An English-Spanish Reading Network

Kyla Cheung
Department of Computer Science
Columbia University
kyc2114@columbia.edu

Advisor: Eliana Colunga
Department of Psychology
Institute of Cognitive Science
University of Colorado at Boulder
Eliana.Colunga@colorado.edu

Abstract

Although education in the United States predominantly assumes a monolingual population, the large presence of multilinguals put pressure upon the existing model to accomodate students who were not raised in households that speak English. Simply thrusting children who speak another language into an English-medium school has been shown not to be a viable solution; while students may eventually master the ability to decode written English words, their reading comprehension in the long run falters in comparison to their monolingual classmates. We suggest a computational approach using neural networks to simulate the process of learning to read and write in two languages, in the hopes of uncovering a more effective way to answer this question: does the greater problem lie with the children's lack of L2 mastery, or does it lie with the children's lack of familiarity with the semantic-orthographic relationship? This report does not include our answers, being merely a write-up of our current progress after ten weeks. However, we will delineate our methods and conjectures based on the experiments we have run.

1 Introduction

There is a growing number of children entering American schools as English-language Learners (ELs), supplemented by a dearth of research on their particular development in language and literacy. By 2030, they will make up 40 percent of the K-12 population (The Center for Research on Education, Diversity, Excellence, 2002; as cited by Ayre, 2010). Despite this, there is a surprising shortage of research that can point to the best educational practices to handle these children's language and literacy development (Genesee, Paradis, Crago, 2004; Gutierrez-Clellen, 2002; McCardle, 2006; Slavin Cheung, 2005; as cited by Ayre, 2010). This is made more concerning by evidence that suggests ELs do not keep up with their English L1 speaking peers in reading comprehension (August et al., 2005; Verho- even, 2000; as cited by Gottardo, 2009). In fact, according to the National Assessment of Education Progress, English L1 speakers outperform EL peers in reading comprehension and writing at the 4th-, 8th-, and 12th-grade levels (2006, as cited by Gottardo, 2008).

The sparse literature on how to address this problem points to several predictors of L2 oral and written proficiency. In our research, we have chosen to focus on one basic dilemma in bilingual literacy: whether it is best to first develop L2 oral proficiency or to first develop L1 literacy. [1] There is research to suggest it is not necessarily the best route to first develop L2 oral proficiency, finding that oral language proficiency does not predict word reading performance (Chiappe et al., 2007; Dickinson et al., 2004; Durgonglu et al., 1993; Gottardo, Yan, Siegel, Wade-Woolley, 2001; Roberts, 2005; as cited by Ayre, 2010). While this suggests SpELL schoolchildren do not need to immediately begin L2 learning, there is enough research to suggest otherwise. Gottardo et al. (2008) found that L2 phonological awareness in first grade SpELLs predicted second grade word attack and word identification. These findings seem to extend into the upper grades as well. Proctor et al. in 2006 had findings

[1] Besides practical applications, there are theoretical reasons to test these two hypotheses as well. If L1 literacy can indeed goad on progress in L2, then there may be reason to argue for certain language universal processes. If not, then it could point to the possibility that literacy is a language specific process. Of course, either, neither, or both could be possible.

Word	Phonetic Rep	Stress	1	2	3	4	5	6	7 (Stressed)	8	9	10	11	12	13
pingüino	p i n g u i n o	g u i	p	p	p	i	n	g	u	i	n	o	o	o	o
penguin	p e ng g w I n	p e n	p	p	p	p	p	p	e	ng	g	w	I	n	n

Figure 1: Two phonological representations. The vowel of the stressed syllables 'gui' and 'pen' are first put into the 7th column, then the rest of the phonetic representation is written out, and finally it's filled out to the ends of the matrix. (Note the repeated p's, o's, and n's.) This is depicted in a matrix too large to show here, but each of the thirteen columns has 49 possible phonemes, which means 49 rows, making for a very sparse matrix.

that suggested reading comprehension performance in SpELL fourth graders was more language-specific, seeing that initial language of instruction and assessment predicted within-language oral language proficiency and reading comprehension (as cited by Ayre, 2010).

Yet in the vein of fostering *bilingual* literacy, there still lies the need to develop L1 reading comprehension. Lingsey et al. (2003) measured L2 English reading comprehension in SpELLs in kindergarten and first grade and found that both Spanish and English measures of phonological processing and listening comprehension were related to English reading comprehension, measured by varying tests of word-level knowledge and decoding (as cited by Gottardo, 2009). By the end of first grade, these SpELL children not only were taught to read Spanish, but had also reached approximately average levels of English word reading (Lindsey et al., 2003; as cited by Gottardo, 2009). On the whole, it remains unclear which to develop first, L2 oral skills or L1 literacy, since L2 reading comprehension seems to be related to both L1 and L2 skills.

Both approaches can be supported by evidence, but there are limitations to traditional data collection. Namely, it is difficult to control for differing linguistic abilities or varying environments. Children pick up words from a variety of sources, and to limit their exposure to language, in order to control the linguistic environment, would be unethical to enact in an observational study. What a computational model offers is the ability to "turn back time", so to speak, in a uniquely controlled setting where both the learner and the words being learned are kept constant and where the time required to learn a skill, as measured in epochs, is quantified and readily comparable. Our model builds off the 'triangle' model (Plaut et al., 1996), which is seen as an authoritative neural network on reading aloud.

2 Methods

2.1 Architecture

We chose to use Emergent to build and implement the model, going off a previous model that had used the same software. The original model is what is known as the "triangle model," based on the work started by Seidenberg and McClelland (1989), extended by Plaut et al. (1996), and further implemented for Emergent software. Our modifications consisted solely of parameter and representation adjustments, without renovating the number of layers. The modifications most significantly allowed the phonology layer to represent trisyllabic words, while the model previously only accommodated monosyllabic words; this was mostly motivated by the proliferation of multisyllabic words in Spanish, wherein a limitation to monosyllabic words would be unrealistic to a child's (or anybody's) experience learning the language.

For each word, there are three mappings: orthographic, phonological, and semantic. The semantic layer, which is 5x20, uses the aforementioned Howell data, which surveyed adult subjects on the semantics of terms in the English MCDI. The orthographic matrix was 11x6 and used the minimum number of letters to fully represent any Spanish or English term in the training data. The matrix entries themselves were organized alphabetically. Finally, the phonological matrix, by far the largest, was 49x13. Within each of the columns, only one phoneme was represented, so the matrix sum was 13, while each column sum was 1 and each row sow sum was either 0 or 1. The decision to use 13 columns was to use the minimum number in order to sufficiently stress- and vowel-center the phonetic representations. Fig 1. depicts the phonological representation in better detail.

2.2 Training

Our training material was largely based off the MacArthur-Bates Communicative Development Inventory (MCDI) for both Spanish and English. The MCDI is a standardized assessment of the language development of children from 8 to 36 months-old. The versions we used had a list of over 300 words (329 in the Spanish MCDI and 395 words in the English MCDI), from which parents report which words their children have produced. In addition to the MCDI's, we also used supplementary semantic material provided by Howell for the English MCDI. We found the words common to both MCDI's, as well as Howell's data, and made the assumption that the same word in English and Spanish (e.g., 'penguin' and 'pinguino') had the same semantics. In total, we used 396 words, half of which were English and the other half in Spanish, both sets direct translations of one another.

2.3 Testing

Our procedure consisted of several different tests, which Figure 2 depicts. Part A had four trainings which kept the model monolingual: (1) training the network to speak, then read; (2) training the network to speak halfway to criterion then read; (3) train on all tasks at the same time; and (4) train only on reading. Part B contained the same four trainings as part A, but prepended each with training the model to speak Spanish fully to criterion. Part C consisted of two parts, which focus on testing our dual hypotheses: (1) train the model to speak Spanish, read Spanish, speak English, and read English, in that order; and (2) train the model to speak English, then read English, speak Spanish, and read Spanish, in that order. (This second part of the experimental design came out of a desire to see if the results seen in C1 were merely an artifact of Spanish somehow making English "easier" to learn after the fact. If the results in both C1 and C2 were comparable, then there might be basis to test their conclusions for other multilingual education, not just Spanish-English and English-Spanish.)

Monolingual:				
A1	SE 100%	RE 100%		
A2	SE 50%	RE 100%		
A3	SE & RE 100%			
A4	RE 100%			
A1	SS 100%	RS 100%		
A2	SS 50%	RS 100%		
A3	SS & RS 100%			
A4	RS 100%			
Bilingual:				
B1	SE	SE 100%	RE 100%	
B2	SE	SE 50%	RE 100%	
B3	SE	SE & RE 100%		
B4	SE	RE 100%		
C1	SS	RS	SE	RE
C2	SE	RE	SS	RS

Figure 2: The testing procedure. "Speaks English" abbreviates to "SE"; "Speaks Spanish" to "SS"; "Reads English" to "RE"; and "Reads Spanish" to "RS".

3 Results

We ended up setting criterion to be an average sse of 2. Figures 2 and 3 show the model's approach to criterion of various tasks within sequences previously described: SS being "Speaks Spanish," RS being "Reads Spanish," SE being "Speaks English," and RE being "Reads English."

Figure 3 draws a comparison between three different models, each of which learned to speak Spanish, then went on to different paths that ultimately arrived at reading English to criterion. The most significant observation, perhaps, is that it took all the models the same number of epochs (2) to reach criterion to speak English. This is more likely a result of the implementation problems (described in Limitations) than the more exciting possibility that reading English is its own, and relatively easy, linguistic process that takes all children a set amount of time to learn, *regardless* of phonological awareness of English *or* of orthographic awarness of Spanish. This might suggest that if the aim for students is to learn English, then they should forgo preliminary education in Spanish; yet, given the relatively short amount of time required for this model to learn to read English, it would not incur a great cost to learn Spanish.

Figure 4 depicts an English-Spanish and Spanish-English model that both learned to read L1 before learning to speak or read L2. What is significant to note is the two models take relatively the same number of epochs (57 for English-Spanish and 56 for Spanish-English) to reach bilingual literacy in both languages. Since the two models followed the same trajectory (Speak L1, Read L1, Speak L2, Read L2), it is an important observation that they reached each milestone at similar times. This suggests that this type of language and literacy learning may, in fact, be more language-independent than originally anticipated. This could still be due to the great similarity between English and Spanish in phonology and orthography, pointing to the necessity to carry out these experiments with different language combinations.

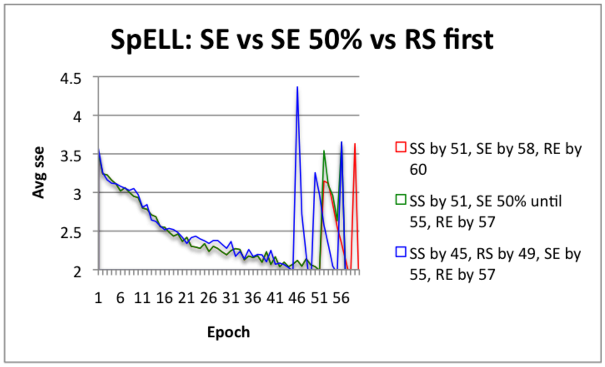

Figure 3: Comparison of three SpELL models, each on different sequences towards reading English.

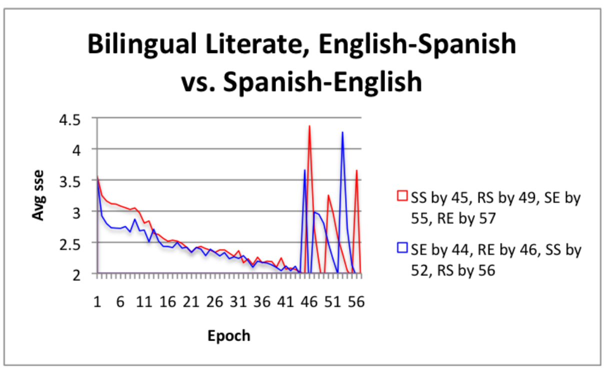

Figure 4: Comparison of two bilingual models, one Spanish-English and the other English-Spanish.

4 Limitations

By the end of ten weeks of work, the model was unable to output phonological codes, instead it only took phonology as input. So the processes given as "reading" and "speaking" are more akin to "dictation" and "listening," given phonology was the input. There were several other lingering issues with the model, as well. It was unable to learn both the Spanish "reading" and "speaking," and the B C phase of our procedure were only able to be run if each task was run separately. This meant that the network is not able to know, at the same time, how to do certain groups of tasks. This casts severe doubt on our already preliminary results shown above, since the purpose of the model is to see how multiple tasks (like, speaking English and speaking Spanish) can be carried out by the same model.

The likely cause is the extreme sparseness of the phonological input (13 1's in a sea of 637 total inputs). Our previous attempts tinkered with the parameters and the learning rate of the model, in order to stop thrashing and to properly wrangle a prefabricated model for our purposes. We are currently trying to find a solution to the problem, using binary representations of the phonology and expanding the size of the hidden layers.

Beyond the limitations of our particular model, there also still remains the quandary of how to apply the results to real-life situations, in which an "epoch" does not easily translate into specific units of time.

5 Conclusion and Future Work

Since our model was not properly functional by the end of ten weeks, the obvious next step would be to make it functional for our purposes, by making the necessary modifications so that the network can function with the phonological layer set to target.

Another extension that we have already begun is re-doing the model in Hindi-Urdu, and potentially South Indian languages to make the model trilingual. Being home to over 100 languages, India has been the site of intense debate over which languages to teach and which to make official. Even the idea that Hindi is the official language of India is sometimes met with incredulity, though an English-medium education is prized by many at the same time. Questions about multilingual education in India are pertinent and ones that our model may be apt to answer.

This proposal to extend the model in Hindi-Urdu is a specific possibility within a grand scope of different orthographies and languages. Other language pairs, which might differ more in script and sound than English and Spanish do, could truly test the hypothesis of whether orthographic awareness is the actual "key" to literacy, rather than phonological familiarity with the language to be learned. Recent research by Kaushanskaya et al. (2011) points to the possibility that this is not the case, and the difference between orthographies can help or hinder multilingual literacy.

It also seems potentially fruitful to test the rates of attrition and retention of both L1 and L2, since it is more realistic to model language learning process as a continuous one and to observe how a multilingual's different linguistic facilities affects the others. Does mastery of L2 erode skill in L1? Does the presence of L1 irreversibly hold back mastery of L2? The mixture of languages exists, without a doubt (e.g., 'Spanglish', 'Franglais'), and with an eye towards arguments about language shift, which more frequently concern demographic and not psychological linguistic changes, how we answer these questions stands to shed light on more than just the educational possibilities available to multilinguals, but also the linguistic costs incurred within the multilingual mind.

References

[1] Ayre, A., Haynes, C., Hook, P., Macaruso, P. (2010). Predictors of English Reading Skills in Spanish-Speaking English-Language Learners (SpELLs). Selected Proceedings of the Hispanic Linguistics Conference, 298-311.

[2] Gottardo, A., Collins, P., Baciu, I., Gebotys, R. (2008). Predictors of Grade 2 Word Reading and Vocabulary Learning from Grade 1 Variables in Spanish-Speaking Children: Similarities and Differences. Learning Disabilities Research Practice, 23(1), 11-24.

[3] Gottardo, A., Mueller, J. (2009). Are First- and Second-Language Factors Related in Predicting Second-Language Reading Comprehension? A Study of Spanish-Speaking Children Acquiring English as a Second Language From First to Second Grade. Journal of Educational Psychology, 101(2), 330-344.

[4] Kaushanskaya, M., Yoo, M., Marian, V. (2011). The Effect of Second-Language Experience on Native-Language Processing. Vigo International Journal of Applied Linguistics, 8, 55-78.

[5] Plaut, D. C., McClelland, J. L., Seidenberg, M., Patterson, K. E. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. Psychological Review, 103(1), 56-115.

[6] Seidenberg, M. S., McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. Psychological Review, 96(4), 523-568.