Computational Modeling of English-Spanish Bilingual Reading: Effects of Oral Language Profiles and Initial Literacy Instruction

Nuria Gutiérrez[1], Matt Cooper Borkenhagen[1], Valeria Rigobon[2], Philippe Miron[1], and Nathan Crock[1,3]

[1]Florida State University, Australian Centre for the Advancement of Literacy[2], Emelex[3]

Introduction

The purpose of this poster is to present a multilingual extension of the *BRIDGE* model, a computational model focused on learning the connection between orthographic and phonological mappings. This multilingual model incorporates bilingual oral language profiles and varying word learning environments. By simulating learners with different English–Spanish language backgrounds, we explored how pre-existing oral language knowledge interacts with early reading instruction in either language.

Experimental Design

3 x 2

(Language Profile: Spanish-Dominant, English-Dominant, Balanced) × (Literacy Instruction Language: Spanish, English)

Step 1: Phonological Pretraining – Phonology to Phonology training (P2P)

Goal:

Pre-train models on spoken language phonology before reading instruction

Datasets:

- English: 4000 frequent words from SUBTLEX-US
- Spanish: 4000 frequent words from EsPAL

Phonology:

Each phoneme in the word is represented by a 33-dimensional vector. Each dimension represents the presence of particular speech sounds, like fricative, rhotic, alveolar, palatal, glottal, etc.

Learners:

Bilingual Learner	% Spanish Vocabulary	% English Vocabulary
Spanish Dominant	80% (3200 words)	20% (800 words)
English Dominant	20% (800 words)	80% (3200 words)
Balanced Bilingual	50% (2000 words)	50% (2000 words)

Step 2: Literacy Training – Orthography to Phonology training (O2P)

Goal:

Simulate phonological recoding (orthography-to-phonology mapping)

Datasets:

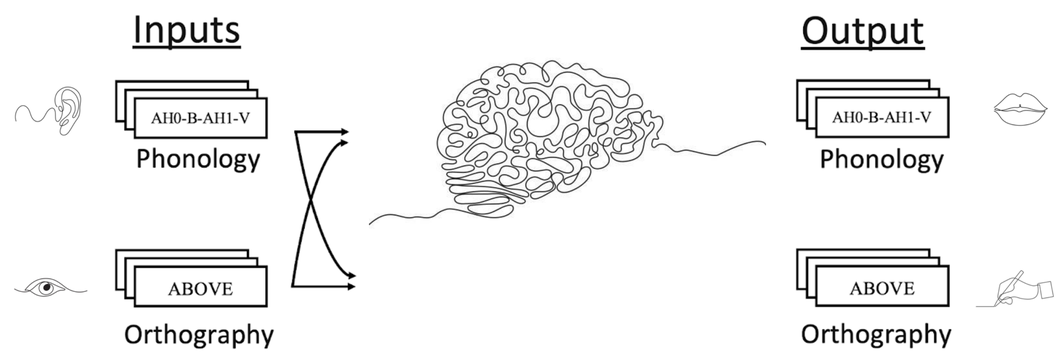
- English: 6000 words from 1st and 2nd grade books from TASA (Touchstone Applied Science Associates database, Zeno, 1995)
- Spanish: 6000 most frequent words from the dictionary of Frequencies for written language in children 6–12 years of age by Martínez & García (2004)

Training:

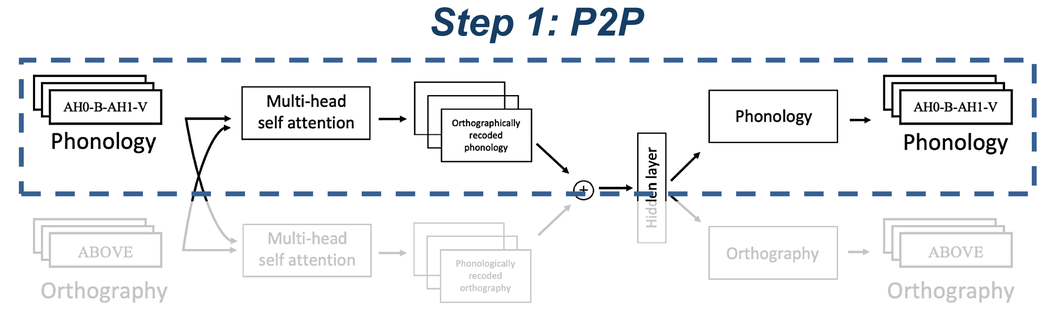
All 3 learners will be trained on both datasets:

- English O2P training (95% English, 5% Spanish- high frequency words)
- Spanish O2P training (95% Spanish, 5% English- high frequency words)

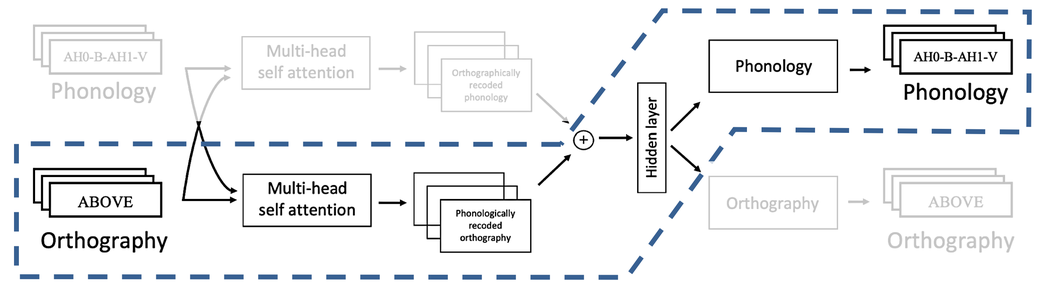
The BRIDGE Model: Architecture



Step 1: P2P

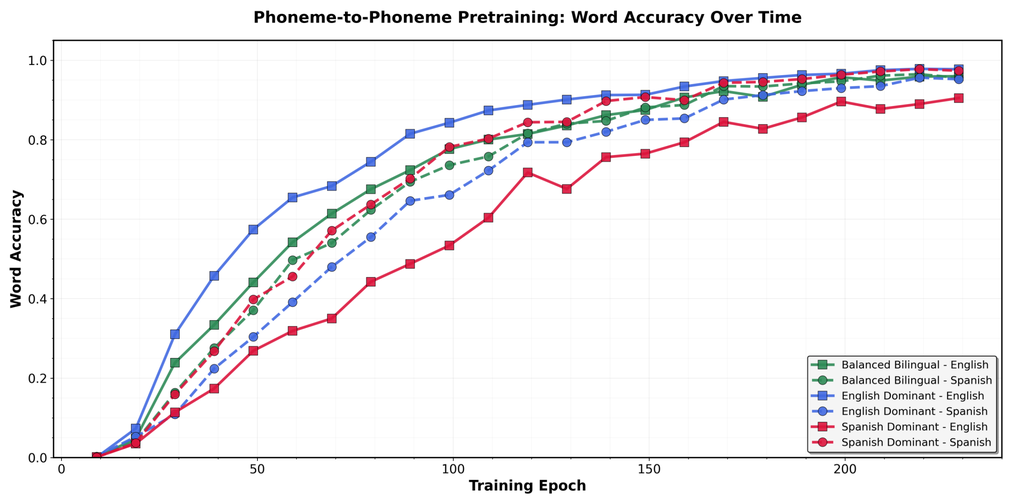


Step 2: O2P

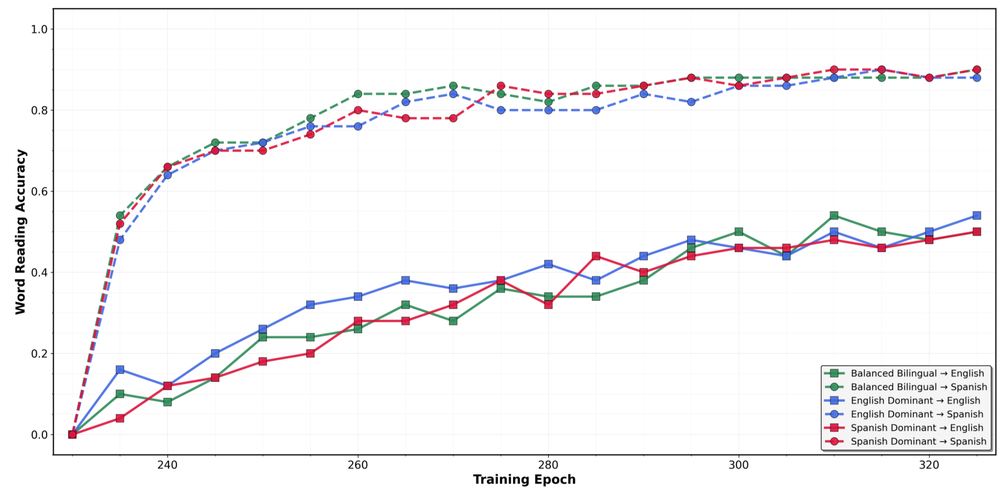


Preliminary Results

Step 1: Phonological Pretraining –P2P



Step 2: Orthography to Phonology training (O2P)



Next steps

- Finalize epoch selection criteria for P2P phase
- Analyze cross-linguistic transfer patterns
- Compare training sequences: first English → then Spanish vs. first Spanish → then English
- Use distance-based accuracy metrics instead of binary scoring
- Conduct word-level analysis based on feature characteristics

Questions? Interested in collaborating? Let's keep in touch!

ngutierrez@fsu.edu