BILinMID
A Spanish-English Corpus of the U.S. Midwest
Irati Hurtado · University of Illinois, Urbana-Champaign

1 Introduction

Aim of the corpus
Document the Spanish and English spoken in the U.S. Midwest by various types of bilinguals

Motivation
Current corpora documenting Spanish–English bilingualism in the U.S. focus mostly on the Southwest

Implications
Pedagogical resource for learners
Research resource to study language contact, variation, and acquisition

2 Speakers

Groups of speakers
Early simultaneous bilinguals (N = 7)
Early sequential bilinguals (N = 40)
Late second language learners (N = 25)

3 Corpus Data

Type of data
Oral narratives of the Little Red Riding Hood story
Data elicited through pictures
Story told once in Spanish and once in English by each speaker
The corpus contains the written transcriptions only (no audios)

4 Data Collection Process

Session 1
Demographic questionnaire
Bilingual Language Profile (BLP)
Narrative 1

Session 2
Narrative 2

There were at least two weeks in between sessions
The language of the narrative was counterbalanced

5 Data Processing

Oral transcriptions in CHAT format using CLAN software

CHAT transcriptions converted to plain text files

Data annotated with udpipe in R (lemmatization, POS-tagging)

Manual check

Output

Dataframe with full transcriptions
Dataframe with individual tokens

6 User Interface

https://go.illinois.edu/BILinMID-corpus

Web application
Built with R–Shiny
Customized with HTML, CSS, JavaScript

Meta-Data
Tab with information about the speakers

Supported queries
Search by KWIC
Search by lemma
Search full transcriptions