6

Computational Approaches to Bilingual Phonetics and Phonology

Frans Adriaans

6.1 Introduction

Computational approaches have over the past two decades become an integral part of language acquisition research, establishing itself as an important methodological counterpart to theoretical and experimental approaches. The benefit of using computational models lies in the explicitness of those models. Researchers formulating a linguistic theory in a computational context need to make that theory explicit enough that it can be implemented in a computer program. This typically means that the researcher develops an *algorithm* that prescribes how linguistic input is processed in a step-by-step fashion, transforming it into some sort of output. In the context of language acquisition the model provides an exact characterization of the relation between linguistic input and linguistic knowledge that is somehow learned from that input. There is no room for vagueness or ambiguity in describing the process, as the entire input-output mapping has to be spelled out in a formal (programming) language.

Due to their explicitness, computational models provide insights into aspects of formal learnability. That is, models can be used to prove that a certain theory (of acquisition) is learnable or not. Simply put, when input is presented to the model, and the model cannot compute the intended output, the theory must be incorrect or incomplete. For example, Yang (2004) used computational modeling to show that a word segmentation model based on syllable

transitional probabilities as described in Saffran, Newport, and Aslin (1996) cannot segment monosyllabic words. Such a limitation could mean that the model needs to be revised, or perhaps the model needs to operate on a different type of representation.

In addition to the explicitness advantage, computational models crucially lead to *testable predictions* and hence provide an essential piece in the empirical testing of the theoretical proposal. Models are generally tested either against natural language corpora, or against data from experiments with human participants. The former allows one to scale up the empirical testing of a theoretical proposal by testing it in a *whole-language* simulation (as opposed to a relatively small set of items used in experiments with human participants). The latter allows one to test if the model performs in a way similar to human learners, and thus addresses the cognitive plausibility of the model.

The past two decades have seen great advancements in the modeling of phonetic and phonological acquisition in a monolingual setting. There has been a wide body of research focused on the question of how infants learn phonetic categories from acoustic input (e.g., Adriaans, 2018; Adriaans & Swingley, 2017; de Boer & Kuhl, 2003; Dillon, Dunbar, & Idsardi, 2013; Feldman et al., 2013; McMurray, Aslin, & Toscano, 2009; Swingley & Alarcon, 2018; Vallabha et al., 2007). Research here has focused on two things: (i) identifying the learning mechanism that allows infants to form categories, and (ii) examining properties of the input that may help or hinder the learning mechanism in achieving its goal. The general consensus of these studies has been that the learning mechanism involves some type of clustering along acoustic dimensions (e.g., Maye, Werker, & Gerken, 2002), but it has not been clear how this mechanism operates on realistic input data. Computational and corpus studies have advanced our understanding of phonetic category learning by showing that categories are hard to learn from isolated phonetic tokens (Swingley, 2009), and that learning can be supported by contextual information, such as the lexical or phonological contexts in which phonetic tokens

occur (e.g., Adriaans, 2018; Dillon et al., 2013; Feldman et al., 2013; Swingley & Alarcon, 2018).

In phonology, computational studies have focused on the learnability of phonological constraints and their rankings. The learnability perspective has been an integral aspect of Optimality Theory (OT; Prince & Smolensky, 2004; Tesar & Smolensky, 2000). In OT (and more recent approaches using Harmonic Grammar; Legendre, Miyata, & Smolensky, 1990), the phonological learning problem is defined as finding the appropriate ranking (or weighting) for a universal set of constraints, and various computational approaches have been proposed to determine the language-specific ranking of constraints (e.g., Boersma & Hayes, 2001; Potts et al., 2010; Prince & Tesar, 2004; Tesar & Smolensky, 2000). More recently, studies in computational phonology have focused on the induction of the constraints themselves (e.g., Adriaans & Kager, 2010; Gouskova & Gallagher, 2020; Hayes & Wilson, 2008). Computational approaches have thus helped to address questions regarding the origins of phonological constraints, to what extent they are learned from input data, and what type of input data they are learned from.

Despite these advancements in the modeling of phonetic and phonological acquisition in a monolingual setting, only very few studies have begun to address the computational modeling of bilingual acquisition. This chapter will put the problem of simultaneous bilingual phonetic and phonological acquisition in a computational perspective. First a general introduction to computational modeling will be provided, using a simplified model of phonotactic learning as an example to illustrate the main methodological issues. The chapter will then discuss recent studies that have used computational modeling to study bilingual phonetic and phonological acquisition in three main areas: phonetic and phonological cues for bilingual input separation, bilingual phonology in computational models of speech comprehension, and computational models of L2 speech perception. The chapter will conclude

by discussing several key challenges in advancing the development of computational models of bilingual phonetic and phonological acquisition.

6.2 The Computational Perspective

Figure 6.1 A computational model provides an explicit description of an input-output mapping.

Before we discuss computational approaches to bilingual phonetics and phonology, we need to establish a general computational perspective, which will then be used to frame the problem of bilingual acquisition. Computational approaches present formal characterizations of the learning problem by distinguishing three basic components, illustrated in Figure 6.1. The *input* is the data available to the learner. Representing the input is a non-trivial issue, and choices that are made regarding the representation of input data affect how the model in the end will perform (Marr, 1982). Ultimately, the input of a model should be the same as the input to a human learner. In the case of early language development, this would be speech data. However, due to the complexities of working with unprocessed speech data, computational models typically operate on a simplified representation of the input.

In phonetic learning, the input is often represented along one or more particular phonetic dimensions of interest, specified by the researcher. For example, a model learning a voicing contrast might be presented with input tokens represented along a single Voice Onset Time dimension (e.g., McMurray et al., 2009). A model of vowel learning might be presented with tokens along two or three dimensions, such as vowel formants and duration (e.g., Vallabha et al., 2007). A more detailed representation of the speech signal can be obtained using

techniques from Automatic Speech Recognition, where it is common to represent the speech signal as a sequence of vectors of 39 (delta) Mel-Frequency Cepstral Coefficients (MFCCs), extracted from the signal at 10 ms intervals (Jurafsky & Martin, 2009), but the link to linguistic properties is not immediately apparent in such cases.

In phonological learning, the input to a computational model may vary from relatively simple sequential symbolic representations of phonological categories to more complex linguistically annotated forms. There is no one right way to represent the input to the learner, because different models aim to achieve different goals, and simplifications are necessary to keep the computational problem solvable. One must therefore critically examine whether the input representation is appropriate to answer the research question at hand.

The middle part in Figure 6.1 is the *learning model* that operates on the input. The model specifies how linguistic knowledge is learned from input data. This is the researcher's theory of the learning process, implemented as a computer program. The model performs computations on the input, which then ultimately leads to some form of *output*. The output usually takes on the shape of some new set of linguistic representations, such a set of categories, or a full-fledged grammar. In the context of acquisition, the model's output should match either what infants or children know about a particular language, or some general property of the language to be learned.

There are two things that should be noted at this point. First, all forms of modeling involve some level of *simplification* of the real-life scenario. That is, the input is simplified, the model is simplified, or (usually) both are simplified. This is inherent to the modeling approach, at least for the types of models that are currently available. Ideally, a computational model of language acquisition would explain the entire mapping from the first speech input to output in the form of an adult grammar. In practice, models focus on particular subproblems,

each with their own input and output representations. One current issue in computational modeling is to uncover how these different learning problems are connected, as it seems increasingly unlikely these different learning problems are solved in a strictly sequential way (e.g., Adriaans, 2018; Dillon et al., 2013; Feldman et al., 2013; Swingley, 2009). Each modeling study should thus be seen as providing a piece of the acquisition puzzle, and those varying pieces need to be connected to obtain a more complete picture of acquisition.

Second, it is important to realize that computational models can provide explanations at different levels. Marr (1982) identifies three different levels of modeling. At the *computational* level, models are focused on the goal of a computation, and the logical ways in which a problem can be solved. Such models may identify types of information which are necessary to solve a problem, but they do not aim to describe the process by which a human learner would solve a problem. In contrast, models at the *algorithmic* (or *mechanistic*) level aim to explain how a particular input representation is transformed into an output representation. Finally, the *implementational* level is concerned with the physical realization of the transformation.

These different levels mean that models should be interpreted and evaluated differently in the context of language acquisition. The main question regarding evaluation at the computational level is whether the model solves the problem or not. At this level, the best model is one that obtains maximal accuracy on some learning task. At the algorithmic level, the question is whether the model approaches the problem in the same way human learners approach the problem. Since humans make errors, a model with maximal accuracy is not necessarily the best model. The model crucially needs to predict human errors, which in acquisition might take on the form of overgeneralization, U-shaped learning curves, intermediate developmental stages, etc. At the implementational level, the question is whether

the model reflects the physical structure of the learning process in the human brain. Data from neural studies are needed to evaluate such models.

Computational models typically operate on large amounts of input data, and the computations performed by the model can be quite complex. This can make it difficult to develop a thorough understanding of the modeling approach. The computational approach will therefore be illustrated below using a fairly simplistic phonotactic learning model, which operates on a small sample of input data. Due to these simplifications, the computations can be tracked by hand, and it quickly becomes apparent how different methodological considerations affect the performance of the model. Later on in this chapter, we will illustrate how this phonotactic learning problem could be modeled for bilingual learners.

6.2.1 A Computational Model of Phonotactic Learning

The simplified model of phonotactic learning will be one that learns sequential biphone-based phonotactic probabilities (e.g., Bailey & Hahn, 2001; Cairns et al., 1997; Jusczyk, Luce, & Charles-Luce, 1994; Vitevitch & Luce, 1999). The model is simplified in the sense that it has a very limited interpretation of phonotactics: probabilities of adjacent segments. The model does not refer to prosodic structure, phonological features, or non-local dependencies. The input to the model is also simplified. We will assume the input to the learner consists exclusively of the following two arbitrary American English utterances taken from the Buckeye Corpus (Pitt et al., 2007):

(6.1) `well i work in the accounting department`

`i'm an accounting assistant`

The entire corpus contains close to 300,000 words, and the corpus has been used as a phonetically transcribed approximation of spoken American English in various computational studies (e.g., Daland & Pierrehumbert, 2011). It should be noted, however, that this is an adult-directed conversational speech corpus, and a more ecologically valid set of acquisition input data would ideally involve transcriptions of child- or infant-directed speech. The 'toy' corpus of two utterances shown here nevertheless suffices to illustrate the workings of the computational model.

6.2.1.1 Representing the Input

We first need to establish an appropriate input representation for this learning task. For the purposes of phonotactic learning, we require a phonemic transcription of the two utterances. Already at this early stage we are faced with making certain modeling decisions that will affect the performance of our model. Phonemic transcriptions provided in corpora are typically either *canonical* transcriptions, which means they are the result of looking up the orthographic word in a pronunciation dictionary, or they are *variable* transcriptions, which means they have been manually coded (or corrected) to reflect some of the variability found in spoken language, such as reductions and assimilations. A machine-readable canonical transcription for the toy corpus can be obtained using *The CMU Pronouncing Dictionary, version 0.7b* (2014), which transforms the utterances into a computerized representation of the International Phonetic Alphabet:

(6.2) `w eh l # ay # w er k # ih n # dh ah # ah k aw n t ih ng # d ih p aa r t m ah n t`

`ay m # ae n # ah k aw n t ih ng # ah s ih s t ah n t`

This canonical transcription is notably different from the manually coded variable transcription that has been included in the corpus:[1]

(6.3) `w ah # aa # w er k # ih n # n ih # ah k aw iy #`

`ih p aa r t m ih n`

`aa m # ah # ah k aw n iy ng # ih s ih s t eh n t`

The difference between these two different input representations can be seen in the word *accounting* which is transcribed as `ah k aw n t ih ng` in the canonical transcription, and as either `ah k aw iy` or `ah k aw n iy ng` in the variable transcription. The variable transcription provides a closer match to the actual speech signal, as it codes the outcome of natural speech production processes that occur in everyday spoken language. (The word *accounting* has various other forms in the corpus.) Nevertheless, canonical transcriptions are commonly used in modeling studies. In fact, phonological learning is commonly assumed to operate on word types rather than on word tokens (e.g., Albright, 2009; Hay, Pierrehumbert, & Beckman, 2004; Pierrehumbert, 2003; Richtsmeier, 2011), which presupposes that the learner has internalized a vocabulary of canonical phonological forms.

Additional input representations are possible, depending on assumptions that are made regarding the input that is used in learning. Several studies on phonotactic learning have suggested that early phonotactics might be learned from continuous speech, rather than from the lexicon (e.g., Adriaans & Kager, 2010, 2017; Brent & Cartwright, 1996; Daland & Pierrehumbert, 2011; Sundara & Breiss, 2020). In these models, phonotactic probabilities are used for the detection of word boundaries in continuous speech. If we assume that words (or word boundaries) are not available to the learner, then this would change the representation of

[1] In this example, the glottalization label for stops has been omitted.

the input. Below is an adapted version of the variable input representation, where word boundaries have been removed:

(6.4) `w ah aa w er k ih n n ih ah k aw iy ih p aa r t m ih n`

`aa m ah ah k aw n iy ng ih s ih s t eh n t`

A final issue regarding input representations is that the learner might be biased to process or attend to different parts or properties of the input. For example, one might hypothesize that in early phonological learning the learner might focus more on consonants than on vowels in the input (e.g., Bonatti et al., 2005; Hochmann et al., 2011). Such biases can be taken into account in modeling by filtering out particular tokens from the input representation (e.g., Kastner & Adriaans, 2018). For example, if we assume that phonotactics is learned exclusively from a consonantal tier, then we could represent the toy corpus as follows:

(6.5) `w w k n n k p r t m n`

`m k n ng s s t n t`

The right representation for training and testing computational models thus depends on the researcher's assumptions (and evidence) regarding the nature of the input that is used for learning. The choice of input representation has direct consequences for the linguistic knowledge that is learned from it. As we will see, a bilingual learning environment will complicate this input issue further.

6.2.1.2 The Learning Model

The model represents the researcher's theory regarding the human learning mechanism (in the case of modeling at the algorithmic level), or general problem solving strategy (in the case of

modeling at the computational level). In this example, we will assume that probabilistic phonotactics is learned through a mechanism of statistical learning (e.g., Saffran, Aslin, & Newport, 1996; Saffran, Newport, & Aslin, 1996). Specifically, phonotactic probabilities are learned by computing transitional probabilities for each sequence of two phonemes in the input (e.g., Albright, 2009; Cairns et al., 1997; Vitevitch & Luce, 1999). We adopt this learning mechanism not because it provides a full account of a language's phonotactics (it doesn't), but because it is a straightforward model that is easy to compute directly from the toy corpus. Formally it is an *N*-gram (in this case bigram) language model applied to phonemic representations (see Jurafsky & Martin, 2009). Applying this learning algorithm to the toy corpus of variable transcriptions in (6.4) results in the phonotactic probabilities shown in Table 6.1.

Table 6.1 A model of phonotactic learning based on biphone transitional probabilities (TP) in the toy corpus.

Biphone	TP	Biphone	TP	Biphone	TP	Biphone	TP	Biphone	TP
aa m	0.33	aw n	0.50	iy ih	0.50	n iy	0.25	s t	0.50
aa r	0.33	eh n	1.00	iy ng	0.50	n n	0.25	t eh	0.50
aa w	0.33	er k	1.00	k aw	0.67	n t	0.25	t m	0.50
ah aa	0.25	ih ah	0.17	k ih	0.33	ng ih	1.00	w ah	0.50
ah ah	0.25	ih n	0.33	m ah	0.50	p aa	1.00	w er	0.50
ah k	0.50	ih p	0.17	m ih	0.50	r t	1.00	-	-
aw iy	0.50	ih s	0.33	n ih	0.25	s ih	0.50	-	-

It should be noted that the model could be applied to any of the representations in (6.2)-(6.5), and each of the representations would lead to a different set of phonotactic probabilities.

The model processes the input data either *incrementally* (one biphone is presented to the model at a time, and at each step the model updates its probabilities), or in *batch* (all biphones are presented to the model at once, and the model performs computations over the entire set). The outcomes of these two ways of input processing may or may not be the same, depending on how the model employs memory. If an incremental model has some form of

memory decay, it will gradually forget older tokens, and the resulting model will be different from the batch model because it relies more heavily on more recently processed biphones. This also means that the order in which data is presented to the model matters. However, if the incremental model makes the assumption of perfect memory, the outcome of the incremental and batch versions will be the same: a set of probabilities computed over the entire input (such as the ones shown in Table 6.1).

The probabilities shown in Table 6.1 are of course not representative of English phonotactics, since they are based on only two utterances. Normally one would apply the model to a larger, more realistic data set (such as the whole corpus), after which the performance of the model can be evaluated.

6.2.1.3 The Output and its Evaluation

We have now applied a formalized learning procedure to some input data, and this generated output in the form of a set of phonotactic probabilities. For this to be a 'good' model of phonotactic learning, it should perform well on some sort of evaluation task. Evaluation of a model depends on the research question it is meant to answer. Some straightforward questions that could be asked are: to what extent do these probabilities capture English phonotactics? To what extent do these probabilities match English speakers' knowledge of phonotactics? To what extent do the probabilities match infants' knowledge of phonotactics? To what extent are these probabilities useful for further language development? Each question requires a different type of data to be used as a gold standard in evaluation. For example, if we are interested in modeling English speakers' knowledge of phonotactics, the probabilities could be matched with human judgements of the well-formedness of nonwords (e.g., Albright, 2009). If we are interested in the usefulness of these probabilities for further language development, we could

assess the effectiveness of the model in predicting the locations of word boundaries in continuous speech (e.g., Adriaans & Kager, 2010; Daland & Pierrehumbert, 2011).

Setting up a test case for the model requires us to make another formalization: how exactly are the probabilities used to predict well-formedness, or segmentation behavior? For well-formedness judgements, nonword probabilities could be obtained by multiplying biphone probabilities within the nonword. Using this approach, our toy model in Table 6.1 would predict that, for example, `m ah k` is more well-formed than `m ih n`, because it has a higher probability (0.5 * 0.5 = 0.25 for `m ah k`, versus 0.5 * 0.33 = 0.17 for `m ih n`).

For word segmentation, two different formalizations have been proposed regarding the application of transitional probabilities to the task of boundary detection. One is segmentation based on 'troughs': a boundary occurs whenever the probability of some bigram sequence *xy* is lower than both of its neighboring bigrams *wx* and *yz*. While this implementation is straightforward, it has some learnability consequences. As Yang (2004) pointed out, one consequence of this formalization is that unigram words cannot be segmented, as they would require two minima next to each other, which is not possible by definition. If one assumes syllables as the basic unit over which to compute transitional probabilities, then monosyllabic words cannot be segmented using the trough-based approach. If one assumes phonemes as the basic unit, then monophonemic words cannot be segmented. The latter is less consequential because English only has a small number of monophonemic words (e.g., 'a', 'I').

A different approach is to set a threshold on the probabilities: whenever a probability is below some threshold *t*, a boundary is inserted. This approach does not face the same limitation as the trough-based approach, but it raises another issue. *t* is a *parameter* of the model, and it is not clear a priori what its value should be. Changing the value of the parameter will change the behavior of the model, so this is a non-trivial issue. It also points to a more

general issue in modeling. Computational models often have various parameters that can be set to improve the fit of the model to the data. This is a potential criticism of the modeling approach, since it could be argued that with the right parameter settings, any phenomenon or data set could be modeled. What does the model itself then prove, exactly? It is therefore important that computational studies give an explicit description of the parameters in the model, which values they take, and how they affect the model's performance.

Below in (6.6) is an example of how the model in Table 6.1 would segment the first utterance from the training set based on the two different approaches. We use an utterance from the training set here for illustrative purposes, but it should be noted that normally one would use novel test utterances that did not occur in the training set. The gold standard ('correct') segmentation is included to illustrate where the boundaries should have been.

(6.6)

i. Threshold: *(t = 0.3)*

```
w ah # aa w er k ih n # n # ih # ah k aw iy ih # p aa r t m ih n
```

ii. Trough:

```
w ah # aa w er k ih n n ih # ah k aw iy ih # p aa # r t m ih n
```

iii. Gold standard:

```
w ah # aa # w er k # ih n # n ih # ah k aw iy # ih p aa r t m ih n
```

We can now ask which of the two approaches leads to a better segmentation, by employing commonly used evaluation metrics such as Precision (= number of hits / (number of hits + number of false alarms)) and Recall (= number of hits / (number of hits + number of misses)). While such metrics indicate which approach is better at solving the problem of word segmentation, it doesn't necessarily mean human learners use the same approach. A stronger

test of a computational model would therefore be to derive novel predictions from the model and design experiments specifically aimed at testing these predictions. But even without such a test, the example shown in this section illustrates how computational approaches force the researcher to posit an explicit theory, and to assess what exact approach would be more successful in solving a particular problem.

6.3 Computational Modeling of Bilingual Phonetic and Phonological Acquisition

Computational approaches can be extremely valuable in understanding the complexities of the bilingual speech input that infants receive, and the mechanisms they employ to learn two languages from that input. Using computational models, concrete solutions can be formulated by which infants could navigate their complex language environment. Despite this potential, computational approaches to bilingual phonetics and phonology have only recently started to emerge. This section will first discuss existing studies that have used computational approaches in three different research areas. We will then zoom in on certain aspects of bilingual modeling by returning to our toy example. These examples show the potential of computational modeling to address essential questions in bilingualism that are hard to address directly with infant experiments.

6.3.1 Phonetic and Phonological Cues for Bilingual Input Separation

One major 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 (Curtin, Byers-Heinlein, & Werker, 2011; Sundara & Scutellaro, 2011). Because the ability to detect and/or separate languages in the input is crucial for the development of two languages,

computational studies have focused on identifying cues that might be used by learners to solve this problem. Several studies have assessed the degree to which either acoustic or phonological cues can be used to separate two languages in the input.

Several studies have attempted to identify ways in which the two languages that are present in a child's input might be separated on the basis of acoustic properties of the signal (Carbajal, 2018; Carbajal et al., 2016; Carbajal, Fér, & Dupoux, 2016; Dehak et al., 2011; de Seyssel & Dupoux, 2020). These studies use an i-vector approach which was originally developed for automatic speaker identification (Dehak et al., 2010). In this approach, the entire acoustic space of a speech data set is modeled as a Universal Background Model (UBM), and individual utterances within the data set are represented as deviations from the UBM. Utterances can then be clustered into different languages, and new test utterances are classified according to their similarity to the clusters. This approach has been used for automatic language identification (Dehak et al., 2011), and more recently for the modeling of language separation in bilingual speech input (Carbajal, 2018; Carbajal, Dawud, et al., 2016; Carbajal et al., 2016; de Seyssel & Dupoux, 2020).

Carbajal, Dawud, et al. (2016) used i-vector representations to separate English and Xitsonga speech in training conditions that employed either separated (monolingual) training data or mixed (bilingual) training data. They found that separation using this approach was more successful when trained on monolingual data than on mixed bilingual data. Their results also pointed to the relevance of speaker information in overcoming the complexities of being exposed to mixed data. De Seyssel and Dupoux (2020) extended this approach to include a bilingual condition where speakers spoke each of the two languages (English and Finnish, or English and German). They found that close languages (English and German) are more difficult to separate than more distant languages (English and Finnish), and that bilingual input where

speakers speak both languages is harder to cluster than input where speakers speak only one language (the ‘one parent one language’ approach).

Taking a phonological approach to language separation, Adriaans (2020) assessed segmental and phonotactic cues for input separation of English and Dutch. Computational modeling was used to determine the effectiveness of different cues in predicting the origin language in mixed data, as well as the robustness of these cues when dealing with different degrees of mixed input. Probabilistic models based on either the relative frequencies of segments, or phonotactic probabilities (biphone and triphone probabilities) were trained on combined samples from English and Dutch corpora in a variety of input mixing proportions, ranging from completely separated training data to 50-50 mixed input. The study found that phonotactics (in particular, biphones) provided the model with a cue for language separation which scored well in terms of both accuracy and robustness.

6.3.2 Bilingual Phonology in Computational Models of Speech Comprehension

Some of the earliest computational approaches to bilingual phonetics/phonology were phonological components embedded in larger computational models of spoken word comprehension. For example, Li and Farkas (2002) proposed a Self-Organizing Connectionist Model of Bilingual Processing (SOMBIP) which includes a bilingual lexicon of phonological forms. The bilingual lexicon was trained by presenting the model with the 400 most frequent English and Chinese word types taken from the Hong Kong Bilingual Corpus from CHILDES (MacWhinney, 2014; Yip & Matthews, 2000). No explicit language markers or labels were given to the model. Instead, the model was presented simultaneously with semantic and phonological representations and the representations self-organized on the basis of the similarity between representations (which included a fixed CVVCCVVC prosodic template,

where each C and V was represented by 5 different features, see also Li & MacWhinney, 2002). After training, the lexicon showed two distinct clusters of English and Chinese lexical representations, indicating that a bilingual lexicon can be learned without explicit language labels.

In similar work using Self-Organizing Maps with English and Spanish words, Shook and Marian (2013) noted that their Bilingual Language Interaction Network for Comprehension of Speech (BLINCS) separates the two input languages automatically on the basis of phonotactics, giving further computational support for the potential relevance of phonotactics in language separation during bilingual acquisition.

6.3.3 Computational Models of L2 Speech Perception

Several models of L2 speech perception have been implemented as computational models. These models do not perform simultaneous bilingual acquisition, but rather explore the effects of an established first language on L2 perception. For example, Keidel et al. (2003) trained a neural network on English CV syllables, and tested the model on the discrimination of isiZulu stimuli. Their model showed similar discrimination behavior to the human data reported in Best, McRoberts, and Goodell (2001). (See also Chapter 7, this volume) The model achieved this performance using exclusively acoustic information, without reference to articulatory gestures.

More recently, van Leussen and Escudero (2015) presented a computational implementation of Escudero's Second Language Linguistic Perception (L2LP) model (Escudero, 2005; see also Chapter 8, this volume). Their model gradually learned a two-way Spanish L2 vowel contrast after establishing a three-way Dutch L1 vowel contrast. One interesting feature of their model, in addition to being able to learn subsets of categories, is that

they aim to explain the learning *trajectory* from non-native to native-like perception, as opposed to L2 perception at one given point in development.

6.3.4 A Computational Model of Bilingual Phonotactic Learning

To illustrate the problem of bilingual input, we return to the 'toy' phonotactic learning example from Section 6.2.1, and put the computational model in a simulated bilingual environment. Being simultaneously exposed to two different languages complicates the learning problem, at least from a computational point of view. This can be seen in the example below, where we have expanded the two-sentence English toy corpus to include two more sentences taken from another language: Dutch. The two sentences were taken arbitrarily from the Spoken Dutch Corpus (Goddijn & Binnenpoorte, 2003), a corpus which is similar to the English Buckeye corpus in terms of size and level of transcription.

(6.7) `well i work in the accounting department`

`i'm an accounting assistant`

`laten we dit in de toekomst ook voortzetten`

('let's also continue this in the future')

`vooral als je in september dan pas weer start`

('especially if you only start again in September')

All four sentences come from adult spoken language corpora, so this simulated bilingual toy corpus is not reflective of an actual child's input, but they contain parts of the phonology of their respective languages, which allows us to illustrate the effects of language mixing on phonotactic probabilities.

Using computational modeling we can simulate bilingual phonotactic learning under different input conditions. For example, we can train one model that assumes the learner is able to separate the languages in the input, and another model that assumes the learner is not able to separate the input. Comparing the output of the two models allows us to quantify the effects of language mixing.

The Separated model would generate two language models, one for each language. The English part of this Separated model would be identical to the model presented in Table 6.1, and the Dutch part of the model would be computed in a similar way, but exclusively from the Dutch data. When using phonotactic probabilities in a word segmentation task, the Separated model would segment English test sentences the way it did in (6.6), because there is no influence of the second language.

In contrast, the Mixed model would compute transitional probabilities over the entire 4-sentence data set, and would use the resulting probabilities on any test set it is presented with, regardless of origin language. The phonotactic probabilities in this Mixed model would match the probabilities of neither English nor Dutch. By means of simulation we can quantify how detrimental this would be to a particular acquisition task, such as word segmentation.

Table 6.2 shows the phonotactic probabilities of the English biphones from Table 6.1 when the training set of the model includes the two Dutch utterances. (This is a subset of the total Mixed model, which also contains biphones that occur exclusively in the Dutch data.) Due to the mixed input, the probabilities have changed in two ways. They have changed in an *absolute* sense: most of the probabilities are lower than in the English-only (Separated) model. This is because the addition of a new language has resulted in an increase of the combinatorial possibilities. For example, where the Separated model says that in English `n` can be followed by four different phonemes (`ih`, `iy`, `n`, `t`), the Dutch data has introduced three more possible

successors to n (s, p, d) in the Mixed model. This by itself is not problematic: in both the Separated and the Mixed model all of these n-initial biphones have equal probabilities; the absolute values are simply lower in the Mixed model. (In fact all of these combinations are possible combinations in 'real' English as well, they just do not occur in 'toy' English.)

More interestingly, the *relative* phonotactic probabilities change in the Mixed model as well. For example, w ah and w er were equally probable in the Separated model, but in the Mixed model w ah has become more likely than w er. It is because of these relative changes that the model will behave differently on any task we might give the model, such as predicting well-formedness judgements or predicting word boundaries. This can be seen in the following example where the Separated and the Mixed model make different predictions in terms of word boundaries:

(6.8)

i. Separated:

w ah # aa w er k ih n n ih # ah k aw iy ih # p aa # r t m ih n

ii. Mixed:

w ah aa w er k # ih n n ih # ah k aw iy ih # p aa # r t # m ih n

iii. Gold standard:

w ah # aa # w er k # ih n # n ih # ah k aw iy # ih p aa r t m ih n

The Mixed model makes different predictions at three locations (underlined in the test sentence), and comparison to the gold standard indicates that in two of the three cases the mixed prediction is false. Mixing English and Dutch data in this case thus had a negative impact on the performance of the model. This is of course just a toy simulation, so no conclusions should be drawn about mixing English and Dutch from this example, but the example illustrates how

mixed statistics could lead to problems in solving computational tasks that are central to language development. Computational modeling provides a way to quantify the impact of mixed input, and to explore potential solutions to separate data coming from two different languages (e.g., Adriaans, 2020; Carbajal, Dawud, et al., 2016; Shook & Marian, 2013).

Table 6.2 A biphone model with mixed English-Dutch statistics.

Biphone	TP	Biphone	TP	Biphone	TP	Biphone	TP	Biphone	TP
aa m	0.13	aw n	0.50	iy ih	0.50	n iy	0.14	s t	0.38
aa r	0.25	eh n	0.25	iy ng	0.50	n n	0.14	t eh	0.20
aa w	0.13	er k	1.00	k aw	0.40	n t	0.14	t m	0.10
ah aa	0.13	ih ah	0.11	k ih	0.20	ng ih	1.00	w ah	0.50
ah ah	0.13	ih n	0.44	m ah	0.25	p aa	0.67	w er	0.25
ah k	0.25	ih p	0.11	m ih	0.25	r t	0.50	-	-
aw iy	0.50	ih s	0.22	n ih	0.14	s ih	0.13	-	-

6.4 Future Directions

This chapter discussed current computational approaches to bilingual phonetic and phonological acquisition. There is a sizeable gap between the number of computational studies on monolingual acquisition and the number of computational studies on bilingual acquisition, and it will take many more modeling studies to start closing that gap. The gap is not entirely surprising though. As other chapters in this volume illustrate, bilingualism comes in many different forms. Bilingual learners vary greatly in terms of the age at which the second language is introduced, the input quantities for each of the different languages, the distribution of languages over speakers in the learner's environment, etc. One key challenge in the computational modeling of bilingual phonetics and phonology will be to take all these factors into account. Computational approaches can help us to understand bilinguals' impressive learning mechanisms by making the learning process, and connections between different factors that affect the learning process, explicit, and by testing hypotheses regarding bilingual acquisition against natural language data.

Currently, however, there is a limited availability of ecologically valid training and test materials for bilingual computational models. Modeling studies have used adult spoken language corpora of different languages (often from different sources) to simulate bilingual input. While this gives an impression of the overall complexity of mixed input from two languages, and there is some evidence that suggests that input from a single bilingual speaker might not be very different from the combined input from two monolingual speakers (Danielson et al., 2014), ultimately there is a need for large phonetically and/or phonologically transcribed bilingual corpora to be used in simulations of bilingual learning.

Another challenge lies in modeling the individual learning trajectories of bilingual learners. Modeling has so far mostly focused on identifying overall solutions to learning a language. A major challenge with modeling bilingualism, perhaps even more so than with monolingual modeling, is to account for individual differences that may vary due to factors such as the different degrees of parental language mixing (e.g., Byers-Heinlein, 2013). Studies that test the behavior of a model at different input mixing proportions (e.g., Adriaans, 2020) are a first step towards understanding how differently shaped input can lead to different outcomes.

Regarding the bilingual learning mechanisms used in phonetic and phonological acquisition, much remains unknown. Computational models can help here by exploring how a particular problem could *in principle* be solved. Highly debated questions such as whether bilingual learning mechanisms are fundamentally different from monolingual mechanisms, or the extent to which bilingual learners separate their languages, can be approached from a computational perspective. In order to assess computational models that simulate human bilingual learning and learning trajectories, the output of the models can be evaluated against developmental findings. Ultimately a cycle between computational modeling and experiments

with human participants is needed to uncover the full complexities involved in bilingual phonetic and phonological acquisition.

References

Adriaans, F. (2018). Effects of consonantal context on the learnability of vowel categories from infant-directed speech. *The Journal of the Acoustical Society of America*, *144*(1), EL20–EL25.

Adriaans, F. (2020). The effectiveness of phonological cues for bilingual input separation. *Paper presented at the 45th Boston University Conference on Language Development (BUCLD45)*.

Adriaans, F. & Kager, R. (2010). Adding generalization to statistical learning: The induction of phonotactics from continuous speech. *Journal of Memory and Language*, *62*, 311–331.

Adriaans, F. & Kager, R. (2017). Learning novel phonotactics from exposure to continuous speech. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, *8*(1), 1–14.

Adriaans, F. & Swingley, D. (2017). Prosodic exaggeration within infant-directed speech: Consequences for vowel learnability. *Journal of the Acoustical Society of America*, *141*(5), 3070–3078.

Albright, A. (2009). Feature-based generalisation as a source of gradient acceptability. *Phonology*, *26*, 9–41.

Bailey, T. M. & Hahn, U. (2001). Determinants of wordlikeness: Phonotactics or lexical neighborhoods? *Journal of Memory and Language*, *44*, 568–591.

Best, C. T., McRoberts, G. W., & Goodell, E. (2001). Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listener's native phonological system. *The Journal of the Acoustical Society of America*, *109*(2), 775–794.

Boersma, P. & Hayes, B. (2001). Empirical tests of the Gradual Learning Algorithm. *Linguistic Inquiry*, *32*(1), 45–86.

Bonatti, L. L., Peña, M., Nespor, M., & Mehler, J. (2005). Linguistic constraints on statistical computations. *Psychological Science*, *16*, 451–459.

Brent, M. R. & Cartwright, T. A. (1996). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition*, *61*, 93–125.

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*(1), 32–48.

Cairns, P., Shillcock, R., Chater, N., & Levy, J. (1997). Bootstrapping word boundaries: A bottom-up corpus-based approach to speech segmentation. *Cognitive Psychology*, *33*, 111–153.

Carbajal, M. J. (2018). *Separation and acquisition of two languages in early childhood: A multidisciplinary approach.* [Doctoral dissertation, Université de recherche Paris Sciences et Lettres].

Carbajal, M. J., Dawud, A., Thiolliere, R., & Dupoux, E. (2016). The "language filter" hypothesis: A feasibility study of language separation in infancy using unsupervised clustering of i-vectors. *2016 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, 195–201.

Carbajal, M. J., Fér, R., & Dupoux, E. (2016). Modeling language discrimination in infants using i-vector representations. *Proceedings of the 38th Annual Conference of the Cognitive Science Society,* 889–896.

The CMU pronouncing dictionary, version 0.7b. (2014). Retrieved from `http://www.speech.cs.cmu.edu/cgi-bin/cmudict`

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.

Daland, R. & Pierrehumbert, J. B. (2011). Learning diphone-based segmentation. *Cognitive Science*, *35*, 119–155.

Danielson, D. K., Seidl, A., Onishi, K. H., Alamian, G., & Cristia, A. (2014). The acoustic properties of bilingual infant-directed speech. The *Journal of the Acoustical Society of America*, *135*(2), EL95–EL101.

de Boer, B. & Kuhl, P. K. (2003). Investigating the role of infant-directed speech with a computer model. *Acoustics Research Letters Online*, *4*, 129–134.

Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., & Ouellet, P. (2010). Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*(4), 788–798.

Dehak, N., Torres-Carrasquillo, P. A., Reynolds, D., & Dehak, R. (2011). Language recognition via i-vectors and dimensionality reduction. *Proceedings Interspeech 2011*, 857–860.

de Seyssel, M. & Dupoux, E. (2020). Does bilingual input hurt? A simulation of language discrimination and clustering using i-vectors. *Proceedings of CogSci 2020 - 42nd Annual Virtual Meeting of the Cognitive Science Society*, 2791–2797.

Dillon, B., Dunbar, E., & Idsardi, W. (2013). A single-stage approach to learning phonological categories: Insights from Inuktitut. *Cognitive Science*, *37*, 344–377.

Escudero, P. (2005). *Linguistic perception and second language acquisition: Explaining the attainment of optimal phonological categorization*. Netherlands Graduate School of Linguistics.

Feldman, N. H., Griffiths, T. L., Goldwater, S., & Morgan, J. L. (2013). A role for the developing lexicon in phonetic category acquisition. *Psychological Review*, *120*, 751–778.

Goddijn, S. & Binnenpoorte, D. (2003). Assessing manually corrected broad phonetic transcriptions in the Spoken Dutch Corpus. *Proceedings of the 15th International Congress of Phonetic Sciences*, 1361–1364.

Gouskova, M. & Gallagher, G. (2020). Inducing nonlocal constraints from baseline phonotactics. *Natural Language & Linguistic Theory*, *38*(1), 77–116.

Hay, J., Pierrehumbert, J., & Beckman, M. (2004). Speech perception, well-formedness, and the statistics of the lexicon. *Papers in Laboratory Phonology vi*, 58–74.

Hayes, B. & Wilson, C. (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry*, *39*, 379–440.

Hochmann, J.-R., Benavides-Varela, S., Nespor, M., & Mehler, J. (2011). Consonants and vowels: different roles in early language acquisition. *Developmental Science*, *14*(6), 1445–1458.

Jurafsky, D. & Martin, J. H. (2009). *Speech and language processing*, 2nd ed., Pearson Prentice Hall.

Jusczyk, P. W., Luce, P. A., & Charles-Luce, J. (1994). Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language*, *33*, 630–645.

Kastner, I. & Adriaans, F. (2018). Linguistic constraints on statistical word segmentation: The role of consonants in Arabic and English. *Cognitive Science*, *42*, 494–518.

Keidel, J. L., Zevin, J. D., Kluender, K. R., & Seidenberg, M. S. (2003). Modeling the role of native language knowledge in perceiving nonnative speech contrasts. *Proceedings of the 15th International Congress of Phonetic Sciences*, 2221–2224.

Legendre, G., Miyata, Y., & Smolensky, P. (1990). Harmonic grammar—a formal multi-level connectionist theory of linguistic well-formedness: Theoretical foundations. *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society*, 884–891.

Li, P. & Farkas, I. (2002). A self-organizing connectionist model of bilingual processing. *Advances in Psychology, 134*, 59–85.

Li, P. & MacWhinney, B. (2002). PatPho: A phonological pattern generator for neural networks. *Behavior Research Methods, Instruments, & Computers*, *34*(3), 408–415.

MacWhinney, B. (2014). The CHILDES project: Tools for analyzing talk, Volume II: The database. Psychology Press.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. San Francisco, CA: W.H. Freeman.

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*, B101–B111.

McMurray, B., Aslin, R. N., & Toscano, J. C. (2009). Statistical learning of phonetic categories: insights from a computational approach. *Developmental Science*, *12*, 369–378.

Pierrehumbert, J. B. (2003). Probabilistic phonology: Discrimination and robustness. In R. Bod, J. Hay, & S. Jannedy, eds., *Probabilistic Linguistics*. Cambridge, MA: The MIT Press, pp. 177–228.

Pitt, M. A., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E., & Fosler-Lussier, E. (2007). *Buckeye corpus of conversational speech (2nd release).* Columbus, OH: Department of Psychology, Ohio State University, 265–270.

Potts, C., Pater, J., Jesney, K., Bhatt, R., & Becker, M. (2010). Harmonic grammar with linear programming: from linear systems to linguistic typology. *Phonology*, *27*(1), 77–117.

Prince, A. & Smolensky, P. (2004). *Optimality Theory: Constraint Interaction in Generative Grammar. Malden, MA/Oxford: Wiley–Blackwell.*

Prince, A. & Tesar, B. (2004). Learning phonotactic distributions. In R. Kager, J. Pater, & W. Zonneveld, eds., *Constraints in Phonological Acquisition*. Cambridge, UK: Cambridge University Press, pp. 245–291.

Richtsmeier, P. T. (2011). Word-types, not word-tokens, facilitate extraction of phonotactic sequences by adults. *Laboratory Phonology*, *2*, 157–183.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928.

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, *35*, 606–621.

Shook, A. & Marian, V. (2013). The bilingual language interaction network for comprehension of speech. *Bilingualism: Language and Cognition*, *16*(2), 304–324.

Sundara, M. & Breiss, C. (2020). Infants are sensitive to phonotactic patterns in their native language at 5-months. *Paper presented at the 45th Boston University Conference on Language Development (BUCLD45)*.

Sundara, M. & Scutellaro, A. (2011). Rhythmic distance between languages affects the development of speech perception in bilingual infants. *Journal of Phonetics*, *39*(4), 505–513.

Swingley, D. (2009). Contributions of infant word learning to language development. *Philosophical Transactions of the Royal Society B*, *364*, 3617–3622.

Swingley, D. & Alarcon, C. (2018). Lexical learning may contribute to phonetic learning in infants: A corpus analysis of maternal spanish. *Cognitive Science*, *42*(5), 1618–1641.

Tesar, B. & Smolensky, P. (2000). *Learnability in Optimality Theory*. Cambridge, MA: The MIT Press.

Vallabha, G. K., McClelland, J. L., Pons, F., Werker, J. F., & Amano, S. (2007). Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Sciences of the United States of America*, *104*, 13273–13278.

van Leussen, J.-W. & Escudero, P. (2015). Learning to perceive and recognize a second language: the L2LP model revised. *Frontiers in Psychology*, *6*, 1000.

Vitevitch, M. S. & Luce, P. A. (1999). Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory and Language*, *40*, 374–408.

Yang, C. D. (2004). Universal Grammar, statistics or both? *Trends in Cognitive Sciences*, *8*, 451–456.

Yip, V. & Matthews, S. (2000). Syntactic transfer in a Cantonese–English bilingual child. *Bilingualism: Language and cognition*, *3*(3), 193–208.