Introducing Nativization to Spanish TTS Systems

Tatyana Polyákova, Antonio Bonafonte

Introducing Nativization to Spanish TTS Systems

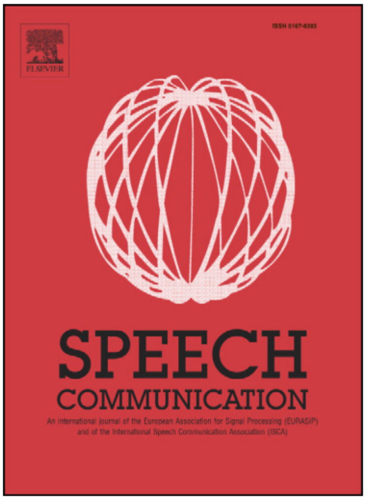
Tatyana Polyákova, Antonio Bonafonte

PII: S0167-6393(11)00074-4
DOI: 10.1016/j.specom.2011.05.009
Reference: SPECOM 1998

To appear in: *Speech Communication*

Received Date: 21 June 2010
Revised Date: 20 May 2011
Accepted Date: 24 May 2011

Please cite this article as: Polyákova, T., Bonafonte, A., Introducing Nativization to Spanish TTS Systems, *Speech Communication* (2011), doi: 10.1016/j.specom.2011.05.009



Introducing Nativization to Spanish TTS Systems

Tatyana Polyákova, Antonio Bonafonte

Universitat Politècnica de Catalunya,
Signal Theory Department and Communications
Jordi Girona 1-3,
08034, Barcelona, Spain

Abstract

In the modern world, speech technologies must be flexible and adaptable to any framework. Mass media globalization introduces multilingualism as a challenge for the most popular speech applications such as text-to-speech synthesis and automatic speech recognition. Mixed-language texts vary in their nature and when processed, some essential characteristics must be considered. In Spain and other Spanish-speaking countries, the use of Anglicisms and other words of foreign origin is constantly growing. A particularity of peninsular Spanish is that there is a tendency to nativize the pronunciation of non-Spanish words so that they fit properly into Spanish phonetic patterns. In our previous work, we proposed to use hand-crafted nativization tables that were capable of nativizing correctly 24% of words from the test data. In this work, our goal was to approach the nativization challenge by data-driven methods, because they are transferable to other languages and do not drop in performance in comparison with explicit rules manually written by experts. Training and test corpora for nativization consisted of 1000 and 100 words respectively and were crafted manually. Different specifications of nativization by analogy and learning from errors focused on finding the best nativized pronunciation of foreign words. The best obtained objective nativization results showed an improvement from 24% to 64% in word accuracy in comparison to our previous work. Furthermore, a subjective evaluation of the synthesized speech allowed for the conclusion that nativization by analogy is clearly the preferred method among listeners of different backgrounds when comparing to previously proposed methods. These results were quite encouraging and proved that even a small training corpus is sufficient for achieving significant improvements in naturalness for English inclusions of variable length in Spanish utterances.

Keywords: Nativization, pronunciation by analogy, multilingual TTS, grapheme-to-phoneme conversion, phoneme-to-phoneme conversion

1. Introduction

Spain is a country of a remarkable linguistic patrimony, which is a cultural treasure but also represents an additional challenge in terms of speech technologies. In the framework of the rapidly expanding field of applications the speech

Email address: `tatyana.polyakova@upc.edu,antonio.bonafonte@upc.edu` (Tatyana Polyákova, Antonio Bonafonte)

tools must be adapted to the multilingual scope allowing a higher level of flexibility and answering the needs of modern users. Currently in Spain, hearing proper names from all over the world has become commonplace. Text-to-speech synthesis finds many important applications in the emerging market of speech technologies. Voices that can embrace more than one language are highly demanded in the era of mass media globalization. We wish to assign the term *nativization* to the pronunciation adaptation process.

To maintain an up-to-date synthesizer, we need an ultimate automatic method for the derivation of the nativized pronunciation. The final goal of nativization is to be able to produce highly intelligible synthesized speech that would be well accepted by native speakers of Spanish with some knowledge of English as well fluent English speakers. Previously in Polyákova and Bonafonte (2008a), we used a table-based method for nativization of foreign words in Spanish that produced significant improvements in comparison with the results obtained by applying a Spanish grapheme-to-phoneme (G2P) converter to pronounce English words. The nativization process can be compared to the task of G2P conversion for out-of-vocabulary words. Pronunciation by analogy, previously used in Marchand and Damper (2000) and Polyákova and Bonafonte (2009), proved to be one of the most efficient methods for G2P tasks. Other data-driven techniques have also produced good results (see for instance Taylor, 2005; Bellegarda, 2005; Bisani and Ney, 2008).

We believe that analogy between the nativized pronunciation and the original pronunciation can be inferred in an even more reliable and simple way because nativization of English words in Spanish is an easier task than finding the pronunciation of unknown English words. In fact, all human attempts to nativize foreign words depend on the analogy between known and unknown words. Transformation-based learning method can be applied to improve the results obtained by other methods and can easily combine different information sources. Very few databases containing non-native pronunciation are available, while the nativization corpora are simply non-existent. The rest of the paper is organized as follows: in Section 2 we further discuss the issues that arise when foreign words enter a language. In Section 3, we review the multilingual G2P conversion system. In Section 4 we explain the differences between English and Spanish phonetic systems. In Section 5 we describe the corpora creation for training and evaluation of the proposed automatic nativization system which is described in Section 6. Section 7 gives a detailed summary of the experimental results. The subjective evaluation is given in Section 8. Finally, Section 9 concludes the paper and addresses future directions of study.

2. Issues with pronunciation of foreign words in a language

Every language receives a constant incoming flow of new words. In addition to the obvious acquisition of neologisms during morphological and semantic word formation, many new words enter the current language from foreign languages (Real Academia Española, 1992). There are several ways that words of foreign origin can become incorporated into a receptor language. On many occasions words are translated through semantic borrowing or calque, e.g., *computer mouse* to *ratón*, or *weekend* to *fin de semana*. Another source of foreign-derived neologisms is lexical borrowing where the lexical form and the semantic meaning are adopted directly from the donor language. This form of borrowing implies adaptation of the pronunciation of the new word to the receptor language and almost always that of the orthographic form as well. For example: *football* to *fútbol*; *whiskey* to *güisqui*; *scanner* to *escáner*. This adaptation has two steps. In the first step, the pronunciation is altered to imitate the pronunciation of the language of origin in regards to the limitations of the phonological system of the receptor language. Then, after the word has been used frequently in everyday life, it loses its original foreign form and its orthography is transformed according to the pronunciation of the receptor language, which is Spanish in our case (see Real Academia Española, 1992). Usually, this involves deletion of the unpronounced consonants, one of the double consonants, the unpronounced final *e*, or other changes. The lexical stress presents an even bigger challenge than the orthographic representation. In English, the stress position is quite irregular. As a matter of fact, many words have primary and secondary stresses that sometimes makes the auditive recognition of where they should be placed an issue, especially for non-native speakers. Due to these factors and also to the receptor language accentuation patterns the stress in the assimilated word does not always match its original position. For example, the pronunciation of the French word *élite* [e ˈl i t] in Spanish varies from [ˈe l i t e] to [e ˈl i t e]. Phonetic representations are given in IPA (Handbook, 1999) throughout the document. Please note that the notation / / is used to indicate the phonemic transcription, and [ ] to indicate allophonic and nativized transcriptions. The apostrophe indicates the stressed syllable. In the first case, the stress is shifted as the consequence of the Spanish interpretation of the French graphic accent (which is used to designate if the vowel is open or closed); the second pronunciation, however, is also accepted by Real Academia Española (1992). Every language has its own accentuation patterns and specific characters that cannot be copied to the new language. This is one of the reasons why it is such a delicate matter to decide the best graphic representation and pronunciation for the new word. In this work, we focus on pronouncing English words in Spanish, before they undergo any graphic assimilation, in the scope of multilingual texts.

2.1. Mixed-language texts

Texts written in several languages present a rapidly spreading phenomenon that should not be ignored when talking about high quality speech applications. Worldwide globalization is responsible for an entirely new form of multilingualism in all types of communications resources. Types of mixed-language inclusions vary from word parts to entire sentences. Pfister and Romsdorfer (2003) classify foreign language inclusions into three classes:

1. words containing a foreign stem but following receptor language morphology
2. full words following foreign morphology that do not always agree syntactically
3. syntactically correct sentence chunks

Single foreign words or phrases such as movie titles and proper names already present language identification and pronunciation issues, while more complex language mix-ups that can be found in chats, forums, and other sources make the disambiguation even more problematic. Multilingual texts vary in their nature and their degree of multilingualism depending on the document source. For example, text extracted from a newspaper obeys the strict style determined by the editor, which dictates whether and how the foreign words should be used and when they should be translated to the official language of the issue. Peoples' names and geographical names would be the only inevitable foreign words in this case. Some popular free of charge newspapers, such as international *Metro* and European *20 minutes*, are usually not very restrictive in their use of foreign terms; there are numerous foreign words and phrases in articles on culture and entertainment events. However, texts originating from sources such as blogs, online forums, emails, and short messages reach the highest degree of multilingualism. In such texts, orthographic errors are abundant and unusual abbreviations are frequent, making the search for the correct pronunciation rather challenging.

2.2. Information sources

We can obtain information about correct foreign word pronunciation from different sources. For instance, the *book of styles* used by the television channels and radio stations provide a general idea of how different foreign words should be adapted to the official language (see for instance Llorente and Díaz Salgado, 2004). This book is consistent, and although it does not give enough detail on the nativization of foreign words in all cases, it sets the main guidelines to follow. The tendencies for the pronunciation of frequently used words are rather clearly defined, yet the degree of multilingualism for spoken programs is considerably inferior to that of written texts. Usually, the only foreign words that appear during a news flash are the well-known proper names and orthographically assimilated foreign words. Nonetheless, to synthesize high quality intelligible speech from multilingual texts, it is necessary to be able to pronounce any new word that one may encounter. The criteria to be applied for nativization should depend on the frequency of use of the word in the language and the target audience. Unfortunately, only a small percentage of Spanish TV viewers are fluent in English (Education First, 2011).

2.3. Phoneset extension

A phoneset or phoneme inventory is a set of symbols that defines the sounds of a language. Extension of the phoneset phenomenon occurs more often in bilingual communities or in cases when a speaker is at least bilingual; 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. Spain has five officially bilingual autonomic regions: Catalonia, Valencian Community, Balearic islands, Basque country, and Galicia. English phonemes /ʃ/, /z/, /ʒ/, /ʤ/, /ə/, and /ŋ/ (as a phoneme but not as an allophone) are absent from Spanish but exist in Catalan; others exist in Galician. English dental fricative /ð/, for example, finds its analog only in Basque. Better coverage of the English phoneset allows speakers from these autonomous regions to use all the sounds from their phonemic inventory in addition to Spanish sounds, bringing their pronunciations of English words closer to the actual English pronunciations.

In the particular case of Catalan, the phenomenon of nativization of foreign words also takes place; the Catalan phoneset as mentioned previously is much closer to English compared to that for Spanish. Therefore, nativization has to cope mostly with the adaptation of vowel pronunciations. 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. An example of the latter is the pronunciation of Spanish name *Jorge* in Catalan being [ˈx o ɾ x e] and not [ˈʤ o ɾ ʤ ə] as Catalan phonetics would stipulate. The phoneme /x/ is absent from Catalan, but is used for Spanish words.

2.4. Previous approaches to nativization

When dealing with a variety of foreign words, the problem of language identification certainly arises. In Font Llitjos and Black (2001), an *n*-gram based technique for identification of the language of proper names is described. The same method can be used for common words. The main goal in Font Llitjos and Black (2001) was to find the correct pronunciation of the proper names from the viewpoint of American English pronunciation, or in other words, to Americanize them. In fact, the problem described by Font Llitjos and Black is similar to the one that we are attempting to solve for Spanish, with the goal of extending it to all types of foreign inclusions. An example of a problem similar to nativization is the development of a cross-language synthesizer described in Black and Lenzo (2004). A Basque synthesizer was developed using an existing diphone Spanish synthesizer. The resulting voice was Spanish accented and sounded like one of the many speakers of Basque whose native language is Spanish. The phonemes in Basque were mapped to the phonemes in the available language (Spanish). Even if the mapping was imperfect, it maintained the vowel-consonant relationships across the languages. This type of mapping can only make sense if there is a significant percentage of phoneme overlap between languages. Spanish and Chinese, for example, do not share enough phonemes for this type of mapping.

In Pfister and Romsdorfer (2003), the language identification issue was approached with a text analyzer. The text analysis was decomposed in two steps: a set of monolingual text analyzers was elaborated with their own lexica and grammar; and then for each pair of languages $\{L_i, L_j\}$ an inclusion grammar defined which elements of the language L_j were allowed as foreign inclusions in the language L_i. This work solved the problem regarding the use of German in Switzerland where there is a tendency to pronounce foreign words or even word parts according to the donor language phonetics. Moreover, the text analyzer provided precise word and morpheme language identification for this narrow problem. The pronunciation of foreign proper names has also been addressed for the case of English and German names in Swedish (Lindström, 2004), but once again, the authors determined that Swedish speakers extended both their phonemic inventories and their phonotactics when pronouncing foreign names. Of course, the intelligibility of such names does not only depend on the speakers but also on the listeners and their linguistic and cultural backgrounds. On the other hand, for both English and European Spanish languages there is a clear tendency to adapt foreign proper name pronunciation to the phonetics of the receptor language. Indeed, in the two languages, and especially in Spanish, due to the smaller coverage of its phoneset, it would sound very unnatural to have foreign sounding inclusions. The nativization issue was mentioned and the factors influencing nativized pronunciations were analyzed in the framework of the Onomastica project dedicated to the creation of a multilingual lexicon (Trancoso, 1995). Later, a rule-based approach was applied to the derivation of alternative pronunciations with different degrees of nativization; both full and null knowledge of foreign language were considered for this purpose. These alternatives were used in voice-controlled navigation system queries for German and French (Trancoso et al., 1999). In French, as in Spanish, foreign words and proper names are nativized to French pronunciation and the phonemes are restricted to the French inventory. However, the lexical accent is placed on the last syllable in 99% of the cases as French pronunciation would suggest. The nativization phenomenon is very common for monolingual regions of European countries. In bilingual regions as well as in countries with significant English-speaking influence such as Sweden or Switzerland, tendencies for phoneset extension and closer proximity to foreign pronunciations can be observed (Trancoso et al., 1999; Pfister and Romsdorfer, 2003; Lindström, 2004; Polyákova and Bonafonte, 2008a).

Nevertheless, the problem of foreign word nativization is relatively new to speech synthesis researchers. The multilingualism problem in general has been given much more attention in the framework of automatic speech recognition (Trancoso et al., 1999; Van den Heuvel et al., 2009), where non-native and dialect variations are reported to be the cause of a great number of recognition errors. In synthesis, when dealing with the problem of non-standard pronunciations, we can divide it into two components: foreign pronunciation of native words or non-native speech, and native pronunciation of foreign words. The first component of the problem is highly variable, given the large number of different accents and corresponding phonesets. Moreover, foreign pronunciations hardly obey any regular pattern because they comprise pronunciation, intonation, and semantic and word-morphing errors, whereas, native or nativized pronunciation of foreign words follows traceable patterns. Prosody, intonation, speech rate, and other components, are defined by the phonetics of the receptor language, and only the pronunciation is influenced by the donor language. Several social linguistic conventions based on the frequency of use of a particular word and its degree of phonetic assimilation in the receptor language help to define pronunciation adaptation criteria. In this work, we strive for a balance between an acceptable pronunciation for a native Spanish speaker and correct pronunciation from the point of view of English pronunciation.

3. Overview of a multilingual grapheme-to-phoneme system

For more accurate multilingual grapheme-to-phoneme conversion knowing the language of each word in the text can be rather helpful. However, it is also important to have a tool capable of efficiently determining the language of the paragraph in mixed-language texts extracted from newspapers, online forums, emails, scientific articles, technical support manuals, web pages, and other sources where the language can suddenly change. Our multilingual G2P conversion system is configured to determine the language of the paragraph and then of each isolated word in that paragraph. By defining correctly the source and the target languages for nativization the synthesis quality can be improved considerably. For clarity of definitions, we will use the term *source language* to define the language of origin of a foreign word and the term *target language* to indicate the language to which that foreign word should be adapted. In this paper, we assumed that all foreign words are labeled with their source language (English) and the paragraph language is always set to Spanish. Some results on identification of the language of the paragraph and

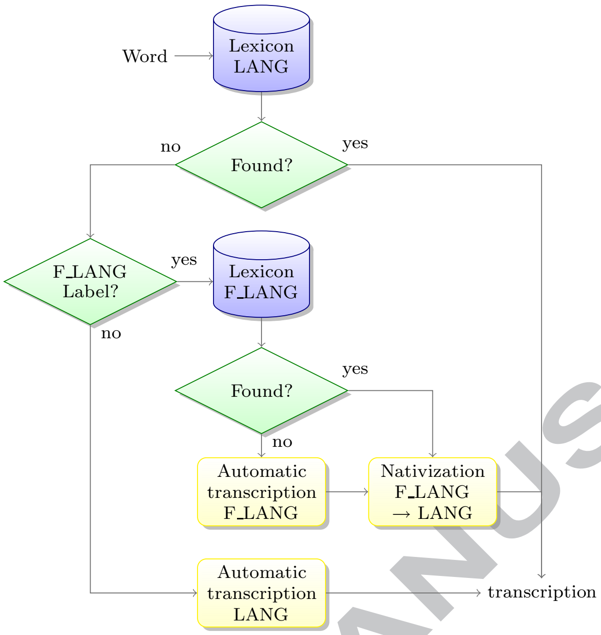

Word
Lexicon LANG
Found?
no
yes
F_LANG Label?
yes
no
Lexicon F_LANG
Found?
yes
no
Automatic transcription F_LANG
Nativization F_LANG → LANG
Automatic transcription LANG
transcription

Figure 1: Scheme of a multilingual G2P system.

isolated words were reported in our previous work (Polyákova and Bonafonte, 2008a). Figure 1 shows our nativization system, where each foreign word had been assigned a special label, ; the default language of the system is or target language.

Our pronunciation module consists of system dictionaries in several languages and corresponding language-specific grapheme-to-phoneme converters. The pronunciation is sought separately for every isolated word. In the post-processing, some phonotactic rules are applied to improve the naturalness of the synthesized speech. The first step is to determine whether the word in question is found in the system dictionary of the target language. If this occurs, that pronunciation is validated. It is important to emphasize that if a foreign word is found in the target language dictionary, we consider that it is already nativized. That is why there is no need to check the language before the first step. If the word is not in the target language dictionary, the next step is to determine if its language of origin is different from the target language (does it have an F_LANG label?). If no label is found, the pronunciation is derived using the automatic transcription system for the target language. For the words identified as foreign, the search continues in the corresponding source language dictionary. Before validating the pronunciation, if it is found in the dictionary, the nativization phoneme-to-phoneme converter is applied to the source. The output of the nativization module is the nativized pronunciation adapted to the target language. In the last case, if the word is also absent from the source language dictionary, its pronunciation is derived using the automatic transcription system for the source language, after which nativization is applied before validating the pronunciation.

4. Spanish phonetics vs. English phonetics

There are numerous phonetic differences between English and Spanish. We sought to examine consonants and vowels separately. The discrepancy between consonants and their orthographic representation in English is less significant than in the case of vowels.

Peninsular Spanish lacks English consonants such as /ʃ/, /v/ /ð/, /ʤ/, /ʒ/, /z/ and /ŋ/, and Latin Spanish also lacks the unvoiced /θ/. Bear et al. (2003), Yavas (2006) and Raynolds and Uhry (2009) reported that the most common substitutions for the missing consonant sounds in English by native Spanish speakers are: /θ/→/t/, /f/ (e.g., *thin/tin, bath/baf*), /ð/→/d/ (e.g., *they/day, lather/ladder*), /v/→/b/ (e.g., *vote/boat*), /z/→/s/ (e.g., *zip/sip, prize/price*), /ʃ/→/ʧ/,/s/ (e.g., *shop/chop, wash/watch, she/see*), /ʤ/→/ʧ/ (e.g., *jeep/cheap*), and /ŋ/→/n/ (e.g., *hanged/hand, sung/sun*). In both English and Spanish phoneme repertoires, we find unvoiced stop consonants /p/, /t/, and /k/ and voiced /b/, /d/, and /g/. However, they have significant differences at the time of articulation. In English, voiced stop consonants /b/, /d/, and /g/ present loss of voicing during their production. In Spanish, however, /b/, /d/, and /g/ are fully voiced because voicing begins before the start of the vowel. In English, there is a small delay after unvoiced stop consonants /p/, /t/, and /k/ before the following vowel in stressed syllable-initial positions that is known as aspiration. Spanish stop consonants, on the contrary, are not aspirated. The phoneme /p/ in the Spanish word *pesos* sounds more like /b/ in the English word *basis* than /p/ in *paces* (Ladefoged, 2003), although this particular difference was not considered in the present work. English has two different phonemes to represent the letters *b* and *v*, /b/ and /v/, respectively. Spanish also contains these letters, however, they are pronounced either with a bilabial approximant sound [β] or a stop /b/ that occurs at the beginning of the word. No labiodental /v/ is produced (Hammond, 2001). The English phoneme /ŋ/ finds its twin in the Spanish velar nasal allophone [ŋ] occurring before velar consonants in words or at word boundaries, e.g., *increíble* or *un gato*. English alveolar-voiced fricative /z/ also exists in Spanish only as an allophone [z] occurring at the end of a syllable before a voiced consonant, e.g., *abismo, desdén*. English dental fricative /ð/, is similar to Spanish dental appoximant [ð] that occurs inside a word when it is not preceded by nasals /m/,/n/, lateral alveolar /l/ or a pause.

English and Spanish vowels are quite different. Spanish has 5 pure vowels while American English has 11 pure vowel sounds. Vowel transcription in English presents a special difficulty due to its deep orthography. For consonants, the length of the preceding vowel contains important information that helps to distinguish voiced consonants from unvoiced stop consonants at the end of a word; this is crucial for making distinctions between words. In Spanish, vowel length is not as variable and these small differences do not cause semantic changes (Fox et al., 1995). The list of Spanish and American English pure vowels is given in Table 1.

Native speakers of Spanish usually have trouble in perceiving and producing the variety of English vowels. For example, no distinctions are made between *ship/sheep* or *fool/full*. Besides, Spanish speakers tend to prefix English words beginning with *s*− consonant cluster with an /e/ sound, so that *school* becomes [e s ˈk u l]. Furthermore, some sound swallowing is typical when three or more consonants occur together, as in *next* [ˈn e k s] (Swan and

IPA symbol	Description	Example
/i/	close front	*pico* /ˈp i k o/
/e/	mid front-central	*pero* /ˈp e ɾ o/
/o/	mid back-central	*toro* /ˈt o ɾ o/
/u/	close back	*duro* /ˈd u ɾ o/
/a/	open central	*valle* /ˈb a ʎ e/

(a) Spanish (Conde, 2001)

IPA symbol	Description	Example
/i/	close front	*tree* /ˈt r i/
/ɪ/	near-close front	*rich* /ˈr ɪ ʧ/
/ei/	close-mid front	*cake* /ˈk ei k/
/ɛ/	open front	*bed* /ˈb ɛ d/
/æ/	near-open front	*had* /ˈh æ d/
/u/	close back	*lose* /ˈl u z/
/ʊ/	near-close back	*put* /ˈp ʊ t/
/oʊ/	close-mid back	*home* /ˈh oʊ m/
/ɔ/	open-mid back	*pause* /ˈp ɔ z/
/ʋ/	open-mid back	*cut* /ˈk ʋ t/
/ɑ/	near-open mid back	*dot* /ˈd ɑ t/

(b) American English (Wells, 1982)

Table 1: Pure vowels in Spanish and American English.

Smith, 2001). These are the main observations that helped us to define the nativization criteria detailed in Section 5.3.

5. Nativization database creation

In this section, we describe the nativization lexicon created for training and evaluation of nativization methods. Rule-based approaches to phonemization require significant linguistic engineering, and they are always language-dependent, thus lacking flexibility. Data-driven approaches were proven to be more efficient than those based on the explicit linguistic modeling and they are undoubtedly superior in adaptability (van den Bosch and Daelemans, 1993). The main purpose of this work was to train a nativization model capable of converting English pronunciation to nativized Spanish. Data-driven techniques require training corpora, so a need for nativization training was apparent. For typical G2P conversion tasks, large pronunciation corpora of 100,000 words and their corresponding pronunciations are available. Since we did not find any existing nativization database, we chose to create a minimalistic corpus that would not require hiring a highly qualified expert in linguistics.

5.1. Training data

For our task, due to the reduced sized of the training lexicon or *TrainingSet*, it was necessary to have it orthographically balanced. A greedy corpus-balancing tool was used for selecting words to be nativized from the available

LC-STAR dictionary of American English (Hartikainen et al., 2003) with more than 50,000 entries, previously used by the authors in G2P conversion experiments (Polyákova and Bonafonte, 2009). To have all possible letter bi-grams in the corpus, we selected 1000 words. The original English transcriptions of these words were manually nativized according to the criteria described in 5.3. It is necessary to emphasize that the phoneme inventory used for nativization was limited to the Spanish phoneset including three allophones [ŋ], [ð] and [z]. The proportion of rare words in the resulting corpus was noticeable; however, a few non-English words were removed because their pronunciations did not obey English phonetics. Therefore, their presence in the nativization corpus could have introduced additional ambiguity. The TrainingSet consisted exclusively of common words. They were manually aligned during the nativization process.

5.2. Test data

To evaluate the nativization methods a test corpus was required. A specific test corpus was created in order to keep the full coverage of the TrainingSet. The test data was divided into two sets. The main one, named *CommonSet*, consisted of common words only. The words selected for CommonSet from the available online free daily newspaper *www.20minutos.es*, were rather frequently used common words. Such a choice was motivated by the fact that the nativized pronunciation of the frequently used words is less ambiguous than that of the rare ones. In addition to the common words, it was interesting to evaluate the nativization algorithm on a set of frequently used people's names. Therefore a secondary evaluation set was defined. *ProperSet* contained people's names of English origin. The database entries for ProperSet were also collected from free online sources. None of the test words were present in the training lexicon. Both CommonSet and ProperSet contained 100 words each.

5.3. Nativization criteria and examples

In this work, we attempted to find a meeting point between a totally incorrect pronunciations of English words by Spanish speakers unfamiliar with English phonetics and almost correct pronunciations by those who are fluent in English. Since the goal of this project was to improve both naturalness and intelligibility of the synthesized speech, nativization was oriented on general Spanish-speaking auditory conventions. Nonetheless, the evaluation of the synthesized speech is a difficult task because its quality can be only defined by a listener and it varies from one listener to another (Black and Lenzo, 2004). With that goal TrainingSet, CommonSet and ProperSet were nativized using the criteria described further in this section. These criteria were based on the principles described in Llorente and Díaz Salgado (2004), however it was necessary to extend them to be able to transcribe the entire corpus and consider each case separately. Table 2 illustrates how some of the criteria were applied in particular cases. As it was already mentioned the non-English words were deleted from all sets since they could have hampered the generalization. In all cases the frequency of usage of a particular English word in Spanish was taken into account seeking better adaptation of its pronunciation to the language.

Absence of certain source language phonemes in the target language poses phonetic challenges for non-native speakers. For example, the English word *these* would be pronounced as [ˈd iː s] because the Spanish phoneset lacks the voiced dental fricative /ð/. When determining the best way to nativize a word the level of assimilation of the word to the target language plays a major role as well as the complexity of its orthography. Another question to be asked is "Is the word consonant-vowel pattern similar to that in the target language?" For instance, in Spanish, it is unnatural to have more than two consonants in a row at the beginning of the word. Additionally, Spanish does not typically allow more than three consonants sounds in a row in any position in the word, while Czech allows no-vowel words consisting of up to 4-5 consonants. It was also important to ensure that no unusual consonant agglomerations in any of the word parts were encountered, even though sometimes it was inevitable due to the lack of vocalization in English. The case of two consonants *st−* at the beginning of the word particularly stands out because in Spanish a vowel is added before this consonant cluster to smooth the agglomeration. The English [s t] is pronounced [e s t] in Spanish as mentioned already in Section 4.

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. Most certainly, it was found inappropriate to apply the same criteria to well-known words and to words with much lower occurrence rates. In the word *jazz*, the phoneme /ʤ/ was nativized to /jj/, while in *Egyptian*, the same phoneme was transformed to /ʧ/. In the word *logjam*, it was transcribed as [ð j] because the latter is a rare word and complete omission of the initial sound /d/ of the English phoneme /ʤ/ would cause important drawbacks in comprehension of the word (see Table 2 for more examples). The database nativization task was conducted using both source language orthographic form and pronunciation. In English to Spanish nativization, vowels were found much harder to transcribe systematically because their nativized pronunciation in Spanish is highly related to the word frequency-dependent English-to-Spanish orthographic analogy. Phonemes representing double sounds such as /ei/, /oʊ/, /aɪ/, /ɔɪ/ and /aʊ/ were transformed into corresponding double phonemes /e j/, /o w/, /a j/, /o j/, and /a u/. The stressed vowels were mapped to the closest match in the Spanish phone table, e.g., *agency* [ˈæ ʤ ə n s i] to [ˈe ʧe n s i]. Most of the unstressed vowels and especially *schwa* /ə/ in the majority of cases were transcribed with a vowel closest to the letter as in *aimless* [ˈei m l ə s] to [ˈe j m l e s]. Additionally, we considered a specific extension of the Spanish phoneset. This decision was based on the hypothesis that conserving vowel length and word stress would contribute to the intelligibility of the nativized pronunciation. Thus, the /ɪ/ in *dip* was mapped to a short vowel [iˑ], /i/ in *deep* to a long [iː], /ɑ/ to a long vowel [aː], /ɚ/ to [eˑ r], /ɝ/ to [eː r], and /ə/ was mapped to the vowel corresponding to the letter but marked as short. For the consonants, as previously mentioned in Section 4, some difficulties were found when transcribing English /ʒ/ /ʤ/ and /ʃ/. The nasal /ŋ/, the voiced /ð/ and /z/ were conserved as they were present as allophones [ŋ],[ð], and [z] in our Spanish text-to-speech (TTS) system. The unvoiced /ʃ/, in most cases, was transcribed to /s/. The letter sequence *rr* corresponding to the Spanish vibrating phoneme /r/ in all nativized words was mapped to a Spanish alveolar tap /ɾ/ with reduced vibration, as well as the Spanish phoneme /r/ corresponding to the letter *r* at the beginning of the word or after a pause (Llorente and Díaz Salgado, 2004). An illustrative review of the criteria used for nativization together with some exceptions is shown in Table 2.

6. Nativization methods

In comparison with non-native speech, nativized speech is easier to manage in many aspects. Non-native speech is different from the native speech in articulation points, pause distribution, and diphone behavior at word boundaries, and moreover, it is characterized by frequent pronunciation errors. Nativized speech, on the other hand, is more consistent in its definition, conserves the articulation point of the target language and does not contain important pronunciation errors, because its only purpose is to mold the pronunciation of a foreign word to fit smoothly into target language utterances where foreign accented pronunciation would be unacceptable. Nativization is based either on a set of manually crafted or data-driven rules, all of which follow coherent criteria. Nativized speech does not contain mispronounced phonemes.

Section 6.1 describes the first method, based on nativization tables, that consists in defining a set of rules in order to transform source phoneme to target phoneme. Section 6.2 gives a summary of the pronunciation by analogy algorithm and its application to the nativization problem. In this case, two nativization approaches were proposed: using information about the orthographic form and the original English pronunciation. Finally, Section 6.3 justifies the application of the transformation-based learning algorithm to improve the results obtained by the preceding methods using both orthographic and phonetic representations.

6.1. Nativization tables (NatTAB)

In our previous work (Polyákova and Bonafonte, 2008a), we developed a system based on nativization tables (NatTAB). Pronunciations were derived according to the scheme shown in Figure 1. When an out-of-dictionary word was labeled as foreign (label F_LANG), its transcription was sought in the dictionary of the corresponding language (F_LANG dictionary). If the word was not in that dictionary, it was fed to a language-specific G2P system. In both cases, after the word pronunciation in a source language was determined, the nativization procedure was applied. Nativization was performed in a phoneme-to-phoneme manner using nativization tables for source-to-target phoneme transformations. For English-to-Spanish nativization, all English phonemes were mapped to their closest Spanish analogs. In the case of ambiguities, such as when the source pronunciation contained a /ə/, the target language G2P system was triggered and the phoneme suggested by this system was chosen. For the word *talent* /ˈt æ l ə n t/, the table suggests that /ə/ should be nativized to a Spanish phoneme /a/, while the Spanish G2P system gives an /e/ for that position. Therefore, the resulting nativized pronunciation had an /e/ in the 4th position. This imperfect system, that considered no contexts and only a few exceptions that were left up to the Spanish G2P converter, showed a significant improvement when compared to the transcriptions generated using Spanish G2P converter alone. This method will be used as our baseline system.

Word	Original pronunciation	Nativized pronunciation	Comment
airways	/ˈɛ r w e z/	[ˈe j ɾ w e j z]	In Spanish /e j r/ instead of [e r] is frequent.
basketball	/ˈb æ s k ə t b ɑ l/	[ˈb e s k e˔ t b oː l]	British pronunciation of *-ball* is widely used.
water	/ˈw ɑ t ɚ/	[ˈw oː t e˔ r]	/o/ in the 2nd position displays British tendency.
Egyptian	/ə ˈʤ ɪ p s ə n/	[e˔ ˈʧ i˔ p s j a˔ n]	/ʤ/ to /ʧ/ between vowels; /s j/ used to imitate /ʃ/.
comfortable	/ˈk ʊ m f t ɚ b ə l/	[ˈk a m f o˔ r t e˔ β o˔ l]	A short vowel inserted between 2 consonants.
dogfight	/ˈd ɑ g f aɪ t/	[ˈd oː ɣ f a j t]	British tendency for a frequent word part.
Aleutian	/ə ˈl j u ʃ ə n/	[a ˈl j u s j a˔ n]	/s j/ used to imitate /ʃ/.
awkward	/ˈɑ k w ɚ d/	[ˈoː k w e˔ ɾ ð]	British tendency observed for a frequent word.
bank's	/ˈb æ n k s/	[ˈb e n ɣ s]	Final /k s/ in Spanish tends to be converted to [ɣ s].
thanksgiving	/ˈθ æ ŋ k s g ɪ v ɪ ŋ/	[ˈθ e ŋ _ s ɣ i˔ β i˔ ŋ]	Deletion of /k/ to break-up 4 consonants.
American	/ə ˈm ɝ r i k ə n/	[a ˈm eː ɾ i k a˔ n]	[ɝ] turns into [eː] in frequent word.
bowman	/ˈb oʊ m ə n/	[ˈb o w m e˔ n]	*-man* transcribed as [m e˔ n] not [m a˔ n] for more intelligibility.
length	/ˈl ɛ ŋ θ/	[ˈl e ŋ k θ]	Insertion of /k/ after a nasal before fricative.
rainforest	/ˈr ei n f ɔ r ə s t/	[ˈr e j m f o ɾ e˔ s t]	/n/ before /f/ is converted to /m/, according to Canellada and Madsen (1987).
straightjacket	/ˈs t r ei t ʤ æ k ə t/	[e s ˈt ɾ e j t j e k e˔ t]	[s] followed by a consonant at the word beginning to /e s/.
webcam	/ˈw ɛ b k æ m/	[ˈw e β k a m]	Very frequent usage of *cam* with an /a/.
jazz	/ˈʤ æ z/	[ˈjj a z]	Frequent word, /jj/ in 1st and /a/ in 2nd positions.
logjam	/ˈl ɑ g ʤ æ m/	[ˈl oː ɣ ð j e m]	/ʤ/ to /ð j/ after a consonant in a rare word and /ɑ/ to [oː] for more naturalness.
headquarters	/ˈh ɛ d k w ɑ r t ɚ z/	[ˈx e ð k w oː ɾ t e˔ ɾ s]	*-quart-* follows British tendency.
work	/ˈw ɝ k/	[ˈw oː ɾ k]	Frequently used form.
Haitian	/ˈh ei ʃ ə n/	[ˈx e j s j a˔ n]	/s j/ used to imitate /ʃ/.
Australian	/ə ˈs t r ei l i ə n/	[a˔ w s ˈt ɾ e l j a˔ n]	/a w/ corresponds to the orthographic form.
Nigerian	/n aɪ ˈʤ ɪ r i ə n/	[n a j ˈʧ i˔ ɾ j a˔ n]	/ʤ/ to /ʧ/ between a diphthong and a vowel.
Norwegian	/n ɔ r ˈw i ʤ ə n/	[n o ɾ ˈβ iː ʧ a˔ n]	/ʤ/ between vowels to /ʧ/.
Argentinean	/ɑ r ʤ ɛ n ˈt ɪ n i ə n/	[aː ɾ j e˔ n ˈt i˔ n j a˔ n]	/ʤ/ to /j/ after a consonant.
backgrounds	/ˈb æ k g r aʊ n d z/	[ˈb e k ɣ ɾ a w n _ z]	Deletion of /d/ before /z/ for more naturalness.
blindfold	[ˈb l aɪ n d f o l d]	[ˈb l a j m _ f o w l ð]	Deletion of [ð] before [f] for more naturalness; [n] before [f] to [m] (Canellada and Madsen, 1987).
brainpower	/ˈb r ei n p aʊ ɚ/	[ˈb ɾ e j m p a w e˔ r]	/n/ before /p/ to /m/ (Canellada and Madsen, 1987).
boyfriend's	/ˈb ɔɪ f r ɛ n d z/	[ˈb o j f ɾ e˔ n _ z]	Deletion of /d/ to avoid 3 consonants at the end of the word.
jeep	/ˈʤ i p/	[ˈd j i p]	[ʤ] to /d j/ at the beginning of the word.
Persian	/ˈp ɝ ʒ ə n/	[ˈp eː ɾ s j a˔ n]	/ʒ/ also transforms to /s j/, like /ʃ/.
father	/ˈf ɑ ð ɚ/	[ˈf aː ð e˔ r]	/ð/ is approximated by Spanish allophone [ð]

Table 2: Some examples of the nativization criteria application.

6.2. Pronunciation by analogy (PbA)

For the first time, pronunciation-by-analogy (PbA) was proposed for reading studies by Glushko (1981), and later, Dedina and Nusbaum (1991) introduced this method to TTS applications. The latest and most successful implementation of the algorithm was published by Marchand and Damper (2000), which we have reimplemented for our experiments. This system as well as the initial one, called PRONOUNCE (Dedina and Nusbaum, 1991) consists of four major components.

- Aligned lexicon (in one-to-one manner)
- Word matcher
- Pronunciation lattice (a graph that represents all possible pronunciations)
- Decision maker (chooses the best candidate among all present in the lattice)

Below we review the entire algorithm because it is necessary for understanding of the new strategies and introducing new terminology.

6.2.1. Alignment

Pronunciation by analogy algorithm requires a one-to-one match between the orthographic and phonetic strings. In other words, each letter has to be aligned with a corresponding phonetic representation. Finding the correct alignment presents a challenge for languages such as English because the orthographic and phonetic representations of a word frequently have different lengths. Due to their rather complex orthography, English words usually have more letters than sounds. In that case, a null phone /_/ needs to be inserted into the phoneme string, e.g., *thing* /θ _ i ŋ _/. Otherwise, if the number of phonemes is greater than the number of letters, the phonemes corresponding to the same letter were joined together in one, e.g., *fox* /f ɑ k_s/. The alignment used here is based on the EM algorithm, and it is similar to that described in Damper et al. (2004). However, the alignment is not always perfect and it can influence negatively on the results.

6.2.2. Description of the algorithm

After the training dictionary has been aligned, the matcher starts to search for common substrings between the input word and the dictionary entries. Every input word is then compared to all the words in the lexicon to find common “arcs”. We called the substrings in the grapheme context *letter arcs* and the corresponding substrings in the phoneme context *phoneme arcs*. All possible letter arcs with a minimum length of two letters and a maximum length equal to the input word length are generated and then searched in the dictionary. For every letter arc from the input word, that matched 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 for each arc and its length. As an example, assume that the word `#top#` is absent from our dictionary; the

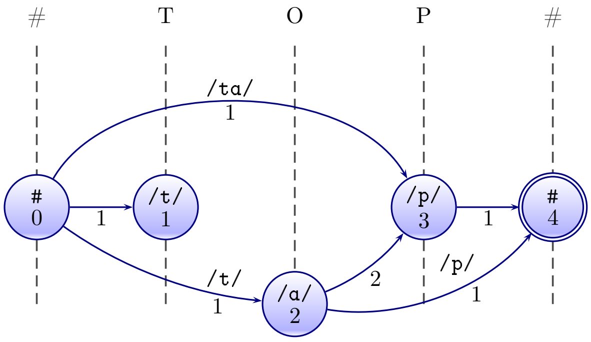


Figure 2: Pronunciation lattice for the word *top* using the arcs extracted from the words *topping* and *cop*.

list of all possible letter arcs for this word can be given as "#t, #to, #top, to, top, top#, op, op#, p#". Now, suppose that in the lexicon, we have the word "#topping#" with the pronunciation /# t ɑ p _ ɪ _ ŋ #/. Here the matcher finds the letter arcs #t, #to, #top, to, op and top with their corresponding phoneme arcs /# t/, /# t ɑ/, /# t ɑ p/, /t ɑ/, /ɑ p/ and /t ɑ p/. Let us assume that the next word in the lexicon is #cop#. It gives us three more matching letter arcs with the word top #top#, which are op, op#, and p# with their corresponding phoneme arcs /ɑ p/, /ɑ p #/ and /p #/. Each time that the same phoneme arc is found for same letter arc, the frequency of the phoneme arc is incremented. After the word *cop* is processed the frequency count for the phoneme arc /ɑ p/ becomes equal to 2, see Figure 2. 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 i, the nodes L_i and L_j are added to the graph and an arc is drawn between them. All nodes are labeled with the corresponding "juncture" phoneme and its position in the word. The arcs are labeled with the remaining phonemes and their frequencies of appearance. An example of the lattice construction for the word *top* using the arcs found in the word *topping* is illustrated in Figure 2. These arcs and their frequency counts are updated when the search continues through all the words of the dictionary. After the pronunciation lattice is completed the decision maker chooses the best pronunciation. Each complete path through the lattice is called a "pronunciation candidate". In this work, we considered only the shortest paths through the lattice (Marchand and Damper, 2000). If there is a unique shortest path through the lattice, it is automatically 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. Please note that no single letter matches were considered. To solve the silence problem, when no complete path through the lattice was found, concatenation of phoneme arcs was allowed disregarding the phoneme overlap that otherwise would be required.

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, with *k* being the pronunciation length. Marchand and Damper (2000) proposed to use five scoring strategies in order to choose the best pronunciation. Also, two methods of strategy combination were introduced. For the NETtalk dictionary (Sejnowski and Rosenberg, 1993), the best accuracy obtained was equal to 65.5% for words and 92.4% for phonemes using all 5 strategies (Marchand and Damper, 2000), which is better than using any one of the mentioned strategies alone. The sum and the product rules of strategy combination gave similar results. In our previous work (Polyákova and Bonafonte, 2009), we proposed six additional strategies for choosing the best candidate that in combination with the others outperformed the original strategies. 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 that constitute the candidate. Different strategies work with different aspects of analogy. High arc frequency is considered a major advantage over low arc frequency. The frequency of suffixes and prefixes are prioritized by different strategies. The final score for the candidate is directly proportional to the number of phonemes it shares with the others. If two candidates share the same pronunciation, both of them are prioritized. These measures are used separately or combined across the strategies. The strategies are explained in detail in Polyákova and Bonafonte (2009) and briefly in Table 3. The pronunciation by analogy algorithm, previously applied to G2P conversion (Marchand and Damper, 2000; Polyákova and Bonafonte, 2009), in this paper is applied to the task of nativization of English words in Spanish.

Nativization by analogy was attempted from two different viewpoints. The first approach, grapheme-to-phoneme nativization (G2Pnat), is very similar to G2P conversion, with the only difference being that the phonemes are the nativized ones. It makes sense to perform grapheme-to-phoneme nativization. In fact, most of the Spanish listeners are only familiar with the orthographic form of English words. However, if there is a phonetic transcription available in the source language, finding automatic correspondences between source and target (nativized) phonemes is a more consistent task than in the case of letters, being that G2P conversion is already a difficult task for English. Therefore, the second approach adapted is the phoneme-to-phoneme nativization (P2Pnat). The pronunciation by analogy method can be also applied to the phoneme-to-phoneme nativization with the differences being that the input data consisted of the English phonemes and there was a need for slight modifications in the dictionary processing part.

6.3. Transformation-based error-driven learning (TBL)

Previous results obtained for grapheme-to-phoneme conversion using TBL to correct the errors (Polyákova and Bonafonte, 2006, 2008b), encouraged us to consider this approach for our current work as well.

In order to further exploit the possibilities for improvement of the nativized pronunciation using TBL, the algorithm was applied to the results obtained by PbA and NatTAB. With the purpose of determining the generalization potential of the TBL algorithm itself, it was also applied to correct the results of a most-likely target phoneme prediction (ML). For this purpose, based on a lexicon aligned in a one-to-one manner each source phoneme was assigned the most-frequent target phoneme in the mapping.

The TBL algorithm, originally invented by Brill (1995), consists in learning transformation rules from the training

strategy mask	strategy meaning
10000000000	maximum arc frequency product
01000000000	minimum standard deviation of arc length
00100000000	highest same pronunciation frequency
00010000000	minimum number of different symbols
00001000000	weakest arc frequency
00000100000	weighted arc product frequency
00000010000	strongest first arc
00000001000	strongest last arc
00000000100	same symbols multiplied by arc frequency
00000000010	lowest count of different phonemes
00000000001	max. freq. product with most frequent same pron.

Table 3: Eleven scoring strategies for pronunciation by analogy.

corpus that is labeled with some initial classes. The TBL algorithm uses templates to generates rules to generalize the transcription errors obtained by the initial prediction method. The templates consist of several features, that for this particular task can be the predicted phoneme, letter context, etc. Some examples of rule templates are given below

```
let_-1 let_0 let_1 → ph
let_1 let_0 let_1 ph_0 → ph
let_-1 let_0 let_1 ph_-1 ph_0 ph_1 → ph
```

where let_0 represents the letter corresponding to the current phoneme, while let_-1 and let_1 define the surrounding orthographic context. In this case, ph_-1 and ph_1 represent the surrounding predicted phoneme context and ph_0 represents the predicted phoneme itself. ph is the correct phoneme to which ph_0 should be transformed.

The erroneous tags in the training corpus serve as the basis for deriving error-correcting transformation rules. During the learning process the TBL algorithm learns rules iteratively with the goal of correcting as many errors as possible in the training corpus. Rules are generated and applied to the current state of the training corpus at each iteration. The number of errors corrected is also called the number of good applications. The number of bad applications is defined by the number of times that application of a rule has introduced a new error. The score of each rule equals to the difference between the number of good and bad applications. The rule capable of correcting the largest number of errors (the one with the highest score) at each iteration is applied to the entire training corpus and appended to the final rule list. The scores of the other rules affected by the application of the best rule at current iteration are also updated. The rule learning process continues until no rule that improves the accuracy of the training prediction could be found or a rule with a score lower than the preset threshold is generated.

Using the TBL algorithm to correct the prediction previously obtained by another classifier allows capturing the imperfections of previous approaches into a set of context-dependent transformation rules, where the context serves as a conditioning feature. During the test phase, each rule from the list is applied whenever a match between the input set of features and those defined in the rule is found. The rules are applied to correct the errors in the initial prediction

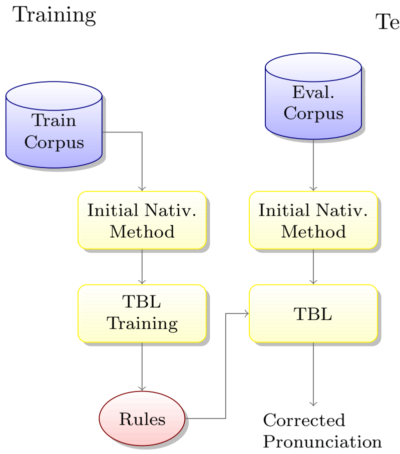


Figure 3: Scheme of combination of TBL with other nativization methods.

previously obtained for the test corpus in the same order that they were generated.

Figure 3 shows the scheme of combination of the data-driven nativization methods with TBL. The transformation rules are derived from the errors in the initial prediction needing improvements obtained by a previous classifier for the training data. The TBL algorithm not only allows correction of errors in the previous predictions but also an easy combination of different conditioning features for error correction. In our work, we used both orthographic and phonetic forms in the source language to improve accuracy of nativized pronunciations in the target language. The performance of the TBL algorithm highly depends on the size of the training corpus and the number of prediction errors obtained by the initial classifier. A higher error ratio in the training phase and a larger training corpus lead to better correction results. This was analyzed for the case of Mandarin polyphones prediction in Zheng et al. (2005).

7. Experimental results

In this section, we present the experimental results obtained for the nativization task using different methods. First, in 7.1, we discuss the baseline results obtained with nativization tables (NatTAB). Next, in 7.2 and 7.3 we evaluate the proposed analogy-based approaches. Later, in 7.4, we describe an attempt to improve the best results obtained so far by applying the transformation-based error-driven learning (TBL) algorithm. Furthermore, we evaluate the combination of TBL with simpler methods, such as nativization tables (NatTAB) or most-likely phoneme assignation (ML). The latter combination allows to validate the performance of the TBL algorithm itself, given that the initial prediction in this case is very simple. Finally in 7.5, we compare the errors obtained by different nativization methods qualitatively.

7.1. Baseline results (NatTAB)

As we explained in Section 6.1, the NatTAB method carries out the nativization in a phoneme-to-phoneme manner, using hand-crafted nativization tables for the source-to-target phoneme transformations. The method based on the nativization tables was able to predict only 73.9% of phonemes and 23.8% of words correctly from CommonSet.

However, these results are much better than those obtained on the same test data without using nativization, applying the Spanish G2P to derive the pronunciation of English words. Spanish G2P nativized only 61.2% of phonemes and 8.6% of words from CommonSet correctly. The only nativizations that this system predicted correctly were those pronounced very close to their orthography, e.g., *bed* → [b e ð] or *car* → [k a r]. The objective results obtained applying the table-based phoneme-to-phoneme mapping (NatTAB) for English to Spanish nativization were quite low in comparison with those reported for the G2P conversion in many languages. Nevertheless, the results of the perceptual evaluation described in Polyákova and Bonafonte (2008a) showed that even such a simple nativization method had better acceptance among listeners than synthesized speech that implied no nativization at all and treated all words as if they were Spanish.

7.2. Grapheme-to-phoneme nativization (G2Pnat)

The first prediction method to be tested was the prediction of nativized pronunciation focusing on analogy in the orthographic word forms. Out of 11 strategies available for the PbA algorithm for choosing the best pronunciation candidate, it was necessary to determine the best strategy combination for our data. As we do not have any development data, an *n*-fold cross-validation was carried out on TrainingSet, leaving out each word at a time and using the remaining words for pronunciation lattice construction as described in 6.2. All possible strategy combinations were considered and compared. For G2Pnat, the resulting best strategy combination for TrainingSet was 10001001011 (1 meaning that the strategy corresponding to that position was included and 0 that it was omitted). The best *n*-fold results obtained for TrainingSet were 85.7% in phoneme and 45.6% in word accuracy. As already mentioned in 5, both training and test data contained lexical stress and vowel length information. However, the vowel length was not predicted at this time but will be addressed in the future. Firstly, it was important to evaluate the nativization accuracy without introducing any additional complexity to the task. For this reason, for the first experiment, the stress markers were removed. The results obtained with CommonSet using the best strategy combination were 84.2% phonemes and 43.8% words correct (Table 4). If we compare these results to the baseline results obtained with NatTAB, we can see that the word accuracy rate has almost doubled. See Figure 4 for an overview of the results.

The follow-up experiment, also carried out with CommonSet was aimed at prediction of the stress and nativized phonemes together. Stress inclusion increased the number of errors considerably, resulting in accuracy rates of 74.7% for phonemes and 20.0% for words. This further demonstrated that in English, stress prediction uniquely from the orthographic form is a difficult task (Black et al., 1998).

Since stress prediction results were slightly discouraging, experiments on the ProperSet were performed discarding this additional feature. The word accuracy obtained for G2Pnat on ProperSet was about 12 percentage points lower than that for CommonSet (see Table 4). Such a loss in accuracy can be explained by the fact that even if the proper names test set contained the most frequent and simple proper names of strictly English origin, their orthography is deeper than that of the common words.

method	test set	phon. acc. [%]	word acc. [%]
G2Pnat	common	84.3	43.8
	proper	74.8	31.5
P2Pnat	common	91.6	63.8
	proper	87.2	55.6

Table 4: Results obtained for G2P and P2P nativization by analogy with CommonSet and ProperSet.

7.3. Phoneme-to-phoneme nativization (P2Pnat)

For P2Pnat experiments the PbA algorithm was also applied. The training lexicon used was the phoneme-to-phoneme version of TrainingSet, with source phonemes on the left-hand side and nativized phonemes on the right-hand side. Similarly as for G2Pnat, the best strategy combination (11011000010) was determined performing *n*-fold cross-validation of all possible strategy combinations. The best *n*-fold results obtained for TrainingSet were 91.8% of phonemes correct and 61.3% of words correct. The accuracies obtained for CommonSet were 91.6% for phonemes and 63.8% for words respectively (Table 4). These results showed that P2Pnat outperforms G2Pnat by 20 percentage points in word accuracy terms. For ProperSet, the phoneme-to-phoneme results were also promising: 87.2% in phoneme and 55.6% in word accuracy beat by 23 percentage points the grapheme-to-phoneme nativization results for the same dataset (Table 4). Furthermore, this method is advantageous because it allows copying of the original accent to the nativized form with 99% accuracy for CommonSet. As both CommonSet and ProperSet datasets are rather small, the confidence interval of these results is relatively large. However, even such small sets allow obtaining statistically significant results at the $p = .05$ level on the basis of a binomial significance test.

7.4. Applying transformation-based learning to nativization

In view of the improvements previously obtained using TBL for the G2P task (Polyákova and Bonafonte, 2008b), our next approach was to apply the transformation-based learning to improve the results of other nativization methods as mentioned in Section 6.3. The experimental results were evaluated on CommonSet. As can be seen from Figure 3, to learn error-correcting rules the TBL algorithm requires an initial prediction both for the training and test sets. In this work, TBL was applied to correct the initial nativization prediction in three cases: for nativized pronunciations obtained by phoneme-to-phoneme nativization (P2Pnat), nativization tables (NatTAB) and most-likely phoneme assignation (ML).

Before applying the TBL algorithm, it was necessary to obtain the initial predictions for training and test data for all methods. For the P2Pnat method the initial prediction for TrainingSet was generated using *n*-fold cross-validation, leaving out each word at a time and using the rest of the lexicon to derive the nativization of the word by analogy with the remaining words. The initial prediction for the test data, CommonSet in our case, was obtained using the entire TrainingSet. To obtain the initial TrainingSet and CommonSet predictions with NatTAB, English phonemes were mapped to the closest corresponding Spanish phonemes given in the nativization table for this pair of languages.

And finally, for the last experiment, the most-likely nativized phoneme was assigned to TrainingSet and CommonSet as explained in Section 6.3.

The correction rules for all methods depended on such features as source letter, source phoneme, and predicted phoneme, therefore, allowing combination of orthographic and phonetic knowledge from the source language. Different context types and lengths were considered (see Section 7.4). Table 5 shows phoneme and word nativization accuracies obtained for different initial predictions and contexts. The source letter context varied from 3 to 5, while the source phoneme context was considered in all cases but the first case, and its length varied from 1 to 3 phonemes. The predicted phoneme context was set to 3 for all experiments. The results are given for different methods (P2Pnat, NatTAB, most-likely phoneme (ML)) combined with TBL and for two different stopping thresholds: $t_1 = 1$ and $t_2 = 5$. The algorithm terminated when no rule with a score lower than the specified termination threshold was generated.

methods	P2Pnat+TBL acc. [%]		NatTAB+TBL acc. [%]		ML+TBL acc. [%]	
context type	phoneme	word	phoneme	word	phoneme	word
Stopping threshold = 1						
orig_let = ± 3 orig_ph= 0	90.4	60.0	88.4	59.1	83.8	43.8
orig_let = ± 3 orig_ph= ± 1	91.1	62.9	88.4	59.1	83.5	40.0
orig_let = ± 3 orig_ph= ± 3	91.3	62.9	90.5	64.8	88.8	49.5
orig_let = ± 4 orig_ph= ± 3	91.5	64.8	90.2	63.8	88.5	47.6
orig_let = ± 5 orig_ph= ± 3	91.5	64.8	90.2	63.8	88.5	47.6
Stopping threshold = 5						
orig_let = ± 3 orig_ph= 0	92.0	63.8	87.7	59.1	78.6	32.4
orig_let = ± 3 orig_ph= ± 1	92.0	63.8	87.7	59.1	78.9	34.4
orig_let = ± 3 orig_ph= ± 3	92.7	66.7	90.0	64.8	87.0	43.8
orig_let = ± 4 orig_ph= ± 3	92.5	66.7	90.0	64.8	87.0	43.8
orig_let = ± 5 orig_ph= ± 3	92.5	66.7	90.0	64.8	87.0	43.8

Table 5: Phoneme and word accuracy obtained by TBL in combination with different nativization methods as a function of letter and phoneme context used by the rules.

The best results of 66.7% words correct, were obtained for the larger source phoneme context and P2Pnat prediction. This was more than 2 percentage points higher than the result obtained using P2Pnat alone. The second best results, 64.8% of correct words, were obtained by applying the TBL method to the NatTAB prediction, and in this case, the initial word accuracy was improved by 20 percentage points. The results obtained by NatTAB+TBL are quite good since they are slightly better than the performance of P2Pnat alone. TBL by itself proved capable of generalizing the nativization criteria when applied to correct the most-likely phone prediction, with a gain of about 24 percentage points in word accuracy in comparison to that obtained with NatTAB without TBL. However, the best TBL results are obtained when the best initial prediction is used, in this case P2Pnat. The results obtained by the combination of NatTAB+TBL and P2Pnat are quite similar and can be considered equal alternatives.

Even though no precise conclusions can be drawn, we can observe that larger letter and phoneme contexts appear to make a greater contribution to error correction. For training data containing less errors, as in the case of P2Pnat a higher stopping threshold seems to be more suitable.

The nativization results obtained on CommonSet using different methods are summarized in Figure 4. The differences are statistically significant in all cases except P2Pnat+TBL: we cannot ensure that TBL combined with P2Pnat gives better performance than P2Pnat alone. Such a small test corpus does not allow us to obtain statistically significant differences between best performing methods. Furthermore, for P2Pnat the number of errors available for rule learning is inferior to that obtained by other methods. Usually, good error correction rates are achieved for large lexicons of about 50K words and high error rates in the training prediction. If the initial prediction accuracy is rather high and the training corpus is rather small, the application of TBL may not give significant improvements. All improvements obtained by the TBL algorithm are consistent.

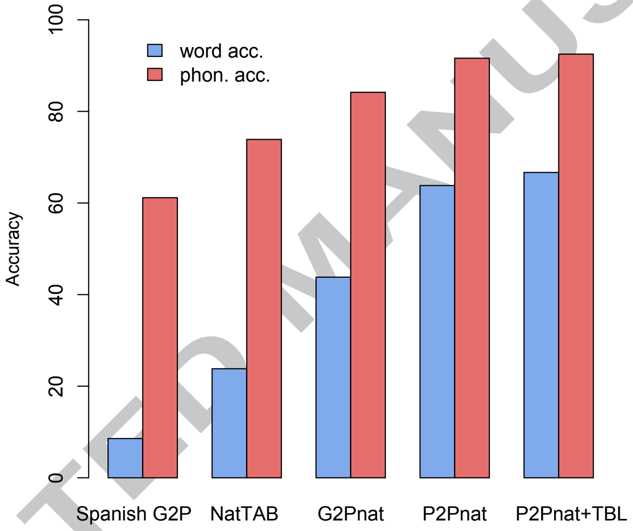


Figure 4: Word and phoneme accuracy obtained with: no nativization; hand-crafted nativization tables; grapheme-to-phoneme by analogy; phoneme-to-phoneme by analogy alone and combined with transformation-based-learning.

7.5. Error analysis

In terms of specific tasks such as nativization, an objective evaluation is insufficient to determine the validity of the results. Test results obtained with PbA using G2Pnat and P2Pnat were compared and exhaustively evaluated by the authors. Three types of errors were suggested. The term *severe errors* referred to the cases when the word was either unrecognizable and/or could be confused with another one. *Medium errors* referred to vowel confusion cases e.g. (a/e) (e/i), or (o/a). Vowel insertions and deletions together with similar consonant confusions (k/g, t/d, etc.) that did not drastically affect the intelligibility of the words were considered to be *light errors*. The results obtained using G2Pnat on CommonSet contained 22 severe errors affecting the intelligibility, while for the same test corpus using P2Pnat, only 10 severe errors were found. An example of a severe error is the pronunciation of the word *agency* nativized to [a ɣ e n s a j] or *general* to [ð j n e ɾ a l]. We considered the following nativization error for the word *agency*: [e j ʧu n s i] to be a medium error. An example of a light error would be the word *beautiful* nativized to

$[\mathrm{b\ j\ u}\ ð\ \mathrm{i\ f\ u\ l}]$. Our experiments were performed using isolated words, making no pronunciation adjustments at word boundaries at this point.

It was also interesting to compare the errors obtained by more sophisticated nativization methods such as pronunciation by analogy with those obtained by the Spanish G2P converter. The most common severe errors obtained by the Spanish G2P converter on CommonSet that rendered words completely unintelligible were the following: in words that contained a combination of *g* with either *e* or *i*, e.g., *girl* and *give*, the first phoneme $/\mathrm{g}/$ turned into $/\mathrm{x}/$, and the final silent *e* in *give* was transcribed as $/\mathrm{e}/$. In words that start with an *h*, the sound $/\mathrm{x}/$ at the beginning was lost and *home* became $[\mathrm{o\ m\ e}]$. Words such as *cool* and *need* were transcribed as $[\mathrm{k\ o\ o\ l}]$ and $[\mathrm{n\ e\ e}\ ð]$, respectively.

8. Subjective evaluation

The last step to validate nativized pronunciations was to carry out a perceptual test with synthesized signals. The perceptual quality of synthesized speech can be only defined by a listener and therefore is a difficult issue.

Thirty-eight volunteers were asked to evaluate 20 utterances synthesized using 3 different nativization methods. The utterances were produced using a concatenative unit-selection synthesizer (Bonafonte et al., 2008). The system concatenates diphones selected from a 10 hour speech database recorded by a professional speaker in a recording studio (Bonafonte et al., 2006). Each of the 20 utterances contained 1 to 6 foreign words, excluding the articles and two-letter prepositions, grouped into maximum of 3 foreign word chunks. A few examples of the sentences offered to the listeners can be found below.

1. Los índices de *Wall Street* abren la sesión con ganancias. (The Wall Street index opens the session with gains).
2. *Microsoft* anunció hoy que sus beneficios cayeron un diez por ciento. (Microsoft announced today that its benefits dropped by ten percent).
3. *New York Stock Exchange* es el mayor mercado de valores del mundo. (New York Stock Exchange is the largest stock market in the world).
4. Su disco *Born to run* vendió quince millones de copias en Estados Unidos. (His album “Born to run” sold 15 millions of copies in the United States).

It is worth mentioning that the sentences varied in their difficulty and uncommon words at the beginning of the sentence could have been found less comprehensible due to the lack of preceding context. Anticipating this possible reaction we inserted a phrase opener that included the word “Frase” (sentence) followed up by its number in the list.

The listeners were given 20 sets of 3 randomly ordered utterances. In the group of 3, the possible choices represented 3 different nativization methods applied to the foreign words. These methods were: no nativization (Spanish G2P); our baseline system NatTAB (Section 7.1); and nativization by analogy P2Pnat (Section 7.3). For each group of 3 utterances the listeners were asked to choose the best and the worst of the 3. However, for cases when a listener could not clearly decide which utterance he/she liked or disliked the most, a “none” option was added.

Listeners who volunteered for the experiment had different backgrounds in speech synthesis as well as in English and Spanish. Thirty out of 38 listeners were native speakers of Spanish, 2 were fluent and 6 claimed to have good knowledge of the language. Only 19 out of 38 were fluent in English, while the remaining half indicated to possess good knowledge of the language. Among the participants, 9 were experts in speech synthesis, 4 were experts in other speech technologies, 8 were occasional users of synthesis and the rest claimed no experience with synthesized speech whatsoever.

Overall evaluation results are shown in Figure 5. The graph shows the average number of times each method was chosen as best or worst independently of the phrase difficulty. From Figure 5 it is easy to see that the Spanish G2P method was voted worst in almost 45% of the cases, while the analogy-based method was voted best with a percentage close to 50%. Finally, the nativization by analogy had the lowest incidence of worst votes in comparison with other methods. The nativization tables (NatTAB) method received a similar number of best and worst votes. The percentage of indecision in both cases oscillated around 10%. The results allow to draw the same conclusions than the objective test: the analogy-based method (P2Pnat) is much better than the table-based method (NatTAB), which in its turn is better than the original Spanish G2P. The results are statistically significant at the level $p = .05$ on the basis of a binomial test.

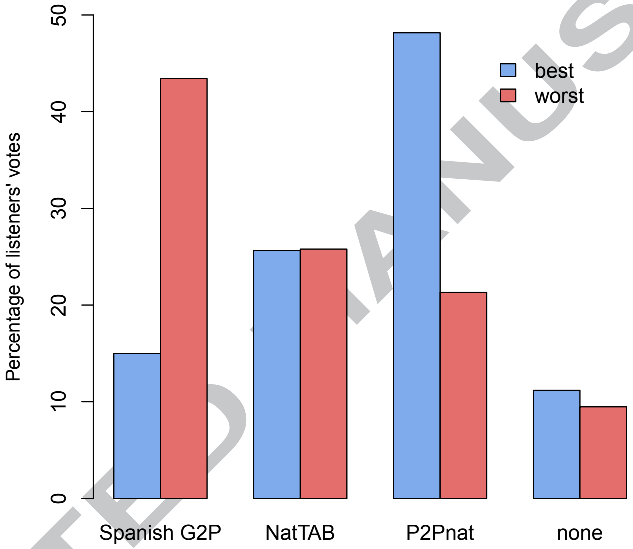


Figure 5: Perceptual evaluation of the TTS system using three different nativization methods.

The separate analysis of the results based on the listeners' background showed that their previous experience with speech synthesis for this particular task had more influence on the results that their knowledge of Spanish or English. Experts in speech synthesis, thus, showed a stronger preference for the analogy-based method and a stronger rejection of the Spanish G2P-based method. Although the goal of this work was to evaluate nativization from the viewpoint of native Spanish speakers, due to the great enthusiasm towards the test shown by the non-native group members, we decided to include and analyze their contribution. The pattern of the non-native Spanish speakers' reactions to the test presented a higher variability as it was more difficult for them to recognize subtle differences between pronunciations generated by different nativization methods. However, they strongly preferred the analogy-based method, choosing it in 53% of the cases. This may be related to the fact that the analogy-based nativization sounded more correct from the point of view of English phonetics. For the rest of the subgroups the general tendency was similar to that shown for the overall results in Figure 5.

Curiously, nativization by analogy was voted as the worst method for inclusions such as *hangover* and *born to run* and the highly assimilated word *Microsoft*. In the first case, the main difference lies in the nativized pronunciation of the English phoneme /r/. The nativization by analogy method disregarded the position-dependent pronunciation of the grapheme *r*, and all English /r/ were converted to Spanish intervocalic /ɾ/ in the training corpus and, therefore, in the resulting synthesis. Consequently, the unit selection synthesizer could not find any /ɾ/ at the beginning of the word or before a consonant using instead the demi-phonemes that sounded very close to the Spanish dental approximant [ð]. In the case of the word *Microsoft*, the most common pronunciations in Spain are [m i k ɾ o ˈs o f t] or even [m i k ɾ o ˈs o f]. Even [ˈm a j k ɾ o s o f t], the alternative offered by the table-based method was found less pleasant and the pronunciation predicted by the analogy-based method [ˈm a j k ɾ o s a f t] was considered too foreign sounding. The assimilation-influenced accent displacement was not accounted for in this work.

At this point, we can conclude that the best received method was the phoneme analogy-based nativization and the worst was the Spanish G2P converter (no nativization). However, the frequency of word usage introduced variability and nuances.

9. Conclusions and future work

In this paper, we propose to use pronunciation by analogy and transformation-based learning for nativization of English words in Spanish. The nativization results obtained using analogy only in the orthographic domain were rather poor (43.8% words correct for the CommonSet) due to deep orthography of the English language. Still we believe that the G2Pnat results are better than G2P results could have been for the same minimalistic corpus of only 1000 words for languages of deep orthography. It is worth mentioning that even in the case of G2Pnat, the results show very significant improvements in comparison to those obtained by direct phoneme-to-phoneme table-based mapping (NatTAB). The method based on analogy in the phonemic domain, P2Pnat, gave an improvement of 20 percentage points with respect to the orthographic analogy, thus showing the tight connection between the pronunciation in the source language and the nativized one. TBL algorithm applied in combination with other methods produced good results. NatTAB+TBL performed slightly better than P2Pnat. The best results were achieved using P2Pnat enhanced by the TBL algorithm (66.7%) that allowed to incorporate additional information about the orthography in the source language. In the perceptual test, P2Pnat also showed a significant improvement over the nativization tables.

Nativized pronunciations are more tolerant to vowel and consonant substitutions, previously referred to as light errors. There is no gold standard for nativization, and some exceptions that occur in highly assimilated pronunciations increase the difficulty of the problem. However, these exceptions are created by humans and obey the analogy both in orthographic and phonetic forms. Simple mapping rules were proven to be insufficient for the task. In the future, it would be interesting to tackle the reverse problem (Soonklang et al., 2008) because Spanish inclusions in English utterances would result in quite unintelligible pronunciations simply by applying an English G2P, being that Spanish is a vocalic language with transparent letter-to-sound rules.

References

Bear, D., Templeton, S., Helman, L., Baren, T., 2003. English learners: Reaching the highest level of English literacy. Ch. Orthographic development and learning to read in different languages, pp. 71–95.

Bellegarda, J., 2005. Unsupervised, language-independent grapheme-to-phoneme conversion by latent analogy. Speech Communication 46 (2), 140–152.

Bisani, M., Ney, H., 2008. Joint-sequence models for grapheme-to-phoneme conversion. Speech Communication 50 (5), 434–451.

Black, A., Lenzo, K., May 2004. Multilingual text-to-speech synthesis. In: Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). Vol. 3. pp. 761–764.

Black, A., Lenzo, K., Pagel, V., Nov. 1998. Issues in building general letter to sound rules. In: SSW3. 3rd ESCA Workshop on Speech Synthesis, pp. 77-80, Jenolan Caves, Australia, pp. 77–80.

Bonafonte, A., Höge, H., Kiss, I., Moreno, A., Ziegenhain, U., van den Heuvel, H., Hain, H., Wang, X., Garcia, M., May 2006. TC-STAR: Specifications of language resources and evaluation for speech synthesis. In: Proc. of the International Conference on Language Resources and Evaluation (LREC). pp. 311–314.

Bonafonte, A., Moreno, A., Adell, J., Agüero, P., Banos, E., Erro, D., Esquerra, I., Pérez, J., Polyakova, T., Sep. 2008. The UPC TTS system description for the 2008 blizzard challenge.

Brill, E., 1995. Transformation-based error-driven learning and natural language processing: a case study in part-of-speech tagging. Computational Linguistics 21 (4), 543–565.

Canellada, M., Madsen, J., 1987. Pronunciación del español. Editorial Castalia.

Conde, X., 2001. Introducción la Fonética y fonología del Español. Ianua. Romance Philology Journal Sup04.

Damper, R., Marchand, Y., Marseters, J., Bazin, A., Jun. 2004. Aligning Letters and Phonemes for Speech Synthesis. In: Proc. of the 5th ISCA Speech Synthesis Workshop (SSW5). Pittsburgh, USA, pp. 209–214.

Dedina, M., Nusbaum, H., 1991. PRONOUNCE: a program for pronunciation by analogy. Computer speech & language 5, 55–64.

Education First, 2011. English Proficiency Index.
URL http://www.ef-uk.co.uk/sitecore/__/ /media/efcom/epi/pdf/EF-EPI-2011.pdf

Font Llitjos, A., Black, A., Sep. 2001. Knowledge of Language Origin Improves Pronunciation Accuracy of Proper Names. In: Proc. the of European Conference on Speech Communication and Technology. Aarborg, Denmark, pp. 1919–1922.

Fox, R., Flege, J., Munro, M., 1995. The perception of English and Spanish vowels by native English and Spanish listeners: a multidimensional scaling analysis. The Journal of the Acoustical Society of America 97 (4), 2540–2551.

Glushko, R., 1981. Interactive processes in reading. Lawrence Erlbaum, Ch. Principles for pronouncing print: The psychology of phonography, pp. 61–84.

Hammond, R., 2001. The Sounds of Spanish: Analysis and Application (with Special Reference to American English). Cascadilla Press.

Handbook, I., 1999. Handbook of the International Phonetic Association: a guide to the use of the International Phonetic Alphabet.

Hartikainen, E., Maltese, G., Moreno, A., Shammass, S., Ziegenhain, U., Sep. 2003. Large Lexica for Speech-to-Speech Translation: From Specification to Creation. In: Proc. the of European Conference on Speech Communication and Technology. Geneve, Switzerland, pp. 1529–1532.

Ladefoged, P., 2003. Vowels and consonants. Blackwell Publishing.

Lindström, A., 2004. English and other foreign linguistic elements in spoken Swedish: studies of productive processes and their modelling using finite-state tools. Ph.D. thesis, University Linköping, Linköping, Sweden.

Llorente, J., Díaz Salgado, L., 2004. Libro de estilo de Canal Sur TV y Canal 2 Andalucía. Radiotelevisión de Andalucía.

Marchand, Y., Damper, R., 2000. A multistrategy approach to improving pronunciation by analogy. Computational Linguistics 26 (2), 195–219.

Pfister, B., Romsdorfer, H., Sep. 2003. Mixed-Lingual Text Analysis for Polyglot TTS Synthesis. In: Proc. the of European Conference on Speech Communication and Technology. Geneva, Switzerland, pp. 2037–2040.

Polyákova, T., Bonafonte, A., Sep. 2006. Learning from errors in grapheme-to-phoneme conversion. In: Proc. of the International Conference on Spoken Language Processing (ICSLP). Pittsburgh, USA, pp. 2442–2445.

Polyákova, T., Bonafonte, A., Nov. 2008a. Transcripción fonética en un entorno plurilingüe. In: Actas de las V Jornadas en Tecnologías del Habla. Bilbao, Spain, pp. 207–210.

Polyákova, T., Bonafonte, A., Nov. 2008b. Further improvements to pronunciation by analogy. In: Actas de las V Jornadas en Tecnologías del Habla. Bilbao, Spain, pp. 149–152.

Polyákova, T., Bonafonte, A., Apr. 2009. New strategies for pronunciation by analogy. In: Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). Taipei, Taiwan, pp. 4261–4264.

Raynolds, L., Uhry, J., 2009. The invented spellings of non-Spanish phonemes by Spanish–English bilingual and English monolingual kindergarteners. Reading and Writing, 1–19.

Real Academia Española, 1992. Diccionario de la lengua Española. Espasa Calpe.

Sejnowski, T., Rosenberg, C., 1993. Nettalk corpus.
URL `<ftp://svrftp. eng.cam.ac.uk/pub/comp.speech/dictionaries>`

Soonklang, T., Damper, R., Marchand, Y., 2008. Multilingual pronunciation by analogy. Natural Language Engineering 14 (04), 527–546.

Swan, M., Smith, B., 2001. Learner English: A teacher's guide to interference and other problems. Cambridge University Press.

Taylor, P., Sep. 2005. Hidden Markov Models for Grapheme to Phoneme Conversion. In: Proc. the of European Conference on Speech Communication and Technology. Lisboa, Portugal, pp. 1973–1976.

Trancoso, I., Nov. 1995. 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, pp. 193–209.

Trancoso, I., Viana, C., Mascarenhas, I., Teixeira, C., Sep. 1999. On deriving rules for nativised pronunciation in navigation queries. In: Proc. the of European Conference on Speech Communication and Technology. Budapest, Hungary, pp. 195–198.

van den Bosch, A., Daelemans, W., Apr. 1993. Data-oriented methods for grapheme-to-phoneme conversion. In: Proceedings of the sixth conference on European chapter of the Association for Computational Linguistics. Utrecht, The Netherlands, pp. 45–53.

Van den Heuvel, H., Réveil, B., Martens, J., Sep. 2009. Pronunciation-based ASR for names. In: Proc. of Interspeech. Brighton, UK, pp. 2991–2994.

Wells, J., 1982. Accents of English: an Introduction. Cambridge Univ Pr.

Yavas, M., 2006. Applied English Phonology. Blackwell Publishing.

Zheng, M., Shi, Q., Zhang, W., Cai, L., 2005. Grapheme-to-phoneme conversion based on a fast TBL algorithm in mandarin TTS systems. Fuzzy Systems and Knowledge Discovery, 600–609.

Abstract

In the modern world, speech technologies must be flexible and adaptable to any framework. Mass media globalization introduces multilingualism as a challenge for the most popular speech applications such as text-to-speech synthesis and automatic speech recognition. Mixed-language texts vary in their nature and when processed, some essential characteristics must be considered.

In Spain and other Spanish-speaking countries, the use of Anglicisms and other words of foreign origin is constantly growing. A particularity of peninsular Spanish is that there is a tendency to nativize the pronunciation of non-Spanish words so that they fit properly into Spanish phonetic patterns. In our previous work, we proposed to use hand-crafted nativization tables that were capable of nativizing correctly 24% of words from the test data.

In this work, our goal was to approach the nativization challenge by data-driven methods, because they are transferable to other languages and do not drop in performance in comparison with explicit rules manually written by experts. Training and test corpora for nativization consisted of 1000 and 100 words respectively and were crafted manually. Different specifications of nativization by analogy and learning from errors focused on finding the best nativized pronunciation of foreign words. The best obtained objective nativization results showed an improvement from 24% to 64% in word accuracy in comparison to our previous work. Furthermore, a subjective evaluation of the synthesized speech allowed for the conclusion that nativization by analogy is clearly the preferred method among listeners of different backgrounds when comparing to previously proposed methods. These results were quite encouraging and proved that even a small training corpus is sufficient for achieving significant improvements in naturalness for English inclusions of variable length in Spanish utterances.

Highlights

>We propose a novel approach for foreign word nativization in TTS. > Nativization is inferred by analogy in the phonemic and orthographic domains.>Objective results showed significant improvements in comparison to our previous work.> Perceptual evaluation showed that the proposed method is preferred by the listeners.