INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Who do you think will speak next? Perception of turn-taking cues in Slovak and Argentine Spanish

Agustín Gravano[1,2], *Pablo Brusco*[1,2], *Štefan Beňuš*[3,4]

[1] Departamento de Computación, FCEyN, Universidad de Buenos Aires, Argentina
[2] Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Buenos Aires, Argentina
[3] Constantine the Philosopher University in Nitra, Slovakia
[4] Institute of Informatics, Slovak Academy of Sciences, Slovakia
gravano@dc.uba.ar, pbrusco@dc.uba.ar, sbenus@ukf.sk

Abstract

We investigate perceptual cues in human-human dialogue management related to signalling the change of speaker and the interlocutor's wish to backchannel or contribute with propositional content. We are interested primarily in the relevance of prosodic cues in relation to textual ones, and their cross-linguistic validity by comparing unrelated languages Slovak and Argentine Spanish. Results of a perception study indicate that 1) in addition to textual cues, prosodic cues also play a clear role in perceiving how the dialogue will unfold; and 2) there exists a non-empty intersection of temporal and intonational prosodic turn-taking cues in the two languages, despite their belonging to separate families.
Index Terms: turn-taking, dialogue, prosody, cross-linguistic.

1. Introduction

Understanding (and predicting) ***when*** a next turn takes place and ***what*** pragmatic function it will have are indispensable in everyday human-human spoken interactions and are thus crucial also for building natural Spoken Dialogue Systems (SDSs). On the recognition side, an SDS must identify cues with which humans signal possible and proper places for it to say something. Moreover, an SDS should be also able to predict the pragmatic nature of the expected utterance; for example should the SDS provide only a backchannel indicating its continued attention or a full utterance with some propositional content. Similarly, on the production side, an SDS must properly cue the transition relevance places (TRP [1]), and if possible, also signal what kind of response (e.g., backchannel or a turn switch) is expected to be produced by the human.

The ***when*** question has been investigated mostly through identifying cues for turn-end prediction. While the importance of syntactic and semantic/pragmatic completion cues on turn-end prediction has been well established, the role of prosody remains less clear. Some studies probing on-line processing argue that lexical and syntactic information may be sufficient for turn-end projection, and that information encoded in the prosody of the utterances is neither necessary nor sufficient for this task [2]. Another set of studies, analyzing records of natural spoken interactions as well as on-line scenarios such as button press, suggest that prosody also provides systematic cues for turn-management in spoken interactions [3, 4, 5, 6, 7, 8].

The ***what*** question is equally complex and challenging. There is evidence that temporal (e.g., lengthening, speech rate) and intonational (e.g., pitch accents, phrase tones) prosodic cues are useful in predicting the metacognitive states of conversational partners in human-human interactions. For example, [9] showed that the intonation of the answer (rising or falling), the duration of silent pause between the end of a question and the initiation of an answer, and the presence of a filled pause (*um*, *uh*) before that answer all facilitated the feeling of another's knowing, that is, the prediction that the speaker knew the answer to the question. Recent studies show that humans use prosodic cues to predict the pragmatic functions of utterances well before the end of the turn and that both temporal and intonational cues play a role ([10] and references therein).

Given the still unclear role of prosody in both ***when*** and ***what*** questions, the current study examines how speakers perceive prosodic cues conveying that the current speaker will hold the floor and continue speaking (HOLD), the current speaker's interlocutor will take the floor to contribute to the dialogue (SWITCH), or that the interlocutor will indicate her continued attention and/or understanding by providing a backchannel such as *mhm* or *okay* (BACKCHANNEL). Hence, we study the ***when*** question by exploring cues for whether the turn-transition occurs (Hold vs. Switch/Backchannel) and the ***what*** question by examining the predictability of the upcoming Switch or Backchannel.

Furthermore, we are interested in better understanding the language-specific and cross-linguistically universal prosodic patterns in conveying these pragmatic functions associated with turn-taking. We thus compare two languages from separate families with minimum language contact – Slovak (SK, Slavic) and Argentine Spanish (ES, Romance), and analyze the way they employ prosody for signalling their communicative intentions and for coordinating joint actions through turn-taking. We study these questions based on data from human-human interactions and seek better understanding of human cognition, but by doing this we hope to facilitate the design of better models for representing how prosody participates in the cognitive model for interactional turn-taking among humans and machines in SDSs with improved cognitive alignment between the user and the system [11].

2. Materials and method

To investigate these questions, we designed a novel scenario for a perception study in which subjects are given an utterance from a corpus of conversational speech and are asked to guess if the speech following this utterance will be a Hold (H), a Switch (S), or a Backchannel (BC). We manipulate three variables: 1) the

http://dx.doi.org/10.21437/Interspeech.2016-585

language of the subject (ES/SK) in a between-subject design, and 2) the language of the speaker (ES/SK) and 3) the speech presentation (text, audio, or masked de-lexicalized audio) in an adjusted within-subjects design. We now describe our methodology in more detail, together with a summary of our hypotheses and research questions.

2.1. Data preparation

Stimuli were taken from two comparable corpora of task-oriented conversational dialogues in Argentine Spanish and Slovak [12]. In both corpora, subjects were asked to play a series of computer games designed to require cooperation and communication in order to achieve a high score. All games were played on separate laptops whose screens were not visible to the other player; the players were separated by a curtain so that all communication would be vocal. Currently, the Slovak corpus includes 9 dyadic sessions with a total of 11 speakers (5F, 6M); the Spanish corpus, 7 dyadic sessions with a total of 12 native speakers of Argentine Spanish (7F, 5M). In both cases, subjects participated in either one or two sessions. Finally, an INTER-PAUSAL UNIT (IPU) is defined as a pause-free chunk of speech from a single speaker. The identification of inter-pausal units (IPUs) in both corpora has been done before [12].

2.1.1. Stimuli selection

We first extracted from each dialogue all pairs of adjacent IPUs (from same or different speakers) such that a) there is no overlap between them, and b) there is no further speech from either speaker between the beginning of IPU1 and the end of IPU2. Figure 1 illustrates two IPU pairs; the first one is formed by two IPUs from a same speaker, and is thus an instance of a Hold transition; the second one contains IPUs from two different speakers, and is thus an instance of either a Switch or a Backchannel. IPUs have a mean duration of 1.47s (SD 1.02s) in the Spanish data, and 1.32s (SD 0.93s) in the Slovak data.

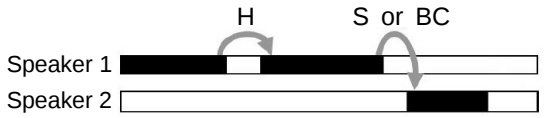

Figure 1: *Two instances of IPU pairs. Black segments represent speech; white segments, silence. The first pair is a Hold transition; the second is either a Switch or a Backchannel.*

Next, our pool of IPU pairs was shuffled to random order. A subset of IPU pairs were manually labeled by the authors using a scheme for turn-exchange annotation [6], and those labeled as H, S and BC were retained as the ground truth for the current study. This procedure yielded a total of 124 S, 88 H and 47 BC pairs for Spanish, and 134 S, 90 H and 38 BC pairs for Slovak. Subsequently, only the first IPUs in these pairs (i.e., IPU1) was used as stimuli in our perception study. The second IPUs were never presented to subjects.

2.1.2. Stimuli presentation

To examine the role of prosody in the potential cuing of the turn-management of upcoming speech, we employed three conditions for stimuli presentation – TEXT TRANSCRIPT, ORIGINAL AUDIO, and MASKED DE-LEXICALIZED AUDIO.

Text transcripts did not include any capitalization or punctuation marks. Word fragments were signalled with a dash after the last recognizable phone (e.g., *est-*); and filled pauses were transcribed phonetically (e.g., *mm*, *eh*). For stimuli in the original audio condition, we first normalized the intensity of all dialogue files in both corpora to 60dB, and extracted individual IPUs. Stimuli were saved as 16-bit, 16kHz wav files.

The masked stimuli were generated in the following steps. The audio for each stimulus IPU was low-pass filtered using the thresholds of $1.75 \times \overline{f0}$ for females and $2.75 \times \overline{f0}$ for males, where $\overline{f0}$ is the speaker's mean pitch computed over the whole corpus session. These thresholds were determined through experimenting with several values over a subset of the stimulus tokens to minimize the intelligibility of the speech and maximize the perception that the sound clip is speech. Finally, the volume of each stimulus was increased 1.5 times since low-pass filtering decreased the volume perceivably. Subsequently, we checked the intelligibility of the stimuli tokens by asking naive listeners (two for each language) to listen to all stimuli in their respective language and mark any words they recognized. The recall was extremely low (0.4% in Slovak and 0.5% in Spanish).

2.2. Web interface

With these stimuli, we created a web-based perception study in which subjects self-identified their native language and other demographic information (age, gender, education level, and language background) and were then asked to proceed to the test.

The test was formed by two separate surveys. In each survey, subjects were asked to classify into H, S or BC 21 stimuli picked at random from the stimulus set described in the previous section. All 21 stimuli in a survey were in the same language (ES or SK) and were presented in the same condition (text, original audio or masked audio). These two dimensions, combined with the subject's native language (ES or SK) yields a total of 12 groups, as shown in Table 1. For example, in a given test the first survey could contain text transcriptions in Slovak, and the second survey, masked audios in Spanish.

Once a survey started, subjects were asked to read or listen to each of 21 stimuli and select which type of turn-taking transition they thought would take place (H, S or BC). The H transition was defined to subjects as *"The same person will continue talking after a short silence"*; S as *"The other person will start talking after a short silence"*; and BC as *"The other person will utter a short expression such as 'uh-huh', 'ok' or 'yeah', and afterwards the first person will continue talking."* Subjects could read or listen to the stimuli as many times as they wanted, and there were no time constraints.

The stimulus language and the presentation condition of

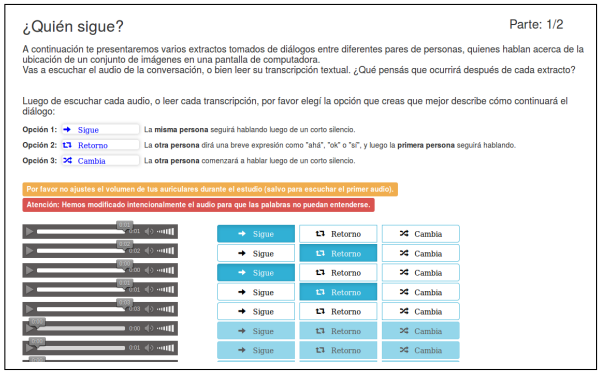

Figure 2: Web interface for one of the audio conditions.

Stimulus language:	Argentine Spanish (ES)			Slovak (SK)		
Stimulus presentation:	Text	Original audio	Masked audio	Text	Original audio	Masked audio
Subject language: ES	ES-ES-text	ES-ES-audio	ES-ES-masked	ES-SK-text	ES-SK-audio	ES-SK-masked
Subject language: SK	SK-ES-text	SK-ES-audio	SK-ES-masked	SK-SK-text	SK-SK-audio	SK-SK-masked

Table 1: The 12 groups defined for our perception study. Each group is a combination of a subject language (ES / SK), a stimulus language (ES / SK) and a stimulus condition (text / audio / masked). (Colors match those in Figure 3.)

each survey were selected at random, with two restrictions. First, the two surveys assigned to each subject must be different. Second, the total number of subjects that fell into each group in Table 1 should be balanced. To achieve statistical significance we aimed at 20 subjects in each group. Since each subject was to be assigned two surveys, this means we needed 60 native speakers of each language to complete the test.

The 21 stimuli in each survey were randomly selected from the IPU pool described in Section 2.1.1. Seven stimuli were chosen for each turn-taking category (H, S and BC) according to our ground truth, shuffled and presented to the subject as depicted in Figure 2. The random selection of these stimuli was performed so as to balance the gender of the speaker, and maximize the number of different speakers (that is, minimize the number of speaker repetitions).

2.3. Statistical analysis

With the data collected using the methodology described above, the hypotheses of the current paper relate to the relevance of prosody in predicting the upcoming speech and the extent to which prosodic cues are (in)dependent of a particular language or culture of the interlocutors. The primary means of testing these hypotheses is the comparison of subjects' mean accuracies in each of the 12 groups defined in Table 1.

First, using one-sample t-tests, we compare the mean accuracy of a particular group against the accuracy of random guessing, 0.333 (since stimuli were balanced in each group, with $1/3$ stimuli for each turn-taking label). Significant deviances from random guessing would suggest that stimuli's turn-taking cues aided subjects in discriminating the type of the upcoming turn-taking transition.

The second type of test is intended to compare the mean accuracy of two particular groups. In this case, significant differences would suggest that turn-taking cues contained in the stimuli in one group were more informative than those in the other group. We used two-sample t-tests for this purpose.

3. Results

Our perception test was completed by 59 native speakers of Argentine Spanish (20F, 39M; age mean 27.7, stdev 7.1) and by 61 native speakers of Slovak (48F, 13M; age mean 24.9, stdev 8.5), between November, 2015 and February, 2016. 19 subjects in the latter group reported some knowledge of Spanish, and thus were excluded from our cross-linguistic analyses.

3.1. Turn-taking cues

Figure 3 shows the overall mean accuracy obtained by subjects in each of the 12 groups defined in Table 1. Statistical tests were performed to compare individual groups against random accuracy (0.333), and accuracies of relevant pairs of groups. N is the number of subjects in each group, and statistical significance is signalled with a star ($\star$) when $p < 0.05$, or with a dot ($\bullet$) when $0.05 \leq p < 0.10$; "n.s." stands for "non-significant". These results are discussed next.

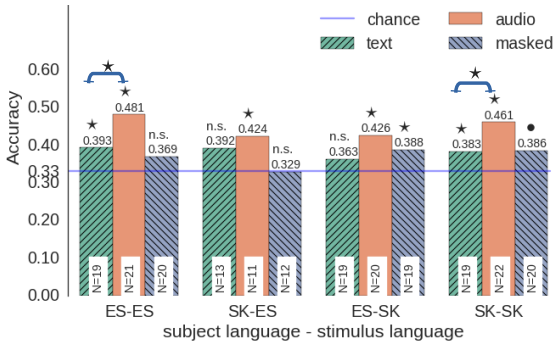

Figure 3: Overall accuracy of subjects in each group.

We first observe that subjects were able to distinguish the turn-taking category of the upcoming transition (H, BC or S) in their native language significantly better than chance, when they listened to the audio versions of our stimuli (ES-ES-audio, $p < 0.001$; SK-SK-audio, $p < 0.001$). This indicates that despite the short duration of stimuli, these do contain turn-taking cues perceptible by listeners.

Subjects' accuracies were also significantly better than chance when presented with orthographic transcriptions of the stimuli (ES-ES-text, $p < 0.05$; SK-SK-text, $p < 0.05$). This is in line with previous studies which show the importance of lexical and syntactic information as turn-taking cues for other languages [2, 6, 7].

Next we assess the relevance of **prosodic** cues by comparing the performance of subjects when judging stimuli in audio and text conditions. We assume that whatever decline found in performance may be attributed to information present in the audio signal that is lost when making only the transcriptions available. For both languages, such a decline is significant (ES-ES-audio vs. -text, $p < 0.05$; SK-SK-audio vs. -text, $p < 0.05$), thus verifying the existence of prosodic cues.

A different approach to studying the relevance of prosodic cues consists in analyzing the accuracies for the masked audios, assuming that in this condition all lexical/syntactic information has been removed, leaving just some prosodic information in the signal. In this case, Spanish subjects were not able to significantly outperform chance (ES-ES-masked, $p \approx 0.15$) and Slovak subjects only approached significance (SK-SK-masked, $p \approx 0.07$). Temporal features such as speaking rate and phrase-final lengthening are well known turn-taking cues in other languages [6, 7, 13]. While pitch and intensity are largely preserved in our masking procedure, segmental information is removed. This results in significant degradation of the temporal cues such as pre-final lengthening or speech rate variation. Therefore, the masked-audio results seem to suggest that temporal features (either alone or in combination with pitch/intensity) play an important role in cueing turn-taking in Slovak and Spanish.

Now we shift our attention to cross-linguistic perception of turn-taking cues.[1] First, as a sanity check, we observe that subjects failed to significantly outperform chance when judging the text transcriptions of stimuli in a foreign language unknown to them (ES-SK-text, $p \approx 0.2$; SK-ES-text, $p \approx 0.19$). This is understandable since we assume that subjects lack any syntactic, lexical, or prosodic cues in this task and can use only features such as IPU length in words.

Interestingly, when subjects listened to audio stimuli in a foreign language unknown to them, they still managed to discriminate turn-taking types significantly better than chance (ES-SK-audio, $p < 0.005$; SK-ES-audio, $p < 0.05$). In this scenario, subjects did not understand the words in the audios, but still managed to form a better-than-random guess of how the dialogue would continue. This suggests that there exists a non-empty intersection of temporal and intonational prosodic turn-taking cues in the two languages, despite their belonging to separate families.

Last, we analyzed how subjects performed when listening to masked stimuli in a foreign language, and we obtained somewhat puzzling, asymmetrical results. While speakers of Spanish were rather successful in classifying Slovak masked stimuli (ES-SK-masked, $p < 0.05$), the same was not true for Slovak speakers judging Spanish masked stimuli (SK-ES-masked, $p \approx 0.92$). Recall from above that Spanish stimuli in the masked condition were also harder to discriminate than Slovak stimuli by native speakers of either language (ES-ES-masked and SK-SK-masked groups, respectively). Although more research is needed to better understand these results, they suggest that temporal information (which is removed by our masking procedure) has different roles as a turn-taking cue in these two languages.

3.2. Identification of individual turn-taking categories

To investigate how well subjects identified the three turn-taking categories under study (H, S, BC), we computed the F_1 score (or F-measure) for each label.[2] Figure 4 shows that upcoming H transitions were the easiest to identify by native speakers of both languages in all three study conditions. S and BC appear to have been equally difficult to identify in most cases, with the exception of the ES-ES-audio group, in which the performance for S was clearly better.

There are two additional partial observations. First, Slovak Switches seem to be cued by prosody to a significant extent since there is no deterioration from audio to masked audio. But in Spanish, there is sizeable deterioration and the identification of Switches in Spanish possibly needs both textual and prosodic information. Second, upcoming Backchannels are surprisingly difficult to predict in the Spanish full audio condition, and the only type for which perception actually improves from full to masked audio.

4. Discussion and conclusions

In this paper we investigated the perceptual cues in human-human dialogue management related to signalling the change of speaker (Hold vs. Switch/Backchannel) and the interlocutor's wish to contribute with propositional content (Switch vs. Backchannel). We were interested primarily in 1) the relevance of prosodic cues in relation to textual ones, and 2) their cross-linguistic validity by comparing unrelated languages Slovak and Argentine Spanish.

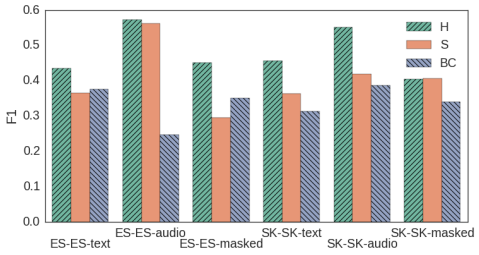

Figure 4: F_1 score for individual turn-taking categories.

Regarding the first issue, we showed that in addition to the textual cues (lexical, syntactic, pragmatic), prosodic cues also play a role in perceiving how the dialogue would unfold. This was shown most clearly in a significant difference between native subjects' accuracy in the audio and text conditions, and better-than-chance performance of subjects with no understanding of the prompts in the audio condition. Our result thus expands on previous studies ([3, 4, 5, 6, 7, 8]) with data on two languages with hardly any prior work on turn-taking cues. Additionally, these data argue against the conclusions of studies arguing for the non-relevance of the prosodic cues [2].

Regarding the second issue, our data support the conclusion that the links between prosodic cues and pragmatic meanings in both languages overlap to some extent, despite the fact that they belong to distinct language families and are geographically distant. This is because speakers that do not understand the content were able to predict the type of upcoming speech (H, S, BC) with accuracy exceeding chance. Hence, at least some prosodic cues to dialogue management are cross-linguistically valid. Our data also show, however, that the results for the two languages differ. Especially the perception of non-native masked audio and accuracy on individual turn types showed salient differences and will be investigated further in follow up analyses.

Finally, studies suggest that turn-initial prosodic cues are also useful for cueing pragmatic intentions; Sicoli et al. found, for example, that speakers use a boosted initial pitch to signal questions [14]. Our subjects only had a turn-final IPU and thus turn-initial prosodic information might be missing from those stimuli where there were more IPUs in a turn.

The relevance of prosody in dialogue management and the existence of both cross-linguistic and language-specific cues suggest a promising line of future research both in the application to Spoken Dialogue Systems (SDS) and in better understanding in human-human dialogue. For example, we plan to study of the actual cues found in the stimuli and explore the links between the perception and production, or examine the timing of responses of our subjects, as well as the effect of the stimulus and subject language (SK, ES) had on this timing.

5. Acknowledgements

This material is based upon work supported by CONICET, ANPCYT PICT 2014-1561, UBACYT 20020120200025BA, 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 Ramiro H. Gálvez and Julia Hirschberg for valuable suggestions and comments.

[1] In the three SK-ES-* groups we excluded Slovak subjects who reported some knowledge of the Spanish language.

[2] $F_1 = 2 \times \text{precision} \times \text{recall} \;/\; (\text{precision} + \text{recall})$.

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, 50(4).

[2] J. P. De Ruiter, H. Mitterer, and N. J. Enfield, "Projecting the end of a speaker's turn: A cognitive cornerstone of conversation," *Language*, pp. 515–535, 2006.

[3] S. Duncan, "On the structure of speaker-auditor interaction during speaking turns," *Language in Society*, vol. 3, no. 2, pp. 161–180, 1974.

[4] C. Ford and S. Thompson, "Interactional units in conversation: Syntactic, intonational and pragmatic resources for the management of turns," in *Interaction and Grammar*, E. Ochs, E. Schegloff, and S. Thompson, Eds. Cambridge, 1996, pp. 134–184.

[5] S. Benus, A. Gravano, and J. Hirschberg, "The prosody of backchannels in American English," in *Proceedings of ICPhS*, 2007.

[6] A. Gravano and J. Hirschberg, "Turn-taking cues in task-oriented dialogue," *Computer Speech and Language*, pp. 601-634, 2011, 25(3).

[7] A. Hjalmarsson, "The additive effect of turn-taking cues in human and synthetic voice," *Speech Communication*, pp. 23–25, 2011, 53(1).

[8] S. Bögels and F. Torreira, "Listeners use intonational phrase boundaries to project turn ends in spoken interaction," *Phonetics*, vol. 52, pp. 46–57, 2015.

[9] S. E. Brennan and M. Williams, "The feeling of another's knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers," *Journal of Memory and Language*, vol. 34, no. 3, pp. 383–398, 1995.

[10] S. C. Levinson, "Turn-taking in human communication, origins, and implications for language processing," *Trends in Cognitive Sciences*, vol. 20, no. 1, pp. 6–14, 2016.

[11] G. Skantze, M. Johansson, and J. Beskow, "Exploring turn-taking cues in multi-party human-robot discussions about objects," in *Proceedings of the ACM International Conference on Multimodal Interaction*, 2015.

[12] R. Levitan, Š. Benuš, A. Gravano, and J. Hirschberg, "Acoustic-prosodic entrainment in slovak, spanish, english and chinese: A cross-linguistic comparison," in *Proceedings of SIGdial*, 2015.

[13] M. Zellers, "Duration and pitch in perception of turn transition by Swedish and English listeners," in *Proceedings of Fonetik*, 2014.

[14] M. A. Sicoli, T. Stivers, N. J. Enfield, and S. C. Levinson, "Marked initial pitch in questions signals marked communicative function," *Language and Speech*, pp. 204–223, 2015.