Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 5511–5516
Marseille, 20-25 June 2022

BILinMID: A Spanish-English Corpus of the US Midwest

Irati Hurtado
Department of Spanish & Portuguese, University of Illinois, Urbana-Champaign
707 S Mathews Ave., 4142 FLB, 61801 Urbana, IL, USA
ihurta3@illinois.edu

Abstract

This paper describes the Bilinguals in the Midwest (BILinMID) Corpus, a comparable text corpus of the Spanish and English spoken in the US Midwest by various types of bilinguals. Unlike other areas within the US where language contact has been widely documented (e.g., the Southwest), Spanish-English bilingualism in the Midwest has been understudied despite an increase in its Hispanic population. The BILinMID Corpus contains short stories narrated in Spanish and in English by 72 speakers representing different types of bilinguals: early simultaneous bilinguals, early sequential bilinguals, and late second language learners. All stories have been transcribed and annotated using various natural language processing tools. Additionally, a user interface has also been created to facilitate searching for specific patterns in the corpus as well as to filter out results according to specified criteria. Guidelines and procedures followed to create the corpus and the user interface are described in detail in the paper. The corpus is fully available online and it might be particularly interesting for researchers working on language variation and contact.

Keywords: Bilingualism, Language contact, Bilingual corpora

1. Introduction

The Hispanic population in the United States makes up to 15-20% of the nation's total population currently, and it is expected to continue growing in the next few decades (Pew Research Center, 2008). Due to its proximity to the US-Mexico border, Hispanics have more presence in the Southwest of the country (US Census, 2015), a situation that has inspired numerous studies on language contact and bilingualism in that area (e.g., Silva-Corvalán, 1994; Teschner, 2009; Travis et al., 2017). The predominance of the Southwest is also reflected in the few available corpora that document Spanish-English bilingualism in the United States (Gironzetti, 2021, 2022), which leave other US areas underrepresented in those datasets.

The Bilinguals in the Midwest (BILinMID) Corpus[1] constitutes an attempt to document the Spanish and English spoken in an area that has not received much attention in the literature but where the presence of Hispanics is certainly considerable, especially in recent years (Potowski, 2020). The creation of a text corpus such as BILinMID has important implications. On the one hand, the corpus serves as a pedagogical resource, since it can make learners value local language varieties and view them as a relevant source of learning. This is especially important in the case of Spanish, as US varieties of this language are often stigmatized (Hill, 1998). On the other hand, because the Midwest has a smaller bilingual population than other US regions, the corpus allows researchers interested in language contact, variation, and acquisition to better understand how the degree of societal bilingualism influences linguistic phenomena. For example, researchers could compare data from BILinMID to data from speakers living in highly bilingual regions (documented in other corpora from the US). This need to compare Spanish-English speakers across the nation living in different bilingual settings has already been pointed out by some scholars (Fuller and Leeman, 2020; Lynch, 2017). Likewise, the use of corpora for sociolinguistic research has also been underscored (Díaz-Campos and Torres, 2018). To this end, the BILinMID Corpus has a functional user interface that allows researchers without a technical background to easily explore the corpus.

This paper describes the process of creating the BILinMID Corpus as well as its user interface. It also reviews other available Spanish-English corpora from the US.

2. Related Work

Even though some corpora have been created to document the speech of Spanish-English bilinguals in the US (Table 1), most of them were compiled to investigate the use of Spanish in the country. Thus, in these corpora, English is only present in those cases where speakers code-switch.

Corpus	US Region (State)
Spanish in Texas Corpus	Texas
Corpus of Spanish in Southern Arizona	Arizona
Miami Corpus	Florida
Corpus of Mexican Spanish in Salinas	California
New England Corpus of Heritage and Second Language Speakers	Various locations
Polinsky Language Sciences Lab Dataverse	Various locations

Table 1: Corpora of Spanish-English bilinguals in the US

Out of the available corpora, the largest is the Spanish in Texas Corpus (Bullock and Toribio, 2014), which includes 96 individual sociolinguistic interviews with bilinguals from Texas. The corpus contains the video, audio, and transcription from each interview. Additionally, POS-tagged annotations are provided separately and can be explored using Google Data Studio. The second largest corpus is the Corpus of Spanish in Southern Arizona (Carvalho, 2012), comprised of 76 individual sociolinguistic interviews. Both the audios and transcriptions are available to researchers, as well as the meta-data. A third important corpus is the Miami Corpus (Deuchar, 2011), which includes 56 audio recordings of 84 Spanish-English bilinguals having conversations. Both the audios and the transcriptions can be downloaded in different formats for further analysis. A smaller corpus is the Corpus of Mexican Spanish in Salinas (Brown, 2022), containing only 11 individual interviews with first-

[1] The BILinMID Corpus can be accessed here: https://go.illinois.edu/BILinMID-corpus

generation immigrants from Mexico who currently live in California. Transcriptions have been annotated using TreeTagger and can be explored using a simple user interface.

Besides these corpora, there are two others which are more general but that also include speech samples from the speakers under consideration here. The first is the New England Corpus of Heritage and Second Language Speakers (Amaral and Gubitosi, 2013), which also includes data from second language learners, contrary to other corpora. This corpus includes both oral and written speech samples. Lastly, the Polinsky Language Sciences Lab Dataverse (Polinsky, 2015) is a data repository of speech samples by multiple types of speakers and languages. Some of them are Spanish-English bilinguals whose oral narratives were collected for different research projects and which are now available online to other researchers.

3. Speakers and Data Collection

The BILinMID Corpus is a bilingual comparable corpus of approximately 35,000 tokens which contains short stories narrated in English and in Spanish. Even though the stories were narrated orally by the speakers, only the texts corresponding to the transcriptions are part of the corpus (and not the audio recordings). To create a corpus that was representative of the bilingual population in the Midwest, three groups of bilinguals were targeted: early simultaneous bilinguals, early sequential bilinguals, and late second language learners. The first two groups are usually referred to in the literature as 'heritage speakers' (Montrul, 2016; Valdés, 2000). However, this term is not used in the corpus due to being too broad. The terminology chosen instead better reflects the specific language developmental pattern of each group of speakers, which is more informative. In order to collect data for the corpus, speakers were contacted through ads on social media, local organizations, and personal contacts. The data collection process took place in two sessions with at least two weeks in between sessions (Figure 1). This was done to prevent the first narrative from influencing the second narrative.

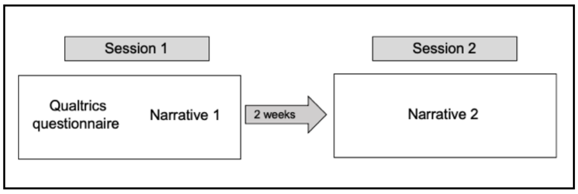


Figure 1: Data collection process

Speakers were recorded in a quiet and familiar space, usually in their office or at home. Speakers saw images depicting a famous children's fairy tale, Little Red Riding Hood. As they saw the images, they had to narrate the story orally in their own words. This fairy tale was chosen because it has been frequently used in research projects examining Spanish-English bilingualism in the US (Cuza et al., 2013; Montrul, 2004, 2010). Storytelling tasks such as this one promote natural speech and are a good source of linguistic phenomena of interest (Kisselev, 2021; Schmid, 2011). In the first session, speakers narrated the story in one language, and in the second session, they narrated the story in the other language (this was counterbalanced across participants). However, there were several instances of code-switching in the narratives. Lastly, even though a researcher was always present in the room, speakers were instructed not to interact with them at all while narrating the story. All narratives were recorded on the researcher's laptop using the Audacity software (Audacity Team, 2021) and speakers had full control of the images while narrating the story, since they could go forward or backward whenever they wanted to.

In addition to narrating the story in one of the languages, in the first session, speakers were also asked to fill out a questionnaire to gather more information about the demographics and linguistic background of each group. Furthermore, the questionnaire included a section where speakers completed the Bilingual Language Profile (BLP) (Birdsong et al., 2012), a standardized test to measure language dominance in the two languages. The BLP test consists of several multiple-choice questions that produce a final score ranging from -218 to 218. A score of 0 indicates balanced bilingualism (i.e., the speaker is equally dominant in Spanish as in English). A negative score indicates the speaker is more dominant in English than in Spanish, and a positive score indicates the speaker is more dominant in Spanish than in English. The inclusion of the BLP test responds to the need of reporting some sort of language proficiency or dominance metric in corpora containing speech from bilinguals (Kisselev, 2021). The full questionnaire was hosted online on Qualtrics (Qualtrics, 2005) so that speakers could fill it out conveniently from their smartphone or tablet.

Based on the answers to the questionnaire, there were 7 early simultaneous bilinguals (5 females, 2 males), 40 early sequential bilinguals (31 females, 9 males), and 25 late second language learners (18 females, 7 males). Speakers from the first two groups were all second-generation immigrants (i.e., either born in the US to at least one parent from a Spanish-speaking country, or born in a Spanish-speaking country and moved to the US with their parents before the age of 5). Simultaneous bilinguals had only one parent of Hispanic origin, which explains why they started speaking both Spanish and English simultaneously from birth. All sequential bilinguals, on the other hand, started speaking Spanish before English. In this latter case, either both their parents were from a Spanish-speaking country and continued speaking only Spanish at home while in the US, or the speaker was born in a Spanish-speaking country and moved to the US at a very young age. Regarding the group of late second language learners, all speakers were first-generation immigrants. Speakers in this group were raised monolingually in a Spanish-speaking country and moved to the US after puberty. They learned English as adults and, at the time they were recorded, they all had lived in the US for at least ten years (mean length of residence in the US: 17.68 years) and used English frequently on a daily basis.

The linguistic background of speakers is also reflected in their BLP scores, with the group of early simultaneous bilinguals being clearly more dominant in English than in Spanish, and the group of late second language learners being more dominant in Spanish than in English, overall (Table 2).

Group	N	Mean age (SD)	Mean BLP (SD)
Early simultaneous bilinguals	7	20 (2.29)	-58.42 (23.13)
Early sequential bilinguals	40	20 (1.57)	-12.54 (31.89)
Late second language learners	25	45 (6.30)	38.58 (25.35)

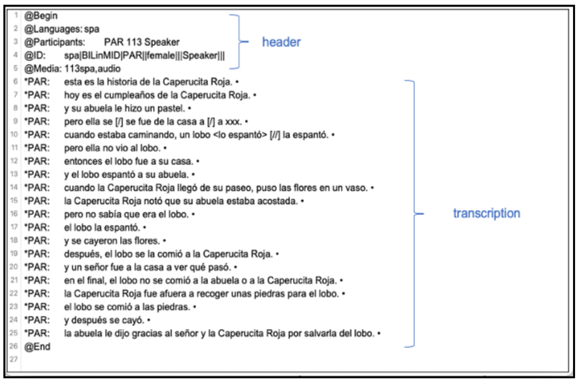
Table 2: Information about the speakers

It is also important to mention that there is a high variability of scores among the second-generation speakers, as indicated by the large standard deviations.

Information about the speakers' linguistic background and demographics was stored in a table format in a .csv file. This file was later used for the user interface (see section 6).

4. Transcription Process

All audio recordings with the short stories were transcribed using CLAN (MacWhinney, 2000). CLAN is an open-source software widely used in the language sciences. It consists of an editor where audio files can be imported, segmented, and easily transcribed thanks to its several features. Some of the corpora previously reviewed have also used CLAN in their transcription process (e.g., the Miami Corpus). The software uses the CHAT format for transcription (producing .cha files), which involves adding some special symbols to the transcribed text in order to provide more detail (e.g., pauses, hesitations, repetitions) as well as a header with information from the speaker (Figure 2). For the BILinMID Corpus, the April 2021 version of the CHAT manual was used.



```
@Begin
@Languages: spa
@Participants:	PAR 113 Speaker
@ID:	spa|BILinMID|PAR||female|||Speaker|||
@Media: 113spa,audio
*PAR:	esta es la historia de la Caperucita Roja. •
*PAR:	hoy es el cumpleaños de la Caperucita Roja. •
*PAR:	y su abuela le hizo un pastel. •
*PAR:	pero ella se [/] se fue de la casa a [/] a xxx. •
*PAR:	cuando estaba caminando, un lobo <lo espantó> [//] la espantó. •
*PAR:	pero ella no vio al lobo. •
*PAR:	entonces el lobo fue a su casa. •
*PAR:	y el lobo espantó a su abuela. •
*PAR:	cuando la Caperucita Roja llegó de su paseo, puso las flores en un vaso. •
*PAR:	la Caperucita Roja notó que su abuela estaba acostada. •
*PAR:	pero no sabia que era el lobo. •
*PAR:	el lobo la espantó. •
*PAR:	y se cayeron las flores. •
*PAR:	después, el lobo se la comió a la Caperucita Roja. •
*PAR:	y un señor fue a la casa a ver qué pasó. •
*PAR:	en el final, el lobo no se comió a la abuela o a la Caperucita Roja. •
*PAR:	la Caperucita Roja fue afuera a recoger unas piedras para el lobo. •
*PAR:	el lobo se comió a las piedras. •
*PAR:	y después se cayó. •
*PAR:	la abuela le dijo gracias al señor y la Caperucita Roja por salvarla del lobo. •
@End
```

Figure 2: A transcription in the CHAT format

The CHAT format also allows to mark errors as such. However, for this corpus, errors were left unmarked unless the transcribed word (or group of words) was difficult to interpret (e.g., if the speaker made up a new word). The motivation behind this choice is that the notion of 'error' is prescriptive and might offend some of the speakers whose narratives were recorded for this corpus. When working with these populations, this terminology should be avoided[2]. Furthermore, since the goal of this corpus is to simply provide a descriptive account of how Spanish-English bilinguals speak in the Midwest, error marking is not necessary.

Three researchers participated in the transcription process, all of them Spanish-English bilinguals familiar with the language varieties of the Midwest and with a solid background in linguistics. Even though all researchers followed the same guidelines and often discussed transcription issues together, a quantitative analysis was carried out to ensure inter-rater reliability. The quantitative analysis chosen was similar to the one used for the Miami Corpus (Deuchar, 2011) and which is described in Deuchar et al. (2014). In this case, 15% of all audio recordings were independently transcribed by two of the three researchers. Then, inter-rater reliability was assessed for each transcription file by calculating percent of agreement, which reached an average of 92%. Most discrepancies had to do with the use of CHAT special symbols and not with language issues (e.g., a researcher marked something as a pause whereas another marked it as a hesitation).

When the transcription process was over, each file was manually checked before starting the annotation process by being proofread by a researcher who did not work on it. This last check was done to make sure there were no spelling or formatting issues that could interfere with the annotation process.

5. Data Annotation

All .cha files with the transcriptions were converted into .txt files and were then imported into R (R Core Team, 2021). An R function was created to go over the transcription files and generate a dataframe or table linking the full transcriptions to basic information about the speakers. This was done as a first step to annotate the transcriptions, as dataframes allow us to work with data in a more efficient manner. The dataframe contained 6 columns:

- id: a number to identify each row
- text: the speaker's transcription as a single paragraph
- speaker: a unique code to identify each speaker anonymously
- generation: whether the speaker was a first- or second-generation immigrant
- gender: whether the speaker was a male or a female
- language: the language of the narrative (either English or Spanish)

All this information was extracted from the .txt files, either from the header or the transcription lines. The dataframe was exported as a .csv file.

Once all the transcriptions were conveniently stored in a dataframe, the annotation process began. The .csv file was imported into R and the *udpipe* R package (Straka et al., 2016; Wijffels, 2021) was loaded. *udpipe* is an R package designed for doing natural language processing directly in R without having to rely on other programming languages such as Python or Java. It allows tokenization, POS-tagging, lemmatization, and dependency parsing. *udpipe* uses pre-trained models which are available in many languages and which follow the Universal Dependencies (UD) framework (Nivre et al., 2016), a project that aims to develop consistent treebank annotation for multiple languages. Together with the R package, the pre-trained

[2] For a discussion on this, see Klee and Lynch (2009), Potowski and Lynch (2014), and Valdés (1995), among others.

Spanish and English models were downloaded and loaded into R. A new R function was created to iterate over each row (i.e., each full transcription) in the main dataframe and annotate the text using the appropriate language model. This generated a new, larger dataframe with one word per row. This dataframe contained 10 columns:

- id: a number to identify each row
- sentence: individual sentences from the transcriptions
- token_id: a number to identify each token in a transcription
- token: each token
- lemma: lemma corresponding to each token
- upos: POS-tag for each token based on the UD framework
- speaker: a unique code to identify each speaker anonymously
- generation: whether the speaker was a first- or second-generation immigrant
- gender: whether the speaker was a male or a female
- language: the language of the narrative (either English or Spanish)

This dataframe containing the annotations was also exported as a .csv file. Since the process of annotation was automatically done relying on the *udpipe* package and the pre-trained models, the .csv file was manually checked by three researchers to ensure all the information was correct. These were the same researchers that had previously worked on the transcription process. For this manual check, a set of guidelines was created describing how to deal with some of the most common issues found in the annotated dataframe (e.g., how to annotate passages with code-switching, how to annotate CHAT special transcription symbols). All researchers followed the guidelines to ensure the annotations were consistent and reliable, and any discrepancies were discussed as a group.

This revised dataset constitutes the basis for the user interface that was developed to explore the corpus.

6. User Interface

The user interface developed for the BILinMID Corpus was built as an R-Shiny app using the *shiny* R package (Chang et al., 2021). R-Shiny apps are interactive web applications that can be built directly in R and which can be further customized with HTML, CSS, and JavaScript. More specifically, these apps are useful to interact with data in tabular format. Since the datasets (i.e., the .csv files) followed that format, developing an R-Shiny app was a good option.

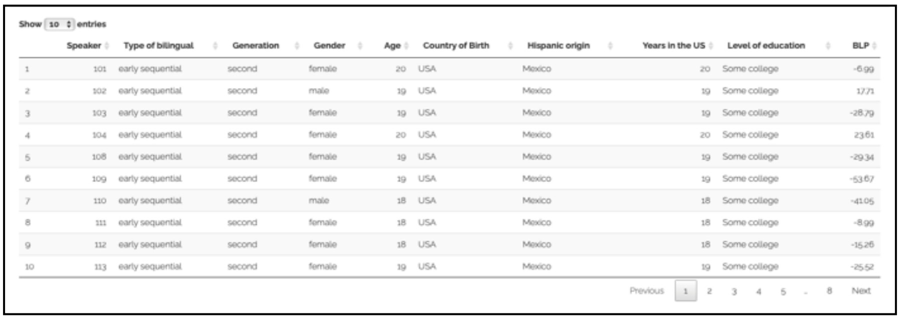
The user interface developed for the BILinMID Corpus is simple and intuitive. It has a horizontal navigation bar at the top with several tabs, as follows:

- Home: the BILinMID Corpus homepage
- About this corpus: a general description about the corpus and the research team
- How to use this corpus: information about what users can find in the different tabs
- The speakers: information about the speakers based on responses from the Qualtrics questionnaire
- Search by KWIC: to search for a specific keyword in the corpus
- Search by lemma: to search for a specific lemma in the corpus
- Search full transcriptions: to search for a transcription given a speaker and a language

To navigate the app, users simply have to click on a tab and they will be taken to that page. Important for our purposes are the four last tabs. The 'the speakers' page displays a table through the *DT* R package (Xie et al., 2021), which is an R interface for the JavaScript library *DataTables*. The table displayed on the app comes from the .csv file containing the information about the speakers' background and demographics (e.g., gender, age, generation, BLP score) (Figure 3). Thus, this table provides the relevant meta-data for the BILinMID Corpus.

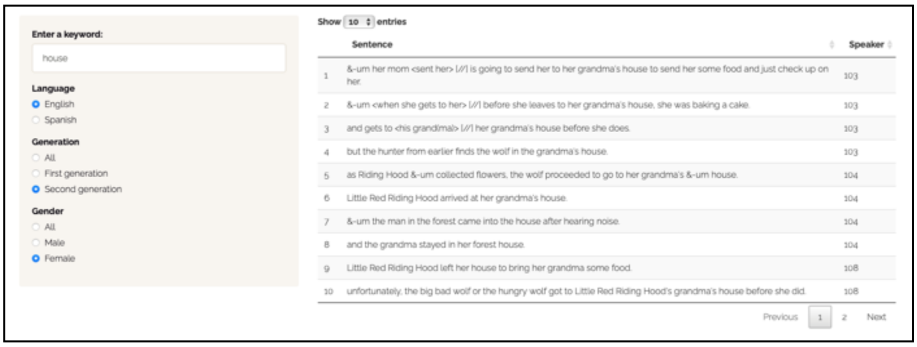
The pages 'search by KWIC' and 'search by lemma' are very similar to one another and provide an easy way to query the corpus. These pages contain a browser on the left together with filters for language, generation, and gender. Users can type in a word (either a keyword or a lemma) and adjust the filters, and a JavaScript table will appear on the right of the page containing any sentences from the corpus that match the query. The table will also list the speaker who produced the sentence. The dataset searched for during these queries is the annotated .csv dataframe that was created with *udpipe*. When the user executes a query, the keyword or lemma is found in the dataframe and the sentences containing the match are selected together with their speakers (Figure 4).

Lastly, the 'search full transcriptions' page has a language filter and a speaker filter on the left. Here, users will see the full transcription of the speaker and language they choose (Figure 5). This transcription can be copied and exported elsewhere for further analysis.



	Speaker	Type of bilingual	Generation	Gender	Age	Country of Birth	Hispanic origin	Years in the US	Level of education	BLP
1	101	early sequential	second	female	20	USA	Mexico	20	Some college	-6.99
2	102	early sequential	second	male	19	USA	Mexico	19	Some college	17.71
3	103	early sequential	second	female	19	USA	Mexico	19	Some college	-28.79
4	104	early sequential	second	female	20	USA	Mexico	20	Some college	23.61
5	108	early sequential	second	female	19	USA	Mexico	19	Some college	-29.34
6	109	early sequential	second	female	19	USA	Mexico	19	Some college	-53.67
7	110	early sequential	second	male	18	USA	Mexico	18	Some college	-41.05
8	111	early sequential	second	female	18	USA	Mexico	18	Some college	-8.99
9	112	early sequential	second	female	18	USA	Mexico	18	Some college	-15.26
10	113	early sequential	second	female	19	USA	Mexico	19	Some college	-25.52

Figure 3: Page with information from the speakers



Enter a keyword:

house

Language

● English

○ Spanish

Generation

○ All

○ First generation

● Second generation

Gender

○ All

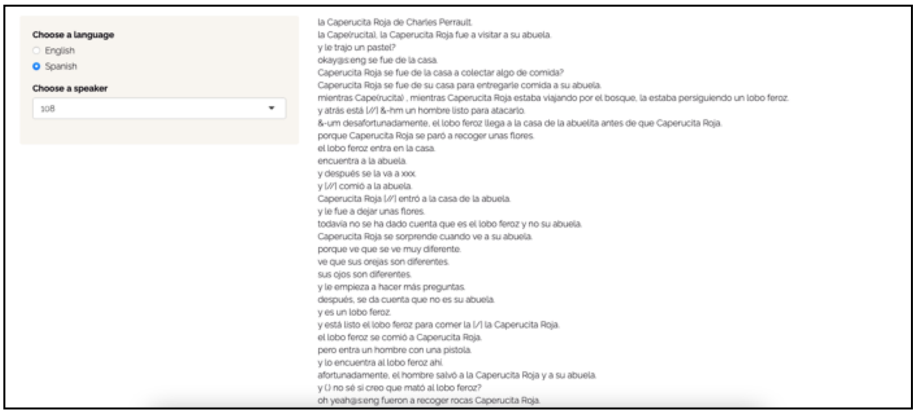
○ Male

● Female

Show 10 entries

	Sentence	Speaker
1	&-um her mom <sent her> [//] is going to send her to her grandma's house to send her some food and just check up on her.	103
2	&-um <when she gets to her> [//] before she leaves to her grandma's house, she was baking a cake.	103
3	and gets to <his grandma> [//] her grandma's house before she does.	103
4	but the hunter from earlier finds the wolf in the grandma's house.	103
5	as Riding Hood &-um collected flowers, the wolf proceeded to go to her grandma's &-um house.	104
6	Little Red Riding Hood arrived at her grandma's house.	104
7	&-um the man in the forest came into the house after hearing noise.	104
8	and the grandma stayed in her forest house.	104
9	Little Red Riding Hood left her house to bring her grandma some food.	108
10	unfortunately, the big bad wolf or the hungry wolf got to Little Red Riding Hood's grandma's house before she did.	108

Figure 4: The ‘search by KWIC’ page



Choose a language

○ English

● Spanish

Choose a speaker

108

la Caperucita Roja de Charles Perrault.
la Cape(rucita), la Caperucita Roja fue a visitar a su abuela.
y le trajo un pastel?
okay@s:eng se fue de la casa.
Caperucita Roja se fue de la casa a colectar algo de comida?
Caperucita Roja se fue de su casa para entregarle comida a su abuela.
mientras Cape(rucita) , mientras Caperucita Roja estaba viajando por el bosque, la estaba persiguiendo un lobo feroz.
y atrás está [//] &-hm un hombre listo para atacarlo.
&-um desafortunadamente, el lobo feroz llega a la casa de la abuelita antes de que Caperucita Roja.
porque Caperucita Roja se paró a recoger unas flores.
el lobo feroz entra en la casa.
encuentra a la abuela.
y después se la va a xxx.
y [//] comió a la abuela.
Caperucita Roja [//] entró a la casa de la abuela.
y le fue a dejar unas flores.
todavía no se ha dado cuenta que es el lobo feroz y no su abuela.
Caperucita Roja se sorprende cuando ve a su abuela.
porque ve que se ve muy diferente.
ve que sus orejas son diferentes.
sus ojos son diferentes.
y le empieza a hacer más preguntas.
después, se da cuenta que no es su abuela.
y es un lobo feroz.
y está listo el lobo feroz para comer la [/] la Caperucita Roja.
el lobo feroz se comió a Caperucita Roja.
pero entra un hombre con una pistola.
y lo encuentra al lobo feroz ahí.
afortunadamente, el hombre salvó a la Caperucita Roja y a su abuela.
y () no sé si creo que mató al lobo feroz?
oh yeah@s:eng fueron a recoger rocas Caperucita Roja.

Figure 5: The ‘search full transcriptions’ page

7. Conclusions and Future Work

This paper has introduced the BILinMID Corpus, a comparable corpus of the Spanish and English spoken in the US Midwest by different types of bilinguals. The corpus contains the transcriptions of short stories narrated by the speakers together with relevant meta-data. The corpus also includes a practical user interface developed with R-Shiny to facilitate exploring the datasets. To my knowledge, this is the first corpus documenting the speech of bilinguals in this region of the US.

In terms of future work, the BILinMID Corpus is still in progress. Therefore, more speakers will be recorded in the near future, especially from the early simultaneous bilingual group and the late second language learner group. The goal is to have roughly the same number of speakers per group (~40-50 speakers). Likewise, the user interface will also be enhanced to support other types of queries and to provide some analytics. Even though this requires more processing of the *udpipe* output, these new features will make the user interface more functional.

8. Acknowledgements

I thank A. Andino and A. Seielstad for their assistance transcribing the audio recordings and annotating the corpus' main dataset; P. Renteria and J. Vazquez for their help during the data collection process; and J. Canty for his assistance developing the user interface.

9. Bibliographical References

Audacity Team (2021). *Audacity(R): Free audio editor and recorder (computer application)*. Available at https://audacityteam.org/

Birdsong, D., Gertken, L., and Amengual, M. (2012). *Bilingual Language Profile: An easy to use instrument to assess bilingualism*. COERLL, University of Texas, Austin, TX.

Fuller, J., and Leeman, J. (2020). *Speaking Spanish in the US: The sociopolitics of language*. Bristol, UK: Multilingual Matters.

Chang, W., Cheng, J., Allaire, J. J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., and Borges, B. (2021). *Shiny: Web application framework for R*. CRAN. Available at https://CRAN.R-project.org/package=shiny

Cuza, A., Pérez-Tattam, R., Barajas, E., Miller, L., and Sadowski, C. (2013). The development of tense and aspect morphology in child and adult heritage speakers. In J. Schwieter (Ed.), *Innovative research and practices in second language acquisition and bilingualism*. Amsterdam, The Netherlands: John Benjamins, pp. 192-220.

Deuchar, M., Davies, P., Herring, J. R., Parafita Couto, M. C., and Carter, D. (2014). Building bilingual corpora. In E. Thomas, and I. Mennen (Eds.), *Advances in the study of bilingualism.* Bristol, UK: Multilingual Matters, pp. 93-110.

Díaz-Campos, M., and Torres, J. E. (2018). Corpus approaches to the study of language, variation, and change. In K. Geeslin (Ed.), *The Cambridge handbook* *of Spanish linguistics*. Cambridge, UK: Cambridge University Press, pp. 121-141.

Gironzetti, E. (2021). Pragmática y multimodalidad en el español como lengua de herencia. In D. Pascual y Cabo, and J. Torres (Eds.), *Aproximaciones al estudio del español como lengua de herencia*. London, UK: Routledge, pp. 66-78.

Gironzetti, E. (2022). Corpus del español como lengua de herencia. In G. Parodi, P. Cantos-Gómez, and C. Howe (Eds.), *The Routledge handbook of Spanish corpus linguistics.* London, UK: Routledge.

Hill, J. (1998). Language, race, and white public space. *American Anthropologist, 100*(3): 680-689.

Kisselev, O. (2021). Corpus-based methodologies in the study of heritage languages. In S. Montrul, and M. Polinsky (Eds.), *The Cambridge handbook of heritage languages and linguistics*. Cambridge, UK: Cambridge University Press, pp. 520-544.

Klee, C., and Lynch, A. (2009). *El español en contacto con otras lenguas*. Washington, DC: Georgetown University Press.

Lynch, A. (2017). The 'in-between' paradigm in Spanish as a heritage language. Talk given at the *2017 Heritage Spanish Workshop*. University of Texas, Austin, TX.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk*. Mahwah, NJ: Lawrence Erlbaum Associates.

Montrul, S. (2004). Subject and object expression in Spanish heritage speakers: A case of morpho-syntactic convergence. *Bilingualism: Language and Cognition, 7*(2): 125-142.

Montrul, S. (2010). How similar are adult second language learners and Spanish heritage speakers? Spanish clitics and word order. *Applied Psycholinguistics, 31*(1): 167-207.

Montrul, S. (2016). *The acquisition of heritage languages.* Cambridge, UK: Cambridge University Press.

Nivre, J., de Marneffe, M-C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., and Zeman, D. (2016). Universal dependencies v1: A multilingual treebank collection. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis (Eds.), *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016).* Paris, France: European Language Resources Association (ELRA), pp. 1659-1666.

Pew Research Center (2008). *US Population Projections: 2005-2050*. Available at https://www.pewresearch.org/hispanic/2008/02/11/us-population-projections-2005-2050/ (last accessed January 3, 2022).

Potowski, K., and Lynch, A. (2014). La valoración del habla bilingüe en los Estados Unidos: Fundamentos lingüísticos y pedagógicos en 'Hablando bien se entiende la gente'. *Hispania, 97*(1): 32-46.

Potowski, K. (2020). Spanish in the Midwest: Hablando in the Heartland. In F. Salgado-Robles, and E. Lamboy (Eds.), *Spanish across domains in the United States: Education, public space, and social media*. Boston, MA: Brill, pp. 65-93.

Qualtrics (2005). *Qualtrics*. Provo, UT. Available at https://www.qualtrics.com/

R Core Team (2021). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at https://www.R-project.org/

Schmid, M. (2011). *Language attrition*. Cambridge, UK: Cambridge University Press.

Silva-Corvalán, C. (1994). *Language contact and change: Spanish in Los Angeles.* New York City, NY: Oxford University Press.

Straka, M., Hajič, J., and Straková, J. (2016). UDPipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis (Eds.), *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016).* Paris, France: European Language Resources Association (ELRA), pp. 4290-4297.

Teschner, R. (1995). Beachheads, islands, and conduits: Spanish monolingualism and bilingualism in El Paso, Texas. *International Journal of the Sociology of Language, 114*: 93-105.

Travis, C., Torres-Cacoullos, R., and Kidd, E. (2017). Cross-language priming: A view from bilingual speech. *Bilingualism: Language and Cognition, 20*(2): 283-298.

US Census (2015). *2011-2015 ACS 5-year estimates.* Available at https://www.census.gov/programs-surveys/acs/technical-documentation/table-and-geography-changes/2015/5-year.html (last accessed January 3, 2022).

Valdés, G. (1995). The teaching of minority languages as academic subjects: Pedagogical and theoretical challenges. *Modern Language Journal, 79*: 299-328.

Valdés, G. (2000). Introduction. In AATSP (Ed.), *Spanish for native speakers.* Fort Worth, TX: Harcourt College, pp. 1-20.

Wijffels, J. (2021). *Package 'udpipe'*. CRAN. Available at https://CRAN.R-project.org/package=udpipe

Xie, Y., Cheng, J., and Tan, X. (2021). DT: A wrapper of the JavaScript library 'DataTables'. CRAN. Available at https://CRAN.R-project.org/package=DT

10. Language Resources

Amaral, P., and Gubitosi, P. (2013). *New England Corpus of Heritage and Second Language Speakers*. Available at https://digitalhumanities.umass.edu/projects/new-england-corpus-heritage-and-second-language-speakers

Brown, E. (2022). *Corpus of Mexican Spanish in Salinas, California.* Available at http://itcdland.csumb.edu/~eabrown

Bullock, B., and Toribio, J. (2014). *Spanish in Texas Corpus*. Available at https://spanishintexas.org

Carvalho, A. (2012). *Corpus del Español en el Sur de Arizona (CESA)*. Available at https://cesa.arizona.edu

Deuchar, M. (2011). *The Miami Corpus*. Available at http://bangortalk.org.uk

Polinsky, M. (2015). *The Polinsky Language Sciences Lab Dataverse*. Available at https://dataverse.harvard.edu/dataverse/polinsky