Acoustic-prosodic entrainment in Slovak, Spanish, English and Chinese: A cross-linguistic comparison

Rivka Levitan [1,2], **Štefan Beňuš** [3], **Agustín Gravano** [4,5], **Julia Hirschberg** [2]

[1] Department of Computer and Information Science, Brooklyn College CUNY, USA

[2] Department of Computer Science, Columbia University, USA

[3] Constantine the Philosopher University in Nitra & Institute of Informatics, Slovak Academy of Sciences, Slovakia

[4] National Scientific and Technical Research Council (CONICET), Buenos Aires, Argentina

[5] Departamento de Computación, FCEyN, Universidad de Buenos Aires, Argentina

levitan@sci.brooklyn.cuny.edu, sbenus@ukf.sk,
gravano@dc.uba.ar, julia@cs.columbia.edu

Abstract

It is well established that speakers of Standard American English entrain, or become more similar to each other as they speak, in acoustic-prosodic features of their speech as well as other behaviors. Entrainment in other languages is less well understood. This work uses a variety of metrics to measure acoustic-prosodic entrainment in four comparable corpora of task-oriented conversational speech in Slovak, Spanish, English and Chinese. We report the results of these experiments and describe trends and patterns that can be observed from comparing acoustic-prosodic entrainment in these four languages. We find evidence of a variety of forms of entrainment across all the languages studied, with some evidence of individual differences as well within the languages.

1 Introduction

In general, entrainment is a ubiquitous tendency observed in human-human dialogues in which interlocutors adapt their communicative behavior to the behavior of their conversational partners in several modalities. Empirical evidence of entrainment in human-human conversations has been documented for numerous acoustic-prosodic features, including intensity (Natale, 1975; Gregory et al., 1993; Ward and Litman, 2007), speaking rate (Street, 1984), and pitch (Gregory et al., 1993; Ward and Litman, 2007). Humans have been shown to entrain to their interlocutor's *language* as well, at the lexical level (Brennan, 1996), syntactic level (Branigan et al., 2000; Reitter et al., 2010), and on what (Niederhoffer and Pennebaker, 2002) called linguistic style, which includes, among other features, the use of pronouns and verb tenses (Niederhoffer and Pennebaker, 2002; Danescu-Niculescu-Mizil et al., 2011; Michael and Otterbacher, 2014). Motivated by Communication Accommodation Theory (CAT) (Giles et al., 1991), which holds that speakers converge to or diverge from their interlocutors in order to attenuate or accentuate social differences, numerous studies have looked for links between entrainment and positive social behavior. Entrainment on various features and at all levels of communication has been linked, respectively, to liking (Chartrand and Bargh, 1999; Street, 1984), positive affect in conversations between "seriously and chronically distressed" married couples discussing a problem in their relationship (Lee et al., 2010), mutual romantic interest in speed dating transcripts (Ireland et al., 2011), cooperation in a prisoner's dilemma (Manson et al., 2013), task success (Nenkova et al., 2008; Reitter and Moore, 2007; Friedberg et al., 2012; Thomason et al., 2013), and approval-seeking (Natale, 1975; Danescu-Niculescu-Mizil et al., 2012). Given that these social aspects are assumed to be culture-specific, and the fact that research on entrainment has been done mainly on English and other Germanic languages, the types and degree of entrainment in other languages and cultures should be explored. Although there are numerous studies documenting entrainment in different aspects of spoken dialogue in particular languages collected in particular circumstances, it has been difficult to compare entrainment across languages due to differences in the corpora examined and and analytical approaches employed. Recently, this gap has been addressed in (Xia et al., 2014; Beňuš et al., 2014) who report on commonalities observed across languages as well as systematic differences in *global* measures of acoustic-prosodic entrainment (i.e. over entire dialogues)

in comparable corpora of conversational speech in Chinese, English and Slovak.

In this study we expand on these findings by focusing on *local* acoustic-prosodic entrainment (i.e. dynamic adjustments at turn exchanges) on a session-by-session basis and present results from a comparative study of four very different languages, English, Chinese, Slovak, and Spanish, collected from subjects engaged in deliberately similar conversational tasks for the purpose of comparison: the Columbia Games Corpus (English), the SK-Games Corpus (Slovak), the Porteño Spanish Games Corpus, and the Tongji Games Corpus (Chinese), and employ identical tools and methods for their analysis. We present the results of analyses of these corpora for positive and negative (complementary) entrainment using a variety of metrics (proximity, synchrony and convergence), and a variety of acoustic and prosodic features (pitch, intensity, speaking rate, and several measures of voice quality).

Section 2 describes the four corpora used in our analysis, the features we examined in the study and the units of analysis over which they were calculated. Section 3 discusses three methods of measuring entrainment at the local level, *proximity*, *synchrony*, and *convergence*, and reports the results of applying each of these measures to the four corpora. Section 4 summarizes our results and discusses patterns that emerge from our analysis.

2 Data and features

This section describes the comparable task-oriented corpora that are analyzed in this study.

2.1 Columbia Games Corpus

The Columbia Games Corpus is a collection of 12 spontaneous dyadic conversations between native speakers of Standard American English (SAE). Thirteen subjects participated in the collection of the corpus. Eleven returned on another day for another session with a different partner. Their ages ranged from 20 to 50 years ($M = 30.0$, $SD = 10.9$). Six subjects were female, and seven were male; of the twelve dialogues in the corpus, three are between female-female pairs, three are between male-male pairs, and six are between mixed-gender pairs. All interlocutors were strangers to each other.

In order to elicit spontaneous, task-oriented speech, subjects were asked to play a series of four computer games of two kinds: Cards games and Objects games. The games were designed to require cooperation and communication in order to achieve a high score. Participants were motivated to do well by a monetary bonus that depended on the number of points they achieved in each game. 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. During game play, keystrokes were captured and were later synchronized with the speech recordings and game events.

There are approximately 9 hours and 13 minutes of speech in the Games Corpus, of which approximately 70 minutes come from the first part of the Cards game, 207 minutes from the second part of the cards Game, and 258 minutes from the Objects game. On average, each session is approximately 46 minutes long, comprised of three Cards games of approximately 8 minutes each and one Objects game, which is approximately 22 minutes long.

The corpus has been orthographically transcribed and manually word-aligned by trained annotators. In addition, disfluencies and other paralinguistic events such as laughs, coughs and breaths were marked by the annotators. The corpus has also been annotated prosodically according to the ToBI framework (Silverman et al., 1992); all turns have been labeled by type; affirmative cue words have been labeled according to their pragmatic functions; and all questions have been categorized by form and function. The annotation of the Games Corpus is described in detail in (Gravano, 2009).

2.2 Sk-Games Corpus

SK-games is a corpus of native Slovak (SK) conversational speech that is identical to the Objects games of the Columbia Games Corpus for SAE barring adjustments to some of the screen images and their positioning. Subjects were seated in a quiet room facing computer screens without visual contact with each other. The corpus currently includes 9 dyadic sessions with a total of 11 speakers (5F, 6M). Seven of the speakers (4F, 3M) participated in two sessions and thus we can compare their behavior in identical communicative situations when they are paired with a different interlocutor. Of the nine sessions, two are between female-female pairs, two between male-

male pairs, and five between mixed-gender pairs. The analyzed material makes roughly six hours of speech, and consists of 35,758 words and 3,189 unique words. The audio signal was manually transcribed, and the transcripts were automatically aligned to the signal using the SPHINX toolkit adjusted for Slovak (Darjaa et al., 2011), which forces the alignment of both words and individual phonemes. This forced alignment was then manually corrected.

2.3 Porteño Spanish Games Corpus

The Spanish data were taken from a larger corpus of Porteño Spanish (Sp) that is currently under construction. Porteño is a variant of the Spanish language spoken by roughly 20-25 million people in East-Central Argentina and Uruguay. It is characterized by substantial differences with other variants of Spanish at the lexical, phonological and prosodic levels (e.g. (Colantoni and Gurlekian, 2004)). The portion of the corpus used in this study is also similar to the Objects games of the Columbia Games Corpus, and currently includes 7 dyadic sessions with a total of 12 native speakers of Porteño Spanish (7F, 5M); only two female speakers participated in two sessions, with different partners in each session. Of the seven sessions, three are between female-female pairs, one between male-male pairs, and three between mixed-gender pairs. The analyzed material makes roughly two hours of speech, and consists of 17,571 words and 1,139 unique words. The audio signal was manually transcribed, and the transcripts were manually aligned to the signal by trained annotators.

2.4 Tongji Games Corpus

The Tongji Games Corpus (Xia et al., 2014) is a corpus of spontaneous, task-oriented conversations in Mandarin Chinese (MC). The corpus contains approximately 12 hours of speech, comprising 99 conversations between 84 unique native speakers (57 female, 27 male), some of whom participated in more than one conversation with a different partner. Conversations average 6 minutes in length. Participants in the corpus were randomly selected from university students who had a National Mandarin Test Certificate level 2 with a grade of A or above. This restriction enforced that the elicited speech would be standard Mandarin, with minimal effect of regional dialect. As in the collection of the Columbia Games Corpus, recordings were made in a sound-proof booth on laptops with a curtain between participants so that neither could see the other's screen and so that all communication would be verbal.

Two games were used to elicit spontaneous speech in the collection of the corpus. In the **Picture Ordering** game, one subject, the information *giver*, gave the other, the *follower*, instructions for ordering a set of 18 cards. When the task was completed, the same pair switched roles and repeated the task. In the **Picture Classifying** game, each pair worked together to classify 18 pictures into appropriate categories by discussing each picture. Seventeen pairs played the Picture Ordering game, 39 pairs played the Picture Classification game, and 14 pairs played both games (each time with the same partner).

The corpus was segmented automatically using SPPAS (SPeech Phonetization Alignment and Syllabification) (Bigi and Hirst, 2012), a tool for automatic prosody analysis. The automatic segments were manually checked and orthographically transcribed. Turns were identified by two PhD students specializing in Conversation Analysis.

For our analysis, we include one randomly chosen conversation from each of ten female-female pairs, ten male-male pairs, and ten female-male pairs, for a total of 30 conversations.

2.5 Features

In each corpus, we look for evidence of entrainment on eight acoustic-prosodic features: intensity mean and max, pitch mean and max, jitter, shimmer, noise-to-harmonics ratio (NHR), and speaking rate. Speaking rate for English was determined from the orthographic transcriptions of the data using an online syllable dictionary. For Slovak, the syllable count for each word was determined algorithmically utilizing the availability of phonemes (from grapheme-to-phoneme conversion required for alignment) and a known set of phonemes forming syllable nuclei. For Spanish, syllable counts were computed automatically using the toolkit developed by Hernández-Figueroa et al. (2013). All other features were extracted using the open-source audio analysis tool Praat (Boersma and Weenink, 2012).

To allow for meaningful comparisons between female and male pitch values, female pitch values in the English, Spanish and Slovak corpora were linearly scaled to lie within the male pitch

range. This gender normalization was not done for the Chinese data. However, a linear scaling of a given speaker's feature values does not affect the analysis, since all comparisons are relative to the speaker's own speech.

The analysis of the Chinese data did not consider the voice quality features (jitter, shimmer, or NHR).

The details of the feature extraction and audio analysis of the Columbia Games Corpus can be found in (Gravano, 2009); the same methods were used for the other three corpora, without any corpus-specific refinements.

2.6 Units of analysis

We compute and compare features from the following units of analysis:

An **inter-pausal unit (IPU)** is a pause-free chunk of speech from a single speaker. The threshold for pause length for three of the corpora was derived empirically from the average length of stop gaps in each corpus (50ms for English and Spanish, 80ms for Chinese); for the Slovak data, pauses were detected with a minimum threshold of 100ms and then manually adjusted.

A **turn** is a consecutive series of IPUs from a single speaker. We include in our definition of "turns" utterances that are not turns in the discourse sense of the term, such as backchannels or failed attempts to take the floor.

A **session** is a complete interaction between a pair of interlocutors.

3 Local entrainment

Local entrainment is defined as similarity between interlocutors at well-defined points in a conversation. Two speakers may be *globally* similar—for example, having similar feature means—while diverging widely at most given points in a conversation, as in Figure 1.

Local entrainment can be thought of as dynamic entrainment: a continuous reaction to one's interlocutor and updating of one's own output in response to what has just been heard. Such entrainment can be **convergent**, adjusting toward greater similarity to the interlocutor, or **complementary**, adjusting away from the interlocutor. Complementary entrainment is often called **disentrainment** or **divergence** (Healey et al., 2014), with the connotation that this behavior reflects a speaker's desire to distance herself from her interlocutor, but

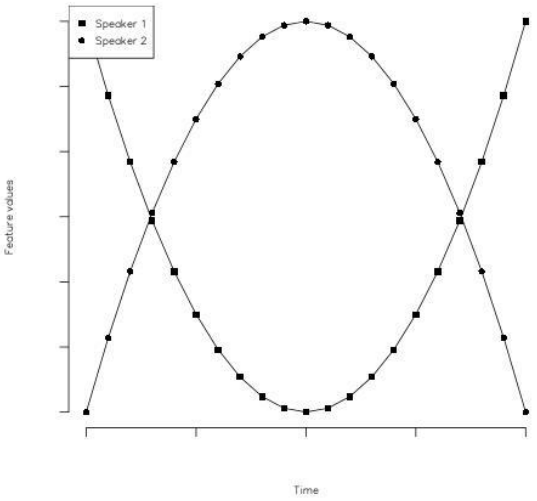

Figure 1: *Global vs. local entrainment*

it can also be viewed as a cooperative behavior in which a speaker completes or resolves the prosody of her interlocutor's previous turn. Independently of either interpretation, local entrainment denotes dynamic responsiveness to an interlocutor's behavior.

Following (Levitan and Hirschberg, 2011), and using the measures described there, we look for evidence of three aspects of local entrainment:*proximity*, *synchrony*, and *convergence*. (Xia et al., 2014) analyzed local entrainment in English and Chinese, and found similar patterns over entire corpora. Here, for a more nuanced view of the prevalence of local entrainment, we apply our analysis separately to each session.

Multiple statistical tests are conducted in the course of this analysis. All significance tests correct for family-wise Type I error by controlling the false discovery rate (FDR) at $\alpha = 0.05$. The kth smallest p value is considered significant if it is less than $\frac{k \times \alpha}{n}$ (where n is the number of p values).

3.1 Proximity

Proximity describes entrainment by value. A session that displays proximity on a given feature will have turns which are more similar to their preceding turns than they are to others of the interlocutor's turns in the session. To measure proximity, we look at the differences in feature values between adjacent IPUs at turn exchanges. For each turn t, and for each feature f, we calculate an *adjacent* difference—the absolute value of the dif-

ference between the value of f in the first IPU of t and the value of f in the last IPU of $t-1$—and a *non-adjacent* difference—the averaged absolute values of the differences between the value of f in the first IPU of t and the values of f in the last IPUs of 50 other turns chosen randomly from the turns of the other speaker.[1]

These non-adjacent differences serve as a baseline for the degree of similarity we might expect to see at turn exchanges if there is no effect of local entrainment. For each session and each feature, if adjacent differences are smaller than non-adjacent differences, we conclude that the speakers in that session are locally entraining to each other.

Table 1 shows the results of paired t-tests between adjacent and non-adjacent differences for each of the nine Slovak sessions we analyze. We see little evidence of local proximity in our Slovak data. Only two sessions show evidence of local proximity on intensity mean, and only one shows negative proximity for intensity max. No other feature shows evidence of local proximity, positive or negative.

Table 2 shows the results of the test for proximity in the seven Spanish sessions. Spanish shows even less evidence of local proximity: Only one session shows evidence of negative proximity of intensity max; there is no evidence of convergent proximity.

Feature	Session								
	1	2	3	4	5	6	7	8	9
IntMean			+	+					
IntMax						-			
PchMean									
PchMax									
Jitter									
Shimmer									
NHR									
Spkrt									

Table 1: *Local proximity by session in Slovak (+: significant positive proximity; -: significant negative proximity; ' ': no significant proximity)*

English, in contrast, shows significant positive local proximity on intensity mean and max in four out of 12 sessions. There is no evidence of positive local proximity on any other feature in English, and no evidence of negative local proximity at all.

The Chinese data also shows evidence of positive local proximity on intensity mean and max in several sessions (three out of 30 for intensity mean), but there is also evidence of negative proximity on those features in multiple sessions. In addition, nearly all sessions show strong negative proximity on the pitch features. Finally, one session (out of 30) shows negative proximity on speaking rate.

Feature	Session						
	1	2	3	4	5	6	7
IntMean							
IntMax			-				
PchMean							
PchMax							
Jitter							
Shimmer							
NHR							
Spkrt							

Table 2: *Local proximity by session in Spanish (+: significant positive proximity; -: significant negative proximity; ' ': no significant proximity)*

3.2 Synchrony

Synchrony describes entrainment by *direction* rather than value, measuring how the dynamics of an individual speaker's prosody relate to those of his or her interlocutor. We take the Pearson's correlation between feature values from adjacent IPUs at turn exchanges to see whether speakers' values at turn exchanges vary together, in synchrony, even if they are not similar in absolute values.

As Table 3 shows, synchrony is a much more significant factor in entrainment in Slovak than proximity is. Nearly every feature shows evidence of synchrony in multiple sessions. Strikingly, nearly every feature in fact shows *negative* synchrony, or *complementary* synchronous entrainment. Only intensity mean shows positive synchrony in three sessions (and negative synchrony in a fourth); synchrony on the other seven features is consistently negative.

Table 3 also makes it clear that this aspect of entrainment is highly individualized. Session 1, for example, shows no evidence of synchrony at all; Session 5 shows significant negative synchrony in

[1] Since some of the Spanish sessions did not have as many as 50 turns from the other speaker, the non-adjacent differences in the Spanish analysis were averaged over 20 turns from the other speaker.

Feature	Session								
	1	2	3	4	5	6	7	8	9
IntMean			+	+		+			-
IntMax					-	-			-
PchMean		-	-		-		-		-
PchMax		-	-				-		-
Jitter					-	-			
Shimmer					-				
NHR		-			-		-	-	
Spkrt					-				

Table 3: *Local synchrony in Slovak by session (+: significant positive synchrony; -: significant negative synchrony; ' ': no significant synchrony)*

Feature	Session						
	1	2	3	4	5	6	7
IntMean							
IntMax				-			
PchMean					-		
PchMax							
Jitter							
Shimmer		-			-		
NHR		-		-	-		
Spkrt			-				

Table 4: *Local synchrony in Spanish by session (+: significant positive synchrony; -: significant negative synchrony; ' ': no significant synchrony)*

everything except intensity mean and pitch max; Session 4 shows only positive synchrony in intensity mean; and Session 9 shows negative synchrony in intensity and pitch mean and max. Further research will be needed to explore the relationships between entrainment on different aspects of prosody.

Table 4 reveals similar trends in the Spanish data. Synchrony is evident for a plurality of features and sessions, and all observed synchrony is negative. One notable difference is in synchrony on pitch features, which is present in five of nine Slovak dialogues, and only one of seven Spanish dialogues. Intensity mean, which shows positive synchrony in three Slovak dialogues and negative synchrony in one, shows no evidence of synchrony in any Spanish dialogue.

In the English data, positive synchrony is evident for intensity mean in six of the 12 dialogues. Intensity max shows positive synchrony in three sessions and negative synchrony in another three. There is also some evidence of positive synchrony on pitch mean, pitch max, and shimmer (one session each), and negative synchrony on pitch mean (three sessions); pitch max, jitter, shimmer, and NHR (two sessions each); and speaking rate (one session).

The most notable aspect of entrainment by synchrony in the Chinese data is the strong negative synchrony on pitch that is present in many of the sessions (19 out of 30 for pitch mean, 15 for pitch max). One session shows positive synchrony on pitch max; none show positive synchrony on pitch mean. The results on intensity are more evenly split: for intensity mean, six sessions show positive synchrony and five show negative, while the count is 3-4 for intensity max. Three sessions show negative synchrony on speaking rate.

3.3 Convergence

We add another dimension to our analysis of local entrainment by looking at *convergence*: whether interlocutors become increasingly similar over the course of a conversation. Where previously we looked at the degree to which interlocutors react and adapt to each other at each turn exchange, now we look at how that degree of adaptation changes with time. This is measured by the Pearson's correlation between adjacent differences (absolute differences in feature values in adjacent IPUs at turn exchanges) and time. A significant negative correlation over a session (differences become *smaller* with time) is evidence of convergence.

Feature	**Session**								
	1	2	3	4	5	6	7	8	9
IntMean	-						+		
IntMax							+		
PchMean									
PchMax		-							
Jitter				+					
Shimmer									
NHR					-				
Spkrt									

Table 5: *Local convergence in Slovak by session (+: significant convergence; -: significant divergence; ' ': no significant convergence)*

Table 5 shows little evidence of local convergence in Slovak. Only two sessions show evidence of convergence: one on intensity mean and max, and one on jitter. Three others show evidence of *divergence*, differences that *increase* with

Feature	Session						
	1	2	3	4	5	6	7
IntMean	+		+				
IntMax							
PchMean							
PchMax							
Jitter							
Shimmer							
NHR				+			
Spkrt							

Table 6: *Local convergence in Spanish by session (+: significant convergence; -: significant divergence; ' ': no significant convergence)*

time: one on intensity mean, one on pitch max, and one on NHR. The diversity of these results, with individual interlocutor pairs converging or diverging on specific features, suggests a strong speaker-dependent component to this aspect of entrainment. The same is true for the Spanish data (Table 6) — two sessions show convergence on intensity mean, and one on NHR — and the Chinese data: one session shows convergence on intensity mean and one on pitch max. English is the outlier here, with evidence of local convergence on intensity mean (two sessions), intensity max (five sessions), pitch mean (six sessions), pitch max (three sessions), and NHR (three sessions).

The significant correlation strengths (for all languages) are not high, ranging in absolute value between 0.13 and 0.32. The effect of convergence, even when significant, is only one of numerous factors affecting speakers' prosodic expression.

4 Discussion

This analysis explored three kinds of local entrainment on eight features over a total of 58 sessions in four languages. Table 7 summarizes our findings. Out of all this data certain patterns emerge:

Negative (complementary) synchrony is more prevalent than positive (convergent) synchrony. In each of the four languages under analysis, negative synchrony is present in a greater number of dialogues and for a greater number of features than is positive synchrony. This seems to indicate that at a local level, and to some extent independently of the specific prosodic characteristics of the language being spoken, human interlocutors adjust the prosodic features of their speech in the *opposite* direction from the dynamics of their partner's speech. That is, if speaker A produces a turn ending in the low part of her range for turn endings, speaker B will produce a turn beginning in the high part of his range for turn beginnings. (This is, of course, simplistic; the correlation strengths are mainly low to moderate, and entrainment is only one of many factors influencing the prosody of a given production.) It should be noted that this relationship cannot be attributed to the prosodic differences inherent in turn beginnings and turn endings, since the Pearson's correlation compares fluctuations within a series rather than the actual values. These results do not show that a low IPU tends to be followed by a high IPU, but that an IPU that is low for a turn ending tends to be followed by one that is high for a turn beginning.

This finding is in line with recent research questioning the ubiquity of entrainment in the syntactic and semantic domains and calling for more refined analyses of entrainment behavior (Healey et al., 2014). As discussed above, negative synchrony may be termed "disentrainment" and interpreted as a distancing behavior. However, its prevalence in cooperative dialogues is an argument for a more neutral interpretation. This can be explored in future work by determining whether negative synchrony is associated with objective and subjective measures of partner engagement and liking, as in (Levitan et al., 2012).

Another consistency found across languages is that mean intensity is the only feature to show significant *positive* synchrony in a plurality of sessions. In English, in fact, it *only* shows positive synchrony, the only feature to do so; in the other three languages it is more evenly split between instances of positive and negative synchrony.

Synchrony is more prevalent than proximity. In measuring local entrainment, we have distinguished between *proximity*, the similarity of a pair of feature values, and *synchrony*, the similarity of the dynamics of two sets of feature values. Our results show that synchrony is a more useful measure for characterizing the way in which human interlocutors adjust to each other at the local level. This is especially true for Slovak and Spanish, which show almost no evidence of proximity, but show evidence of synchrony in multiple sessions for almost every feature. Comparing proximity and synchrony, however, should be taken with caution since their prevalence has been assessed with different statistical tests.

Feature	Proximity (% sessions)								Synchrony (% sessions)								Convergence (% sessions)							
	SAE		MC		Sk		Sp		SAE		MC		Sk		Sp		SAE		MC		Sk		Sp	
	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-
IntMean	33		10	20	22				50		20	17	33	11			17		3		11	11	29	
IntMax	25		7	17		11		14	25	25	10	13		33		14	42				11			
PchMean				67					8	25		63		56		14	50							
PchMax				63					8	17	3	50		44			25		3			11		
Jitter			-	-						17	-	-		22					-	-	11			
Shimmer			-	-					8	17	-	-		11		28			-	-				
NHR			-	-						17	-	-		44		43	25		-	-		11	14	
Spkrt				3						8		10		11		14								

Table 7: *Cross-linguistic summary of results on local acoustic-prosodic entrainment as percentages of sessions with significant positive (+) and negative (-) entrainment type (proximity, synchrony, convergence) A '-' indicates that the corresponding statistical test was not done for that language.*

Chinese shows the strongest evidence of pitch synchrony. While all four languages show evidence of negative synchrony on pitch, Chinese has the strongest and most prevalent negative pitch synchrony: it is present in a majority of the sessions, with correlation strengths of about 0.90. The reason for this is unknown, but it is reasonable to hypothesize that it is linked to the importance of pitch in Chinese, a tonal language.

Pitch is also the feature displaying the strongest and most prevalent negative synchrony in Slovak: negative pitch synchrony is present in a majority of sessions, as in Chinese, with correlation strengths of about 0.50.

English shows the strongest evidence of local convergence. In English, we observe positive local convergence in a plurality of sessions, on all features except jitter, shimmer, and speaking rate. There is no evidence of negative convergence. The only other language with significant evidence of local convergence is Spanish, which displays local convergence on intensity mean in two sessions and on NHR in one (out of seven). Chinese and Slovak have scattered instances of convergence; Slovak is the only language to show negative convergence, though the evidence is sparse (one session each for intensity mean, jitter, and NHR).

Individual behavior varies. While the patterns we have identified are apparent when looking at the data in the aggregate, none can be said to apply to all the sessions they describe, or even almost all. Clearly, a session's entrainment behavior is significantly influenced by the particular dynamics of its speaker pair. Gender, power, liking, personality, and similar factors have all been shown to influence the degree of entrainment to some extent (Levitan et al., 2012; Š. Beňuš et al., 2014; Gravano et al., 2014). Exploring how these factors correlate with entrainment in different languages and cultures is an interesting area for future work.

5 Conclusion

We have presented the results of applying an identical analysis of acoustic-prosodic entrainment to comparable corpora in four different languages. This approach allows us to identify trends that are characteristic of human behavior independently of language and culture, and behaviors that seem to be characteristic of a given language.

This study can be considered an exploratory contribution to what is currently a very small body of work concerning language differences in entrainment. Since three of the corpora we analyze have a relatively small number of participants, it is possible that the differences we identify may be the products of individual behavior rather than the characteristics of the given language. In future work, these results will be confirmed or refined by further research on a larger scale.

Acknowledgments

This material is based upon work supported by the Air Force Office of Scientific Research, Air Force Material Command, USAF under Award No. FA9550-15-1-0055 and by UBACYT 20020120200025BA.

References

Š. Beňuš, R. Levitan, J. Hirschberg, A. Gravano, and S. Darjaa. 2014. Entrainment in Slovak collaborative dialogues. In *Proceedings of the 5th IEEE Conference on Cognitive Infocommunications*, pages 309–313.

Brigitte Bigi and Daniel Hirst. 2012. SPeech phonetization alignment and syllabification (SPPAS): a tool for the automatic analysis of speech prosody. In *Speech Prosody*, pages 19–22. Tongji University Press.

Paul Boersma and David Weenink. 2012. Praat: doing phonetics by computer [computer program].

Version 5.3.23, retrieved 21 August 2012 from http://www.praat.org.

Holly P. Branigan, Martin J. Pickering, and Alexandra A. Cleland. 2000. Syntactic co-ordination in dialogue. *Cognition*, 75(2):B13–B25.

Susan E Brennan. 1996. Lexical entrainment in spontaneous dialog. *Proceedings of ISSD*, pages 41–44.

T. L. Chartrand and J. A. Bargh. 1999. The chameleon effect: The perception-behavior link and social interaction. *Journal of Personality and Social Psychology*, 76(6):893–910.

L. Colantoni and J.A. Gurlekian. 2004. Convergence and intonation: Historical evidence from Buenos Aires Spanish. *Bilingualism: Language and Cognition*, 7(2):107–119.

Cristian Danescu-Niculescu-Mizil, Michael Gamon, and Susan Dumais. 2011. Mark my words! linguistic style accommodation in social media. In *Proceedings of WWW*.

Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. 2012. Echoes of power: language effects and power differences in social interaction. In *Proceedings of WWW*.

S. Darjaa, M. Cerňak, M. Trnka, M. Rusko, and R. Sabo. 2011. Effective triphone mapping for acoustic modeling in speech recognition. In *Proceedings of Interspeech*.

Heather Friedberg, Diane Litman, and Susannah BF Paletz. 2012. Lexical entrainment and success in student engineering groups. In *Spoken Language Technology Workshop (SLT), 2012 IEEE*, pages 404–409. IEEE.

Howard Giles, Nikolas Coupland, and Justine Coupland. 1991. Accommodation theory: Communication, context, and consequence. *Contexts of accommodation: Developments in applied sociolinguistics*, 1.

A. Gravano, S. Benus, R. Levitan, and J. Hirschberg. 2014. Three tobi-based measures of prosodic entrainment and their correlations with speaker engagement. In *IEEE Spoken Language Technology Workshop (SLT)*.

Agustín Gravano. 2009. *Turn-taking and affirmative cue words in task-oriented dialogue*. Ph.D. thesis, Columbia University.

Stanford Gregory, Stephen Webster, and Gang Huang. 1993. Voice pitch and amplitude convergence as a metric of quality in dyadic interviews. *Language & Communication*, 13(3):195–217.

Patrick GT Healey, Matthew Purver, and Christine Howes. 2014. Divergence in dialogue. *PloS one*, 9(6):e98598.

Z. Hernández-Figueroa, F.J. Carreras-Riudavets, and G. Rodríguez-Rodríguez. 2013. Automatic syllabification for Spanish using lemmatization and derivation to solve the prefix's prominence issue. *Expert Systems with Applications*, 40(17):7122–7131.

Molly E Ireland, Richard B Slatcher, Paul W Eastwick, Lauren E Scissors, Eli J Finkel, and James W Pennebaker. 2011. Language style matching predicts relationship initiation and stability. *Psychological Science*, 22(1):39–44.

Chi-Chun Lee, Matthew Black, Athanasios Katsamanis, Adam Lammert, Brian Baucom, Andrew Christensen, Panayiotis G. Georgiou, and Shrikanth Narayanan. 2010. Quantification of prosodic entrainment in affective spontaneous spoken interactions of married couples. In *Proceedings of Interspeech*.

Rivka Levitan and Julia Hirschberg. 2011. Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. In *Proceedings of Interspeech*.

Rivka Levitan, Agustín Gravano, Laura Willson, Stefan Benus, Julia Hirschberg, and Ani Nenkova. 2012. Acoustic-prosodic entrainment and social behavior. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 11–19, Montréal, Canada, June. Association for Computational Linguistics.

Joseph H Manson, Gregory A Bryant, Matthew M Gervais, and Michelle A Kline. 2013. Convergence of speech rate in conversation predicts cooperation. *Evolution and Human Behavior*, 34(6):419–426.

Loizos Michael and Jahna Otterbacher. 2014. Write like i write: Herding in the language of online reviews. In *Proceedings of the Eigth International AAAI Conference on Weblogs and Social Media*.

Michael Natale. 1975. Convergence of mean vocal intensity in dyadic communication as a function of social desirability. *Journal of Personality and Social Psychology*, 32(5):790–804.

Ani Nenkova, Agustín Gravano, and Julia Hirschberg. 2008. High frequency word entrainment in spoken dialogue. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, pages 169–172. Association for Computational Linguistics.

Kate G. Niederhoffer and James W. Pennebaker. 2002. Linguistic style matching in social interaction. *Journal of Language and Social Psychology*, 21(4):337–360.

David Reitter and Johanna D. Moore. 2007. Predicting success in dialogue. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 808–815.

David Reitter, Johanna D. Moore, and Frank Keller. 2010. Priming of syntactic rules in task-oriented dialogue and spontaneous conversation.

K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg. 1992. ToBI: A standard for labeling English prosody. In *International Conf. on Spoken Language Processing*, volume 2, pages 867–870.

Richard L Street. 1984. Speech convergence and speech evaluation in fact-finding interviews. *Human Communication Research*, 11(2):139–169.

Jesse Thomason, Huy V Nguyen, and Diane Litman. 2013. Prosodic entrainment and tutoring dialogue success. In *Artificial Intelligence in Education*, pages 750–753. Springer.

Š. Beňuš, A. Gravano, R. Levitan, S.I. Levitan, L. Willson, and J. Hirschberg. 2014. Entrainment, dominance and alliance in Supreme Court hearings. *Knowledge-Based Systems*, 71:3–14.

Arthur Ward and Diane Litman. 2007. Measuring convergence and priming in tutorial dialog. Technical report, University of Pittsburgh.

Zhihua Xia, Rivka Levitan, and Julia Hirschberg. 2014. Prosodic entrainment in Mandarin and English: A cross-linguistic comparison. In *Speech Prosody*.