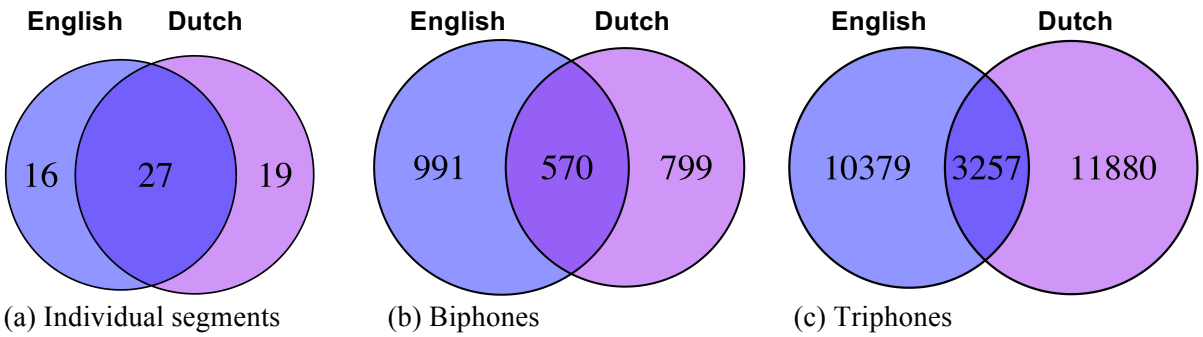
Phonological cues for language separation in bilingual development: a computational approach

Frans Adriaans (f.w.adriaans@uu.nl)
Utrecht University

A key challenge in early bilingual acquisition is to distinguish two different languages in the input speech stream. The ability to detect and separate languages could, for example, allow infants to develop separate statistical distributions to learn sound categories for two languages [1, 2]. What type of cues might infants use to detect different languages? It has been suggested that rhythmic cues could be used to separate rhythmically distinct languages such as Spanish and English [2]. For rhythmically similar languages (such as Spanish and Catalan) differences in the inventory and frequency distributions of sound categories, as well as lexical and phonotactic constraints could support the development of two language systems [3, 4]. While various phonological cues have been proposed as a potential basis for language separation, it is unclear how effective these cues are in accomplishing this task. Moreover, parents often mix languages [5] and it is unclear to what extent mixed input affects the reliability of phonological cues to separate languages in the input. This paper begins to address these issues by investigating the effectiveness of segmental and phonotactic cues for input separation of two rhythmically similar languages: English and Dutch. Specifically, computational modeling is used to determine (i) to what extent segmental and phonotactic cues can predict the origin language in mixed input data, and (ii) how robust these cues are when confronted with different degrees of language mixing.

Bilingual input was simulated by combining transcriptions from two different speech corpora (English: [6], Dutch: [7]). These corpora were chosen because of their comparable size and transcription level. Figure 1 illustrates the overlap between the corpora in terms of segmental inventory and phonotactics (biphones and triphones). The segmental overlap between English and Dutch is 44%. (27 segments occur in both corpora, 16 occur only in English, and 19 occur only in Dutch.) The overlap in terms of phonotactics is 24% for biphones and 13% for triphones. The smaller overlap indicates that phonotactics might be a more useful cue for language separation. This prediction was tested in a series of computer simulations aimed at determining which cue is most effective at predicting the origin language in a mixed test set. Three probabilistic phonological models were implemented, one based on the relative frequencies of individual segments, and two based on phonotactic probabilities (biphone and triphone transitional probabilities). The models were trained on samples from both corpora in a variety of input mixing proportions, ranging from completely separated training data to 50-50 mixed input training data. Figure 2 shows the language prediction accuracies in different input mixing proportions. When languages are completely separated all models perform with high accuracy, ranging from 0.87 (segments) to 0.95 (triphones). When confronted with mixed training data the performance of the triphones model drops substantially (likely due to data sparsity, e.g. [8]), and the most accurate predictions are made by the biphones model (accuracy $\approx$ 0.90).

The results show that biphone-based phonotactics could provide English-Dutch bilingual infants with a relatively accurate and robust cue for language separation. Importantly, biphones provide a more effective separation cue than a model that is based on independent segments. Bilinguals as a population are understudied in computational modeling work, and the approach presented here adds to recent efforts using computational methods to investigate the complexities of mixed input [9]. Such methods may ultimately help us to understand bilinguals' impressive learning mechanisms.



(a) Individual segments (b) Biphones (c) Triphones

Figure 1. *Overlap between English (blue) and Dutch (purple) in terms of (a) individual segments (overlap = 44%), (b) biphones (overlap = 24%), and (c) triphones (overlap = 13%).*

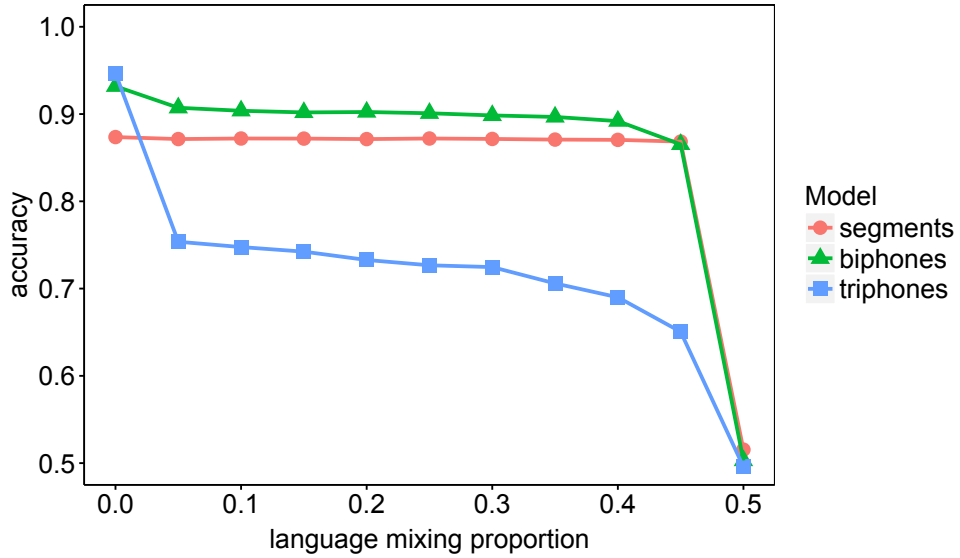


Figure 2. *Language prediction accuracies for models trained on segments, biphones, and triphones.*

[1] Curtin, S., Byers-Heinlein, K., & Werker, J. F. (2011). Bilingual beginnings as a lens for theory development: PRIMIR in focus. *Journal of Phonetics*, 39(4), 492-504.

[2] Sundara, M., & Scutellaro, A. (2011). Rhythmic distance between languages affects the development of speech perception in bilingual infants. *Journal of Phonetics*, 39, 505-513.

[3] Bosch, L., & Sebastián-Gallés, N. (2003). Simultaneous bilingualism and the perception of a language-specific vowel contrast in the first year of life. *Language and Speech*, 46, 217-243.

[4] Sundara, M., Polka, L., & Molnar, M. (2008). Development of coronal stop perception: Bilingual infants keep pace with their monolingual peers. *Cognition*, 108(1), 232-242.

[5] Byers-Heinlein, K. (2013). Parental language mixing: Its measurement and the relation of mixed input to young bilingual children's vocabulary size. *Bilingualism: Language and Cognition*, 16, 32-48.

[6] Pitt, M. A., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E., & Fosler-Lussier, E. (2007). *Buckeye Corpus of Conversational Speech*. Ohio State University.

[7] Goddijn, S., & Binnenpoorte, D. (2003). Assessing manually corrected broad phonetic transcriptions in the Spoken Dutch Corpus. In *Proceedings ICPhS* (p.1361-1364).

[8] Pierrehumbert, J. (2003). Probabilistic phonology: Discrimination and robustness. *Probabilistic linguistics*, 177-228.

[9] Carbajal, M. J., Dawud, A., Thiollière, R., & Dupoux, E. (2016). The "language filter" hypothesis: A feasibility study of language separation in infancy using unsupervised clustering of I-vectors. In *ICDL-EpiRob* (pp. 195-201).