29 Transcription and annotation of a Japanese accented spoken corpus of L2 Spanish for the development of CAPT applications

Mario Carranza[1]

Abstract

This paper addresses the process of transcribing and annotating spontaneous non-native speech with the aim of compiling a training corpus for the development of Computer Assisted Pronunciation Training (CAPT) applications, enhanced with Automatic Speech Recognition (ASR) technology. To better adapt ASR technology to CAPT tools, the recognition systems must be trained with non-native corpora transcribed and annotated at several linguistic levels. This allows the automatic generation of pronunciation variants, new L2 phoneme units, and statistical data about the most frequent mispronunciations by L2 learners. We present a longitudinal non-native spoken corpus of L2 Spanish by Japanese speakers, specifically designed for the development of CAPT tools, fully transcribed at both phonological and phonetic levels and annotated at the error level. We report the results of the influence of oral proficiency, speaking style and L2 exposition in pronunciation accuracy, obtained from the statistical analysis of the corpus.

Keywords: non-native spoken corpora, spontaneous speech transcription, L1 Japanese, L2 Spanish, standards for transcription and annotation.

1. Universitat Autònoma de Barcelona, Barcelona, Spain; mario.carranza@uab.cat

How to cite this chapter: Carranza, M. (2016). Transcription and annotation of a Japanese accented spoken corpus of L2 Spanish for the development of CAPT applications. In A. Pareja-Lora, C. Calle-Martínez, & P. Rodríguez-Arancón (Eds), *New perspectives on teaching and working with languages in the digital era* (pp. 339-349). Dublin: Research-publishing.net. http://dx.doi.org/10.14705/rpnet.2016.tislid2014.446

1. Introduction

Several studies have pointed out the possibility of efficiently adapting ASR systems to pronunciation assessment of non-native speech (Neri, Cucchiarini, & Strik, 2003) if technology limitations are compensated with a good design of language learning activities and feedback, and the inclusion of repair strategies to safeguard against recognition errors.

An ASR system can be adapted as an automatic pronunciation error detection system by training it with non-native speech data that generates

> "new acoustic models for the non-native realizations of L2 [phones], and by the systematization of L1-based typical errors by means of rules […]. In order to do so, phonetically transcribed non-native spoken corpora are needed; however, manual transcription of non-native speech is a time-consuming costly task, and current automatic transcription systems are not accurate enough to carry out a narrow phonetic transcription" (Carranza, 2013, p. 168).

In this paper we will introduce a corpus of non-native Spanish by Japanese speakers that contains spontaneous, semi-spontaneous and read speech. The corpus is transcribed at the orthographic, phonological and phonetic levels, and annotated with an error-encoding system that specifies the error type and its phonological context of appearance.

This database was compiled and annotated considering its future adaptation as a training corpus for developing ASR-based CAPT tools and applications for the teaching of Spanish pronunciation to Japanese speakers.

In section 1, we will present the general features of the corpus. Section 2 deals with the levels of transcription, the annotation standards, and the phone inventory used in the transcriptions. Finally, the results of the statistical analysis of errors are presented in section 3, followed by a discussion concerning our findings.

2. Corpus data description

The corpus features 8.9h of non-native speech, divided into semi-spontaneous speech (91'), spontaneous speech (214'), read speech (9') and conversational speech (201'). Spontaneous speech represents more than 80% of the recordings. The data was obtained from 10 male and 10 female Japanese students of L2 Spanish at the Spanish Department of the Tokyo University of Foreign Studies. They were selected according to their dialectal area (Kanto dialect) and none of them had previous academic contact with Spanish. The corpus contains the oral tests of the 20 informants throughout their two first academic years of Spanish study (from 1/4/2010 to 31/3/2012), which corresponds to the A1 and A2 levels in the Common European Framework of Reference for Language Learning (Council of Europe, 2001). Oral tests took place every six months, and consisted of different types of tasks that involved spontaneous, semi-spontaneous, and read speech. Semi-spontaneous speech was obtained from oral presentations prepared before-hand (in the 1st and 2nd semesters) and spontaneous speech was gathered from conversations between the student and the examiner and role-plays with no previous preparation (in all semesters). Oral proficiency was also taken into account by computing the mean of all the oral-test scores of each informant. Three proficiency levels were established according to this score: low (N=6), intermediate (N=8) and high (N=6).

The recordings were made with portable recorders and were segmented into individual audio files. The audio files were converted into WAV format and labelled with information regarding the student, the task type and the period of learning (semester). This allows the automatic computation of error rates according to proficiency level, learning stage and speaking style after the transcription and annotation of the corpus.

3. Levels of transcription

Transcription of non-native spontaneous speech is a complex activity due to its high degree of variability and the interference of the L1 and constant

presence of vocalizations and other extra-linguistic phenomena. For this reason, transcribers should follow a set of rules to interpret and represent speech, aimed at maintaining consistency across all levels of transcription (Cucchiarini, 1993). Moreover, transcription is always bounded to 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 factors regarding the speech materials to be transcribed, such as word intelligibility and length of the utterance.

Training corpora for ASR need to be transcribed in a very detailed way, preferably at a narrow phonetic level; acoustic non-linguistic phenomena that could interfere in the generation of the acoustic models should be correctly labelled. Furthermore, the narrow phonetic transcription of non-native speech must be compared to a reference transcription (i.e. a 'canonical' transcription) that represents the expected pronunciation of the utterance by native speakers. This will allow the system to automatically detect discrepancies between both levels and generate rules for pronunciation variants and acoustic models for non-native phones.

The corpus was transcribed and annotated using *Praat* (Boersma & Weenink, 2014). Two levels of representation – canonical phonemic and narrow phonetic transcriptions – were considered, and the resulting tiers were aligned with the orthographic transcription. Vocalizations and non-linguistic phenomena were also marked in two independent tiers. Finally, mispronunciations were encoded in a different tier, and every error label was aligned with the linguistic transcriptions.

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 that tend to be employed when the speaker is confronted with some syntactic or lexical difficulty. The cases of fragmented speech are problematic for the orthographical transcription, especially truncations, when the word is never completed and the transcriber must guess the actual word that the informant intended to say. TEI-conformant (XML-like) tags were used for labelling these phenomena (Gibbon, Moore, & Winski, 1998; TEI Consortium, 2014), as well as unclear cases, missing words, foreign words and erroneous words (like regularized irregular verbs). Hesitations and interjections were also transcribed at this level according to their standardized forms in dictionaries. Only the speech of the informant is transcribed. The commentaries of the examiner are not considered, except when they overlap with the student's speech; in these cases, the overlapping speech is tagged with an XML label in the incident tier (the list of XML tags employed is shown in Table 4).

3.2. Canonical phonemic transcription

The canonical phonemic tier shows the phonological transcription of each word as pronounced in isolation. Northern Castilian Spanish (Martínez Celdrán & Fernández Planas, 2007; Quilis, 1993) was adopted as the standard reference for the transcription, considering that Japanese students had been taught mainly in this variety. Consequently, at this level, the phonemic opposition /s/–/θ/ is preserved, but not the opposition /ʝ/–/ʎ/, which is neutralized in favor of /ʝ/ (Gil, 2007). An adaptation of SAMPA to Spanish (Llisterri & Mariño, 1993) was chosen for the inventory of phonological units (see Table 1), since the obtained transcription had to be machine-readable.

3.3. Narrow phonetic transcription

The narrow phonetic level represents the actual pronunciation of the speaker in the most accurate way. In order to avoid transcriber's subjectivity, the transcription was based primarily on acoustic measurements and visual examination of the spectrogram and the waveform. We avoided perceptual judgment, except in cases where the decision cannot be taken from the methods stated before and should be reached upon auditorial perception. Coarticulatory phenomena (nasalization

and changes in place or articulation) are considered here, as well as the Spanish allophonic variants of the phonemes presented in Table 1. We added a new set of symbols and diacritics taken from X-SAMPA (Wells, 1994) to account for these phenomena (see Table 2). Further symbols were also added to account for the L2 Spanish pronunciation of Japanese speakers. In total, 11 new symbols and 7 diacritics were needed for the narrow phonetic transcription (see Table 3).

3.4. Vocalizations and non-linguistic phenomena

Vocalized or semi-lexical elements, such as laughters, hesitations, and interjections were labelled in a separate tier. The 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.

Table 1. Our SAMPA inventory for phonemic transcription, based on Llisterri and Mariño (1993)

IPA	SAMPA	Description	IPA	SAMPA	Description
/a/	a	central open vowel	/m/	m	voiced bilabial nasal
/e/	e	front mid vowel	/n/	n	voiced alveolar nasal
/i/	i	front close vowel	/ɲ/	J	voiced palatal nasal
/i̯/	j	front close vowel (used in glides)	/t͡ʃ/	tS	voiceless palatal affricate
/o/	o	back mid rounded vowel	/f/	f	voiceless labiodental fricative
/u/	u	back close rounded vowel	/θ/	T	voiceless interdental fricative
/u̯/	w	back close rounded vowel (used in glides)	/s/	s	voiceless alveolar fricative
/p/	p	voiceless bilabial stop	/x/	x	voiceless velar fricative
/b/	b	voiced bilabial stop	/l/	l	voiced alveolar lateral
/t/	t	voiceless dental stop	/ʝ/	jj	voiceless palatal stop
/d/	d	voiced dental stop	/ɾ/	r	voiced alveolar flap
/k/	k	voiceless velar stop	/r/	rr	voiced alveolar trill
/g/	g	voiced velar stop			

Table 2. SAMPA inventory of Spanish allophones used in the narrow phonetic transcription

IPA	SAMPA	Description	IPA	SAMPA	Description
[β̞]	B	voiced bilabial approximant	[z]	z	voiced alveolar fricative
[ð̞]	D	voiced dental approximant	[d͡ʒ]	dZ	voiced palatal affricate
[ɣ̞]	G	voiced velar approximant	**Diacritics (X-SAMPA)**		
[ʝ]	J\	voiced palatal stop	[ã]	_~	nasalized
[j]	j	voiced palatal approximant	[ă]	_X	extra short
[ŋ]	N	voiced velar nasal	[i̯]	_^	non-syllabic (used in combination with full vowels in glides)

Table 3. X-SAMPA inventory of symbols used to represent Japanese and other sounds in the narrow phonetic transcription

IPA	X-SAMPA	Description	IPA	X-SAMPA	Description
[ɯ]	M	unrounded central-back vowel	[d͡z]	dz	voiced alveolopalatal affricate
[ə]	@	central mid vowel	[v]	v	voiced labiodental fricative
[ʃ]	S	voiceless postalveolar fricative	**Diachritics (X-SAMPA)**		
[ɸ]	p\	voiceless bilabial approximant	[ḁ]	_0	devoiced
[ç]	C	voiceless palatal fricative	[a̰]	_k	creaky voiced
[ʔ]	?	glottal stop	[aʲ]	_j	palatalized
[h]	h	voiceless glottal fricative	[aʰ]	_h	aspirated
[ʝ]	j\	voiced palatal fricative	[a̤]	_t	breathy voiced

This is why vocalizations were separated from the rest of speech. They were marked using XML tags 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 laughs, breathing, external noise and overlapping speech of the examiner in this

group. All tags used in the orthographic, vocalization and incident tiers are shown in Table 4.

4. Results and discussion

All the data from the *Praat* transcription tiers was recovered using *Praat* scripts, and data tables were generated for the statistical analysis. The resulting tables contain every mispronounced sound and all the information annotated in the transcriptions. Since the audio files varied in their duration, longer speech can make the possibility of committing errors rise. Consequently, we adopted a metric (error ratio) that takes into account the length of the file by counting the total number of mispronunciations and dividing it by the total number of words, after subtracting the number of hesitations and interjections. The adopted formula for obtaining the error ratio is shown in Figure 1. This metric indicates the total number of mispronunciations per linguistic word, and serves to better evaluate the speaker's performance in spontaneous non-prepared speech, as the duration of the audio files varies drastically from speaker to speaker.

Table 4. XML tags used for the annotation of extra-linguistic and non-linguistic phenomena, adapted from TEI Consortium (2014)

Tag	Explanation	Transcription tier
<repetition>	The following word is completely repeated at least once.	orthographic
<truncation>	The word is not completely uttered. Also used when repetitions are not complete.	orthographic
<unclear>	The word is recognized but cannot be phonetically transcribed due to problems in the signal.	orthographic
<foreign>	Foreign words articulated differently than target-language conventions.	orthographic
<gap\>	The marked segment cannot be recognized (no need to close).	orthographic
<sic>	Made-up word or non-existing word in target language.	orthographic
<noise>	External noise that interferes with speech.	incident

<breath>	Breathing of the speaker. It can happen alone or interfering with speech.	incident
<overlap>	The interviewer's speech overlaps with the informant's speech.	incident
<hesitation>	Filled-pause.	vocalization
<interjection>	Exclamation due to surprise, annoyance or other feelings.	vocalization
<laugh>	Inserted laughing or speech uttered while laughing.	vocalization

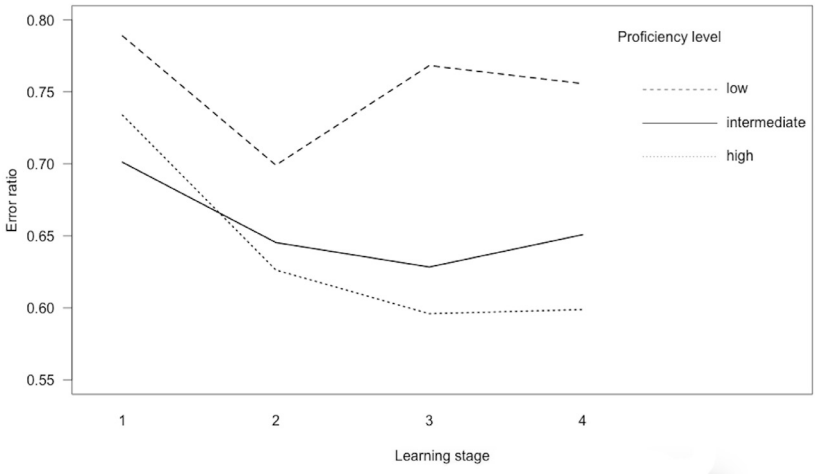
Figure 1. Formula for calculating the error ratio

$$ratio_e = \frac{N_e}{N_w - (N_h + N_i)}$$

Error ratio scores were obtained from each audio file and statistically analyzed considering three variables: oral proficiency, speaking style and period of learning. The statistical tests (ANOVA) showed no significant influence of the speaking style or time on the error ratio, which means that the number of pronunciation errors does not depend on the preparation or spontaneity of the discourse and does not vary throughout the first two years of language teaching (in a non-immersive L2 environment). However, oral proficiency had a clear influence on the error ratio (df=2, F=7.431, p=0.00079), which is lower in the high proficiency group and higher in the low proficiency group.

The mean error ratio actually varies in all the four learning stages when each proficiency group is analyzed separately (see Figure 2). Error ratio decreases specially in the period from 6 to 12 months of learning in all proficiency groups. From the 12th month, this tendency continues in intermediate and high proficiency groups, but not in the low proficiency group, which shows an error increase up to the first period level. These findings suggest that language exposure has a positive influence in intermediate and high proficiency learners, but not in low proficiency learners. Regarding the influence of speaking style on error ratio, it should be highlighted that spontaneous and conversational speech shows much more variability in the results than semi-spontaneous and read speech, as expected. Differences on mean error ratio by the speaking style are minimal.

Figure 2. Mean error ratio by period of learning separated by oral proficiency groups



5. Conclusions

Our results show that the starting oral proficiency level of the student, due mainly to individual abilities, is the only variable that reported a positive impact on Spanish pronunciation acquisition. Although L2 exposure seems to reduce error ratios in intermediate and high proficiency groups – especially from the sixth month of instruction onwards –, the obtained differences did not prove to be statistically significant. Consequently, it seems that exposure to the target language is not enough to expect pronunciation accuracy improvement in foreign language students.

In further reports, we will focus on the specific errors found in the corpus and offer results by frequency of occurrence and error type. Future research will aim at the evaluation of erroneous utterances by means of native-speaker perceptual assessment and automatic evaluation by an ASR system.

References

Boersma, P., & Weenink, D. (2014). *Praat: doing phonetics by computer*. Retrieved from http://www.fon.hum.uva.nl/praat/

Carranza, M. (2013). Intermediate phonetic realizations in a Japanese accented L2 Spanish corpus. In P. Badin, T. Hueber, G. Bailly, D. Demolin, & F. Raby (Eds.), *Proceedings of SLaTE 2013, Interspeech 2013 Satellite Workshop on Speech and Language Technology in Education* (pp. 168-171). Grenoble, France. Retrieved from http://www.slate2013.org/images/slate2013_proc_light_v4.pdf

Council of Europe. (2001). Common European framework of reference for languages: learning, teaching, assessment. Cambridge, U.K: Press Syndicate of the University of Cambridge. Retrieved from https://www.coe.int/t/dg4/linguistic/Source/Framework_EN.pdf

Cucchiarini, C. (1993). *Phonetic transcription: a methodological and empirical study*. PhD dissertation, Radboud Universiteit Nijmegen.

Gibbon, D., Moore, R., & Winski, R. (1998). *Spoken language system and corpus design*. Berlin: Mouton De Gruyter. Retrieved from http://dx.doi.org/10.1515/9783110809817

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

Llisterri, J., & Mariño, J. B. (1993). *Spanish adaptation of SAMPA and automatic phonetic transcription*. ESPRIT PROJECT 6819 (SAM-A Speech Technology Assessment in Multilingual Applications).

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

Neri, A., Cucchiarini, C., & Strik, H. (2003). Automatic speech recognition for second language learning: how and why it actually works. *Proceedings of the 15th International Congress of Phonetic Sciences* (pp. 1157-1160). Barcelona, Spain.

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

TEI Consortium. (2014, January 20). *TEI P5: Guidelines for electronic text encoding and interchange – 8 Transcriptions of speech*. Retrieved from http://www.tei-c.org/release/doc/tei-p5-doc/en/html/TS.html

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

Published by Research-publishing.net, not-for-profit association
Dublin, Ireland; Voillans, France, info@research-publishing.net

New perspectives on teaching and working with languages in the digital era
Edited by Antonio Pareja-Lora, Cristina Calle-Martínez, Pilar Rodríguez-Arancón

Typeset by Research-publishing.net
Cover design and frog picture by © Raphaël Savina (raphael@savina.net)

ISBN13: 978-1-908416-34-6 (Paperback - Print on demand, black and white)
Print on demand technology is a high-quality, innovative and ecological printing method, with which the book is never 'out of stock' or 'out of print'.

ISBN13: 978-1-908416-35-3 (Ebook, PDF, colour)
ISBN13: 978-1-908416-36-0 (Ebook, EPUB, colour)

Legal deposit, Ireland: The National Library of Ireland, The Library of Trinity College, The Library of the University of Limerick, The Library of Dublin City University, The Library of NUI Cork, The Library of NUI Maynooth, The Library of University College Dublin, The Library of NUI Galway.

Legal deposit, United Kingdom: The British Library.
British Library Cataloguing-in-Publication Data.
A cataloguing record for this book is available from the British Library.

Legal deposit, France: Bibliothèque Nationale de France - Dépôt légal: mai 2016.