Gettin' Bi: An unsupervised distributed connectionist model of developmental bilingual speech processing

Tessa Patapoutian
UMass Amherst / LING692c
tpatapoutian@umass.edu

Abstract

As written in Carling and Moore (1988), "Languages are in some respect like maps. If each of us sees the world from our particular perspective, then an individuals language is, in a sense, like a map of their world. Trying to understand another person is like trying to read a map, their map, a map of the world from their perspective. The main question explored in this paper is if self-organizing maps (SOMs) can accurately model the lexical and semantic representations of the developing bilingual mind. To accomplish this, I attempted to replicate the results of a previous unsupervised distributed connectionist model of bilingual speech comprehension implemented by Li and Farkas (2002) using simpler and updated tools, and compared my results to theirs. Interestingly, my results were not the same and in fact sometimes contradicted those of Li and Farkas. If we can use SOMs to represent the bilingual mind is an interesting question to investigate because there hasnt been much work done with connectionist models of the bilingual mental lexicon and not much work with self-organizing maps recently, and they could help expand our knowledge of bilingual representation and processing in the brain.

1 Introduction

There have been several previous connectionist models of bilingual speech comprehension. Not all of these models support the same theories about bilingual processing and representation, and there are still phenomena that are not able, or have not been attempted, to be modeled.

The most well-known localist, adult-state models are BIA+ (Dijkstra and van Heuven, 2002), which is an extension of the monolingual IA model (McClelland and Rumelhart, 1988) and mostly handles processing of visual and orthographic input, and BIMOLA (Lwy and Grosjean, 1997), which is an extension of the monolingual TRACE model (McClelland, 1986) and handles processing of auditory word recognition. BIA+ assumes an integrated lexicon but explicitly marks language membership of words, while BIMOLA separates the two languages at a lexical level but relies on global language information to group words. In both models, representations are manually coded and fixed.

The most well-known distributed, developmental models are BSRN (French, 1998), a simple recurrent network that learns representations through sentence processing, and SOMBIP (Li Farkas, 2002), which uses self-organizing maps (SOMs) and Hebbian learning to create phonological and semantic representations. SOMs use unsupervised learning by competition and similarity-based multidimensional clusters, and can represent them in a visualizable 2-dimensional space. Hebbian learning trains associations between the two phonological and semantic maps. Essentially, clusters are formed in the two maps based on similarity in word forms or word meanings. BSRN supports the theory of a single, distributed lexicon. SOMBIP has been shown to account for distinct patterns of the bilingual lexicon without using language nodes or tags, develop meaningful lexical-semantic categories through self-organizing processes, account for priming and interference effects, and explain lexical representation in bilinguals with different levels of proficiency and working memory capacity. The main advantages of the distributed models are that they can model language learning and acquisition.

The benefits of using a SOM are that they can depict data in a more figurative and a better visual way, they produce a topologically ordered result, and the similarity in a multidimensional space is well preserved in the 2-dimensional, discrete space. When coupled with Hebbian learning, SOMs become a very useful model for developmental phe-

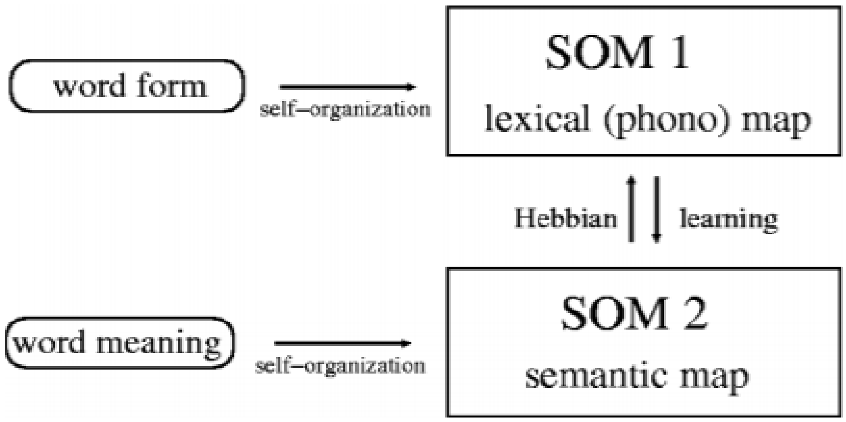

Figure 1: SOMBIP Architecture

nomena. Currently, they are most popular as tools for visualizing data and modeling image-based developmental phenomena.

Previously, most work done with SOMs in the fields of computational linguistics and natural language processing has been focused on creating semantic representations of monolingual data, and usually for the purpose of text or document clustering (Honkela, 1998). There has not been as much work, and especially not much recent work, on using SOMs for modeling the bilingual lexicon, the developmental mind, or a mental phonetic representation.

2 Dataset

Because the intention of this experiment was to replicate the results of Li and Farkas (2002), I used the same dataset for training my model. This dataset was the Hong Kong Bilingual Corpus extracted from the CHILDES dataset (Yip and Matthews, 2007), and specifically only the transcripts from the child Timmy in this corpus. This data included Cantonese and English child-directed speech, and included plenty of code-switching. The weekly recording of Timmy took place from November 1994 to December 1996, from ages 1 to 3. He had exposure to both Cantonese and English from birth, although the language between the parents is mainly Cantonese with much English mixed in.

The transcripts of this corpus were downloaded as CHAT files and accessible by using the CLAN program (MacWhinney, 2000), which was created specifically to work with the CHILDES dataset. Using CLAN, any speech from Timmy was excluded from the transcripts so that only the parental child-directed speech was included. The appropriate transcripts were then converted from CHAT to TXT files.

3 Methods

The architecture of my model was directly inspired by that of SOMBIP (Li and Farkas, 2002), shown in Figure 1. SOMBIP used two self-organizing neural networks, one for phonetic representations and one for semantic representations. Representations of word forms were given to the former map as input and representations of word meanings were given to the latter, and Hebbian links connected the maps with associative pathways.

Word form representations were created by using a syllable-based template coding, in which the phonology of a word is made up of a combination of syllables where each syllable is represented by a 5-dimension feature vector with values between 0 and 1. Specifically, they used the template CVVCCVVC so that every word was encoded into a maximum of eight syllables. If a syllable was not used, then that feature vector was left blank. For example, the Cantonese word *jat* is represented as *jaVtCVVC* and the English word *about* is represented as *C@VCbaUt*. The 5-dimension feature vectors for all 8 syllables were concatenated together to create a 40-dimension vector. An additional 12-dimension tonal vector was appended, although it was left empty when it was representing an English word.

Word meaning representations were created with an RNN, which learned the lexical co-occurrence constraints of the vocabulary. The dimensions of each vector was then reduced to a uniform 100 dimensions.

In my model, I updated how word forms and word meanings were represented. I first used the Python library NLTK to tokenize the transcripts (Loper and Bird, 2002). Instead of using an RNN to learn lexical co-occurrence constraints, I instead directly calculated a semantic co-occurrence matrix and used singular value decomposition to reduce dimensionality to 100 dimensions. I considered using Word2Vec to create word meaning representations, but decided that the corpus did not contain enough data to learn trustworthy embeddings. Instead of representing word forms as hard-coded phonological features, I instead converted each word to IPA using the Epitran Python library created by Mortensen et al. (2018), and calculated a phonological co-occurrence matrix and used singular value decomposition to reduce dimensionality to 56 dimensions. This means each phoneme is represented by a 7-dimension vector, and each word form vector is padded to 56-dimensions to include 8 possible phonemes. And just as in Li and Farkas (2002), only the 400 most frequent words in the corpus were used as training data.

To implement the SOMs and the Hebbian links, I used the pyERA library[1]. Both SOMs have a size of 50 x 50 nodes, which is the same as SOMBIP. A learning rate and radius size (which defines the neighborhood size when calculating similarity) was not defined, so values of 0.1 and 10.0, respectively, were chosen. Li and Farkas (2002) did specify that they used 500 epochs to train SOMBIP, but they did not specify whether or not they used batches, so I decided to train my model with 1000 epochs.

During the training process, a random word was chosen from the 400 most frequent words in the corpus, and its corresponding word form and word meaning were extracted and given to the phonological and semantic map at the same time. Each SOM was trained on this input, and then the Hebbian link between the two SOMs was trained.

To visualize the trained SOMs, the phonological and semantic vector representations of the 400 most frequent words were mapped to the best-matching unit in each SOM, which is also represented by a vector. Each word form or meaning could then be visualized as a point on a 50 x 50 graph, with more similar word forms or meanings clustered together.

[1]https://github.com/mpatacchiola/pyERA

4 Results and Error Analysis

Visualizations of the two maps were created on 50 x 50 grids, with one word present in each cell, and can be seen in Figures 3 and 4. In the phonological map, the Cantonese character and its pinyin were both included in the cell, while in the semantic map, the character and its English translation were both included. To see the spread of languages more easily, English words are shaded blue and Cantonese words are shaded green.

A manual inspection of the two maps created by my model reveals that similar word forms and meanings did indeed cluster together. For example, in the phonological map the Chinese words *mao*, *maomao*, *hao*, *yao*, and *dao* were clustered together. Words from different languages were also clustered together, which can be seen where the Chinese words *zhi*, *zhu*, and *shi* are clustered with the English words *juice* and *this*. Not every clustering makes immediate sense, as in the cluster containing *traingle*, *fly*, *drive*, *kai*, and *guai*, but with more inspection it usually becomes apparent that there is a phonetic commonality between all the words (in this case, the diphthong /aI/). In this map, clusters are small to medium size and are usually fairly clearly defined.

In the semantic map, there are two large clusters and several very small ones, and similarities in the clusters are harder to recognize. For example, one cluster contains the English words *dangerous*, *police*, and *knife*, but also the words *nice*, *boat*, and *ball*. Another cluster contains the Chinese words *vanilla*, *cook*, and *chicken* as well as the English words *egg*, *banana*, and *carrot*, but also the English words *orange* and *red* and the Chinese words *door* and *cat*. It seems that similar meaning words are clustered together, but the clusters are so large that many meanings are included in the same one.

To compare with the results of SOMBIP, I will focus on two of the bilingual speech phenomena that it claimed to accurately represent.

First, language separation in the mental lexicon without language nodes. In Figure 2, it can be seen that SOMBIP seems to have organized both phonological and semantic representations into distinct clusters of Cantonese from those of English. There is a clearer separation between the two languages

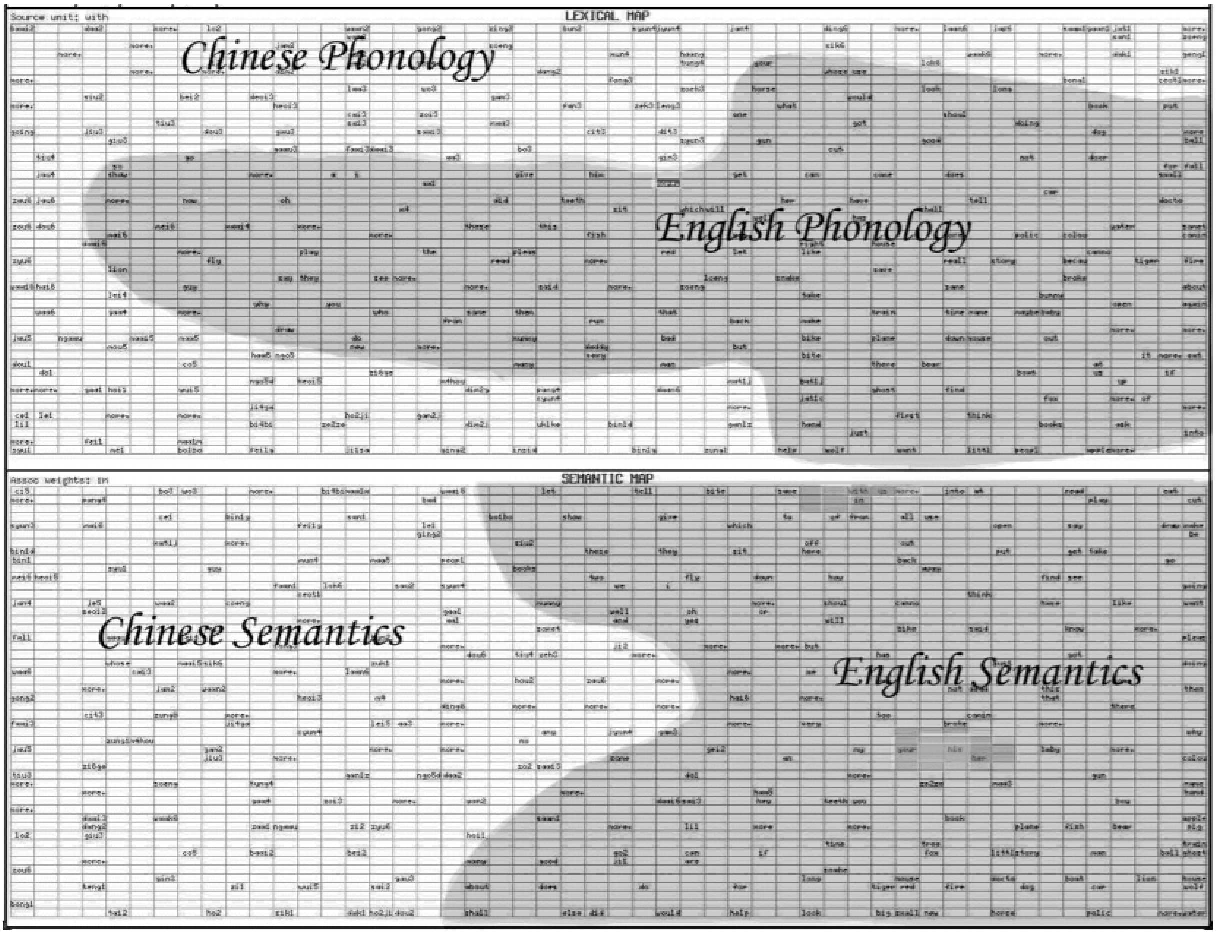

Figure 2: SOMBIP: Phonological and Semantic Organizations

in the semantic map, while the phonological map is more integrated and intermixed. With SOMs, integration can be used as a similarity metric, so these maps could be interpreted as showing that there is indeed separation between languages in the bilingual lexicon, and that phonological separations are more similar than semantic representations.

As can be seen in Figures 3 and 4, although there is some intermixing in both maps, there is also language separation. In the phonological map, clusters will often contain almost entirely words of one language with 1 or 2 words from the other language. English words are mostly contained in the middle and bottom left of the map, while Cantonese words are more frequently around the top edge and sides of the map. In the semantic map, language separation is less clear. Words from both languages seem to be more mixed together, although this may be a result of the fact that there are only two very large clusters so it's more likely that words of different languages will be intermixed. However, it is still possible to see that the top right cluster is mostly English words and the bottom left cluster contains mostly Cantonese words.

These results are interesting because they are contradictory to the ones presented by SOMBIP in that my semantic map is more integrated than my phonological map. Using integration as a similarity metric, my maps can be interpreted as showing that semantic representations in Cantonese and English are more similar to each other than phonological representations in those languages. This is the opposite finding of Li and Farkas (2002), but I believe my maps make more sense because it seems likely that word meanings do not differ drastically between the two languages because the data used is for child-directed speech and so speech in either language likely covers the same topics. However, word forms could be expected to be quite different because different consonants, vowels, and phonological patterns are evident between Cantonese and English. Therefore, there should be greater separation between languages in the phonological map than in the semantic map.

Although Li and Farkas (2002) claim to have language separation without langugae nodes, I would

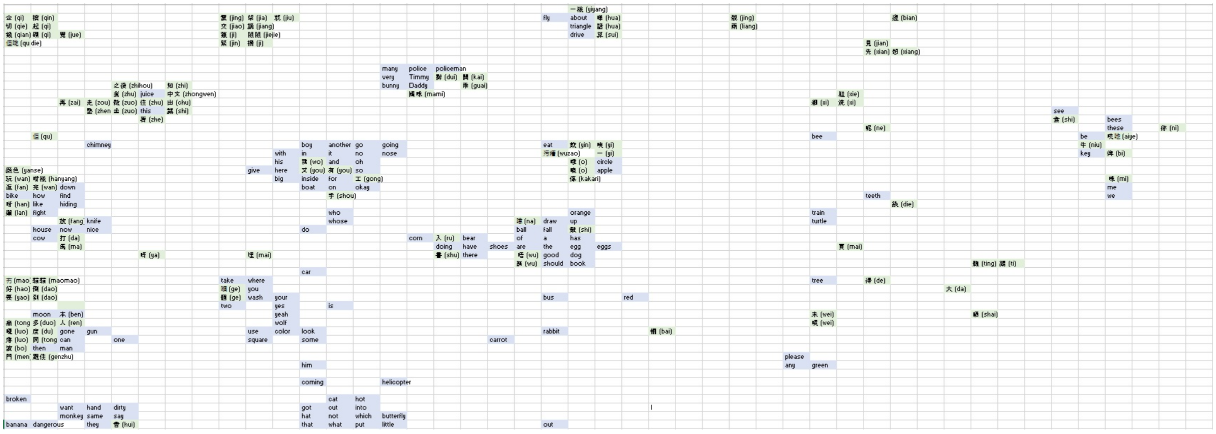

Figure 3: SOM 1: Phonological Organizations

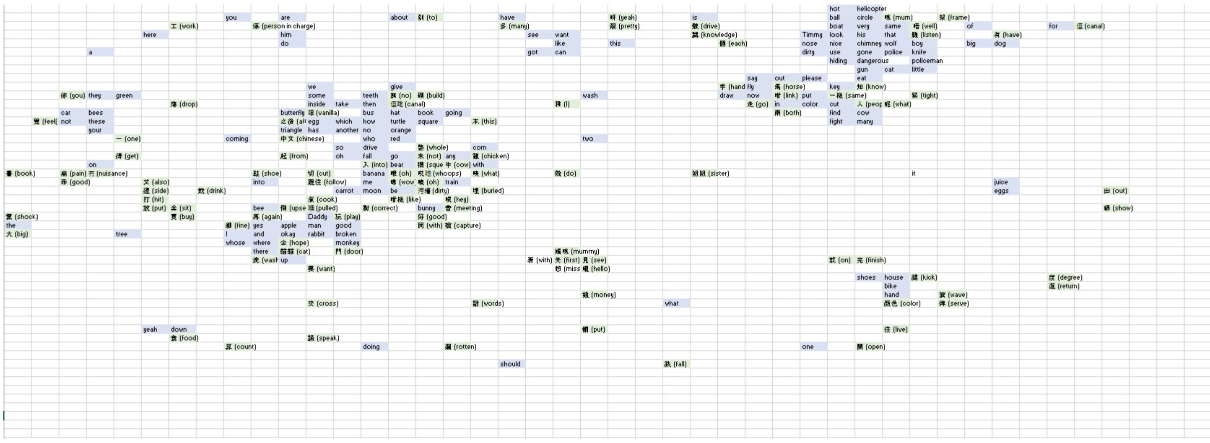

Figure 4: SOM 2: Semantic Organizations

argue with this assertion. In their phonological encoding of the vocabulary they included a tonal vector which was left empty for English words. I believe this is an implicit language tag, and must have assisted their model in distinguishing between words of the two languages. In my model, I did not encode tone, which is a loss of information but also means that there is no explicit or implicit language tag. Because of this, I do not think that SOMBIP necessarily represents language separation in the mental lexicon without language nodes, but I think my model does, at least in the phonological map. This supports the empirical studies by Kirsner et al. (1984) that argue for a language-specific bilingual lexicon in an integrated network.

The second behavior that I will be looking at is the emergence of lexical categories in the semantic map. Li and Farkas (2002) assert that SOMBIP organizes the semantic representations into structurally meaningful clusters, which can be seen in Figure 5. Nouns, verbs, and prepositions are treated distinctly, so that words with shared grammatical categories are clustered together. They also assert that within these categories, semantically similar words occurred together. For example, state verbs like *know*, *like*, and *have* are grouped separately from activity verbs like *draw*, *eat*, and *play*.

In my model, there did not seem to be any derivation of semantic categories in either language. Words of all grammatical categories were clustered together, as well as words of multiple semantic meanings. For example, one cluster contains the English words *do*, *are*, and *him* as well as the Cantonese word for *person-in-charge*. Another contains the Cantonese words for *drive*, *knowledge*, and *each* as well as the English word *is*. Again, this may simply be a result of the fact that there were only two large clusters in my semantic map so it was necessary for multiple meanings and grammatical categories to be mixed together. It may also have been more difficult for lexical categories to be derived given that my semantic map was more integrated than the one produced by SOMBIP.

Overall, my results disagreed with those of Li

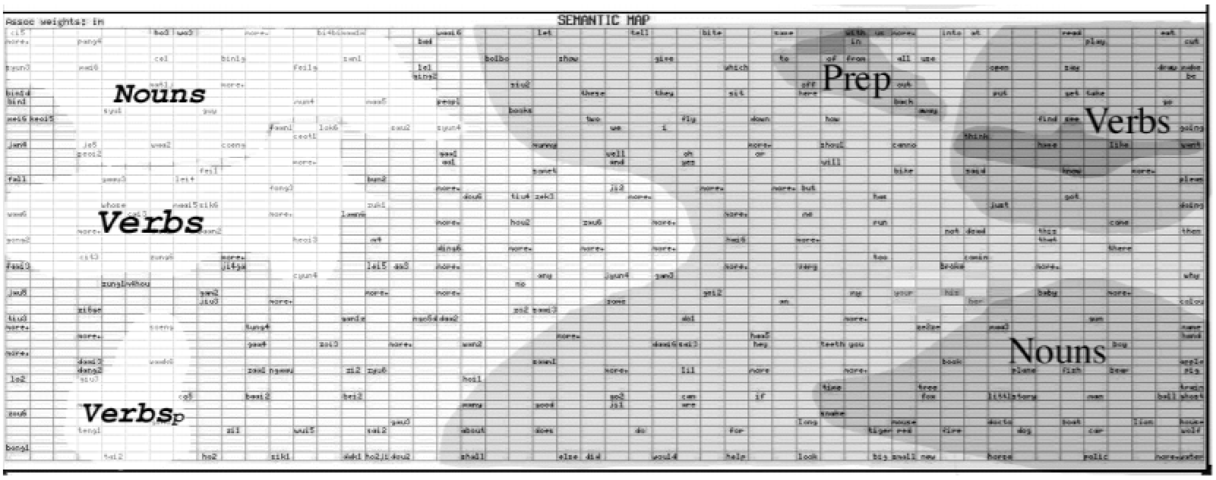

Figure 5: SOMBIP: Lexical Categories

and Farkas (2002). My model did produce language separation without language nodes, but the phonological map it produced showed more integration than the semantic map and no apparent semantic categories were derived. However, my implementation could have differed enough from that of SOMBIP to create these varying results. For example, it is shown that by simply changing the size of the SOMs to 40 x 40 instead of 50 x 50 (meant to simulate a limited working memory capacity), SOMBIP produces maps that had much more diffuse categories. In other words, clusters often contained words representing multiple word meanings or forms, which is similar to what was produced in my semantic map. Changing the size of the SOM from 50 x 50 to 40 x 40 seems like a not very significant difference to result in such a change, as well as an arbitrary difference. All of the hyperparameters chosen to model SOMBIP (e.g. number of epochs, number of words included in the training data, learning rate, etc.) seem to be chosen arbitrarily and as a result of the computational limits available at the time. Because of this and because the implementation of my SOMs and Hebbian links almost certainly differs, I am sure that my model could also be tuned to represent the phenomena represented by SOMBIP.

5 Conclusion and Future Work

In summary, the results produced by my model contradicted most of the results produced by SOMBIP, but I do not think this means that either one of the models is a correct representation of the bilingual lexicon. Rather, I believe that either of these models could be manipulated to represent some phenomena that the creator chooses. Making a change to one of the hyperparameters is fairly arbitrary and not directly comparable to the features of the actual bilingual mind. Because of this, I see this model as more useful for visualizing what could be possible lexical and semantic bilingual representations instead of as proof of certain speech processing behaviors.

If I were to continue with this project in the future, I'd like to first investigate more how the hyperparameters affect the resulting phonological and semantic maps. For example, the radius, learning rate, epochs, and SOM size when training the models, as well as window size when creating co-occurrence matrices.

I would also like to model data from the Hong Kong Bilingual Corpus with children other than Timmy to verify that there isn't a significant difference in maps between bilingual speakers of the same two languages. I would also like to mode data from two language that are more similar than Cantonese and English, such as French and English or Cantonese and Mandarin, to see if the maps are more integrated because the languages are more similar. Additionally, I'd like to model trilingual data, like Mandarin, Cantonese, and English, to see if the two more similar languages integrate with each other more than with the third.

Acknowledgments

I would like to thank Max Nelson for his assistance in updating the word form and meaning representations used by SOMBIP with more current methods, specifically using co-occurrence matrices and SVD, as well as being able to talk through

my project with him to make sure it made sense. Thanks, Max!

References

Christine Carling and Terence Moore. 1988. *The Limitations of Language*. Macmillan Press, Houndmills.

Ton Dijkstra and Walter J.B. van Heuven. 2002. The architecture of the bilingual word recognition system: From identification to decision. *Bilingualism: Language and Cognition*, 5(3):175197.

R. M. French. 1998. A simple recurrent network model of bilingual memory. *Proc. 20th Annu. Cogn. Sci. Soc. Conf. Erlbaum*, pages 368–373.

Timo Honkela. 1998. Self-organizing maps in natural language processing.

Kim Kirsner, Marilyn C. Smith, R.S. Lockhart, M.L. King, and M. Jain. 1984. The bilingual lexicon: Language-specific units in an integrated network. *Journal of Verbal Learning and Verbal Behavior*, 23(4):519 – 539.

Ping Li and Igor Farkas. 2002. A self-organizing connectionist model of bilingual processing. *Advances in Psychology*, pages 59–85.

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*, ETMTNLP '02, pages 63–70, Stroudsburg, PA, USA. Association for Computational Linguistics.

Nicolas Lwy and Franois Grosjean. 1997. *The Lwy and Grosjean BIMOLA model*, page 201210.

B. MacWhinney. 2000. *The CHILDES Project: Tools for Analyzing Talk, 3rd Edition*. Lawrence Erlbaum Associates, Mahwah, NJ.

Elman J. L. McClelland, J. L. 1986. The trace model of speech perception. *Cognitive Psychology*, 18:1–86.

James L. McClelland and David E. Rumelhart. 1988. Neurocomputing: Foundations of research. chapter An Interactive Activation Model of Context Effects in Letter Perception: Part 1.: An Account of Basic Findings, pages 401–436. MIT Press, Cambridge, MA, USA.

David R. Mortensen, Siddharth Dalmia, and Patrick Littell. 2018. Epitran: Precision G2P for many languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

V. Yip and S. Matthews. 2007. *The bilingual child: Early development and language contact*. Cambridge University Press.