IberSPEECH 2018
21-23 November 2018, Barcelona, Spain

TransDic, a public domain tool for the generation of phonetic dictionaries in standard and dialectal Spanish and Catalan

Juan-María Garrido[1], *Marta Codina*[2], *Kimber Fodge*[2]

[1]Phonetics Lab, National Distance Education University, Spain
[2]Department of Translation and Language Sciences, Pompeu Fabra University, Spain
jmgarrido@flog.uned.es

Abstract

This paper presents TransDic, a free distribution tool for the phonetic transcription of word lists in Spanish and Catalan which allows the generation of phonetic transcription variants, a feature that can be useful for some technological applications, such as speech recognition. It allows the transcription in both standard Spanish and Catalan, but also in several dialects of these two languages spoken in Spain. Its general structure, input, output and main functionalities are presented, and the procedure followed to define and implement the transcription rules in the tool is described. Finally, the results of an evaluation carried for both languages are presented, which show that TransDic performs correctly the transcription tasks that it was developed for.

Index Terms: Automatic phonetic transcription, Speech recognition, Dialects, Spanish, Catalan, Tools

1. Introduction

Phonetised dictionaries are language resources that are useful for several purposes in Speech Technologies, but probably its most frequent use is in automatic speech recognition. Phonetised dictionaries included in speech recognition systems are expected to contain not a unique transcription for each entry in the dictionary, but several transcription variants representing alternative pronunciations of the word in the language, as it has been shown that including these variants in the dictionary improves speech recognition accuracy (see for example [1], among many other works). These variants may represent either dialectal, sociolectal or even individual variations of the pronunciation of the word.

Two main approaches to the generation of these dictionaries have been attempted: data-driven and knowledge-based [2]. Data-driven techniques obtain the transcription variants, as well as their relative frequency, from the analysis of a phonetised corpus, which has to be large and representative enough to derive a reliable set of pronunciation variants. This analysis has been carried out frequently using machine learning techniques ([3,4,5,6,7], among others). Despite its evident advantages, the main withdraw of this approach is that it is very dependent on the contents of the used corpus, in which probably rare words in the language won't be represented. Knowledge-based approaches make use of explicit linguistic knowledge, for example, in the form of transcription rules implemented in an automatic phonetic transcriber. In this case, the automatic transcriber generates one or several transcription variants for each word, depending on the number of phonetic processes that can be applied to it. The main advantage of this approach is that it ensures a full coverage of the considered pronunciation variants, but it has also disadvantages, such as the fact that an excessive number of variants may be generated for some words, which may cause a decreasing of the performance of the speech recognizer ([8]). Also, the probability of each transcription variant should be also obtained by rule, a much more difficult task in the current state of linguistic knowledge.

Other possible technological applications of phonetised dictionaries, such as the phonetisation of classical language dictionaries, do not require transcription variants. In this case, only a single transcription, corresponding to the standard pronunciation, is needed.

Most existing phonetic transcribers for Spanish and Catalan have been developed for a specific system of application, usually commercial, and they are not public domain ([9], for example). Most non-commercial transcribers available for these two languages ([10,11]) have limitations to its use (for example, they do not allow the transcription of items that are not words, they only handle a phonetic transcription alphabet—IPA or SAMPA— or they only allow the transcription through a web interface) which makes difficult the task of phonetising word lists. Also, they usually do not allow the transcription in other dialects different to the standard one. Saga [12], for Spanish, and Segre [13] for Catalan, do allow the generation of both standard and dialectal phonetic transcriptions, but they do not allow the simultaneous generation of alternative transcriptions for a single item. TransText [14], a phonetic transcriber for Spanish and Catalan, has the same limitation.

This paper presents TransDic, a free distribution tool for the phonetic transcription of word lists in both standard Spanish and Catalan, but also in several dialects spoken in Spain of these two languages. Its main novel feature is that it allows the generation of dictionaries both with one single transcription per entry or with several phonetic transcription variants, which makes it suitable for the generation of dictionaries for speech recognition systems. Also, it has been designed to generate only 'relevant' dialectal variants, that is, those which are general enough in the corresponding geographical area. It has been developed from TexAFon [15,16], a complete rule-based linguistic processing system for text-to-speech that includes a phonetic transcriber, and which has been improved to include new features, such as the generation of phonetic variants for a single input item and the transcription in several non-standard dialects.

Next sections present the general structure of TransDic and its main functionalities, and explain the procedure followed to define and implement in the tool the transcription rules. The results of an evaluation carried for both languages are also presented.

10.21437/IberSPEECH.2018-61

2. General description

TransDic is a multiplatform tool that can be used either in Linux, Mac OS or Windows. It is a command-line tool, which requires to specify a set of arguments for its execution (for example, the language or dialect or the phonetic alphabet for the transcription). It can be used then both as an independent tool or integrated in other tools or procedures.

TransDic internal structure includes three levels, as illustrated in Figure 1:

- The tool itself, TransDic.
- The processing core, shared by other applications (such as TransText, for the phonetic transcription of texts, or texafon, a full text-processing tool for text-to-speech applications), which includes the letter-to-sound module, other text processing modules, some of them (the text processing module, for example) used also by TransDic.
- A set of language/dialect modules, including language or dialect dependent dictionaries and rules. Every dialect is considered then as an 'autonomous' language, with its own rules and dictionaries.

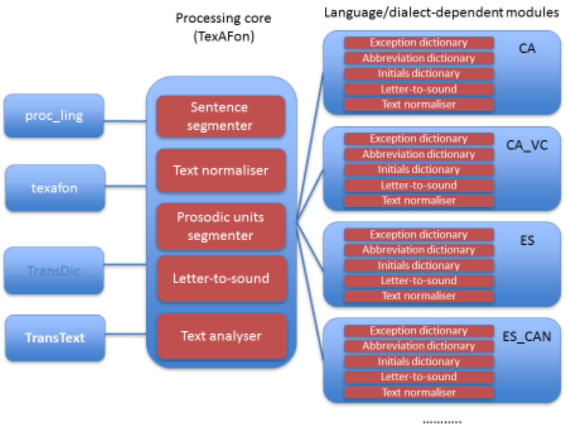

Figure 1: *TransDic structure.*

The transcription procedure in TransDic involves three main stages, which are similar to the ones in other existing transcription tools:

1) Text preprocessing
2) Lexical stress prediction
3) Phonetic transcription

Unlike other available phonetic transcription tools, TransDic is able also to transcribe non-standard tokens (such as symbols, dates, times or other figures), which are converted to readable Spanish items by its text processing module.

TransDic input must be a UTF-8 text file containing the list of tokens (one line per token) to be transcribed. Arguments allow to specify, for example, the dialect in which transcription has to be produced:

- Catalan: Standard (ca), Ribagorza (ca_ri), Pallars (ca_pa), Tortosa (ca_to), Central Western (ca_ac), Northern Valencian (ca_vs), Central Valencian (ca_vc), Southern Valencian (ca_vm) and Alicante Valencian (ca_al)
- Spanish: Peninsular Standard (es), Western Andalucía (es_aoc), Eastern Andalucía (es_aor), Extremadura North (es_exn), Extremadura South (es_exs), Canarias (es_can), Castilla-La Mancha (es_clm), Madrid (es_mad) and Murcia (es_mur)

Other arguments allow also to specify the phonetic alphabet to be used (IPA or SAMPA), if the transcription may include or not pronunciation variants and syllable marks or the format of the output dictionary.

It is important to note that the definition of the pronunciation variants to be considered in transcription is not specified directly using arguments, as in Saga, for example. In this case the selection of a given dialect determines the phonetic phenomena and pronunciation variants that will be taken into account.

The output of TransDic is another UTF-8 text file containing the phonetised dictionary, which can be generated in two different formats: default (Figure 2) or HTK (Figure 3).

```
boxejar    b o k . s e d . ʒ ˈa     b o k . s e j . ʒ ˈa     b o k . s e d . ʒ
ˈa ɾ       b o k . s e j . ʒ ˈa ɾ
boxers     b o k . s ˈe s           b o k . s ˈe ɾ s
branca     b ɾ ˈa n . k a
braçalet   b ɾ a . s a . l ˈɛ t
brioix     b ɾ i . ˈɔ j s
brisa      b ɾ ˈi . z a
```

Figure 2: *Sample of default output dictionary generated by TransDic.*

```
estas [estas] ˈe s t a s
sea [sea] s ˈe a
tenía [tenía] t e n "i a
nunca [nunca] n "u n k a
poder [poder] p o D ˈe r
aquí [aquí] a k "i
```

Figure 3: *Sample of output dictionary in HTK format generated by TransDic.*

3. Development of the dialect transcription modules

Earlier versions of TexAFon [15, 16] included only language modules for standard Catalan and Spanish. The development of TransDic has been possible after the development of a set of new language modules for the TexAFon processing core, covering several Spanish and Catalan dialects spoken in Spain. Each language/dialect module includes a pronunciation exception dictionary and the transcription rules, written in Python, for the corresponding dialect, and they have been developed from the corresponding standard version. The detailed description of the development procedure of these rules exceeds the scope of this paper, but it can be summarised in the following steps for each dialect:

1. Definition of the phonetic phenomena characterising its pronunciation
2. Implementation of the phonetic rules covering the selected phenomena
3. Evaluation of the transcription and fixing of errors

These three steps are described briefly in the following subsections. A detailed description of the development procedure for the different dialect modules in Spanish and Catalan can be found in [17] and [18], respectively.

3.1. Phonetic characterization of dialects

Existing literature on dialectological studies of Spanish spoken in Spain ([19,20,21,22], among many others) and Catalan ([23], a recent work also among many others) usually describes dialectal pronunciations in a very detailed way, paying attention to both general and local phenomena. For this work, however, only those phenomena which are general enough in the geographical area of the dialect were considered. And within these general phenomena, a distinction was made between 'primary' (the most frequent in their corresponding geographical area) and 'secondary' (not as frequent as primary ones, but general enough to be taken into account in a general description of the dialect in question), in order to keep the number of pronunciation variants within a reasonable number. Only geographical variants were considered, not social, stylistic nor individual.

The goal of this phase was then to define a set of 'primary' and 'secondary' pronunciations for each considered dialect, and to establish the relations between them (that is, which secondary pronunciations are alternative pronunciations to the primary ones). This task was carried out through an extensive literature review for both languages. It was not an easy task at all from a linguistic point of view, as the information provided in the literature is frequently incomplete, with limited information about the frequency or extension of the described phenomena.

The detailed description of all the defined dialect sets is out of the scope of this paper, so only one is presented here in some detail, the one for Canarias Spanish, based in the description provided mainly in [19,21,24,25]. Table 1 presents the subset of primary pronunciations which are different from standard Spanish in this dialect, the subset of secondary pronunciations and the relation between both subsets (that is, which secondary realisations are pronunciation variants of the primary ones). As it can be observed in this table, the relation between primary and secondary pronunciations can follow different patterns:

- One primary pronunciation has no secondary pronunciations associated, which means that no alternative realisations are considered. This is the case, for example, of the *seseo*.
- One primary pronunciation has one (or more) secondary pronunciation(s) associated. This means that all pronunciations, primary and secondary ones, are possible in the same context, although the primary one is considered as more frequent than the secondary one(s). For example, elision of syllable-final orthographical <s> (secondary pronunciation) is a possible alternative realisation to the pronunciation as [h], considered primary (that is, most frequent) in Canarias Spanish.
- One secondary pronunciation has no primary pronunciation associated. This means that it is an alternative to a standard pronunciation, which is also a primary pronunciation in that dialect. This is the case of *yeísmo*, the realisation of orthographical <ll> as [j], which according to the literature has been considered as less frequent than the realisation as [ʎ], default pronunciation in standard Spanish.

These lists of phonetic phenomena were used for the implementation phase, explained in the following subsection.

Table 1: *Primary and secondary phenomena considered for Canarias Spanish.*

Primary	**Secondary**
Pronunciation of orthographical <c,z> as [s] (*seseo*) *<caza> → [ˈkasa]*	
Pronunciation of orthographical <s> at the end of a syllable as [h] (*aspiración*) *<los> → [loh]*	Elision of syllable-final orthographical <s> *<los> → [lo]*
Pronunciation of orthographical <g, j> as [h] (*aspiración*) *<ajo> → [ʻaho]*	
Elision (no pronunciation) of orthographical <d> in words ending with <-ado> *<cansado> → [kanˈsao]*	Pronunciation of intervocalic <d> as [ð], as in standard Spanish *<cansado> → [kanˈsaðo]*
	Elision of intervocalic <d> different from those of <-ado> words *<comido> → [koˈmio]*
Elision of word-final <r,l> *<comer> → [koˈme]* *<Raquel> → [raˈke]*	Pronunciation of word-final <r,l> as [r,l], as in standard Spanish *<comer> → [koˈmeɾ]* *<Raquel> → [raˈkel]*
Elision of word-final <d> *<corred> → [koˈre]*	Pronunciation of word-final <d> as [d], as in standard Spanish *<corred> → [koˈred]*
	Pronunciation of orthographical <ll> as [j] (*yeísmo*) *<calle> → [ˈkaje]*

3.2. Implementation

The implementation of the new dialect modules was done in all cases taking as starting point the rules and dictionaries for the corresponding standard dialect and then making the necessary modifications to include the defined primary and secondary phenomena. Some changes in the language-independent core of TexAFon were done also to allow the generation of several pronunciation variants for a single input token.

To implement primary phenomena, new context-dependent Python rules were developed to replace the standard ones in those cases in which the primary pronunciation was different from the standard. Figure 4 presents an example of rule for a primary pronunciation in Andalusian Spanish. In some cases, the inclusion of these rules led to modify also the exception dictionary, to make some entries coherent with the new rule.

The implementation of the secondary phenomena was done in a second phase and involved the modification of the primary context-dependent rules to allow the generation of secondary transcriptions for the same context. Figure 5 illustrates the result of the modification of the same rule presented in figure 4 to include deletion of <s> as secondary variant.

Finally, if the user has specified with the corresponding argument that transcription variants should be generated, TransDic produces all transcription variants for the input word. These transcription variants are derived from the character-by-character pronunciation variants generated by the transcription rules: the language-independent letter-to-sound module takes the obtained pronunciations for each character and combines them to generate the word transcription variants. So, for example, for the word ‘casas’ two different transcription variants would be generated ([kˈasah], [kˈasa]) using the previously described rules, whereas in the case of ‘llover’ the output variants would be four ([ʎoβˈeɾ], [joβˈeɾ], [ʎoβˈe], [joβˈe]), as the input word includes two characters with transcription variants which are combined to create the different alternative word transcriptions.

```
if ch == "s" and nch == "NIL":
            salida.append(["hh",0,False])
return salida
```

Figure 4: *Python implementation of a phonetic transcription rule for the pronunciation of orthographical <s> as [h] (aspiración) at the end of a word in Andalucía Spanish (Fodge, 2014).*

```
if ch == "s" and nch == "NIL":
            salida.append(["hh",0,False])
            # Update deletion of word final syllable final s
            salida.append(["",0,False])
return salida
```

Figure 5: *Python implementation of a secondary pronunciation transcription rule for <s> deletion (in bold) in Andalucía Spanish (Fodge, 2014).*

3.3. Evaluation

The procedure to evaluate the transcription produced by the new dialect modules was similar for Spanish and Catalan: lists of isolated words, representative of the implemented phenomena (307 words for Spanish; 99 for Catalan), were processed using each dialect module to obtain the corresponding output transcription. These output transcriptions were then revised manually to detect possible errors. Both evaluations were carried out using a different tool of the TexAFon package, but the evaluated rules were the same used in TransDic [17,18].

In the case of Spanish, the transcription performance was perfect: no errors were detected. Some errors were found, however, in the case of Catalan dialects. An error measure was computed in this case, using the procedure described in [26]: the sum of all phone substitutions (S*ub*), deletions (*Del*) and insertions (*Ins*) divided by the total number of phones in the reference transcription (*N*). Table 2 presents the obtained error values.

The number of generated variants was also evaluated. Phonetised dictionaries were generated with TransDic for all dialects using two reference word lists in Spanish (1,000 most frequent words in the CREA corpus [27]) and Catalan (a cleaned version of the CesCa corpus [28], 1,648 words) and the mean number of variants per entry was computed. Tables 3 and 4 present the results.

The results of these two evaluations indicate that TransDic performs reasonably well both as for transcription quality and number of generated variants is concerned. Spanish modules provide a more accurate transcription than Catalan ones, but they tend to generate more variants per input entry than Catalan modules. Anyway, mean number of variants per entry is always below 2 in both languages.

Table 2: *Error measures obtained for the Catalan dialects.*

Dialect	Value
Ribagorza	8
Pallarés	11
Tortosa	9
Central area	8
North Valencia	9
Central Valencia	10
South Valencia	10
Alicante Valencian	9

Table 3: *Mean number of variants per entry obtained for the Spanish dialects.*

Dialect	Mean number of variants
Standard	1.021
Western Andalucía	1.76
Eastern Andalucía	1.354
Extremadura North	1,354
Extremadura South	1.76
Canarias	1.468
Castilla-La Mancha	1.207
Madrid	1.207
Murcia	1.713

Table 4: *Mean number of variants per entry obtained for the Catalan dialects.*

Dialect	Mean number of variants
Standard	1
Ribagorza	1.006
Pallarés	1
Tortosa	1
Central area	1
North Valencia	1.269
Central Valencia	1.003
South Valencia	1.003
Alicante Valencia	1.003

4. Conclusions

This paper has presented TransDic, a tool for the generation of phonetised dictionaries for Catalan and Spanish. Its most innovative features are that it allows to transcribe in several Spanish dialects spoken in Spain (Saga [12], for example, was developed considering mainly the American Spanish dialects) and that it allows the creation of phonetic dictionaries containing a reasonable number of pronunciation variants. The knowledge-based approach used in TransDic, based on a careful selection of the phonetic phenomena considered for transcription, does not lead to an overgeneration of variants, a classical problem in this kind of approach. Finally, another interesting feature of TransDic is that it is available for free download, from https://sites.google.com/site/juanmariagarrido/research/resources/tools/transdic. TransText [14], a phonetisation tool which allows the transcription of texts in the same dialects as TransDic, is also available for download from https://sites.google.com/site/juanmariagarrido/research/resources/tools/transtext.

Some expected improvements for the tool in the future are the inclusion of new types of rules for variants (for example, rules for informal pronunciations), and a deeper evaluation of the output by native speakers of each dialect.

5. References

[1] H. Al-haj, R. Hsiao, I. Lane, and A.W. Black, "Pronunciation modeling for dialectal arabic speech recognition". *IEEE Workshop on Automatic Speech Recognition & Understanding ASRU 2009*, pp. 525–528, 2009.

[2] J. Zheng, *Pronunciation Variation Modeling for Automatic Speech Recognition*, Ph. D, Thesis, University of Colorado, Boulder 2014.

[3] P. Taylor, "Hidden Markov Models for Grapheme to Phoneme Conversion", *Proceedings of the European Conference on Speech Communication and Technology, Lisboa, Portugal, September 2005*, pp. 1973-1976, 2005.

[4] M. Bisani and H. Ney, "Investigations on joint-multigram models for grapheme-to-phoneme conversion", *Proceedings of the 7th International Conference on Spoken Language Processing, ICSLP2002 - INTERSPEECH 2002, Denver, Colorado, USA, September 16-20*, pp. 105-108, 2002.

[5] S. Deligne, F. Yvon, and F. Bimbot, "Variable-length sequence matching for phonetic transcription using joint multigrams", *Proceedings of the Fourth European Conference on Speech Communication and Technology, EUROSPEECH 1995, Madrid, Spain, September 18-21*, pp. 2243-2246, 1995.

[6] A. Laurent, P. Deleglise, and S. Meignier, "Grapheme to phoneme conversion using an smt system", *Proceedings of the 10th Annual Conference of the International Speech Communication Association, Brighton, United Kingdom, September 6-10*, pp. 716-719, 2009.

[7] M. Gerosa and M. Federico, "Coping with out-of-vocabulary words: open versus huge vocabulary ASR", *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2009*, pp. 4313-4316, 2009.

[8] T. Holter and T. Svendsen, "Maximum likelihood modeling of pronunciation variation", *Speech Communication*, vol. 29, no. 2-4, pp. 177–191, 1999.

[9] P. Bonaventura, F. Giuliani, J. M. Garrido, and I. Ortín, "Grapheme-to-Phoneme Transcription Rules for Spanish, with Application to Automatic Speech Recognition and Synthesis", *Proceedings of the Workshop 'Partially Automated Techniques Transcribing Naturally Occurring Continuous Speech', Université de Montréal, Montreal, Quebec, Canada*, pp. 33-39, 1998.

[10] X. López, *Transcriptor fonético automático del español*, 2004. http://www.aucel.com/pln/transbase.html

[11] Molino de Ideas, *Transcriptor fonético*, 2012 http://www.fonemolabs.com/transcriptor.html.

[12] TALP-UPC, *SAGA - Phonetic transcription software for all Spanish variants*, 2017, https://github.com/TALP-UPC/saga.

[13] P. Pachès, C. de la Mota, M. Riera, M. P. Perea, A. Febrer, M. Estruch, J. M. Garrido, M. J. Machuca, A. Ríos, J. Llisterri, I. Esquerra, J. Hernando, J. Padrell, and C. Nadeu, "Segre: An automatic tool for grapheme-to-allophone transcription in Catalan", *Proceedings of the Workshop on Developing Language Resources for Minority Languages: Reusability and Strategic Priorities, LREC*, pp. 52-61, 2000.

[14] J. M. Garrido, M. Codina, and K. Fodge. **"**TransText, un transcriptor fonético automático de libre distribución para español y catalán", *Actas del Workshop "Subsidia: herramientas y recursos para las ciencias del habla"*, in press.

[15] J. M. Garrido, Y. Laplaza, M. Marquina, C. Schoenfelder, and S. Rustullet. "TexAFon: a multilingual text processing tool for text-to-speech applications", *Proceedings of IberSpeech 2012, Madrid, Spain*, pp. 281-289, 2012.

[16] J. M. Garrido, Y. Laplaza, B. Kolz, and M. Cornudella, "TexAFon 2.0: A text processing tool for the generation of expressive speech in TTS applications", *Proceedings of LREC 2014, Ninth International Conference on Language Resources and Evaluation, Reykjavik (Iceland)*, pp. 3494-3500, 2014.

[17] K. Fodge, *Introducing Spanish dialects in a linguistic processing module for improved ASR and novel speech synthesis capabilities*, Master Thesis, Barcelona: Pompeu Fabra University, 2014.

[18] M. Codina, *Automatic Phonetic Transcription of dialectal variance in Catalan*. Master Thesis, Barcelona: Pompeu Fabra University, 2016.

[19] P. García Mouton, *Lenguas y dialectos de España*, Madrid: Arco Libros, 1994.

[20] M. Alvar (Director), *Manual de dialectología hispánica. El español de España*. Barcelona: Ariel, 1999.

[21] F. Moreno Fernández, *La lengua española en su geografía*. Madrid: Arco Libros, 2009.

[22] J. A. Samper Padilla, "Sociophonological Variation and Change in Spain", In M. Díaz Campos (Ed.), The Handbook of Hispanic Sociolinguistics. West Sussex: Wiley-Blackwell, pp. 98–117, 2011.

[23] J. Veny and M. Massanell, *Dialectologia catalana. Aproximació pràctica als parlars catalans*. Barcelona: Universitat de Barcelona, 2015.

[24] M. M. Azevedo, *Introducción a la lingüística española*, Upper Saddle River: Prentice Hall, 2009.

[25] J. I. Hualde, A. Olarrea, and A. M. Escobar, *Introducción a la lingüística hispánica*. Cambridge: Cambridge University Press, 2001.

[26] C. Van Bael, L. Boves, H. Van den Heuvel, and H. Strik, "Automatic Phonetic Transcription of Large Speech Corpora". *Computer Speech & Language*, vol. 21, no. 4, pp. 652-668, 2007.

[27] Real Academia Española, *Banco de datos (CREA). Corpus de referencia del español actual.* http://corpus.rae.es/lfrecuencias.html.

[28] A. Llauradó, M. A. Martí, and L. Tolchinsky, "Corpus CesCa: Compiling a corpus of written Catalan produced by school children". *International Journal of Corpus Linguistics, vol.* 17, no. 3, pp. 428–441, 2012.