Carranza, M., Cucchiarini, C., Llisterri, J., Machuca, M. J., & Ríos, A. (2014). A corpus-based study of Spanish L2 mispronunciations by Japanese speakers. In *Edulearn14 Proceedings. 6th International Conference on Education and New Learning Technologies*. July 7th-9th, 2014 - Barcelona, Spain (pp. 3696–3705). IATED Academy.

http://liceu.uab.cat/~joaquim/publicacions/Carranza_et_al_14_Corpus_Spanish_L2.pdf

A CORPUS-BASED STUDY OF SPANISH L2 MISPRONUNCIATIONS BY JAPANESE SPEAKERS

M. Carranza[1], C. Cucchiarini[2], J. Llisterri[1,a], M. J. Machuca[1,a], A. Ríos[1,a]

[1] Universitat Autònoma de Barcelona (SPAIN)
[2] Radboud Universiteit Nijmegen (THE NETHERLANDS)

Abstract

In a companion paper (Carranza et al.) submitted to this conference we discuss the importance of collecting specific L1-L2 speech corpora for the sake of developing effective Computer Assisted Pronunciation Training (CAPT) programs. In this paper we examine this point more deeply by reporting on a study that was aimed at compiling and analysing such a corpus to draw up an inventory of recurrent pronunciation errors to be addressed in a CAPT application that makes use of Automatic Speech Recognition (ASR). In particular we discuss some of the results obtained in the analyses of this corpus and some of the methodological issues we had to deal with.

The corpus features 8.9 hours of spontaneous, semi-spontaneous and read speech recorded from 20 Japanese students of Spanish L2. The speech data was segmented and transcribed at the orthographic, canonical-phonemic and narrow-phonetic level using Praat software [1]. We adopted the SAMPA phonemic inventory for the phonemic transcription adapted to Spanish [2] and added 11 new symbols and 7 diacritics taken from X-SAMPA [3] for the narrow-phonetic transcription. Non linguistic phenomena and incidents were also annotated with XML tags in independent tiers. Standards for transcribing and annotating non-native spontaneous speech ([4], [5]), as well as the error encoding system used in the project will be addressed. Up to 13410 errors were segmented, aligned with the canonical-phonemic tier and the narrow-phonetic tier, and annotated following an encoding system that specifies the type of error (substitutions, insertion and deletion), the affected phone and the preceding and following phonemic contexts where the error occurred. We then carried out additional analyses to check the accuracy of the transcriptions by asking two other annotators to transcribe a subset of the speech material. We calculated intertranscriber agreement coefficients. The data was automatically recovered by Praat scripts and statistically analyzed with R. The resulting frequency ratios obtained for the most frequent errors and the most frequent contexts of appearance were statistically tested to determine their significance values.

We report on the analyses of the combined annotations and draw up an inventory of errors that should be addressed in the training. We then consider how ASR can be employed to properly detect these errors. Furthermore, we suggest possible exercises that may be included in the training to improve the errors identified.

Keywords: Non-native speech corpus, Japanese L1 - Spanish L2, automatic speech recognition.

1 INTRODUCTION

In this paper we report on further research we carried out on a corpus of non-native speech of Spanish L2 by Japanese speakers that has been previously introduced in [6] and [7]. This corpus was specifically compiled for the development of CAPT applications enhanced with ASR technology, and addresses a specific language pair. The compilation of this corpus responds also to the objective of obtaining empirical data about the most frequent mispronunciations made by Japanese students of Spanish L2. The aim is to determine whether the differences between the phonological systems of the two languages, addressed by some previous contrastive studies [8], correspond to actual difficulties in the pronunciation.

With these two purposes in mind, we designed this corpus trying to attain a high degree of representativeness concerning the various factors that can intervene in L2 acquisition [9]. Time of exposure to the L2 is one of the principal factors in the acquisition of a foreign language; therefore, we adopted a longitudinal approach for the corpus compilation. Speaking style is, likewise, another factor that can interfere in the pronunciation of errors [10]. We therefore decided to include

[a] Work supported by grant 2014SGR27 of the Catalan Government

(semi)spontaneous speech, conversational and read-out speech with the aim of collecting learners' utterances in the most real situations that can be found in the language classroom. The participants in our corpus were selected so as to represent a range of different levels of oral proficiency; the participants were grouped according to their scores in oral tests, and the number of male participants and female participants was also balanced.

Other learner spoken corpora featuring Spanish as the target language have been already compiled and most of them are fully accessible on-line. The Spanish Learner Language Oral Corpora (SPLLOC) compiled by the universities of Southampton, York and Newcastle [11] contains Spanish L2 utterances by English speakers. The corpus is transcribed at the orthographical and phonemic level using an orthographical convention based on the CHAT standard [12]. The corpus Language Usage in Spanish LUIS-L2/LE is aimed at the study of Spanish L2 by learners of different L1s backgrounds, all utterances were transcribed using the CHAT convention. This corpus was compiled by the Pompeu i Fabra University and was published in book format [13]. The Spanish Learner Oral Corpus [14], belongs to the C-ORAL-ROM project and was compiled at the Autonomous University of Madrid with the aim of obtaining insight into the most common errors in Spanish L2 in all levels and features interviews with learners of Spanish L2 of various L1 backgrounds. Transcription and annotation follow the CHAT convention for the orthographical transcription and the annotation of vocalizations and extra-linguistic phenomena. The annotations of errors take into account all the linguistic levels of analysis: morphological, syntactical, lexical and phonological; nevertheless, only pronunciation errors (segmental and suprasegmental) were transcribed phonemically. As far as we know, none of these corpora offers a full phonemic or phonetic transcription of the utterances, a requisite for adapting the database as a training corpus that can be used in the field of speech technologies.

This paper is organized as follows. Section 2 contains a description of the corpus, the participants and the variables considered in the design phase, the levels of representation with the standards used for the annotations, and the encoding of the pronunciation errors. We also considered the need to validate the corpus transcription by calculating the agreement score between our transcription and the transcriptions of two extra annotators with no previous expertise in Japanese. Section 3 presents the results for the error analysis, which were obtained semi-automatically by means of Praat scripts and statistically analyzed with R. The most frequent pronunciation errors are grouped into lists according to their type: substitution, insertion and deletion. We then indicate the most salient and persistent errors that might hamper communication and that should be addressed by a CAPT system in the first place. This is followed by a discussion of our findings in section 4.

2 METHODOLOGY

2.1 Participants

The speech material used for this research was recorded from 10 male and 10 female Japanese students of Spanish L2 at the Spanish Department of the Tokyo University of Foreign Studies. Participants were selected by taking into account their dialectal origins (Kanto dialect) and discarding students that had received any previous academic contact with Spanish. The corpus contains the recordings of the oral tests of the 20 students throughout the two first academic years of Spanish study (from April 2010 to March 2012), which corresponds to the A1 and A2 levels in the Common European Framework Reference for Language Learning (CEFR) [15]. Oral proficiency was also taken into account by computing the mean score for the total oral tests of all students. Afterwards, three groups were established: low, intermediate and high proficiency levels, and the participants were distributed into each group according to their mean score (low: N= 6, high: N= 6, intermediate: N=8).

2.2 Database description

The corpus contains 8.9h of non-native speech, divided into 91 minutes of semi-spontaneous speech, 214 minutes of spontaneous speech, 9 minutes of read speech and 201 minutes of conversational speech. In total, spontaneous speech represents more than 80% of the recordings. Oral tests took place every six months –four academic semesters–, and consisted of different types of tasks which involved all speaking styles. Semi-spontaneous speech was obtained from oral presentations prepared beforehand (in 1st and 2nd semesters) and spontaneous speech was obtained from conversations between the student and the examiner, and extemporaneous role-plays with no previous preparation (in 3rd and 4th semesters). All the recordings were made with portable recorders and were segmented into individual audio files. The different recording conditions of the tests resulted

in different levels of quality of the audio files. Then, the audio files were converted into WAV format and labelled by means of a code in which the participant, the task type and the period of learning (semester) was encoded. This encoding system was designed considering the automatic computation of error ratios according to proficiency level, learning stage and speech style.

2.3 Transcription levels and error annotation

Non-native spontaneous speech is intrinsically more difficult to transcribe than read-out speech due to its high degree of variability, phonetic and syntactic interference of the L1 and the constant presence of vocalizations and other extra linguistic phenomena. This limitation requires that the transcribers should follow a set of rules (protocol) to interpret and represent speech trying to maintain consistency and agreement at all levels of transcription, because factors that can increase the degree of variability such as non-nativeness and spontaneity contributes to a high degree of variability in the transcription process [16]. Moreover, transcribing speech contains a certain degree of subjectivity because it is based on individual perception and implies other sources of variation, such as familiarity of the transcriber with the L1 of the student, training and experience received, auditory sensitivity, quality of the speech signal, and other linguistic factors regarding word intelligibility or length of the utterance. Manual transcription is always prone to mistakes, but it is still the only way of obtaining a narrow phonetic transcription of speech, since state-of-the art ASR technology is not yet capable of such degree of detail [17].

Training corpora for ASR technology need to be at least phonemically transcribed, and non linguistic phenomena should be properly annotated so that disfluencies such as filled pauses, hesitations, laughs, creaky voice and breathy voice, would not interfere in the generation of the acoustic models. Moreover, transcription of the actual speech must be compared to a reference model –e.g. a canonical transcription– that represents the correct pronunciation of the utterance as pronounced by native speakers. This will allow the system to automatically detect discrepancies between the two levels and generate rules for pronunciation variants and acoustic models for non-native phones.

This corpus was transcribed and annotated using Praat. The orthographical, canonical phonemic and narrow phonetic transcriptions were annotated separately and the resulting tiers were manually time-aligned to the speech signal on word and phone levels. Vocalizations and non linguistic phenomena were also marked in two independent tiers. Finally, the encoding of pronunciation errors took place in a different tier and every error label was aligned with the linguistic transcriptions. Other levels of representation such as canonical-allophonic or fully acoustical were not considered, due to the possibility of automatically generating the former from the canonical phonemic tier using phonological rules, and the difficulty of performing a detailed acoustic analysis on data with limited acoustic quality, on the latter. A more detailed description of the phone units and XML tags used for the transcription can be found in [7].

2.3.1 Orthographic transcription

In the orthographic tier every word is transcribed in its standardized form, but no punctuation marks are used due to the difficulty of establishing syntactic boundaries in spontaneous speech. Non-native spontaneous speech is characterized by a high number of filled pauses or hesitations, repetitions and truncations. The cases of fragmented speech are problematic for the orthographical transcription, especially truncations when the word is never completed so that the transcriber must guess the actual word that the informant wanted to say. XML tags were used for labeling these phenomena ([4], [5]), as well as unclear cases, missing words, foreign words (in Japanese and English) and erroneous words –like regularized irregular verbs–. Hesitations and interjections were also transcribed at this level according to their standardized form in dictionaries. Only the speech from the student was transcribed. The utterances of the examiner were not considered, except when they overlapped with the student's speech; in these cases, the overlapping speech is tagged with an XML label in the incident tier (see 2.3.4.).

XML tagging is one standardized protocol for transcribing speech material and, as tags and its content can be automatically removed, is employed for obtaining a clean orthographic transcription –without disfluencies– from a verbatim transcription, in other words, an orthographic transcription that represents all sounds produced by the speaker, verbal and non-verbal [4]. The reason for adopting this methodology is that our corpus is mainly composed of spontaneous speech, which is highly affected by verbal disfluencies.

2.3.2 Canonical-phonemic transcription

The canonical-phonemic tier shows the phonological transcription of each word as pronounced in isolation, consequently no coarticulation phenomena in word boundaries or inside words are considered at this level. Northern Castilian Spanish ([18], [19]) was adopted as the standard reference for the transcription since the Japanese students had been taught mainly in this variety. Consequently, at this level the phonemic opposition /s/–/θ/ is preserved, but not the opposition /ɟ/–/ʎ/ which is neutralised in favor of /ɟ/ [20]. Since a computer-readable alphabet was needed for the automatic processing of the corpus and its adaptation as a training corpus for an ASR system, an adaptation of SAMPA to Spanish [2] was chosen for the inventory of phonological units.

2.3.3 Narrow-phonetic transcription

The narrow phonetic level represents the actual pronunciation of the speaker in the most accurate way. In order to decrease the degree of transcriber's subjectivity, our transcription was based mainly upon visual examination of the spectrogram and oscillogram and PRAAT calculations of acoustic features, rather than on auditory perception alone. Although this method might influence transcription, it was regarded as a more reliable procedure based on sound criteria than mere auditory perceptual judgement. We avoided perceptual judgment except in cases where the decision could not be taken from the methods stated and could be reached upon auditory perception. The process of transcribing spontaneous speech implies dealing with disfluencies due to hipoarticulation and extemporaneity, such as creaky voice, breathy voice, noises, laughs, and other phenomena that should be conveniently labeled in order to appropriately specify which segments are not optimal for computing the acoustic models when training the ASR recognizer. Coarticulation phenomena (nasalization and changes in place or articulation) are considered here, as well as the Spanish allophonic variants (β̞], [ð̞], [ɣ̞], [ɟ̞], [ŋ], [z], [dʒ]). We added a new inventory of symbols and diacritics taken form X-SAMPA [3] to the previous SAMPA inventory used in the canonical phonological transcription to account for these phenomena. Further symbols were also added to represent the Spanish pronunciation of Japanese speakers. In total, the inventory of phone units for the narrow phonetic transcription contains 51 symbols and 8 diacritics. Ambiguous or intermediate realizations, in which the realization of the target sound shows features of both the L1 and the L2, were specially problematic to transcribe. To solve these cases, a methodology based on hierarchical criteria was defined by manually creating a decision tree for each type of intermediate realization [6]. The use of decision trees might contribute to avoid biased individual judgements, especially when more than one transcriber is involved.

2.3.4 Vocalizations and non-linguistic phenomena

Vocalized, or semi-lexical elements, such as laughter, hesitations and interjections were labelled in a separate tier. The reason for separating vocalizations from the rest of speech was that acoustic realizations of these elements resemble linguistic sounds –hesitations are usually realized as vowels or nasal sounds and interjections as short vowels– and can interfere in the acoustic modeling when training the recognizer. Vocalizations were marked by using XML labels to explicitly indicate that these segments should not be employed in the ASR training phase. Non linguistic (or non-lexical) phenomena were marked in the incident tier. We considered laugh, breathing, external noise and overlapping speech of the interviewer as non-linguistic. As we stated above, the rationale behind tagging this phenomena with XML tags in separate tiers is to incorporate the option of automatically deriving a clean orthographical transcription from a verbatim transcription by eliminating all segments labelled with a non-verbal tag.

2.3.5 Error annotation

Determining what should be considered as a pronunciation error in non-native speech is a difficult task, and moreover establishing a hierarchy of importance [21]. Non-native speech presents a high degree of variability between and within speakers, due to the fact that in non-native speech the L2 phonological categories are not fixed and the L1 phonological categories can also interfere. The issue of foreign speech assessment is always prone to subjectivity; language teachers tend to be more strict in judging their students pronunciations skills than native speakers [22]. Besides, spontaneous speech is characterized by a high degree of variability even in native speech, so that it is common to find non-canonical articulations by native speakers originated by hipoarticulation in spontaneous style [23].

ASR systems are usually trained with good quality native speech, often produced by professional voice actors; but, in order to automatically detect pronunciation errors made by non-natives, the annotations should take into consideration every utterance which deviates from a canonical

pronunciation, including articulations triggered out by the spontaneity of the discourse. For this reason, we labelled as an error every realization which we considered deviant from the standard expected realization of a prototypical native speaker. Notice that, since the main objective of the annotations is to mark the deviant realizations to avoid interfering in the generation of the non-native acoustic models, this definition of the pronunciation error is far more strict than what it has been generally defined as error in L2 acquisition research.

The error encoding system developed for this research is an adaptation of the encoding used in the Corpus Interphonologie du Français Contemporaine (IPFC) [24], and combines a set of alphanumeric characters to encode information about the type of error (substitution, insertion, deletion), the affected phone and the phonologic context –previous and following phones–. Each error was encoded following this procedure. Then, all error codes were automatically processed and lists of most frequent errors with their context of appearance were obtained in the statistical analysis. Up to 13410 errors were encoded in the corpus.

2.4 Transcription verification

In order to assess the accuracy of our transcription, we asked two external annotators to verify the narrow phonetic tier. The annotators were Spanish native speakers and expert phoneticians with no expertise in Japanese phonetics. Our goal was to determine the consistency of our transcription and finding possible discrepancies between annotators with expertise in the L1 of the student and annotators without this knowledge; considering the high degree of subjectivity involved in the process of manual transcription of non-native speech.

Two different subsets of the corpus, each one containing an approximate 10% of the total recordings (45 min. approximately each), were given to annotators including the textgrids with the annotations, the phone inventories and the instructions for transcription. The correspondent International Phonetic Alphabet (IPA) symbols for the SAMPA and X-SAMPA inventories of phone units used in the corpus were also provided. The two corpus selections included data from different speaking styles and oral proficiency and were not randomly obtained, we selected the files according to the speech quality and in an attempt that both subsets were representative of the whole corpus. We provided the two external annotators with the audio files and the full annotated transcriptions, and asked them to check the narrow-phonetic tier by following the instructions for transcription and make all necessary changes in a new tier, duplicated from the narrow phonetic tier. They followed the same procedure for the annotation of the corpus, verified the narrow phonetic transcription and corrected it if needed.

The revised transcriptions were then automatically compared with the original transcription, the agreement score was calculated and confusion matrices were generated to determine discrepancies.

3 RESULTS

In this section we present first the results of the automatic processing of error annotations, grouped by substitution, insertion and deletion errors, collected with their context of appearance and sorted on frequency. Then, we discuss the discrepancies found between the revised narrow phonetic transcription by the two external annotators and our original transcription, in an attempt to describe the specific segments that caused the lowest degree of agreement between annotators.

3.1 Type and frequency of errors and influence of the context

The most frequent mispronunciations made by the students are categorized in three groups: substitution, insertion and deletion errors. This taxonomy was adopted from the standard evaluation of the output of ASR systems [25] and serves to obtain the word error rate (WER), which is an indicator of the system's recognition performance and robustness.

3.1.1 Substitution errors

Phone substitutions were the most frequent errors made by the students (47%). In the error encoding system we encoded the substituted phone, and the previous and following phones (the phonological context of appearance). Notice that silence (#) is considered also as a context of appearance if the substitution takes place at the beginning or at the ending of an utterance or before/after a pause. Table 1 shows the first ten most frequent substitutions, which represents 74% of the total substitution errors, along with their most frequent preceding and following phones. Whenever the phone context

can be grouped by a more generic class –such as vocalic–, the class is preferred to the individual phones.

The most frequent vocalic substitutions tend to be related to hipoarticulation (creaky voice) and devoicing in voiceless contexts. A tendency to articulate the rounded back high vowel [u] of Spanish as an unrounded central-back vowel [ɯ] of Japanese is also remarkable. Consonant substitutions affect the approximant realizations ([β̞] [ð̞] [ɣ̞]) of Spanish voiced stops /b/, /d/, /g/, which are pronounced as full stops ([b] [d] [g]) or devoiced ([p] [t] [k]). The substitution of [θ] by [s] is very frequent, but it should not be considered a characteristic of non-nativeness, since this substitution is commonly adopted in southern Spanish and American varieties too. Finally, one of the most noticeable substitution error was the mispronunciation of [l] as [ɾ] (but not in the opposite way) in vocalic contexts mainly and before/after silence in some cases. This error can be explained by the different phonological status of these two sounds in Spanish and Japanese, since both sounds function as allophonic variants of one unique rhotic phoneme /ɾ/ in Japanese, whose flap realization [ɾ] is much more common than the lateral realization [l] except in the context between vowels and after a pause [26], which corresponds roughly to the contexts of mispronounced [l] detected in the corpus.

Table 1 : Most frequent substitution errors sorted by frequency (74% of total substitutions)

N	Target	Realized (f)		Previous context (f)	Following context (f)
921	[ð̞]	[d]	.82	vocalic (.79)	vocalic (.89)
		[t]	.16	vocalic (.49) [s] (.40)	vocalic (.94)
815	[l]	[ɾ]	.66	vocalic (.68) # (.08)	vocalic (.76) # (.08)
806	[β̞]	[b]	.89	vocalic (.84)	vocalic (.87)
		[p]	.09	vocalic (.68) [s] (.16)	vocalic (.88)
742	[a]	[a̠]	.57	voiceless cons. (.36) # (.10)	# (.57) [s] (.13)
		[a̟]	.19	voiceless cons. (.50) # (.13)	# (.50) voiceless cons. (.30)
634	[θ]	[s]	.90	vocalic (.74) # (.09)	vocalic (.87) # (.09)
		[ʃ]	.04	[i] (.46)	[ɯ] (.63) [i] (.37)
590	[e]	[e̠]	.57	[t] (.17) # (.13) [ɾ] (.11) [d] (.08)	# (.51) [s] (.19)
		[e̟]	.24	# (.29) [t] (.20) [s] (.15)	[s] (.47) # (.28)
494	[u]	[ɯ]	.85	[t] (.21) [ɣ̞] (.20) [m] (.17) # (.09)	[s] (.30) [tʃ] (.17) [n] (.12) [ð̞] (.10)
		[u̟]	.04	[k] (.26) [ð̞] (.13) [p] (.13)	[t] (.26) [p] (.21)
		[u̠]	.03	# (.29) [m] (.18)	[n] (.35) [s] (.18)
480	[o]	[o̠]	.49	[t] (.16) [k] (.14) [ð̞] (.11) [ɾ] (.08)	# (.68) [s] (.11)
		[o̟]	.25	[k] (.30) [t] (.11) [ɲ] (.09) [θ] (.08)	# (.49) [s] (.20)
		[u]	.06	[p] (.25) [n] (.16) [ɾ] (.16)	[n] (.19) [s] (.19) [ɾ] (.16) [k] (.16)
327	[ɣ̞]	[g]	.78	vocalic (.80) [l] (.07) [s] (.04)	vocalic (.89) [ɾ] (.10)
		[k]	.20	vocalic (.53) [s] (.40)	vocalic (.51) [ɾ] (.46)
286	[d]	[t]	.90	# (.90) [n] (.05)	[e] (.54) [o] (.21) [i] (.10)
		[ð̞]	.06	[n] (.89) [l] (.10)	vocalic (1)

3.1.2 Insertion errors

The error analysis showed two types of insertions: first, epenthetic vowels that appear in phone sequences that are illegal in the L1 (Japanese) an which are inserted to reassign the problematic consonant(s) to a legal syllabic position in the L1 by changing the syllable structure. For instance, [ɯ] is mainly inserted after final [ɾ] or [s], two consonants that are not allowed in coda position according to Japanese combinatory rules, thus a vocalic element [ɯ] is inserted to facilitate the articulation by assigning the consonant to the onset position. However, the most frequent insertion cases can be explained by the addition of articulatory features; in other words, sounds that are pronounced with a new articulatory feature which is not expected in native speech. According to the results (see Table 2), acoustic feature insertions are more common than full phone insertions. The insertion of new acoustic

features could be the result of hipoarticulation, since creacky voice insertion was the most frequent case, but it can be interpreted conversely as a case of hyperarticulation if we consider that an aspirated or palatalized consonant is more tense than the unaspirated or non-palatalized counterparts. Aspiration tends to appear after voiceless stops and before central and back vowels ([a], [o], [u]). Palatalization, on the other hand, is mostly found with the same preceding context but before front vowels ([i], [e]).

Vowel epenthesis with [u], [ɯ] and [ə] are mostly found in word-final position (after silence) and before alveolar sounds [s] and [ɾ], although [ə] has also been detected in the opposite context: in word-beginning position before [ɾ] and [d] and between the sequence [s]+[ɾ]. The insertion of the vowel [o] appears between consonant clusters formed by dental stop+flap ([tɾ], [ðɾ]) and before the sequence [ɾs]. Insertions of [e] have also been detected in similar contexts. Finally, only one type of consonant insertion, a voiced glottal stop [ʔ], has been found in word-beginning position (after pause) and before vowel-starting words.

Table 2 : Most frequent insertion errors sorted by frequency (96% of total insertions)

N	Inserted	Previous context (f)	Following context (f)
1165	creacky voice	vocalic (.64) # (.20)	# (.50) vocalic (.42)
759	aspiration	voiceless stop (.89)	[a] (.32) [o] (.23) [u] (.13) [ɾ] (.09)
541	palatalization	voiceless stop (.88) [s] (.06)	[i] (.61) [e] (.37)
503	[ʔ]	# (.85)	vocalic (.96)
236	[ɯ]	[ɾ] (.49) [s] (.35)	# (.67)
141	[u]	[ɾ] (.41) [s] (.40) [l] (.10)	# (.52) [d] (.10)
86	[ə]	# (.52) [s] (.30)	[r] (.38) # (.15) [d] (.10)
84	#	vocalic (.36) [n] (.20) [ɾ] (.14) [s] (.09)	[t] (.45) [ð] (.32)
30	[o]	[ð] (.30) [t] (.17) [ɾ] (.13) [s] (.13)	[ɾ] (.40) # (.20)
29	[e]	[ɾ] (.34) [s] (.21) [ð] (.10)	# (.31) [l] (.24) [r] (.13)

3.1.3 Deletion errors

According to the data (see Table 3), deletions are the less frequent errors. The majority of deletions occur in word-final position, which is expected in spontaneous speech due to the falling intonation and lower intensity in the sentence boundaries. The most frequent deleted consonants are [ɾ] and [n] and correspond to two of the most frequent sounds in word-ending positions in Spanish; consequently, this error is somehow expected in spontaneous speech. On the other hand, we found frequent cases of [l] deletion between vowels or after [n], this is an unexpected case, since [l] is usually substituted by [ɾ] (see 4.1), the deletion solution could be the consequence of the phonetic context (all voiced sounds) which favors coarticulation lenition by vowel deletion instead of substitution.

Vowel deletion is mostly found in voiceless contexts, where Japanese phonological rules predict vowel devoicing. The deletion solution adopted in these cases could be interpreted again as a lenition of this rule triggered by hipoarticulated spontaneous speech. Finally, glides are simplified by deleting the semi-vocalic element [j̬] or [ɰ̬].

Table 3 : Most frequent deletion errors sorted by frequency (87% of total deletions)

N	Deleted	Previous context	Following context
288	[ɾ]	vocalic (.83)	# (.40) vocalic (.22)
207	[n]	vocalic (.98)	# (.46) [x] (.07) [l] (.06)
167	[i]	[s] (.33) [t] (.14) [θ] (.09)	[k] (.38) [t] (.15) [ð] (.12)
160	[l]	vocalic (.67) [n] (.12)	vocalic (.49) # (.36)
124	[o]	[k] (.23) [n] (.11) [t] (.07)	# (.23) [m] (.14) [s] (.13)
98	[u]	[s] (.34) # (.16) [k] (.13) [t] (.09)	[p] (.34) [n] (.23) [l] (.11)
90	[j̟]	[θ] (.22) [t] (.21) [a] (.17) [s] (.10)	[e] (.54) [o] (.15) [l] (.10)
71	[e]	# (.18) [u̟] (.14) [t̪] (.13) [ɾ] (.08)	[n] (.21) # (.17) [s] (.15)
61	[u̟]	[k] (.36) [a] (.25) [p] (.21)	[e] (.41) [a] (.33) [ɾ] (.16)
49	[a]	# (.14) [x] (.12) [ɾ] (.10) [t] (.08)	# (.29) [p] (.14) [k] (.08)

3.2 Transcription reliability

The two revised transcriptions provided by the external annotators showed a very high coefficient of agreement with our original transcription (Cohen's kappa = .98 for annotator a, = .99 for annotator b). Contrary to the expectations, the lack of the annotators' expertise in Japanese phonetics did not cause a low degree of agreement. The high agreement score could have been caused by the instructions given to annotators, or to the uncertainty in transcribing foreign speech and using new symbols to represent sounds not heard before, as uncertainty can lead to a decrease in the number of changes. A new transcription from scratch would probably have shown a much lower degree of agreement than what it was obtained from the transcription verification.

In spite of the high degree of agreement, there are discrepancies between the transcriptions made by the different transcribers, which are mainly related to the use of symbols for specific Japanese sounds, such as [ɯ], [ʔ] and [ɸ]. This was to be expected, as the annotators were not familiar with these sounds, and the corrected sounds share some similarities with Spanish sounds, e.g. [ɯ] was substituted by [u], but also by [e]. Discrepancies were also found between voiceless sounds and their voiced counterparts, like [tʃ] that was substituted by [dʒ]. In these cases, the annotators seem to opt for the voiced sound. According to the annotators, their choice was primary motivated by their perceptual judgment, whereas our transcription was based mainly on the visual examination of the spectrogram and the automatic calculations of Praat. This can be an explanation for this discrepancy. Intermediate realizations –sounds that share features of both the target and the source language– were also problematic to annotate, due to their in-between nature. Although some differences in the annotation of intermediate realizations have been found, the annotators developed similar strategies to ours, especially for the distinction between [ɾ] and [l], as we proposed in [6]. The choice between voiced stops ([b], [d]) and voiced approximates ([β̞], [ð̞]) was difficult in some cases, as these were also intermediate realizations in-between two distinct sounds, and discrepancies between the transcriptions show that they were problematic to annotate.

4 DISCUSSION

In this paper we have presented a procedure for compiling, transcribing and annotating a non-native speech corpus aimed at the development of CAPT tools. The lists for the most frequent errors obtained from processing the corpus reflect the types of errors that should be addressed in the design of a CAPT application intended for Japanese speakers of Spanish L2. As this corpus was transcribed at a narrow phonetic level, not all the disfluencies annotated as errors should be considered, since many of them are related to acoustic phenomena triggered by hipoarticulation, a characteristic of spontaneous speech (e.g: creaky voice or vowel reduction). If we consider only the mispronunciations related to phonological categories in Spanish, we can establish the following most salient errors.

First, the realization of the lateral alveolar sound [l] as an alveolar flap [ɾ] or deleted, especially in intervocalic position; vocalic insertion or consonant deletion in consonant sequences not allowed by

the Japanese combinatory rules, that is to say: onset formed by consonant groups and codas formed by consonants, especially [ɾ] and [n]; and last, vowel devoicing or deletion in voiceless contexts –voiceless consonants or silence–. Spanish sounds new to Japanese speakers, such as [x], [r], [θ] or [ɲ] do not seem to present difficulties of pronunciation, compared to difficulties relating to the articulation of the approximant allophones [β̞] [ð̞] [ɣ̞] of the voiced stops phonemes /b/ /d/ /g/. Thus, we can conclude that the most frequent mispronunciations by Japanese learners of Spanish L2 that might hamper communication and should be addressed in a Spanish pronunciation course are related not only to the articulation of new sounds but, more importantly, to the combination of already existing sounds in both languages, but with different phonological status, and to coarticulation phenomena. For the aim of correcting learners' mispronunciations, phonological context should be taken into account, considering that the majority of errors are related to phoneme combinations in the syllabic structure, and in more general terms, to differences in rhythm structure. ASR developers might also consider the importance of context as a predictor of common mispronunciations by Japanese learners of Spanish.

Although the agreement between our transcription and the transcriptions made by the two external annotators appear to be very high, this score should be taken with caution; since, as we pointed out in the previous section, the high score could be the result of a high degree of uncertainty in transcribing foreign speech with symbols to represent foreign sounds with which the annotator is not familiarized. The most frequent discrepancies are found in the use of these symbols and in the transcription of sounds that usually present features of both the L2 and the L1, intermediate realizations, such as the lateral flap in Japanese that can be perceived as [ɾ] or [l] in Spanish.

In future research we will analyze these problematic sounds further by selecting a subset of common errors in the corpus and will conduct a perception experiment with native speakers. The examples will be perceptually judged by native Spanish speakers who will asses the quality and non-nativeness of words and utterances. Non-nativeness ranking will allow us to determine which types of mispronunciations show the highest scores. This information will be incorporated into the corpus and the scores of non-nativeness will be used for the development of pronunciation rules. Human-based judgment of non-nativeness can be, likewise, incorporated in the automatic assessment of Spanish L2 pronunciation by Japanese speakers, by training an ASR system with data taken from the corpus in order to determine the distance between each segment and the native acoustic models. The Goodness of Pronunciation (GOP) algorithm will be used to compute the distances; then, thresholds between the distances obtained for the segments tagged as errors and the ones obtained for the remaining segments can be established. Finally, the non-nativeness scores obtained from human judgments and the distances obtained by automatic computation can be evaluated in order to determine the degree of agreement between the two types of assessment of foreign speech. This will be the focus of our research in the near future.

REFERENCES

[1] Boersma, P. & Weenink, D. (2014). Praat: Doing phonetics by computer (version 5.3) [Computer program]. Amsterdam: Department of Language and Literature, University of Amsterdam. Retrieved from http://www.praat.org/

[2] Llisterri, J. & Mariño, J. B. (1993). Spanish adaptation of SAMPA and automatic phonetic transcription. SAM-A/UPC/001/v1. ESPRIT project 6819 (SAM-A Speech Technology Assessment in Multilingual Applications). Retrieved from http://liceu.uab.cat/~joaquim/publicacions/SAMPA_Spanish_93.pdf

[3] Wells, J. C. (1994). Computer-coding the IPA: a proposed extension of SAMPA. Speech, Hearing and Language, Work in Progress, 8, 271-289.

[4] TEI Consortium. (2013). 8 Transcription of speech. TEI P5: Guidelines for electronic text encoding and interchange. Retrieved from http://www.tei-c.org/release/doc/tei-p5-doc/en/html/TS.html

[5] Gibbon, D., Moore, R., & Winski, R. (1997). Handbook of Standards and Resources for Spoken Language Systems. Mouton De Gruyter, New York

[6] Carranza, M. (2013). Intermediate phonetic realizations in a Japanese accented L2 Spanish corpus. In Proceedings of the ISCA workshop on Speech and Language Technology in Education (SlaTE). Grenoble, France, August 30-31 & September 1, 2013, (pp. 168-171).

[7] Carranza, M. (in press). Transcription and annotation of a Japanese accented spoken corpus of L2 Spanish for the development of CAPT applications. In II International Workshop on Technological Innovation for Specialized Linguistic Domains (TISLID). Ávila, May 7-9, 2014 .

[8] Carranza, M. (2012). Errores y dificultades específicas en la adquisición de la pronunciación del español LE por hablantes de japonés y propuestas de corrección. In GIDE (Eds.) Nuevos enfoques en la enseñanza del español en Japón –Concha Moreno y GIDE–. Tokyo: Asahi.

[9] Flege, J. (1991). Perception and production: the relevance of phonetic input to L2 phonological learning. In Huebner & Ferguson (Eds.), Crosscurrents in Second Language Acquisition. Amsterdam: John Benjamins (pp.249-289).

[10] Celce-Murcia, M., Brinton, D. M. & Goodwin, J. M. (1996). Teaching Pronunciation. A reference for teachers of english to speakers of other languages. Cambridge: Cambridge University Press.

[11] Mitchell, R., Domínguez, L., Arche, M. J., Myles, F., & Marsden, E. (2008). SPLLOC: A new database for Spanish second language acquisition research. EUROSLA Yearbook, 8, 287-304.

[12] MacWhinney, B. (2000). The CHILDES Project: Tools for Analyzing Talk. Mahwah, New Jersey: Lawrence Erlbaum Associates.

[13] Díaz Rodríguez, L. (2007). Interlengua española. Estudio de casos. Regael: Barcelona.

[14] Campillos Llanos, L. (2013). Oral expression in Spanish by low-intermediate learners : a computer-aided error analysis. In Learner Corpus Research Conference (LCR2013). Bergen.

[15] Council of Europe. (2001). Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR). Retrieved from:http://www.coe.int/t/dg4/linguistic/cadre1_en.asp

[16] Cucchiarini, C., & Strik, H. (1999). Automatic assessment of second language learner's fluency. In Proceedings of the 14th International Congress of Phonetic Sciences, San Francisco, (pp. 759-762). Retrieved from http://lands.let.kun.nl/literature/strik.1999.1.ps

[17] Witt, S. M. (2012). Automatic error detection in pronunciation training: Where we are and where we need to go, In Proceedings of the International Symposium on Automatic Detection of Errors in Pronunciation Training (IS ADEPT). Stockholm, 6-8 June, 2012 (pp. 1-8).

[18] Quilis, A. (1993). Tratado de fonología y fonética españolas (2nd ed.). Madrid: Gredos.

[19] Martínez Celdrán, E. & Fernández Planas, A. M. (2007). Manual de fonética española: articulaciones y sonidos del español. Barcelona: Ariel.

[20] Gil Fernández, J. (2007). Fonética para profesores de español: de la teoría a la práctica. Madrid: Arco/Libros.

[21] Hammerly, H. (1982) "Contrastive phonology and error analysis", IRAL, 20, pp. 17-32.

[22] Cucchiarini, C., Strik, H., & Boves, L. (2000). Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology. Journal of the Acoustical Society of America, 107(2), pp.989-999.

[23] Machuca, M. J. (1997). Las obstruyentes no continuas del español: relación entre las categorías fonéticas y fonológicas en habla espontánea. (PhD thesis). Universitat Autòmoma de Barcelona, Barcelona. Retrieved from: http://ddd.uab.cat/pub/tesis/1997/tdx-0507108-134610/

[24] Detey, S. (2012). Coding an L2 phonological corpus: from perceptual assessment to non-native speech models — an illustration with French nasal vowels. In Y. Tono, Y. Kawaguchi, & M. Minegishi (Eds.), Developmental and Crosslinguistic Perspectives in Learner Corpus Research. Amsterdam/Philadelphia: John Benjamins, (pp. 229-250).

[25] Goronzy, S. (2002). Robust Adaptation to Non-Native Accents in Automatic Speech Recognition. Berlin: Springer.

[26] Ingvalson, E. M., & Holt, L. L. (2011). Perception-production relationship for / r – l / by native Japanese speakers. In The 17th International Congress of Phonetic Sciences (ICPhS XVII), (pp. 938-941).