="normal" pos="p" />
  <w v="la" p="l a 1" type="normal" pos="d" />
  <w v="manhana" p="m a - n a 1 - n a" type="normal" pos="N" />
  <w v="." type="punc" pos="O" regularText="." />
</words>
```

FIGURE 1

Such fully annotated prompts made the TTS voice font training procedure possible: The approach used to train the font model is called Statistical Parameter Synthesis (SPS) and it is based on standard Hidden-Markov-Model (HMM) 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 observation (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 – the prompt pairs we use 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 (Figure 2) (as well as its duration and excitation).

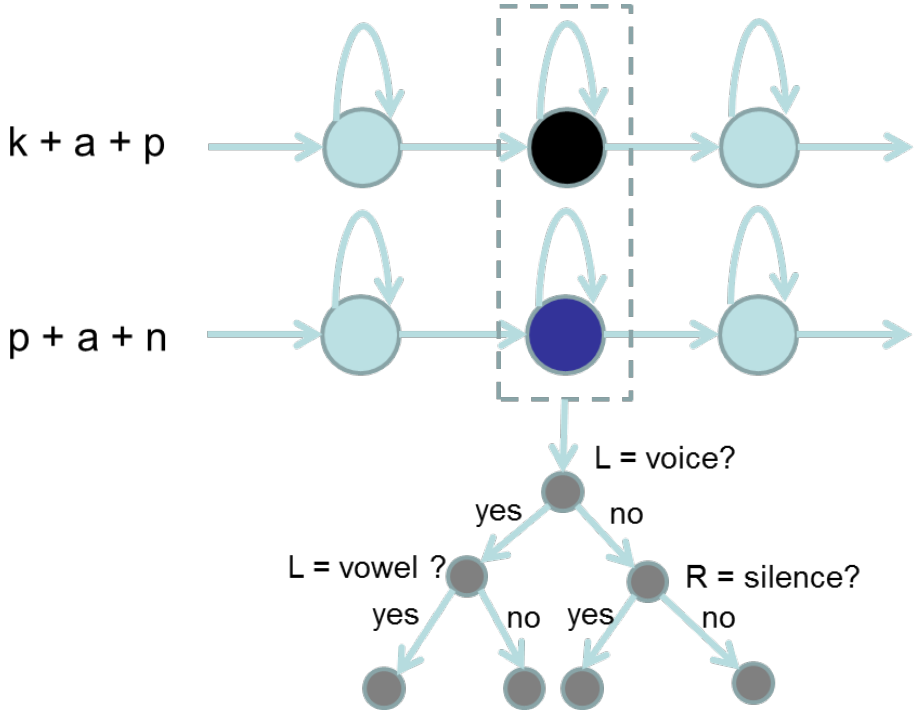

FIGURE 2: Gaussian Models training, organized by Decision Tree

Notice that the SPS procedure not only allows using distinctive phonetic features (linear spectrum pair, LSP model, cf. Zheng 2000) as parameters, but also prosodic cues, like pitch (F0). This allows us to both keep the advantages of having an HTS Voice Font[1] (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, cf. Yi-Jian & Wang 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 of this process is a fully intelligible TTS voice font.

[1] In generation, the model parameter is used to create the wave form signal.

5. CONCLUSIONS AND FUTURE WORK. In this paper, we presented the development of the first TTS system for Mirandese reaching intelligibility. The data and language resources developed to achieve this goal are being made available to the speaking and scientific community through the speech-community-led *Casa de la Lhéngua*, a free access web interface still under development. It rests to be seen if the secondary objective of granting Mirandese a stronger sociolinguistic profile within its speech community will be aided by the availability of these tools.

Future work should include the conversion of the NLP resources we developed to internationally standardized formats.

REFERENCES

Barros Ferreira, Manuela. 2002. O mirandês, língua minoritária. In Maria Helena Mira Mateus (org.), *Uma política de língua para o português*, 137–145. Lisboa: Colibri.

Beckman, Mary E. & Julia Hirschberg. 1994. *The ToBI annotation conventions*. Manuscript, Ohio State University.

Braga, Daniela, Francisco Campillo, Miguel Sales Dias, Carmen García-Mateo, Francisco Méndez, Ana Belén Mourín & Pedro Silva. 2010. Building high quality databases for minority languages such as Galician. In N. Calzolari, Choukri K., Maegaard B., Mariani J., Odijk J., Piperidis S., Rosner M & Tapias D. (eds.), *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 10)*, 113–116. La Valletta: ELRA.

Braga, Daniela, Luís Coelho, Resende, Fernando Gil & Miguel Dias. 2008a. Subjective and Objective Evaluation of Brazilian Portuguese TTS Voice Font Quality. In Kačič, Zdravko & Aleksandra Zögling Markuš (eds.), *Advances in Speech Technology – Proceedings of the 14th International Workshop*, 129–138. Maribor: Faculty of Electrical Engineering and Computer Science.

Braga, Daniela, Pedro Silva, Manuel Ribeiro, Mário Henriques & Miguel Sales Dias. 2008b. HMM-based Brazilian Portuguese TTS. In Daniela Braga et al. (eds), *Propor 2008 Special Session: Applications of Portuguese Speech and Language Technologies, September 10, 2008, Curia, Portugal*, 47–50. http://download.microsoft.com/download/A/0/B/A0B1A66A-5EBF-4CF3-9453-4B13BB027F1F/Braga_Propor08.pdf (10 December, 2015)

Branco, António, Amália Mendes, Sílvia Pereira, Paulo Henriques, Thomas Pellegrini, Hugo Meinedo, Isabel Trancoso, Paulo Quaresma, Vera Lúcia Strube de Lima & Fernanda Bacelar. 2012. *The Portuguese Language in the Digital Age*. Berlin: Springer.

Caseiro, Diamantino, Isabel Trancoso, Céu Viana & Manuela Barros. 2003. A Comparative Description of GtoP modules for Portuguese and Mirandese using Finite State Transducers. In M. J. Solé, D. Recasens & J. Romero (eds.), Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS), 2605-2608. Barcelona.

Cho, Hyongsil, Daniela Braga, Cristiano Chesi, Daan Baldewijns, Manuel Ribeiro, Kaisa Saarinen, Jeppe Beck, Silvia Rustullet, Peter Henriksson, Miguel Dias & Heiko Rahmel. 2010. A Multi-lingual TN/ITN Framework for Speech Technology. In Carmen García Mateo, Francisco Campillo Díaz & Francisco Méndez Pazó (eds.), *Proceedings of FALA 2010*, 213–216. Vigo: Universidad de Vigo.

Ferreira, Amadeu & Ferreira, José Pedro. 2001–. *Dicionário Mirandês-Português*. Lisboa: authors. http://www.mirandadodouro.com/dicionario/ (10 December, 2015).

García, Marcos & Pablo Gamallo. 2010. Análise Morfossintáctica para o Português Europeu e Galego: Problemas, Soluções e Avaliação. *Linguamática* 2(2). 59–67.

Janssen, Maarten. 2012. NeoTag: a POS Tagger for Grammatical Neologism Detection. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 12)*, 2118–2124. Istanbul: ELRA.

Janssen, Maarten. 2011. Computer-Aided Inflection for Lexicography Controlled Lexica. In Iztok Kosem & Karmen Kosem (eds.), *Electronic lexicography in the 21st century: New applications for new users*, 96–105. Ljublijana: Trojina.

Janssen, Maarten & Fabíola Santos. 2012. Building a database of phonetic transcriptions from a speech corpus. *VII GSCP International Conference: Speech And Corpora*, 29 February – 3 March 2012, Belo Horizonte, Brazil.

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

Müller, Karin, Bernd Möbius & Detlef Presher. 2000. Inducing probabilistic syllable classes using multivariate clustering. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, 225–232.

Ratnaparkhi, Adwait. 1996. A maximum entropy model for part-of-speech tagging. In *Proceedings of the Empirical Methods in Natural Language Processing 1*, 133–142. New Brunswick & New Jersey: ACL.

Scannell, Kevin. 2007. The Crúbadán Project: Corpus building for under-resourced languages. In Cédrik Fairon, Hubert Naets, Adam Kilgarriff, Gilles-Maurice de Schryver (eds.), *Building and Exploring Web Corpora, Proceedings of the 3rd Web as Corpus Workshop, Incorporating Cleaneval*, 5–15. Louvain-la-Neuve: Université Catholique de Louvain.

Trancoso, Isabel, Céu Viana, Manuel Barros, Diamantino Caseiro & Sérgio Paulo. 2003. From Portuguese to Mirandese: fast porting of a letter-to-sound module using FSTs. In Nuno Mamede, Isabel Trancoso, Jorge Baptista & Maria das Graças Volpe Nunes (eds.), *Computational Processing of the Portuguese Language. Proceedings of the 6th International Workshop PROPOR 2003*, 49–56. Berlin & Heidelberg: Springer.

Viswanathan, Mahesh & Madhubalan Viswanathan. 2005. Measuring speech quality for text-to-speech systems: development and assessment of a modified mean opinion score (MOS) scale. *Computer Speech & Language* 19(1). 55–83.

Wu, Yi-Jian & Ren-Hua Wang. 2006. Minimum generation error training for HMM-based speech synthesis. In *Acoustics, Speech and Signal Processing. Proceedings of ICASSP*, 89–92. Toulouse: IEEE.

Zen, Heiga, Keiichi Tokuda & Alan W. Black. 2009. Statistical parametric speech synthesis. *Speech Communication* 51(11). 1039–1064.

Zen, Heiga, Takashi Nose, Junichi Yamagishi, Shinji Sako, Takashi Masuko, Alan W. Black & Keiichi Tokuda. 2007. The HMM-based speech synthesis system version 2.0. In *Speech Synthesis Workshop*, 294–299. Bonn.

Zheng, Fang, Zhanjiang Song, Ling Li, Wenjian Yu, Fengzhou Zheng & Wenhu Wu. 2000. The distance measure for line spectrum pairs applied to speech recognition. *Journal of Computer Processing of Oriental Languages* 11. 221–225.

José Pedro Ferreira
jpf@uc.pt

Cristiano Chesi
t-chres@microsoft.com

Daan Baldewijns
v-daanb@microsoft.com

Daniela Braga
d-braga@hotmail.com

Miguel Dias
Miguel.Dias@microsoft.com

Margarita Correia
margarita@campus.ul.pt