INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Cross-linguistic study of the production of turn-taking cues in American English and Argentine Spanish

Pablo Brusco [1,2], *Juan Manuel Pérez* [1,2], *Agustín Gravano* [1,2]

[1] Departamento de Computación, FCEyN, Universidad de Buenos Aires, Argentina
[2] Instituto de Investigación en Ciencias de la Computación, CONICET-UBA, Buenos Aires, Argentina

pbrusco@dc.uba.ar, jmperez@dc.uba.ar, gravano@dc.uba.ar

Abstract

We present the results of a series of machine learning experiments aimed at exploring the differences and similarities in the production of turn-taking cues in American English and Argentine Spanish. An analysis of prosodic features automatically extracted from 21 dyadic conversations (12 En, 9 Sp) revealed that, when signaling Holds, speakers of both languages tend to use roughly the same combination of cues, characterized by a sustained final intonation, a shorter duration of turn-final inter-pausal units, and a distinct voice quality. However, in speech preceding Smooth Switches or Backchannels, we observe the existence of the same set of prosodic turn-taking cues in both languages, although the ways in which these cues are combined together to form complex signals differ. Still, we find that these differences do not degrade below chance the performance of cross-linguistic systems for automatically detecting turn-taking signals. These results are relevant to the construction of multilingual spoken dialogue systems, which need to adapt not only their ASR modules but also the way prosodic turn-taking cues are synthesized and recognized.

Index Terms: turn-taking, dialogue, prosody, cross-linguistic.

1. Introduction

Production and **perception** of turn-transition cues in dialogues is a topic of great interest in the computational linguistics research area. The understanding of **when** and **what kind** of cues speakers produce when driving a conversation is an unsolved problem addressed by different points of view and techniques. These techniques vary from more descriptive analysis with small corpora where results are analysed by examining examples to more robust results with larger corpora permitting more statistically significant results.

Early hypotheses on the mechanics of turn-taking include influential work by Sacks et al. [1], who propose that turn-taking allocation is controlled by a set of fixed rules, and by Duncan [2], who suggests that participants give a number of cues in order to handle turn-taking, which could be of prosodic, syntactic or even gestural nature. Also, these cues are not given in solitude, but combine together to form complex signals. Further studies in this direction have formalized and reinforced these original ideas on turn-taking cues, confirming that acoustic/prosodic and syntactic features contribute importantly to the turn-allocation mechanism [3, 4, 5, inter alia]. Along these lines, in [6] the authors investigate which acoustic, prosodic or syntactic cues can be automatically extracted from the speech signal, and find strong evidence of seven turn-yielding cues and six backchannel-preceding cues in Standard American English.

Turn-taking cues have also been studied in other languages, including Swedish [7], Japanese [8] and German [9], among others. Cross-lingual comparisons of turn-taking behaviors are less frequent, and belong mostly to the anthropological literature. Some studies claim that cultures strongly deviate in these turn-taking systems [10], whereas others argument that there exists something like a 'universal' for turn-taking [11]. A quantitative analysis of response offsets in turn transitions from a sample of ten different languages [12] provides some support to these 'universals', showing the response distributions to be very similar across different languages.

A recent cross-linguistic study investigates the perception of prosodic cues in Slovak and Argentine Spanish [13]. The authors show once more that prosody plays a clear role in the anticipation of turn-taking transitions, and also that these two languages overlap to some extent despite belonging to different linguistic families. Subjects who did not speak one of the languages were still able to predict the upcoming turn-transition type with better-than-chance accuracy. This contributes evidence in favor of the aforementioned turn-taking 'universals'.

In this work we study whether the acoustic/prosodic mechanisms that signal upcoming turn-taking transitions are shared among two different languages – American English and Argentine Spanish. In particular, we are interested in the **production** of such cues in unrestricted conversation. A number of subquestions arise in this context: How similar are the distributions of acoustic/prosodic features in these two languages? Are these features equally important as predictors? Can a machine learning classifier trained in one language be used successfully in another? To answer these questions, we perform a number of machine learning experiments on two similar corpora of task-oriented spontaneous dialogue.

2. Materials and Methods

2.1. Speech corpora

For this work, we used two versions of the Objects Games Corpus (first described in [6]), in Standard American English and in Argentine Spanish. Each session consisted of 15 to 30 instances of the Objects Game, in which one subject was instructed to describe the position of a target object on her screen to the other subject, whose task was to position the same object on her own screen. Subjects alternated in the describer and follower roles. At the end of the session, subjects were paid a fixed amount of money for their participation, plus a bonus based on the number of awarded points.

The Spanish corpus was recorded at the University of Buenos Aires in April, 2014. A total of 20 subjects (10F, 10M) participated in the study in 10 sessions; no subjects repeated the experiment. Their ages ranged from 19 to 43 years ($M = 26.4$, $SD = 6.3$). All subjects were native speakers of Argentine Spanish, lived in the Buenos Aires area at the time of the study, and agreed to join the study by signing a consent

http://dx.doi.org/10.21437/Interspeech.2017-124

Table 1: *IPU counts for each turn-taking transition type, with mean and standard deviation per speaker.*

	English Corpus			Spanish Corpus		
	BC	H	S	BC	H	S
count	**553**	**8125**	**3246**	**663**	**3725**	**2278**
mean	23.0	338.5	135.2	36.8	206.9	126.5
std	16.7	113.9	47.8	28.1	120.4	81.6

form. Together with the audio recordings, electroencephalography (EEG) recordings were taken from the subjects. The English data were obtained from the 258 minutes taken from the Objects games in the Columbia Games Corpus [6]. In this case, 13 subjects (6F, 7M) with ages between 20 and 50 ($M = 30.0$, $SD = 10.9$), all native speakers of Standard American English (SAE), participated in 12 sessions in total [6].

2.2. Unit of analysis and acoustic/prosodic features

Following previous work, we define an INTER-PAUSAL UNIT (IPU) as a maximal speech segment from a single speaker that is surrounded by pauses longer than a specified threshold, 50ms in the English corpus and 100ms in the Spanish one (due to differences in the transcription procedures). IPUs in both corpora were manually aligned to the audio signal by trained annotators.

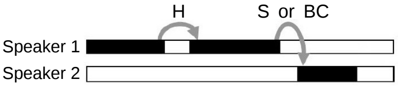


Figure 1: *IPUs transitions*

In the Spanish corpus, all transitions from one IPU to the next were manually labeled by a professional annotator following the labeling scheme described in [6]. Turn-taking transitions from the English database were obtained from the original corpus. Figure 1 illustrates the three turn-taking transitions we study in the present work: a HOLD (H) takes place when the current speaker continues talking after a short pause; in a SMOOTH SWITCH (S) the other speaker starts talking after a short pause; a BACKCHANNEL (BC) is a short utterance such as *uh-huh* used to display attention and invite the current speaker to carry on talking [14]. Note that none of these labels include overlapping speech. Table 1 summarizes the IPU counts in both corpora.

The turn-taking transitions we analyze in this work occur either between two adjacent IPUs from the same speaker (H), or between an IPU from one speaker and a subsequent IPU from the other speaker (S and BC). As in [6], we analyze turn-taking cues that can be automatically extracted from the final portion of the IPU that precedes a transition, assuming that this speech segment contains information useful for predicting the following transition type.

We extracted a number of acoustic/prosodic features that have shown to effectively capture differences between turn-taking events using the same procedures described in [6], from the final 200, 300 and 500 ms of each IPU: F0 slope (as an estimate of final intonation); mean intensity level; mean pitch level; and voice quality features such as jitter, shimmer and noise-to-harmonics ratio (NHR). Additionally, we computed the IPU duration in ms. We did not include speech rate or text-based features, since the orthographic transcription of the Spanish corpus has not been yet completed. To avoid subject-specific characteristics, all features were speaker-normalized using z-scores.

2.3. Classification tasks

We conducted several machine learning (ML) experiments, in which models were trained to classify IPUs in different ways, based on their extracted features. We used the ensemble method RANDOM FOREST CLASSIFIER (RF) [15] due to the simplicity and transparency of the resulting models that has proven to perform well in this kind of tasks. In particular, we were interested in this algorithm's straightforward assessment of the relative predictive power of individual features based on the mean impurity decrease of features over all the constructed trees. For this purpose, we used the implementation of this algorithm included in the Python open-source library *scikit-learn* [16].

Missing values in the data (caused e.g. by undefined pitch values) were filled with the feature's median before running the classification experiments. Since filling missing values with certain numbers may harm a classifier's performance, a binary indicator for each feature was added to the dataset, encoding whether a value was filled with a column median or not [17].

In order to avoid overestimated results, all classification experiments were performed using the leave-one-speaker-out cross-validation method. That is, in each iteration the data from one speaker are used for validation, and the other speakers' data are used for building the model. This process is repeated by varying the left-out speaker; in consequence, all instances from all speakers are used for validation exactly once.[1] We measure model performance with ROC curves, by combining the probability scores from the instances in all validation sets. We also compute the area under the curve (AUC), which summarizes a ROC curve in a single number (AUC $= 1.0$ is a perfect ordering; 0.5 is chance).[2]

3. Research questions and experiments

The main purpose of this study is analyzing the similarities between English and Spanish prosodic turn-taking cues. Here we present a series of experiments to approach this issue from different viewpoints.

3.1. Distribution analysis by language

The first question we address is, **Q1.** *How do the distributions of the features extracted from IPUs preceding each turn-taking category compare in Spanish and English?* That is, if we look at how speakers produce each turn-taking cue in each language, what differences do we find?

Our first experiment **(E1)** consists in visually comparing the distributions of the extracted features preceding each turn-taking transition in both languages, as summarized in Figure 2. Each plot shows the approximated probability density function of a given z-score-normalized feature for each turn-taking label. In English, as reported in [6], clear differences are found for all seven features between S and H (i.e., turn-yielding cues), and also for intonation, pitch and intensity levels, IPU duration and NHR between BC and H (backchannel-inviting cues).

In Spanish, we observe remarkably similar distributions for the S and H labels for IPU duration, intensity level and voice quality features, suggesting that both languages share many aspects of the acoustic/prosodic realization of turn-yielding cues. In both languages, Hold transitions are characterized by a final

[1] Note that a control set was not used, for the same reason we selected a simple model over state-of-the-art techniques such as deep neural networks: Our focus is not on achieving a competitive classification performance on new data, but on drawing conclusions from relative, cross-class comparisons.

[2] We used AUC over metrics such as accuracy, recall or F-score since the ROC measures how the model orders the instances. Then, depending on the requirements of a real system, a threshold must be set so the model finally assigns a label to each instance.

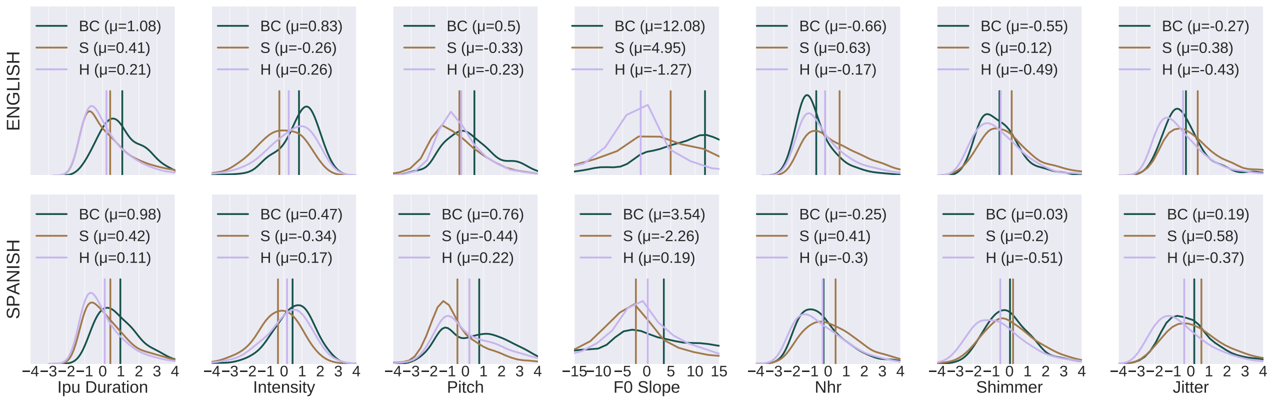


Figure 2: *Distributions of z-score-normalized features. These selected visualizations correspond to features extracted from the IPU-final* $500ms$ *in the case of intensity, pitch, NHR, shimmer and jitter, and from the IPU-final* $300ms$ *in the case of F0 slope.*

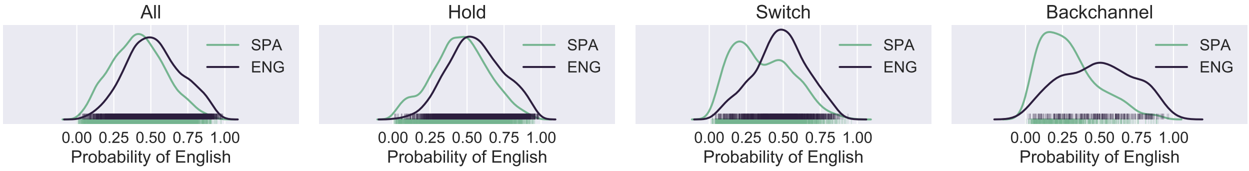


Figure 3: *Language discrimination. In each plot, the x-axis is the posterior probability given by the language classification model; at the bottom, a Raster plot that shows the punctuation given to each individual instance, separated by true language. The curves are the approximated probability density functions computed from the instances' posteriors.*

plateau intonation (F0 slope close to zero), while Switch transitions present a nearly flat distribution in English (meaning that any final intonation is equally likely) but a strong tendency for a falling final intonation in Spanish. The observed differences for final F0 slope in the three conditions are likely a consequence of the distinct prosodic characteristics of these two languages.

For the BC category there are a few differences: in Spanish, backchannel-preceding cues are less likely to have a high intensity and a rising final intonation than in English. Pitch level shows an interesting bimodal distribution in Spanish, suggesting that speech before a backchannel may have a region of either high or low pitch, as opposed to English, in which a low pitch level is prevalent [18]. For intensity and voice quality features, the distributions for BC are similar in both languages, though slightly more distinct in English.

3.2. Language discrimination

So far, we have observed similarities and differences among the languages in the production of individual turn-taking cues. In addition, interactions between some of the variables may exists, adding to the similarities and differences observed for individual features. Our a second question then is, **Q2.** *Are these cues (individual and combined) different enough to distinguish between the two languages?* If the answer to this question is affirmative, we may be able to build an automatic classifier that, given the features extracted from an IPU, predicts if the IPU was produced by a Spanish speaker or an English speaker.

For our second experiment **(E2)**, we merged the data from all subjects in both corpora into one big data set and built a classifier to predict the original language of each IPU based only on the extracted features. We hypothesized that, if the two languages shared identical turn-taking cues, the classification would result in near random output – i.e., a ML algorithm would fail to discriminate the language based on this input. Conversely, a better-than-chance classification would indicate significant differences between the two languages.

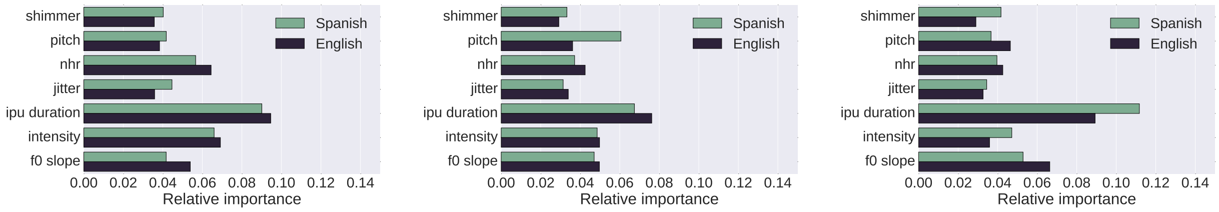
Figure 3 shows a distribution plot of the posterior probabilities of the ML classifiers when predicting the language of all instances (first panel on the left), and separately for each turn-taking transition (the rest of the panels). The two colors indicate the instances' actual language. If the classifications were perfect, we would observe two perfectly separate distributions.

In the first panel of Figure 3, we see that the distributions are not clearly separate, but still some information seems to have been effectively captured by the classifiers, since the resulting AUC is above chance (AUC=0.64). When looking at each transition type in isolation (the remaining panels in Figure 3), we observe that for the Hold transition type, the two distributions are most similar (AUC=0.62). For the S and BC categories, the distributions are clearly different (AUC=0.71 and AUC=0.70, respectively).

These differences found in the classification performance for different turn-taking labels can be explained by the similarities and differences described in section 3.1. Acoustic/prosodic features from IPUs preceding H have relatively similar means and distributions among the two languages. However, the differences found for BC and S seem to indicate that turn-yielding and backchannel-preceding cues do present significant differences in English and Spanish, despite the similarities found in the previous section.

3.3. Cross linguistic comparisons

In the previous two experiments we saw that some of the acoustic/prosodic turn-yielding cues have different distributions in English and Spanish. Next we ask **Q3.** *Is the relative importance of these cues equal in the two languages?* In other words, do the features provide the same amount of information for predicting the type of turn-taking transition in either language? To address this question we conducted a third experiment **(E3)** in which three binary RF classifiers were trained for each language



(a) *S vs. H classifier*

(b) *S vs. BC classifier*

(c) *BC vs. H classifier*

Figure 4: *Comparison of feature importance in binary classification tasks, as determined by the RF classifiers.*

Table 2: *AUC of the ROC curve and performance difference when testing in cross-language settings. The first column in each table represents the performance when training and testing on the same language. The second column represents the performance when training on a different language.*

	Test on Spanish Data			*Test on English Data*		
Classifier	Same	Cross	Diff	Same	Cross	Diff
BC vs H	0.76	0.70	-8.5%	0.83	0.76	-9.2%
S vs H	0.68	0.58	-17.2%	0.71	0.65	-9.2%
S vs BC	0.75	0.73	-2.7%	0.81	0.76	-6.5%

for classifying **S vs. H**, **S vs. BC** and **BC vs. H**. [3]

We performed a grid search varying different combinations of parameters for finding the optimal ones for each RF model. In Table 2, the 'Same' columns show the AUC values for these classifiers. After training the models, we measured the relative importance these classifiers gave to each feature. Figure 4 shows the relative importance of the different features in each binary classification task, for English and Spanish.

We can see that, in general, the relative importance assigned to the features by the classifiers is similar in both languages. Nevertheless, a couple of exceptions are found. For example, in the Spanish classifier for S vs. BC, the pitch level predominates, together with IPU duration over the other features. This kind of dissimilarities may be explained by looking again at the distributions (Figure 2), where the difference between S and BC for pitch level is greater in Spanish than in English.

Since the relative importance of features was comparable across the two languages, as were the feature distributions, we ask our final question, **Q4.** *Are the rules learned in a language general enough to apply to the other language?* That is, are the patterns learned by a classifier in Spanish useful for classifying turn-taking cues in English, and vice versa? To answer this question we run our last experiment **(E4)**, in which we use the trained classifiers described above, and test their performance on instances from the other language.

Table 2 summarizes the performance of classifiers trained on data from a given language, and later tested on data from either the same or the other language. In general, the results show a decrease in performance when changing language. Interestingly, in all cases this decrease is not strong enough to render the classifiers useless, since all results are above chance level (AUC=0.5). Rather unexpectedly, the decrease for the S vs. BC classifier is very small when training on English data and testing on Spanish data. This may seem counter-intuitive since S and BC were the classes that showed greater differences when predicting the language (see Figure 3). Nevertheless, the patterns a language classifier (E2) found in the data could be independent to the patterns that distinguish S vs. BC in both languages. For example, analyzing the relative importance of the features according to the language classifier (not shown here due to space limitations), we note that, for predicting the language of backchannel-preceding IPUs, f0-slope seems to be the most important feature, in contrast with what happens with the S vs. BC classifier, for which f0-slope does not have the same relevance (see Figure 2).

Finally, the decrease in performance when training a classifier on a language and testing on the other is not symmetrical. That is, training on Spanish and testing on English seems to be different from doing the opposite. A possible explanation for this could be that the rules learned by a classifier for a given language may generalize well to the other, because e.g. the former language has a richer inventory of turn-taking cues. In this way, the opposite would not occur, since once a classifier has learned a more specific set of rules for one language, it will not generalize well to the other.

4. Conclusions

We conducted a number of experiments to explore similarities and differences between American English and Argentine Spanish in the production of acoustic/prosodic cues before turn exchanges. After analyzing the speech in two corpora of spontaneous dyadic conversations, we found that when signaling a Hold transition, speakers tend to use the same combinations of cues in both languages. When preceding a smooth switch or a backchannel, cues are also produced in a similar, yet not identical manner. Still, the observed differences are not large enough to prevent our cross-linguistic classifiers (i.e., classifiers trained on a language, tested on the other) from achieving better-than-chance results. These results indicate that American English and Argentine Spanish, despite belonging to different linguistic families, share some of the way acoustic/prosodic turn-taking cues are realized. In the future, these results could be relevant when building spoken dialogue systems for new languages, especially under-resourced ones, since they show that the turn-taking module could be borrowed from a different language as an initial implementation, with better-than-random accuracy. It remains an open question to determine how the observed differences between the two languages will affect the user experience with real spoken dialog system.

5. Acknowledgments

Work partially supported by ANPCYT PICT 2014-1561, Bilateral Cooperation Program CONICET-SAS, and the Air Force Office of Scientific Research, Air Force Material Command, USAF under Award No. FA9550-15-1-0055. The authors thank Luciana Ferrer, Ramiro H. Gálvez, Juan Kamienkowski and anonymous reviewers for valuable suggestions and comments.

[3] Binary comparisons were preferred over multiclass ones because we used discriminative models. These techniques do not model each class separately and thus, our post-hoc analyses of individual turn-transition types would be harder with multiclass classification.

6. References

[1] H. Sacks, E. A. Schegloff, and G. Jefferson, "A simplest systematics for the organization of turn-taking for conversation," *language*, pp. 696–735, 1974.

[2] S. Duncan and D. Fiske, "Face-to-face interaction: research, methods and theory," 1977.

[3] C. E. Ford and S. A. Thompson, "Interactional units in conversation: syntactic, intonational, and pragmatic resources for the management of turns," *Studies in interactional sociolinguistics*, vol. 13, pp. 134–184, 1996.

[4] A. Wennerstrom and A. F. Siegel, "Keeping the floor in multiparty conversations: Intonation, syntax, and pause," *Discourse Processes*, vol. 36, no. 2, pp. 77–107, 2003.

[5] A. Stolcke, L. Ferrer, and E. Shriberg, "Is the speaker done yet? faster and more accurate end-of-utterance detection using prosody," 2002.

[6] A. Gravano and J. Hirschberg, "Turn-taking cues in task-oriented dialogue," *Computer Speech & Language*, vol. 25, no. 3, pp. 601–634, 2011.

[7] A. Hjalmarsson, "The additive effect of turn-taking cues in human and synthetic voice," *Speech Communication*, vol. 53, pp. 23–25, 2011.

[8] H. Koiso, Y. Horiuchi, S. Tutiya, A. Ichikawa, and Y. Den, "An analysis of turn-taking and backchannels based on prosodic and syntactic features in japanese map task dialogs," *Language and speech*, vol. 41, no. 3-4, pp. 295–321, 1998.

[9] D. Schlangen, "From reaction to prediction: Experiments with computational models of turn-taking," *Proceedings of Interspeech 2006*, 2006.

[10] R. Bauman and J. Sherzer, *Explorations in the Ethnography of Speaking*. Cambridge University Press, 1989, no. 8.

[11] E. A. Schegloff, "Interaction: The infrastructure for social institutions, the natural ecological niche for language, and the arena in which culture is enacted," *Roots of human sociality: Culture, cognition and interaction*, pp. 70–96, 2006.

[12] T. Stivers, N. J. Enfield, P. Brown, C. Englert, M. Hayashi, T. Heinemann, G. Hoymann, F. Rossano, J. P. De Ruiter, K.-E. Yoon *et al.*, "Universals and cultural variation in turn-taking in conversation," *Proceedings of the National Academy of Sciences*, vol. 106, no. 26, pp. 10 587–10 592, 2009.

[13] A. Gravano, P. Brusco, and Š. Beňuš, "Who do you think will speak next? perception of turn-taking cues in slovak and argentine spanish," *Interspeech 2016*, pp. 1265–1269, 2016.

[14] A. Gravano, J. Hirschberg, and Š. Beňuš, "Affirmative cue words in task-oriented dialogue," *Computational Linguistics*, vol. 38, no. 1, pp. 1–39, 2012.

[15] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[16] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[17] J. Cohen, P. Cohen, S. G. West, and L. S. Aiken, *Applied multiple regression/correlation analysis for the behavioral sciences*. Routledge, 2013.

[18] N. Ward and W. Tsukahara, "Prosodic features which cue backchannel responses in English and Japanese," *Journal of Pragmatics*, vol. 32, no. 8, pp. 1177–1207, 2000.