FALA 2010
VI Jornadas en Tecnología del Habla and II Iberian SLTech Workshop

Defining analogy for non-native inclusions in Spanish TTS

Tatyana Polyákova, Antonio Bonafonte

Universitat Politècnica de Catalunya, Barcelona, Spain
tatyana.polyakova@upc.edu, antonio.bonafonte@upc.edu

Abstract

Mass media globalization introduces the challenge of multilingualism into most popular speech applications such as text-to-speech synthesis and automatic speech recognition. In Spain as well as in the other countries, the usage of English words is rapidly growing, however due to the linguistic diversity of the languages spoken across the country, Spanish is not less influenced by inclusions from the four official languages. This work is focused on the pronunciation of Catalan inclusions in Spanish utterances. Our goal was to approach the nativization phenomenon by data-driven methods, making it easily transferable to other languages without loss in performance. For this particular task, training and test nativization corpora were manually crafted and the task itself was approached using pronunciation by analogy. The results were encouraging and showed that even small corpus of 1000 words allows to capture the analogy in the nativization process. The resulting pronunciations allowed significant improvements in the intelligibility of Catalan inclusions in Spanish utterances.

Index Terms: nativization of Catalan words, grapheme-to-phoneme conversion, phoneme-to-phoneme conversion, Spanish TTS, pronunciation by analogy

1. Introduction

Speech technologies in the framework of their rapidly expanding usage must be adapted to the multilingual scope allowing a higher level of flexibility and answering the modern users' needs. The text-to-speech synthesis finds many important applications on the emerging market of speech technologies. Voices capable of embracing more than one language are highly demanding in the era of mass media globalization. The TTS systems are used in telephone companies, smart phones, car navigation systems and recently in speech-to-speech translation, a technology that is highly demanded due to the globalization of the world industry and mass media.

Every language receives a constant incoming flow of new words. In addition to the natural process of appearance of neologisms, by morphological or semantic word and word meanings creations, a lot of new words come to the current language from other languages. The are several ways that the words of foreign origin are incorporated into a receptor language.

Very few databases containing non-native pronunciation are available, while the nativization corpora is simply inexisting. This need for training data lead us to a creation a minimalistic nativization corpus described in Section 2.

In order to have a synthesizer always up-to-date we need an ultimate automatic method for the derivation of the nativized pronunciation. The problem of foreign words, more particularly, of proper names of foreign origin was studied in [1]. The goal in [1] was to transcribe proper names of different origins correctly from the point of view of English phonetics. The nativization problem and different influencing factors were also described in [2] and [3]. Summarizing all possible influence factors and the difficulties encountered for the correct nativization of non-native words we are betting on an approach that can combine the knowledge of the orthographic and phonetic forms in the language of origin with pronunciation adaptation rules to the target language. In [4] it was demonstrated that the analogy between the nativized pronunciation and the original one can be inferred in a reliable and simple way since the nativization of English words in a Spanish text given the English pronunciation is an easier task than the native pronunciation of unknown English words and yet all human attempts to nativization are highly dependable on the analogy between known and unknown words. In [4] the final goal of the nativization was to be able to correctly pronounce English phrases in Spanish utterances as well as those out-of-dictionary proper names, commercial trademarks, etc., such as *Bruce* or *PlayStation* that were incorporated into Spanish in an already nativized form. This work is extended to Catalan inclusions in Spanish. The resulting nativized pronunciations should be well accepted by general Spanish audiences.

This paper is organized as follows, Section 2 explains the creation of the training and test datasets, Section 3 explains the algorithm used for nativization of Catalan words. Section 4 followed up by conclusions summarizes the experimental results obtained.

2. Nativization database

The main idea of this work was to train a nativization model to convert Catalan pronunciations to acceptable Spanish ones, adapting in a suitable way the pronunciation of the phonemes that do not exist in Spanish depending on such factors as frequency of usage of the word, and Spanish phonemization rules. Two nativization scopes were exploited: 1)training of a nativization model using the information about the orthographic form and the nativized phonetic transcription and 2) usage of the original Catalan and nativized to Spanish pronunciations for training. In order to apply data-driven techniques to nativization a need for training and test data raises. For usual grapheme-to-phoneme conversion tasks large pronunciation corpora of 100 thousands words and their corresponding pronunciations are available. Since we did not find any existing nativization database we chose to manually create a minimalistic corpus that would not require expert linguistic knowledge. For our task the training corpus was orthographically balancedin order to have all possible letter bi-grams in the corpus, we selected s total of 1000 words. The original phonetic transcriptions of these words were manually nativized according to the criteria described by Llorente in the book of styles for one of the Spanish TV chan-

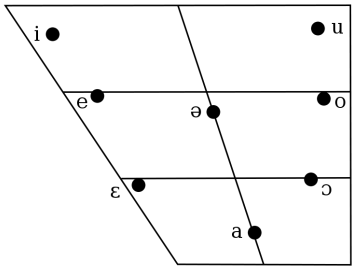

Figure 1: Vowels of standard Eastern Catalan

nels [5]. It is necessary to emphasize that the phoneme inventory used for nativization was limited to the Spanish phoneset. The test data was manually collected from the available on-line sources. Since a thousand words was selected for training, it was found appropriate that the test data comprised 10% of the training corpus. None of the test words were present in the training dictionary. It was intended that the test words were frequently used and with simple meaning in order for the results to be unbiased by other factors. Here are some examples of train *agredolça, boirumós, migjorn* and test words *enllaç, desig, forjar*.

2.1. Phonetic differences

The sounds of a language are defined by a phoneme inventory or phoneset. A phenomenon called extension of the phoneset often occurs in bilingual communities and speakers; however, it is impossible to study foreign word pronunciation on the level of the individual. In bilingual societies, it is much easier to observe general tendencies. In the particular case of Catalan, both of nativization and phoneset extension phenomena occur. It is curious to note that Spanish words in Catalan are pronounced using the regular Spanish phoneset, due to the fact that the majority of Catalan speakers are perfectly fluent in Spanish. For example, the Spanish name *Jorge* in Catalan is pronounced `/x 'o r x e/` and not `/dZ 'o r dZ @/` as Catalan phonetics would stipulate even though the phoneme `/x/` is absent from Catalan it is used for Spanish names. On the contrary, the pronunciation of Catalan words in Spanish is adapted according to Spanish pronunciation rules and the phoneset extension phenomenon is rare. Spanish and Catalan have several major phonetic differences which depend on the dialect of the latter. Most varieties of Catalan possess seven stressed vowels that are: `/a/`, `/e/`, `/o/`, `/u/`, `/i/`, `/E/`, `/O/`,`/@/`. The open vowels `/a/`, `/E/` and `/O/` as well as the unstressed `/@/` do not occur in Spanish. In Spanish medium vowels can be realized as open only in particular contexts while in the rest of the cases all vowels are articulated as closed. In Catalan, however there is an important phonological difference between open and closed vowels which can not be attributed to the context and therefore is not predictable. For example, homographs *seu (yours) vs. seu (headquarters)* `/s 'e u/` vs. `/s 'E u/` have different meanings depending on the vowel articulation point. A diagram of Catalan vowels can be found in Figure 1 As well as in Spanish, in Catalan there are six plosives `/b/`, `/d/`, `/g/`, `/p/`, `/t/`, `/k/` (3 voiced and 3 unvoiced) at three different articulation points. Catalan does not have any dental, uvular or velar fricative consonants sounds, but has two alveolo-palatals `/Z/` (voiced) e.g. *vigent* `/b i Z 'e n/` and `/S/` (unvoiced) e.g. *caixa* `/k 'a S a/`. The labiodental `/v/` exists in Catalan as a result of sonorization of any `/f/` before a voiced consonant or a vowel at the beginning of the word. Besides, in Catalan, all unvoiced fricatives are sonorized if followed by a voiced consonant. The fricative `/z/` which is very frequent in Catalan, exists in Spanish as an allophone but not as a phoneme. Catalan has four affricates which is 3 more than Spanish, the voiced affricates `/dz/`, `/dZ/` and the unvoiced `/ts/`, `/tS/`. The phonemes `/ts/` and `/dz/` arise mainly from compounding such as in *potser* /p u ts 'e/, but may as well occur at any other position as in *dotze* `/d 'o dz @/`. Similarly to Spanish the nasals are adapted to the articulation point of the following consonant, however in Catalan both `/m/` and `/J/` can occur at the end of the word e.g. *any* `/'a J/`. There are two laterals in Catalan as well as in Spanish, the alveolar `/l/` and the alveolo-palatal `/L/`. Additionally Catalan has the double *l.l* that is pronounced as a prolonged alveolar articulation `/l:/`. Both languages possess 2 trills, the simple `/r/` and the multiple `/rr/`. In contrast with Spanish the `/r/` in Catalan can only appear at the intervocalic position or after a plosive or fricative that forms part of the same syllable e.g. *frau* `/ f r 'a uw/` or *cara* `/k 'a r a/` [6]. A complete set of Catalan consonants can be found in Figure 2. Taking into account most of the above

	Bila-bial		Labio-dental		Inter-dental		Dent.-alv.		Alveolar		Alveol.-palatal		Palatal		Velar	
Ocl.	p	b					t	d							k̟	g
Fric.			f	v			s̪	z̪	s	z	ʃ	ʒ				
Afric.									t͡s	d͡z	t͡ʃ	d͡ʒ				
Aprox.		β̞				ð̞								j̞		ɣ̞
Nasal		m		ɱ				n̪		n		nʲ ɲ				ŋ
Later.								l̪		l		lʲ ʎ				ɫ
Vibr. simp.										ɾ ɻ						
Vibr. múlt.										r						

Figure 2: Catalan consonants [6].

mentioned phonetic differences, we developed nativization criteria in order to find the best pronunciation for Catalan words in Spanish utterances.

2.2. Criteria

The challenge of this task consisted of developing solid criteria for nativization, taking into account local specifications of certain words, pronunciation and word popularity factor, among others. Some of the criteria could not be easily formulated that is why using a training corpus clearly has an advantage over the rule-based approach. Several examples of the criteria used are described below. All open vowels were mapped to the closed ones, while the unstressed `/@/` was mapped to `/a/` in most of the cases except for those words that are were similar to Spanish where it was transcribed as `/e/` e.g. *adrearà* from `/a D r a s a r 'a /` to `/a D r e s a r 'a/` in the nativized form. For consonants, some difficulties were found when transcribing `/Z/` `/S/`. Their nativization depended both on letter and phoneme context. The voiced fricative `/Z/` at the beggining of the word was nativized to `/jj/` e.g *jutge*, to `/tS/` before a nasal e.g. *taronja*, and to `/j/` in other cases as in *vorejar*. The unvoiced fricative `/S/` was transcribed or to `/j s/` when it corresponded to the digraph *ix* e.g. *coix*; to `/tS/` when it corresponded to the same phoneme in a similar Spanish word e.g *anxoves*(*cat.*) vs. *anchoas*(*sp.*); or to `/s/` in the rest of the cases. The affricate `/dZ/` was nativized to `/tS/` as in the word *migdia*. Affricates `/ts/` and `/dz/` were mapped to the corresponding double phonemes `/t s/`

and /d z/. The multiple trill /rr/ was conserved only in the cases when it corresponded to the Spanish phonetic rules, in all other cases it was changed to the simple trill /r/. The nasal /N/ and the voiced /z/ were conserved as they were present in our voice database. Silent r at the end of the verb in a compound verb-pronoun construction such as $afegir - n'hi$, was restored in the nativized form for the sake of comprehension. The database nativization task was carried out by the authors using both the source language orthographic form and pronunciation. In the following section we describe the functionality of a multilingual grapheme-to-phoneme system used in this work.

3. Pronunciation by analogy

Data-driven approaches were proven to be more efficient than the ones based on the explicit linguistic modeling and they undoubtedly gain in adaptability [7]. For g2p conversion the best results were obtained using data-driven corpus-based methods. Pronunciation by analogy method previously used in [8, 9] was found to be the most efficient for grapheme-to-phoneme task. In this section we review the pronunciation by analogy algorithm. Our implementation is based on [8] with the new strategies introduced in [9]

3.1. Algorithm description

After the training dictionary has been aligned, the matcher starts to search for common substrings between the input word and the rest of the dictionary entries. Every input word is then compared to all the words in the lexicon in order to find common "arcs". Let us call the substrings in the grapheme context "letter arcs" and the corresponding substrings in the phoneme context "phoneme arcs". All the possible letter arcs with the minimum length of 2 letters and the maximum length equal to the input word length are generated and then searched in the dictionary. For every letter arc from the input word, matching with the same letter arc in a dictionary word, the corresponding pronunciation or the phoneme arc is extracted. The frequency of appearance of each phoneme arc corresponding to the same letter arc is stored along with the starting position and length for each arc.

Each time that for the same letter arc we find the same phoneme arc; the frequency of the phoneme arc is incremented. The matching phoneme arcs are introduced into the pronunciation lattice that can be represented by nodes and connecting arcs. If an arc starts at a position i and ends at a position j, and if there is yet no arc starting or ending at position j, the nodes L_i and L_j are added to the graph. An arc is drawn between them. All the nodes are labeled with the corresponding "juncture" phoneme and its position in the word. The arcs are labeled with the remaining phonemes and their frequency of appearance.

Each complete path through the lattice is called "pronunciation candidate". We considered only the shortest paths through the lattice [8]. If there is a unique shortest path, it is chosen as the best pronunciation and the algorithm stops. Usually there are several shortest paths through the lattice, and a decision function is necessary to choose the best pronunciation candidate among them.

Each candidate can be represented as C_j=$\{F_j, D_j, P_j\}$, where F_j = $\{f_1,\ldots,f_n\}$ are the phoneme arc frequencies along the j^{th} path, D_j = $\{d_1,\ldots,d_n\}$ are the arc lengths and P_j = $\{p_1,\ldots,p_k\}$ are the phonemes comprising the pronunciation candidate, being k the pronunciation length. Marchand and Damper in 2000 [8] proposed to use 5 scoring strategies in order to choose the best pronunciation.

In our previous work [9] we proposed 6 additional strategies for choosing the best candidate which in combination with the others outperformed the original ones. The scoring strategies are based on the following parameters, frequency of appearance of a given phoneme arc in the dictionary, its length and the actual phonemes which constitute the candidate. Different strategies work with different aspects of analogy. High arc frequency is considered to be a major advantage over the low arc frequency. The frequency of suffixes and prefixes are prioritized by different strategies. The more common phonemes the candidate shares with the others the higher will be its final score. If a candidate has exactly the same pronunciation as the other one both of them are prioritized. These measures are used separately or combined across the strategies. All the strategies previously used in grapheme-to-phoneme conversion are described below [8, 9]. The pronunciation by analogy algorithm was previously applied to grapheme-to-phoneme conversion [8, 9]. In this work it was extended to the nativization task.

4. Experimental results

The experimental results are given below for each method.

4.1. Previous results: nativization tables

In our previous work [10] we developed a nativization system based on nativization tables (Ntab). Pronunciations were derived according to the scheme shown in Figure **??**. The nativization was carried out in a phoneme-to-phoneme manner, using nativization tables for source→target phoneme transformations. The source language was Catalan and the target language was Spanish. Therefore all Catalan phonemes were mapped to the closest Spanish ones. The nativization tables were able to convert 79.74% phonemes and 21.78% words correct. These results are given for the same 100 word test corpus described in 2. However, these results are much better than those obtained without using nativization, applying the Spanish g2p to derive the pronunciation of Catalan words, Spanish g2p scored only 33.97% correct in phoneme and 3.96% on word nativization on a 100 word test corpus. The only words that this kind of system can "nativize" correctly are those that are pronounced very closely to Spanish orthography, for example *aqu´ to* /a k 'i/ or *sac to* /s 'a k/.

4.2. Grapheme-to-phoneme nativization (g2p_nat)

The first hypothesis to be tested was prediction of nativized pronunciation by analogy in the orthographic context. Out of eleven strategies available in the PbA for choosing the best pronunciation candidate it was necessary to determine the best strategy combination for our data. All possible strategy combinations were considered and compared. For grapheme-to-phoneme nativization (g2p_nat) the resulting best strategy combination was the following: 0000101101. For each of the eleven strategies described in [9] 1 means that the strategy corresponding to that position was included and 0 means it was left out. The best results obtained on training data equaled to 86.51% in phoneme and 38.61% in word accuracy. When we considered each strategy individually the best results were obtained for the eleventh strategy that combines the frequency product with the frequency of the same pronunciation. The lowest scoring strategy is seventh strategy that prioritizes the candidates with very frequent first arc. The results for each single strategy and the best strategy combination can be found in Table 1

Table 1: Single strategy results for g2p_nat and best strategy combination

strategy mask	ph. acc.	word. acc.
10000000000	85.80	35.64
01000000000	83.96	31.68
00100000000	83.82	31.68
00010000000	83.79	31.68
00001000000	84.62	32.67
00000100000	85.65	33.66
00000010000	83.38	30.69
00000001000	85.53	33.66
00000000100	83.96	35.64
00000000010	83.69	30.69
00000000001	85.86	36.63
00001011011	86.51	38.61

4.3. Phoneme-to-phoneme nativization (ph2ph_nat)

For Catalan, it makes a lot of sense to perform grapheme-to-phoneme nativization, in fact, non-Catalan speakers apply Spanish grapheme-to-phoneme rules when reading Catalan; however, in order to find the best pronunciation for Catalan phonemes absent from Spanish the phonetic transcription available in the source language may be quite helpful. Finding automatic correspondences between source and target (nativized) phonemes is a more consistent task than in the case of letters, being g2p conversion already a difficult task for Catalan especially for such a reduced training corpus. For phoneme-to-phoneme nativization experiments the PbA was modified in order to receive phoneme input. The best strategy combination (11010101010) as in the g2p_nat case was determined performing n-fold evaluation of all possible strategy combinations. The results obtained on 100 word test set of common names are 92.09% phonemes and 56.44% words correct. These results show that p2p_nat nativization outperforms g2p_nat nativization by 22% in word accuracy terms. Performing single strategy experiments for phoneme-to-phoneme nativization we can also observe that the best scoring strategies are the sixth and the eight one, while the worst places belongs to the ninth. For more results see Table 2.

Table 2: Single strategy results for p2p_nat and best strategy combination

strategy mask	ph. acc.	word. acc.
11000000000	91.51	53.47
10000000000	91.51	52.48
01000000000	90.09	48.51
00100000000	90.51	48.51
00010000000	90.95	50.50
00001000000	90.79	49.50
00000100000	91.65	54.46
00000010000	89.94	46.53
00000001000	91.51	54.46
00000000100	89.18	42.57
00000000010	90.38	48.51
00000000001	90.92	50.50
11010101010	92.09	56.44

5. Conclusions

In this paper we proposed to use pronunciation by analogy for the nativization of Catalan words in Spanish utterances in the framework of a multilingual TTS system. The best results were achieved using phoneme-to-phoneme nativization based on the analogy in the phoneme context. The nativization results obtained using analogy only in the letter context were rather poor, due to the reduced corpus size. It is worth mentioning that even in the case of grapheme-to-phoneme nativization the results show very significant improvements in comparison to those obtained by direct phoneme-to-phoneme table-based mapping. Nativized pronunciations are more tolerant to the vowel and consonant substitution and the persisting errors are minor and are not crucial for intelligibility. Even though the test corpus that consisted of 100 hundred frequently used common Catalan names can be considered somewhat tiny, for both g2p_nat and p2p_nat methods n-fold evaluation was performed on the training corpus of 1000 rather infrequent common names (selected by the greedy corpus balancing tool) and the results obtained were quite similar to those obtained on the test data. Simple mapping rules were proven to be insufficient for the task because some of the criteria could not be easily formulated.

6. Acknowledgements

This work has been partially funded by the Spanish Government under grant TEC2009-14094-C04-01 (BUCEADOR project)

7. References

[1] A. Font Llitjos and A. Black, "Knowledge of Language Origin Improves Pronunciation Accuracy of Proper Names," in *Proc. the of European Conference on Speech Communication and Technology*, Genova, Italy, Sep. 2001.

[2] I. Trancoso, "Issues in the pronunciation of proper names: the experience of the Onomastica project," in *In Proceedings of Workshop on Integration of Language and Speech*, Moscow, Russia, 1995.

[3] I. Trancoso, C. Viana, I. Mascarenhas, and C. Teixeira, "On deriving rules for nativised pronunciation in navigation queries," in *Proc. the of European Conference on Speech Communication and Technology*, Budapest, Hungary, 1999.

[4] T. Polyákova and A. Bonafonte, "Nativization of english words in spanish using analogy," in *The Seventh ISCA Tutorial and Research Workshop on Speech Synthesis, Kyoto, Japan, September 22-24*, 2010.

[5] J. Llorente and L. Díaz Salgado, *Libro de estilo de Canal Sur TV y Canal 2 Andalucía*. Radiotelevisión de Andalucía, 2004.

[6] A. Planas, *Así se habla: nociones fundamentales de fonética general y española: apuntes de catalán, gallego y euskara*. Horsori Editorial, Sl, 2005.

[7] A. van den Bosch and W. Daelemans, "Data-oriented methods for grapheme-to-phoneme conversion," in *Proceedings of the sixth conference on European chapter of the Association for Computational Linguistics*, Morristown, NJ, USA, 1993, pp. 45–53.

[8] Y. Marchand and R. Damper, "A multistrategy approach to improving pronunciation by analogy," *Computational Linguistics*, vol. 26, no. 2, pp. 195–219, 2000.

[9] T. Polyakova and A. Bonafonte, "New strategies for pronunciation by analogy," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Taipei, Taiwan, Apr. 2009.

[10] T. Polyákova and A. Bonafonte, "Further improvements to pronunciation by analogy," in *Actas de las V Jornadas en Tecnologas del Habla*, Bilbao, Spain, Nov. 2008, pp. 149–152.