Exploratory use of automatic prosodic labels for the evaluation of Japanese speakers of L2 Spanish

David Escudero-Mancebo[1], *César González-Ferreras*[1],
Lourdes Aguilar[2], *Eva Estebas-Vilaplana*[3], *Valentín Cardeñoso-Payo*[1]

[1]Department of Computer Science, University of Valladolid, Spain
[2]Department of Spanish Philology, Universitat Autònoma de Barcelona, Spain
[3]Department of Modern Languages, UNED, Spain
descuder@infor.uva.es

Abstract

An automatic labeling system using Sp_ToBI annotation conventions has been applied both to a non-native corpus of Japanese speakers using Spanish and to a reference corpus of Spanish speakers. A set of metrics based on conditional entropy is computed by using the output of an automatic labeler which happens to be highly correlated with the rates assigned by a team of subject evaluators. An analysis of the relative frequencies in the use of each of the Sp_ToBI symbols permits to identify the recurrent mistakes in the productions of non-native speakers. It is discussed with the results that the majority of the observed prosodic deficits can be explained by the prosodic transference between the Japanese and Spanish systems as it had been previouly reported in the state of art.

Index Terms: Prosody in language contact and second language acquisition, Prosodic ToBI labeling, Computer assisted pronunciation training.

1. Introduction

It is well known that to achieve a good competence in a second language, a crucial step concerns to the advances in the prosodic domain. Related to this, academic curriculum such as the proposed by the Cervantes Institute for Spanish [1] include prosody among the evaluation criteria that allows to assess the level of foreign language proficiency according to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR). We can agree then that if prosody is an evaluation indicator, the development of ICT tools to diagnose the prosodic limitations is justified. Within the framework of a larger project [1], this work presents an experiment to analyse mispronunciations of Japanese students of Spanish, with an automatic system that profits the ToBI labels to diagnose the deviations of the prosodic productions of non-native speakers. To do this, new metrics on the distance between native and non-native students are proposed.

Prosody is characterized by a high variability which difficults its study. In particular, evaluating the quality of a prosodic profile is a difficult task because two melodic contours that have the same function can have different shapes, and even more, two melodic contours that have a similar shape (for example with differences affecting only a single syllable) can be perceived very differently [2]. Under these conditions, it is risky to follow approximations for evaluating prosody consisting on computing the distances between the prosodic contours of non-native speakers with respect to the contours of reference speakers or golden speakers [3]. Two prosodic contours that are different in shape can be perceived as identical whereas two similar prosodic contours can be perceptually very different. To face up the high variability of prosodic contours we present here an approach that is based in prosodic labels. By labeling the prosodic contours we simplify its representation by means of symbolic information that specifies the relevant aspects in terms of communication. In line with other works in pronunciation assessment [4, 5, 6], the pitch contours of non-native Spanish speakers as well as those of the reference speakers are analysed following the Autosegmental-Metrical framework [7] and in particular the Sp_ToBI labeling conventions.

Labeling utterances with any ToBI system is a costly job both in time and resources as it requires highly qualified personnel [8]. Fortunately, there are automatic tools that simplify the task [9]. This type of tool is also available for Spanish [10] with more than 80% of correct labeling rates. In contrast with manual labeling, these techniques have the advantage of reduced cost and offer repeatable results. They are supported by objective measurements that have to do with the temporal evolution of the signal and with the lexical-syntactic function of the words that compose the message. Labeling a reference corpus of native pronunciation like the Glissando corpus for Spanish [11] or a corpus including the voice of non-native students doing pronunciation exercises can be done automatically with a reduced cost. Even more, the labeling of the utterances of the students to be evaluated in terms of prosodic quality could be done in real time. In spite of this, the quality of the automatic prosodic labels could always be under doubt. In this work we show that these automatic labels permit to offer reliable indicators concerning the quality of prosody that are consistent with the judgments of human evaluators. Moreover, we show that they permit to give cues concerning the types of mistakes of the non-native speakers.

The prosodic labeling does not eliminate all the prosodic variability. Indeed, the same sentence can be uttered differently by associating boundaries and accents with different words resulting on different prosodic realizations for the same sentence, all of them correct. We do not use a single reference for each utterance but a set of them, uttered by different native speakers. An analysis based on conditional entropy is applied so as to

[1] TIN2014-59852-R VIDEOJUEGOS SOCIALES PARA LA ASISTENCIA Y MEJORA DE LA PRONUNCIACION DE LA LENGUA ESPAÑOLA, VA145U14 EVALUACIÓN AUTOMÁTICA DE LA PRONUNCIACIÓN DEL ESPAÑOL COMO LENGUA EXTRANJERA PARA HABLANTES JAPONESES

determine the degree of deviation of the non-native utterances with respect to the references.

In this paper we focus on Japanese speakers of L2 Spanish. It is know that most of the mistakes that non-native students commit when pronouncing L2 are due to prosodic transference [12]. In this work we show that using the Sp_ToBI labeling system, many of the detected mistakes are already reported in the state of art as typical mistakes of Japanese speakers having its origin in L1 pronunciation. We discuss about the possibilities of the proposed method for detecting these predictable mistakes and the implications for diagnostic evaluation of non-native speech.

First, we present the experimental procedure, including details of the corpus, the automatic prosodic labeling and the metrics that have been used. Next, we describe and discuss the results. The paper ends with the conclusions and future work.

2. Experimental procedure

2.1. The Corpus

In the framework of the SAMPLE research project, a corpus of spoken Spanish by non-native speakers was developed as a means to support future CAPT studies. The central part of the corpus includes a set of sentences and paragraphs selected from the news database of a popular Spanish radio news broadcasting station. The texts cover various information domains related to everyday's life. They were obtained from the Glissando corpus, which was developed in connection to another project related to automatic prosodic labelling. The materials used in this study belong to the subset of prosodically balanced sentences in Glissando, which statistically resemble the prosodic variability found in Spanish [11].

The whole corpus is described in [13]. It contains different materials: read sentences, the Aesop's Fable 'The North Wind' and news paragraphs. In this study, fifteen read sentences from the news paragraphs of the Glissando corpus [11] were selected to be read by a group of non-native Spanish speakers. The list of sentences is described in [13] (see table 1 of that paper). All sentences followed a phonetic coverage criterion. In this study we only focus on the Japanese speakers for the sake of simplicity. These speakers are referred as *f11*, *m03*, *f12*, *f14* and *f13* in the database where *f* means female and *m* means male.

The reference sentences of native pronunciation are the corresponding fifteen sentences extracted from the Glissando corpus. As the Glissando corpus recorded eight different professional speakers, we have more than one reference to contrast the non-native pronunciation. The speakers are referred as *f16a*, *f11r*, *f13r*, *f15a*, *m09a*, *m10a*, *m12r* and *m14r*. As before, *f* means female and *m* means male. Furthermore, *r* stands for a radio speaker and *a* indicates an actor.

2.2. Automatic prosodic labeling

For the labeling of the spoken material, the procedure described in [14] was used. An automatic labeling system was formerly trained with a subcorpus of the Glissando corpus consisting of a series of news recorded by two professional speakers (12 news were read by a female radio broadcaster and 12 other news were recorded by a male adversiting professional). These news items include a total of 3202 words (7091 syllables) labeled with 2058 pitch accents, 1115 boundary tones and 1029 breaks.

The automatic system is a pairwise coupling classifier that combines evidences of three complementary types of classifiers such as artificial neural networks (NN), decision trees (DT), and support vector machines (SVM) [15]. In order to combine the three classification modules (DT, NN and SVM), we used the comprehensive fuzzy technique proposed in [10].

The reference unit for the automatic labeling system is the word. Every word is characterized in terms of prosodic information (F0, energy and duration features) and POS tags, as described in [15]. As a result, we obtain up to two Sp_ToBI labels per word: one for the pitch accent and another one for the boundary tone. We use the following Sp_ToBI pitch accents: H*, L* = {L* ∪ L*+H ∪ H+L* }, L+>H*, L+H* ={L+H* ∪ (L+)H*}, L+!H* ={L+!H* ∪ (L+)!H* ∪ !H*}, L+¡H* ={L+¡H* ∪ (L+)¡H* ∪ ¡H*}; and the following boundary tones: L%, H%, =%, !H%, LH% ={LH% ∪ L!H%}. Additionally, the label "none" represents the absence of tone.

2.3. Contrasting prosodic labels

Our proposal for contrasting the prosodic profiles of the different speakers is based on mutual information between the different speakers. Two groups of informants are involved in this study: a group of the native speakers used as a reference group (from now on R) and a group of non-native speakers whose productions have been evaluated (from now on n). We describe the variety observed in the reference by using the entropy as $H(R)$ and the variety observed in the non-native speakers as $H(n)$. We can measure how different n is with respect to R by computing:

$$I(n; R) = H(R) + H(n) - H(n, R) \tag{1}$$

It is small if $H(n, R)$ is approximately equal to $H(R) + H(n)$ and it is large if $H(n, R)$ is much smaller than $H(R) + H(n)$. If we compute the entropies H by using the ToBI symbols, a large $I(n; R)$ would indicate that the non-native speaker n is doing a similar use of the tones with respect to the R reference. On the contrary, a small $I(n; R)$ indicates that they have very little in common.

		Subjective metrics					
Spk	I(n,R:T)	int	flu	pho	acc	rhy	dele
f11	0.0613	2.7	2.0	2.9	2.4	3.0	2.9
m03	0.0554	2.9	2.4	3.1	2.8	2.6	3.1
f12	0.0368	3.0	2.3	3.0	2.7	2.9	3.3
f14	0.0390	3.0	2.4	3.1	2.5	3.4	3.2
f13	0.0264	4.2	3.3	3.8	3.5	3.6	3.9

Table 1: Contrast between the objective and the subjective metrics. In the subjective metrics int means intelligibility, flu is fluency, pho is phonetic correctness, acc means lexical accent correctness, rhy measures to what an extent the non-native prosody resembles that of the native speakers and $dele$ indicates the level of proficiency in Spanish. The figures are mean values of the opinions of the four evaluators. More details can be found in [13]. $I(n, R : T)$ is computed with the samples of the first repetition of the sentences.

As every speaker s is characterized by the distribution of tones $t \in T$ obtained by the automatic labeling system, the computation of $H(s)$ is

$$\mathrm{H(s)} = -\sum_{\mathrm{t}} \frac{\mathrm{n_{st}}}{\mathrm{n_s}} \log_2 \frac{\mathrm{n_{st}}}{\mathrm{n_s}} = -\sum_{\mathrm{s}} \mathrm{p_{t|s}} \log_2 \mathrm{p_{t|s}} \tag{2}$$

where n_{st} indicates the number of appearances of type t tones produced by the speaker s. By replacing s by n and R, $H(n)$

	Sp_ToBI tones												Metrics	
Spk	=%	!H%	H*	H%	L*	L%	L+>H*	L+!H*	L+¡H*	L+H*	LH%	none	#	I(n,R:T)
f11	3	20	24	26	14	29	11	14	1	91	6	53	292	0.061
f16a	2	32	19	10	16	26	16	19	8	60	4	54	266	0.049
f11r	5	22	14	27	2	14	14	16	8	72	0	68	262	0.057
f13r	8	19	14	19	7	11	17	21	10	57	2	69	254	0.043
f15a	6	14	23	23	8	24	5	22	7	72	5	57	266	0.040
m09a	7	23	9	5	7	22	10	22	6	75	3	63	252	0.050
m10a	4	33	17	12	12	16	15	11	5	75	2	59	261	0.033
m12r	2	26	9	15	10	19	7	11	9	79	4	70	261	0.035
m14r	2	12	16	20	6	16	12	23	8	64	5	63	247	0.031

Table 2: Count of the different tones. The speakers f16a, f11r, f13r, f15a, m09a, m10a, m12r, m14r are the native speakers $r \in R$ of the Glissando corpus [11].

and $H(R)$ could be computed. However, in order to compute $H(n, R)$ paired data are required.

In computer assisted pronunciation training, paired data is difficult to obtain because non-native pronunciation can include disfluences with repetititions, so that the number of words in the reference sentences and in the non-native pronunciation can be different. In order to avoid this limitation, we use the *information distance* metric between n and R as proposed by Krippendorf in [16] which is computed as:

$$\mathrm{I(n,R:T)} = \sum_{\mathrm{t}} \mathrm{p_{nt}} \log_2 \frac{\mathrm{p_{nt}}}{\mathrm{p_t p_n}} + \sum_{\mathrm{t}} (\mathrm{p_t} - \mathrm{p_{nt}}) \log_2 \frac{\mathrm{p_t} - \mathrm{p_{nt}}}{(1-\mathrm{p_n})\mathrm{p_t}} \qquad (3)$$

This metric compares the distribution of tones $t \in T$ in n with the distribution of tones in R. The higher the value the stronger the differences between the non-native speakers and the native speakers represented as the R samples.

3. Results and discussion

Table 1 shows the mean values of the scores that the speakers obtained in the subjective test and the objective metric $I(n, R : T)$. In [13], we already discussed that the correlation between the subjective metrics is high. The figures of table 1 show that objective metrics are inversely correlated with the subjective ones. Indeed, *f11* is identified as the worst speaker by the evaluators and the same speaker obtains the worst objective results. Speaker *f13* is the best one both for the evaluator and for the objective metric. This result permits to be optimistic with respect to the potential of the *information distance metric* as an objective indicator of the quality of non-native speech. Nevertheless, more experiments should be performed as only a portion of the whole data has been considered and thus the size of the testing corpus is small.

In table 2 we focus on the speaker that obtained the worst results in table 1 (*f11*) so as to analyze the reasons for higher objective metric values. The number of appearances of some of the tones contrast with the values obtained for the native speakers. Thus, for example, the tone *L+H** appears 91 times in the non-native speaker productions, whereas in the utterances produced by the native speakers, its appearance ranges from 79 to 57 times. Other tones that contrast are *H**, *L%* and *LH%* with a value in the *f11* row that is higher than in the value in the rest of rows. Additionally, for the tones *L+¡H** and *none*, *f11* has less samples than the rest of speakers. The consequence of this atypical distribution is that the metric $I(n, R : T)$ has a value 0.061 that is higher than the value of the metric in rest of rows.

Table 3 shows the most frequent confusions. We have aligned the tones of the sentences that have the same number of words (sentences without repetitions that appears as a consequence of disfluent speech). The words of the non-native utterances are paired with the corresponding ones of the eight native speakers. Due to space restrictions, we only include the results of the two best and the two worst non-native speakers according to the ranking displayed in table 1. The rows *Coincidences* and *Confusions* of table 3, could also be indicators of their quality since the worse the speaker, the lower the percentage of coincidences between the tones of the non-native speaker n and those of the reference speakers $r \in R$. Again, the speaker $f11$ is the worst one if we take into account these rows of the table.

However, the most interesting results can be observed in the last rows of the table. First, there are frequent confusions that are coincident for every Japanese speaker. For example, the pair of tones *L+H*-L+!H** are the ones that have more confusion. It has been reported in the state of the art that a common default of Japanese students of Spanish is that they do not reproduce properly the typical Spanish intonation contours [17] but use a less melodic intonation [18]. The Castilian Spanish speakers of the reference corpus frequently use the *L+!H** tone as the typical falling intonation for the nuclear configuration of declarative sentences. In view of this fact, we could hypothesize that the Japanese students are not reproducing correctly this melodic pattern.

Other recurrent confusions are *H%-none*, and *!H%-none*. Japanese students of Spanish tend to make small pauses between words resulting in a slow and paused discourse. These types of inconsistences have also been reported in the state of the art [18].

Figure 1 shows the prosodic acoustic parameters and the Sp_ToBI labels for an utterance that contains some of the confusions that have been identified in table 3. More boundaries appear in the non-native version (after the word *hoy* and *escuelas*) evidencing a less fluent speech and a lower proficiency in the L2. Additionaly, the native version ends with the typical Castilian Spanish declarative pattern *L+>H* L* L%* while the non-native end includes an atypical higher tone (*L+>H* L+!H* L%*) before the boundary tone.

Along with those observations, another frequent inconsistence that can be observed in table 3 occurs when the non-native speakers use an accent when the native speakers do not (like *H*-none* and *L+H*-none*). Again, this is an expected mistake as the Japanese accent is determined by a high-low pitch in each

f11			m03			f14			f13		
	#	%		#	%		#	%		#	%
Sentences	49	41	Sentences	67	56	Sentences	60	50	Sentences	49	41
Coincidences	583	56	Coincidences	1007	61	Coincidences	888	65	Coincidences	576	60
Confusions	451	44	Confusions	651	39	Confusions	488	35	Confusions	378	40
Others	209	46	Others	308	47	Others	249	51	Others	192	51
none-H%	13	2.9	L*-L+H*	24	3.7	!H%-none	15	3.1	L%-none	12	3.2
L+H*-none	18	4.0	L+>H*-L+H*	25	3.8	L+H*-L+>H*	17	3.5	!H%-none	13	3.4
H%-!H%	19	4.2	L+H*-none	27	4.1	L*-L+H*	17	3.5	L+H*-L+>H*	14	3.7
none-L+H*	19	4.2	H%-!H%	28	4.3	L+H*-none	20	4.1	L*-L+H*	14	3.7
L+H*-L+¡H*	22	4.9	L+!H*-L+H*	28	4.3	none-H*	23	4.7	none-H*	14	3.7
none-H*	22	4.9	H*-none	29	4.5	none-H%	23	4.7	H%-!H%	19	5.0
!H%-none	27	6.0	none-H*	38	5.8	H*-none	26	5.3	H*-none	20	5.3
H*-none	30	6.7	L+H*-L+!H*	46	7.1	none-!H%	26	5.3	H%-none	22	5.8
H%-none	35	7.8	!H%-none	49	7.5	L+!H*-L+H*	31	6.4	L+H*-none	23	6.1
L+H*-L+!H*	37	8.2	H%-none	49	7.5	L+H*-L+!H*	41	8.4	L+H*-L+!H*	35	9.3

Table 3: *Sentences* is the number of pairs of sentences that have been contrasted. *Coincidences* is the number of tones that are equal in the utterances of the non-native speakers n and in the equivalent utterances of the reference native speakers $r \in R$. *Confusions* is the number of tones that are different computed in the same conditions as *Coincidences*. The pair of tones that are computed in *Confusions* are listed in the last rows of the table sorted in term of frequency of appearance. Every confusion is expressed as $t_n - t_r$ so that t_n is the tone in the non-native speaker's utterance and t_r is the tone in the reference speaker's one.

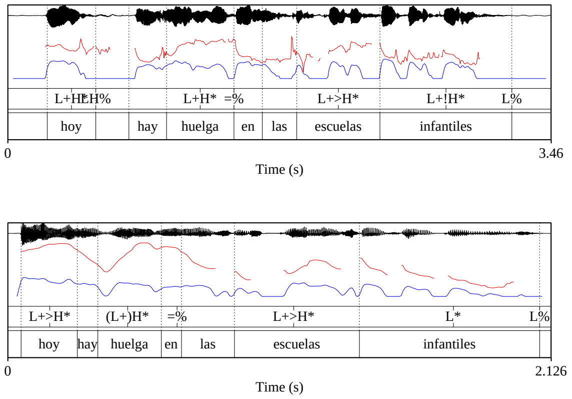

Figure 1: Sample of an utterance spoken by a non-native speaker (upper version) and by a native speaker (lower version). This corresponds to the utterance $s15$ and to speakers $f11$ and $f16a$.

word accent with a fixed morae position as the basic unit of reference [19]. The consequence is that Japanese students tend to overarticulate the pronunciation of Spanish unaccented words by placing a pitch accent in the tonic syllable [20].

4. Conclusions and future work

In this work we have presented an experiment in which automatic prosodic labels have been used to check the prosody of Japanese students of Spanish. The information distance permits to build a ranking of students that is highly correlated with the ranking built with manual assessments. Furthermore, we have analyzed the most frequent potential misuses of the Sp_ToBI tones as a cue of the possible mistakes that appear in the prosodic productions of the non-native speakers. We have discussed the correspondence between these frequent confusions and the typical mistakes of these group of speakers as reported in the state of the art.

The results are promising and encourage performing new experimentation. As future research, we will work on the computation of *mutual information* and *information distance* with paired data so that a new ranking of speakers can be built, and what is more important, the most informative confusions (no the most frequent ones) can be identified. Being able to inform the learners about the type of mistake that they are making could be a contribution on diagnostic evaluation, which is a significant advance with respect to simpler assessment. This is a common approach in nowadays computer assisted pronunciation training technology.

5. References

[1] A. G. Santa-Cecilia, “Plan curricular del instituto cervantes: niveles de referencia para el español,” *MarcoELE: Revista de didáctica*, no. 5, p. 1, 2007.

[2] N. Campbell, “Evaluation of speech synthesis,” in *Evaluation of text and speech systems*, L. Dybkjaer, H. Hemsen, and W. Minker, Eds. Springer Science & Business Media, 2007.

[3] J. P. Arias, N. B. Yoma, and H. Vivanco, “Automatic intonation assessment for computer aided language learning,” *Speech communication*, vol. 52, no. 3, pp. 254–267, 2010.

[4] A. Rosenberg, “Symbolic and direct sequential modeling of prosody for classification of speaking-style and nativeness.” in *INTERSPEECH*, 2011, pp. 1065–1068.

[5] J.-m. Kim, “Annotation of a non-native english speech database by korean speakers,” *Speech Sciences*, vol. 9, no. 1, pp. 111–135, 2002.

[6] J. Tepperman, A. Kazemzadeh, and S. S. Narayanan, “A text-free approach to assessing nonnative intonation.” in *INTERSPEECH*, 2007, pp. 2169–2172.

[7] D. R. Ladd, *Intonational phonology*. Cambridge University Press, 2008.

[8] A. K. Syrdal, J. Hirschberg, J. McGory, and M. Beckman, “Automatic tobi prediction and alignment to speed manual labeling of prosody,” *Speech communication*, vol. 33, no. 1, pp. 135–151, 2001.

[9] A. Rosenberg, “AutoBI-a tool for automatic ToBI annotation.” in *INTERSPEECH*, 2010, pp. 146–149.

[10] D. Escudero-Mancebo, C. González-Ferreras, C. Vivaracho-Pascual, and V. Cardeñoso Payo, “A fuzzy classifier to deal with similarity between labels on automatic prosodic labeling,” *Computer Speech and Language*, vol. 28, no. 1, pp. 326 – 341, 2014.

[11] J.-M. Garrido, D. Escudero, L. Aguilar, V. Cardeñoso, E. Rodero, C. de-la Mota, C. González, C. Vivaracho, S. Rustullet, O. Larrea, Y. Laplaza, F. Vizcaíno, E. Estebas, M. Cabrera, and A. Bonafonte, “Glissando: a corpus for multidisciplinary prosodic studies in Spanish and Catalan,” *Language Resources and Evaluation*, vol. 47, no. 4, pp. 945–971, 2013.

[12] E. E. Vilaplana, *Teach Yourself English Pronunciation: An Interactive Course for Spanish Speakers*. Netbiblo, 2009.

[13] D. Escudero-Mancebo, C. González-Ferreras, and V. Cardeñoso Payo, “Assessment of non-native spoken spanish using quantitative scores and perceptual evaluation,” in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14)*, N. C. C. Chair), K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, Eds. Reykjavik, Iceland: European Language Resources Association (ELRA), may 2014, pp. 3967–3972. [Online]. Available: http://www.infor.uva.es/descuder/investig/pdfs/LREC2014Assessment.pdf

[14] D. Escudero, L. Aguilar, C. González, V. Cardeñoso, and Y. Gutiérrez, “Preliminary results on Sp_ToBI prosodic labeling assisted by an automatic fuzzy classifier,” in *Proceedings of the 7th International Conference on Speech Prosody*, Dublin, Ireland, May 2014, pp. 457–461.

[15] C. Gonzalez-Ferreras, D. Escudero-Mancebo, C. Vivaracho-Pascual, and V. Cardeñoso Payo, “Improving automatic classification of prosodic events by pairwise coupling,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 7, pp. 2045 –2058, sept. 2012.

[16] K. Krippendorff, *Information theory: structural models and qualitative data*. SAGE Publications, 1986.

[17] “Influencia de la entonación oracional en la percepción de la posición del acento castellano por parte de estudiantes japoneses de castellano.” Tech. Rep. [Online]. Available: https://www7.uc.cl/letras/laboratoriodefonetica/html/actividades_realizadas/2011_percepcion_acento_atria/poster.JJAtria_2011_percepcion_ac_japoneses_esp.pdf

[18] M. Carranza, “Errores y dificultades específicas en la adquisición de la pronunciación del español le por hablantes de japonés y propuestas de corrección.” *Nuevos enfoques en la enseñanza del español en Japón.*, p. 51–78, 2012.

[19] M. Sugito, *Word Accent in Japanese and English*. Hituzi Syobo, 2012.

[20] S. Yamaziki, “Supeingo-oto no topikkusu - kokonomi no kyojuho,” in *Supeingo Sekai no Ktoba to Bunka- Conferencias sobre la lengua y cultura del mundo de habla hispana*, 1991, pp. 141–157.