UTEP Technical Report UTEP-CS-15-33

A Corpus for Investigating English-Language Learners' Dialog Behaviors

Nigel G. Ward and Paola Gallardo

Department of Computer Science
University of Texas at El Paso
500 West University Avenue
El Paso, TX 79968-0518

nigelward@acm.org, pgallardo2@miners.utep.edu

April 13, 2015

We are interested in developing methods for the semi-automatic discovery of prosodic patterns in dialog and how they differ between languages and among populations. We are starting by examining how the prosody of Spanish-native learners of English differs from that of native speakers. To support this work, we have collected a new corpus of conversations among college students. This includes dialogs between a nonnative speaker of English and a native, dialogs between native speakers of English, and Spanish conversations.

Index Terms: prosody, speech, dialog, learners, L1, L2, English, Spanish

1 Goals

Learners often have difficulty acquiring the prosody of a new language, especially the prosodic skills needed to be effective and efficient in dialog. This has, unfortunately, not yet been systematically studied.

We are interested in identifying and understanding how non-native prosodic behaviors differ from natives, and in developing tools to support this process. This will enable, among other things, the development of resources and software that will help people learn to communicate more effectively in a new language, or indeed in their own language.

This report documents a data collection designed to support this work. Leveraging our local demographics, we focus on Spanish-native learners of English.

The corpus is available for research purposes upon request.

2 Native and Non-Native Speakers

Our non-natives were advanced learners with high competence, and all had at least one semester of study in an English-speaking University. They were, nevertheless, noticably non-native in some aspects of their pronunciation and interaction styles. All were upperclass or graduate-level computer science students.

Table 6 overviews the linguistic backgrounds of all participants, based on their self-reports. Although we had initially intended to recruit two clearly different and internally homogeneous populations — natives and non-natives — things turned out more complicated.

As a result, we reviewed the data after collection and, after much discussion, selected the two main participant sets, and put the rest in an "other" category.

Non-Natives We decided to count as non-natives only those who clearly appeared to us to be non-native; this correlated with those who learned English after age 12. Six participants met this definition.

Monolinguals Eleven subjects were native English speakers with little or no knowledge of other languages; these we called monolinguals.

Other The remaining seven participants were bilingual speakers, including two tagged Ambiguous because their Spanish was somewhat stronger.

Thus in total the corpus contains 24 participants, 15 male and 9 female, all between the ages of 20 and 34.

We note that the reality of our border community means that the language backgrounds are more varied and complicated than the table indicates. Complications included the fact that many participants had been commuting daily across the border for many years, had lived at various times on different sides of the border, or had grown up in a part of El Paso where the dominant language was Spanish. For example, participant 20 was born in the US but grew up in Mexico. Her English exposure included attending a bilingual kindergarten, attending grades 3-5 in an elementary school in El Paso, periodically visiting El Paso thereafter, and then coming to UTEP for college. She doesn't consider herself a native speaker of English, but certainly doesn't seem like a classic non-native.

3 Data Subsets

There are three key sets of conversations:

1. Non-native speakers talking with a monolingual English speaker. This is our primary collection, as we want to understand the prosody of the non-natives. (Table 1)
2. As a control, monolingual native English speakers talking with each other. These include many of those in our primary collection. (Table 2)
3. For investigating native-language influences, native Spanish speakers speaking together in Spanish. These include most of the non-native speakers who appeared in our primary collection. (Table 3)

There are also two other sets, comprised of conversations initially intended for one of the main collections, but which we later determined were different enough to separate out.

1. Near-native speakers talking with monolingual English speakers. These were speakers who we first thought of non-natives, but later decided were too close to native to include in the primary collection. (Table 4)

2. Bilingual speakers in Spanish. The first two pairs are both essentially perfect bilinguals. The other pair was between someone with rather weaker Spanish and a learner from our primary collection. (Table 5)

The filenames of recordings in these last two categories reflect the intention we had when recording them. There are a total of 28 conversations.

4 Recruitment, Instructions, and Recording

Participants were recruited from among friends and acquaintances. Each was requested to participate in two dialogs. Those who did were compensated with $15. All participants dedicated their recordings as performances to the public domain, without restriction.

We gave them no instructions on what to talk about, as we wanted casual, unscripted, everyday conversations. While the dialogs were solicited, all of them were among people who might have had a conversation anyway that day. Most of the participants were engaged in their conversations, most had to be stopped when the 10 minutes was up, and several remarked that they gladly would have continued talking. Overall the dialogs were quite natural.

Participants were seated in adjacent rooms and talked across a glass window. Each conversant wore a SHURE BRH441M single-sided broadcast headset. The inputs were fed into a TASCAM DR-40 linear PCM recorder set to record 48,000 24-bit samples per second on each channel. Audio from the monitoring jack of the recorder was split and sent it back to the participants' headsets so they could hear each other.

The resulting audio quality is good. There is however some cross-channel bleeding, not noticeable at normal listening volumes, but detectable, for example, by good pitch trackers.

Acknowledgments

We thank the National Science Foundation for funding under grant IIS-0914868 and our dialog participants.

Table 1: Native/Non-Native Conversations (English). "non" indicates the non-natives

Conversation ID	Subject L ID	Subject R ID	Relationship
nn001	13 non	11	Friends
nn002	10 non	11	Strangers (just met)
nn003	7	5 non	Strangers (just met)
nn005	14	9 non	Acquaintances
nn006	13 non	12	Friends
nn007	17 non	15	Strangers (just met)
nn011	18	9 non	Acquaintances
nn012	9 non	12	Friends
nn013	23 non	24	Friends

Table 2: Native/Native Conversations, between Monolinguals (English)

Conversation ID	Subject L ID	Subject R ID	Relationship
eng001	12	11	Friends
eng002	16	11	Friends
eng003	14	15	Acquaintances
eng004	18	19	Acquaintances
eng005	17	4	Friends
eng006	14	11	Friends
eng007	16	18	Acquaintances

Table 3: Spanish Conversations, between native Spanish speakers)

Conversation ID	Subject L ID	Subject R ID	Relationship
esp004	13	17	Friends
esp005	10	9	Acquaintances
esp006	2	3	Friends
esp007	2	5	Friends
esp008	5	9	Friends

Table 4: Other English Conversations. "less" indicates the participant with less fluency.

Conversation ID	Subject L ID	Subject R ID	Relationship
nn004	7	8	Friends
nn008	4	2 less	Friends
nn009	3	4	Friends
nn010	20 less	22	In a Relationship

Table 5: Other Spanish Conversations

esp001	20	21	Friends
esp002	1	21	Siblings
esp003	6	5	Friends

Table 6: Subject Information

Subject ID	Age	Gender	"Type"	Language	Age	Country
1	20-24	F	Bilingual	Spanish	1-3	Mexico
					3-24	United States
				English	9-24	United States
				French	14-24	United States
2	20-24	M	Bilingual	Spanish	0-20	United States
				English	6-20	United States
				French	18-19	United States
3	20-24	M	Bilingual	Spanish	0-17	Mexico
					6-20	United States
				English	6-20	United States
				French	14-17	United States
4	25-29	M	Monolingual	English	0-25	United States
5	20-24	F	Non-Native	Spanish	0-20	United States
					13-20	Mexico
				English	7-20	United States
6	25-29	M	Bilingual	Spanish,English	0-29	United States
7	25-29	F	Near Monolingual	English	0-26	United States
				Portuguese	20-21	Brazil
				Spanish	22-23	Spain
				French	23-24	France
8	20-24	M	Bilingual	Spanish,English	0-24	United States
9	20-24	F	Non-Native	Spanish	1-24	Mexico
				English	18-20	Mexico
					23-24	United States
10	25-29	F	Non-Native	Spanish	0-25	Mexico
				English	17-25	United States
11	20-24	M	Monolingual	English	1-21	United States
12	20-24	M	Monolingual	English	1-21	United States
13	20-24	M	Non-Native	Spanish	0-22	Mexico
				English	18-22	United States
14	25-29	M	Monolingual	English	1-27	United States
15	20-24	M	Monolingual	English	1-2	Italy
					2-21	United States
16	20-24	M	Monolingual	English	1-23	United States
17	20-24	F	Non-Native	Spanish	0-22	Mexico
				English	18-22	United States
18	30-34	M	Monolingual	English	0-34	United States
19	25-29	M	Monolingual	English	0-27	United States
20	20-24	F	Bilingual	Spanish	0-22	Mexico
					0-22	United States
				English	7-22	United States
21	20-24	F	Bilingual	Spanish	0-20	United States
				English	6-20	United States
22	20-24	M	Monolingual	English	0-22	United States
23	20-24	M	Non-Native	Spanish	0-21	Mexico
					18-21	United States
				English	0-21	Mexico
					18-21	United States
24	20-24	F	Monolingual	English	0-22	United States

Interactive Systems Group

To build the next generation of user interface, our research group is collecting data from people communicating in various situations. We would like your help.

Project Name: Prosodic Patterns in Dialog

Purpose: This will enable us to identify differences in the ways people use prosody in communication, and eventually develop software for helping people learn to communicate more effectively in a new language or in their own language.

Activity: You will have two 10-minute conversations with different people.

Data to be Recorded: audio of both participants

Intended Use: We will statistically analyze patterns of prosodic behavior, to identify the patterns used and how they differ among people and languages.

Other Likely Uses: This data may be used for other research in our lab and elsewhere, fragments may be placed on the web or otherwise made publicly available for educational purposes, and the entire dialog may be made publicly available.

Protection: If, during the session, you say something that might embarrass someone or which you otherwise do not wish to be retained, please tell the experimenter. He or she will delete that part of the recording.

If you agree to participate, please cross out any "other likely uses" that you do not accept, then sign and date below.

I acknowledge receipt of $15 for my contribution to this project, and I agree not to assert copyright on the recording of my performance, but rather to dedicate it to the public domain.

_____	_____	_____
name	signature	date
_____	_____	_____
witness	signature	date

--- Participant Information ---

Subject ID: _____

Your age: 18-19, 20-24, 25-29, 30-34, 35-39, 40-49, 50-59, 60+

Your sex: M F

Your linguistic background:

From age	To age	Country	Main languages/dialects used
_____	_____	_____	_____
_____	_____	_____	_____
_____	_____	_____	_____
_____	_____	_____	_____
_____	_____	_____	_____
_____	_____	_____	_____
_____	_____	_____	_____