Unsupervised learning

Modelling the earliest stages of phonological acquisition

Alma de Jonge, 10205462
Thesis RMA Linguistics, 2013
Thesis advisor: dr. Silke Hamann
University of Amsterdam

Contents

Chapter 1

Introduction

Infants learn to distinguish the different phonetic categories of their native language remarkably quickly. While some research suggests that phonetic categorization begins at the same time that infants learn to acquire words (i.e., at around 1 year of age, Best, 1995; Werker & Lalonde, 1988), other research on the perception of sounds (Kuhl et al., 1992; Polka & Werker, 1994) in contrast demonstrates that infants 6 to 10 months of age already show more sensitivity towards the phonetic properties of their language that signal phonemic contrast than towards those properties that are not contrastive. While these experiments show that learning at this point in life has already developed to the extent that the infant can successfully distinguish different phonemic contrasts, actual learning of this ability probably begins at an even earlier age, as the ability to contrast certain language-specific sounds needs to be developed before it can function. Indeed, some studies (e.g. Mehler et al., 1988; Moon et al., 1993) suggest that even extremely young infants of only a few days old already show some sensitivity towards their native language. We may assume, then, that from birth an infant needs approximately 6 to 10 months of input sounds[1] before it has built up enough evidence to justify the existence of language-specific phonetic categories.

Several models have been proposed to account for this early sensitivity of infants. One model is given by Maye et al. (2002), who attribute the learning mechanism to the general ability of distributional learning. Applied to the acquisition of phonetic categories, this model states that infants gradually build phonetic categories from the information that some tokens on a particular dimension they hear are more similar to each other than to other tokens on that same dimension. For example, in a language that contrasts voiced from voiceless consonants, on the Voice Onset Time (VOT) dimension the infant hears that different tokens of /b/ are more similar to each other than to tokens of /p/. Then, when the infant maps the acoustic values of different /p/-/b/ tokens it heard on the VOT dimension, eventually two separate clusters will begin to emerge, as the infant will hear many tokens that are positioned on the edges of the dimension and relatively few that are positioned in the middle. With each token heard, the clusters become more defined, eventually reaching a state for which the infant can with great probability assume that each cluster represents a separate phonetic category. When certain tokens are not contrastive on a particular dimension (e.g. because the language does not contrast voiced from voiceless consonants), all tokens will be perceived as being more or less similar on that particular dimension. This leads to the emergence of one broad cluster in the middle of the dimension, and eventually it leads to the formation of one phonetic category to which

[1] The exact time needed to create reliable phonetic categories partly depends on the acoustic nature of the sound. For example, it has been shown that vowel categories are acquired earlier than consonant categories (Polka & Werker, 1994; Werker & Tees, 1984).

all tokens are mapped.[2] The difference between such unimodal and bimodal distributions is shown in figure 1.1.

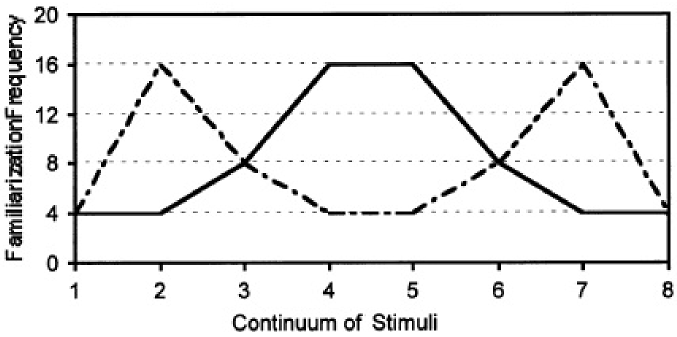

Figure 1.1: The emergence of unimodal and bimodal distributions (Maye et al., 2002, p. B104)).

While I do think that learning in this stage is indeed guided by statistical information given by the input, a problem with purely statistical models like this one is that it cannot be made clear when *exactly* the infant has heard enough tokens to assume the existence of either one or two (or even more) phonetic categories. If we want to be able to give an exact point at which we can say that phonetic categories have emerged (for example if we want to simulate this type of acquisition), then we need a more precise and formal model than simple statistical learning in order to do that. Also, while statistical learning can explain the emergence of multimodal clusters in cases where the tokens of all phonetic categories on one dimension are fully separated, it is not immediately clear if and how this model can explain the correct emergence of phonetic categories in cases where parts of the clusters overlap, such that some input tokens can belong to more than one phonetic category.

Other models try to explain early phonological acquisition in a more formal way. In many of these computational models of phonological acquisition, learning is represented with a Mixture-of-Gaussians (henceforth MoG) model (e.g. Vallabha et al., 2007; Adriaans & Swingley, 2012). With MoG models, learning is represented as a process of estimating the phoneme distribution as a mixture of Gaussian (Normal) functions. These functions are defined by a number of parameters (for more detailed information on the exact nature of these parameters, see e.g. Benders (2013, ch. 5)). Furthermore, it is assumed that each Gaussian function represents exactly one phonetic category. For every input token, the model calculates the probability that this token belongs to a particular Gaussian function. Initially, the number of Gaussian functions and their parameter values are random. Then, with each input token, the functions and parameter values are updated in order to increase the probability that the model generates that token. In other words, with each input value the parameter values of each function are changed to come to lie closer to the input values. The probability parameter value is only increased for the winning function, i.e., the function whose parameter values were closest to the acoustic values of the input token. All functions with low enough probability values are removed from the model. Eventually, the model is supposed to reach a state in which each function represents one phonetic

[2] Even though Maye et al. (2002) only mention the idealized situation where tokens of an unimodal distribution are always situated in the middle of the acoustic dimension, it should in theory also be possible for an infant to acquire unimodal distributions of which the tokens lie more to one of the sides of the dimension.

category, each with the appropriate parameter (or: acoustic) values.[3]

A MoG model is more discrete than a statistical learning model in that the assumed number of phonetic categories is always equal to the number of Gaussian functions. That is, in a MoG model an infant is assumed to already begin with a number of phonetic categories and as soon as one category becomes too improbable (specified by an exact probability value) it is dropped. This way it is always possible to say exactly how many phonetic categories the infant has acquired. However, it is perhaps unlikely to assume that infants already start with a random number of categories, each with random acoustic values. It is also unlikely or at least undesirable to assume that for each input token, the infant precisely calculates how the acoustic values of this token relate to the acoustic values of its phonetic categories. In principle we want learning at such an early stage to be as easy as possible, without the need to make the assumption that infants do complicated calculations for each sound they perceive. Modelling with MoG in that sense is more like a computational model of speech perception than like an actual empirically sound model of how infants really perceive sounds.

There are more theories that try to explain early acquisition, such as Neural Network models (Benders, 2013, ch. 5) or Exemplar Theory (Pierrehumbert, 2003), but most of these theories either do not make explicit at what point exactly categories emerge or they are too complicated to be a probable model of actual learning in infants. The aim of this thesis, therefore, is to develop a model in which early phonetic acquisition is represented in an empirically adequate way while also being able to precisely predict at what point phonetic categories have actually emerged. I will do this by using a model that is based on a (heavily adapted) version of Optimality Theory (henceforth OT). While there is a number of work that uses OT to model supervised (or lexicon-driven) learning (e.g. Tesar & Smolensky, 1998; Boersma & Hamann, 2008), there seems to be virtually no work on unsupervised learning in Optimality Theory (the only work I am aware of is a short proceedings paper by Boersma et al. (2003) and a poster presentation by Van Leussen (2012)). To fill this gap, most of this thesis will be centred around the development of an unsupervised OT-based model and the assumptions one needs to consider when trying to model unsupervised learning.

The rest of this thesis is built up as follows. Chapter 2 in general explains the assumptions we will make in our model of unsupervised learning. Section 2.2 briefly outlines the basic assumptions of traditional OT. Some of the assumptions that are used in traditional OT are not used in our model. Section 2.3 explains which assumptions are dropped in our model and why we chose to do this. The more detailed assumptions about our model will be dealt with in section 2.4.

In chapter 3, the mechanism of our model will be shown in detail by applying the model on a constructed language example. To make the picture complete, chapter 4 will discuss how learning proceeds beyond the first unsupervised stage all the way until phonological acquisition is completely finished, and maybe even beyond that point.

The best way to see if a model actually works is to run virtual simulations of the model using real language data. There is no room to do this in this thesis, but chapter 5 will briefly discuss the things to consider when running such simulations. Finally, chapter 6 lists the conclusions we can draw on the basis of this thesis and some points that are still open for discussion.

[3] It is implied (although not explicitly stated) by e.g. Vallabha et al. (2007) that Gaussian functions can only be removed, such that it is impossible for new Gaussian functions to emerge during the learning process. This would mean that it is essential that the initial number of Gaussian functions is always equal to or higher than the final number of distributional categories.

Chapter 2

Explanation of the model

2.1 Introduction

The development of a new model comes with many assumptions that need to be specified and explained. That is why most of this chapter is dedicated to the specification of both the general and more detailed assumptions in our model. Also, it was already mentioned that the model is based to a large extent on the foundations of traditional Optimality Theory. Therefore, the first section of this chapter outlines the basic notions used in OT.

2.2 Optimality Theory

Optimality Theory was first developed by Prince & Smolensky (1993). It was primarily developed in order to be able to explain the discrepancies between the surface form of an utterance (i.e., the way a certain utterance is pronounced) and its underlying mental representation. For example, in Dutch the word *hond* 'dog' is pronounced [hɔnt]. Still, the underlying phonological form of this word is /hɔnd/, which for example becomes apparent in the plural form /ˈhɔn dən/. While earlier accounts used *derivational rules* to explain such differences (e.g. Chomsky & Halle, 1968; Kiparsky, 1982), Optimality Theory tries to account for this difference between underlying form and surface structure through the introduction of *phonological constraints*. These constraints serve to prevent certain underlying forms from appearing in the surface structure. The constraints are often assumed to be part of an infant's innate knowledge, such that all infants are born with the same set of constraints. All constraints are ranked in a particular order; it is the exact ordering of these constraints that determines the surface structure of a given underlying form.

In OT, underlying representations are often referred to as *input forms* and surface representations as *output candidates*. For every input form, the entity GEN (generator) constructs a list of possible output candidates. The number of candidates in this list is potentially infinite and is at least made up of all possible sound sequences of the given language, although in practice linguists usually only consider those output candidates that are relevant to the input form. Then, EVAL (evaluator) considers which output candidate is most optimal given the set of constraints and their language-specific ranking. In OT, being most optimal does not mean that the output candidate violates no or the least number of constraints, but that this candidate does not violate highly ranked constraints that are violated by all other output candidates. To illustrate this, consider our earlier example of Dutch *hond*. The evaluation mechanism is often illustrated with evaluation tableaux like tableau 1 below. Now, suppose that a language user wants to produce the word *hond*. The mental representation /hɔnd/ in this case serves as the input form. In the tableau, the input form is always represented in the upper left corner. With GEN, a number of output candidates is generated which are shown directly below the input form. Which of these output candidates will be chosen as the actual surface form depends on the ranking of the constraints. In tableau 1, the only three relevant constraints are Ident, *VoicedCoda and *ɔ. The faithfulness constraint Ident states that any surface form should be phonologically identical to the input form, with each alteration between the two forms counting as a constraint violation. The markedness constraint *VoicedCoda states that all coda sounds should be voiceless, such that each utterance with a voiced sound in coda position violates this constraint. *ɔ argues against any forms that include the phoneme /ɔ/. All violations are marked with asterisks. Fatal violations are marked with an exclamation mark behind the asterisk. Whenever one of the output candidates is fatally violated, further violations are no longer considered. This is illustrated by shading of the respective grids. The comparison of the input form with the output candidates continues until there is only one output candidate left. This candidate is then chosen as the winning candidate.

/hɔnd/	*VoicedCoda	Ident	*ɔ
[hɔnd]	*!		*
☞ [hɔnt]		*	*
[hɑnd]	*!	*	
[hɑnt]		**!	
[rɔmb]	*!	***	*

Tableau 1

In the case of our example, we can see that *VoicedCoda is ranked higher than both Ident and *ɔ. This means that in Dutch, it is worse to have a voiced coda sound than to have an output form that is not identical to the input form. The outcome of this particular constraint ranking is that the input form /hɔnd/ will be pronounced in Dutch as [hɔnt], just as we have already seen in this language. Because the evaluation of the first two constraints already yielded exactly one (winning) candidate, the influence of lower-ranked constraints such as *ɔ does not further alter the outcome.[4]

As we assume that all human beings (and thus all human languages) are given the same set of constraints at birth, the only way for languages to differ phonologically is in the way that they rank their constraints. This may not only hold for the most commonly studied phenomena in OT such as the computation of certain underlying representations into surface forms or the occurrence of certain phoneme clusters, but even for the inventory of phonemes a language uses. In phonological theory, it is sometimes assumed that some phonemes are inherently more *marked* than others, with this markedness limiting their frequency of occurrence within languages (for a detailed exploration of the notion of markedness, see Hume (2011)). In generative theories, this might mean that each infant is born with constraints that try to prevent the more marked phonemes from appearing in language. The fact that some marked phonemes may still exist in certain languages is then accounted for by assuming that the constraints arguing against these phonemes are ranked too low to prevent this.

While OT is usually used to model speech production, it is entirely possible to model speech perception with OT as well. In the case of perception, an interesting aspect of OT

[4] In fact, since faithfulness constraints such as Ident always favour one and only one candidate over the other candidates (as there is always one and only one output candidate that is exactly identical to the input form), the influence of constraints that are ranked lower than such faithfulness constraints is by default excluded from further consideration.

is that the constraints can be used to categorize speech tokens coming from a continuous distribution into a discrete number of phonological categories. That way, the constraints nicely mimic the fact that while speech sounds are highly continuous in that each speech sound is at least slightly different from all other speech sounds, human beings still seem to categorize each speech sound into a finite set of phonemes. In other words, in this sense OT can be used to limit the number of perceptual options for each speech sound.

While there are undoubtedly many appealing aspects to OT, some of the implications made within OT cannot be used as easily in our model. In the next section I will list these implications and explain why we will not use them.

2.3 General assumptions

In the introduction, it was explained why we chose to use Optimality Theory to model an infant's phonetic perception and not some other model, such as Mixture of Gaussians and distributional learning. However, we will not be using OT in the 'traditional' sense, as was first introduced by Prince & Smolensky (1993). The reason for this is that traditional OT accounts come with a number of assumptions, of which four especially are problematic to or at least not needed in our account of OT learning. The first assumption has to do with the more typological focus of traditional OT accounts, while the second assumption is aimed at the supposedly innate nature of the OT constraints. The third problematic assumption represents phonological learning as the *discrete* demotion of constraints. Lastly, the fourth assumption has to do with the supervised nature of learning in OT. In the following sections, I will explain in more detail why these assumptions cannot be used in our account of OT learning.

2.3.1 The typological nature of OT

Traditional OT accounts such as the one described by Prince & Smolensky (1993) are more typologically focused than focused on learnability issues. This means that, traditionally, OT was developed to try to account for the phonological variation and phonological processes observed within languages of the world and to try to put restrictions on the set of possible phoneme inventories. For various reasons, it is highly unlikely that these accounts mirror actual speech perception or production. One reason for this is that, in principle, these accounts assume that every input should be compared to an infinite number of output candidates (or: an extremely large number of output candidates that are made up of all possible phonological combinations if we assume a finite set of phonemes and a maximum length of strings) before it is possible to decide which output candidate is most optimal compared to the input form (Idsardi, 2006). Since speakers or listeners have far from infinite time to produce or process a certain input form, traditional OT is not capable of making claims about the way language users produce or perceive strings of sounds. Note, however, that these types of OT accounts do not necessarily consider this to be a problem. For example, Kager (1999, p. 25-26) notes that as a phonological theory that is typically based on the framework of generative grammar, traditional OT focuses solely on the formal similarities between language systems. In other words, its main focus is on the regularities it finds within the linguistic competence. How these regularities should be cognitively implemented into the human brain (the linguistic performance) is considered a research goal of other disciplines (psycholinguistics, neurolinguistics, language acquisition etc.). We can thus assume that traditional OT accounts consider learnability issues to fall outside the scope of their research questions.

When OT is used to explain typological differences, the issue of having an infinite list of output candidates is not necessarily problematic. But since we claim in this thesis that OT can also be used to formalize how children acquire phonetic and phonological categories, we simply cannot assume that every input a child hears will be compared to an infinite number of output candidates. The simple solution, then, is to assume that the number of output candidates is always finite (as well as being reasonably small). For us, this assumption is maybe more easy to make than for those versions of OT that deal with typological variation, as we will only look at the acquisition of single phonetic or phonological categories and not at the rules that define the observed interactions between various phonemes in actual words. To make this more concrete, reconsider the example given in section 2.1. In this example, the input form /hɔnd/ was compared to a number of output candidates. For ease of exposition, only the relevant candidates where shown, but in principle our input form should not only be compared to similar output candidates such as e.g. [hɔnt] and [hɑnd], but also to all other Dutch words such as [boːm] 'tree', [ˈhɑnt sxun] 'glove' and [neː] 'no' and possibly even all words that are made up of phoneme sequences that are possible in Dutch but that do not have any meaning, such as [sxoːp]. Furthermore, since the length of words is not necessarily limited, this would make an infinite list of output candidates. It is unclear how the number of output candidates could be limited in these cases, and to my knowledge there are no real solutions to this 'infinite list' problem. In our model, however, we will not encounter this problem of having a possibly infinite number of output candidates, as every input form will always only be compared to either a finite number of phonological categories or a finite number of acoustic values. In short, we can thus say that our OT model differs from the usual OT models in that it explicitly departs from the (often implicit) assumption that every input should be compared to an infinite number of output candidates. We can do this because we will not focus on phonemes within words, but only on separate phonemes, phonological features or phonetic categories outside their lexical context. By doing this, the number of output candidates in OT-based models of phonological acquisition can be kept both finite and reasonably small (this will be further shown when we define our models in chapters 3 and 4).

2.3.2 The innate nature of OT constraints

A very well-known implication of Chomsky's generative grammar is that parts of a child's language abilities are innate, i.e., that infants are born with some form of linguistic knowledge. This innate knowledge is often termed Universal Grammar. Since many accounts of OT are heavily rooted in the assumptions of the generative framework, traditional OT accounts usually assume that all phonological constraints are both universal and innate. This means that all infants are born with the same set of phonological constraints, and that the phonological differences we find between languages derive from a difference in the relative ranking of the constraints within these languages.

In my opinion, there are a number of drawbacks to the assumption of constraints being innate. In general, the assumption of innate language knowledge requires the presence of a cognitive 'language faculty' unique to human beings (Hauser et al., 2002). Apart from that this might be problematic from an evolutionary perspective, another question is whether this innate knowledge is really needed to explain the linguistic patterns found in various languages. While I do realize that at this point, argumentation perhaps boils down to personal preference, to me most linguistic patterns can adequately be explained without assuming any specific language-oriented innate knowledge. Instead, language patterns are then the outcome of general cognitive, physiological or communicative abilities and preferences. When applied to phonological knowledge, we might for example say that the notion of markedness that causes some phonemes to be more prevalent than others does not derive from a number of innate constraints, but simply from the fact that some phonemes are physiologically harder to pronounce than others. As speakers prefer to use the least possible amount of effort when pronouncing sounds, this might explain the relatively infrequent occurrence of hard to pronounce sounds in languages. Here, the principle of Occam's razor[5] comes into play: when language phenomena can be explained without using language-specific mechanisms, then these explanations are to be preferred over explanations that require human beings to be born with language-specific devices.

Even if we assume an innate set of constraints, it is extremely hard if not impossible to give a list of which constraints should be part of the innate set.[6] We can only do this indirectly by looking at the languages in the world to see which phonological patterns arise less frequently than we would expect on the basis of chance. But even such tendencies do not necessarily point to the existence of a constraint that limits the occurrence of that phonological pattern, as the absence may also be due to other factors. In general, I think it is important to realize that every model, including OT, only *mimics* what happens in reality. The constraints used in OT are constructs invented by linguists. They do not necessarily exist outside the tableaux. When we start assuming that every human being actually has a list of constraints in their head, to me the line between linguistic model and cognitive reality becomes too blurred. Even though in this thesis I do use OT to model phonological acquisition, I want to stress that, essentially, the model is just that: a model. My goal here is to make a model that is largely in line with what we empirically know about early perception, but I do not at all hold the illusion that infants use actual tableaux with constraints, input forms and output candidates every time they hear a sound.

So if we do not assume any innate knowledge, then what do we assume? In this thesis, I will take up the *emergent approach* that assumes that all phonological knowledge emerges from input data. This means that, in our model, infants are born without any innate linguistic knowledge but with some physiological and general cognitive abilities that guide early perception. All further knowledge comes from the input sounds an infant receives from its parents and other speakers. How these assumptions are converted to OT-like constraints will be discussed in section 2.4.

2.3.3 Learning in traditional OT

Despite the fact that many OT accounts only focus on typological similarities between languages, this does not mean that questions on learnability are completely unattested in the OT literature. In fact, a number of researchers did (and do) try to show how OT can be used to explain a child's acquisition of a phonological grammar. Possibly the most well-known account of learnability within OT was developed by Tesar & Smolensky (1998). They developed a theory of language acquisition in which phonological acquisition involves the relative re-ranking of constraints based on the input a child receives. This re-ranking relies on a principle Tesar and Smolensky called the 'Constraint Demotion principle'. This principle is defined as follows (Tesar & Smolensky, 1998, p. 240):

[5] This principle states that when one needs to choose between two competing hypotheses, the hypothesis that needs the least amount of additional assumptions is preferred over the other one.

[6] That is not to say, of course, that it is impossible to make a list of all constraints that have been proposed by linguists up to now. This has for example been done by Ashley et al. (2010). However, such lists do not necessarily directly correspond to the universal list of constraints that human beings are supposed to be born with.

Constraint demotion

"For any constraint C assessing an uncanceled winner mark, if C is not dominated by a constraint assessing an uncanceled loser mark, demote C to immediately below the highest-ranked constraint assessing an uncanceled loser mark."

What Tesar and Smolensky mean here, is that every time the then-current ranking of constraints predicts another output candidate than what the child considers to be the correct candidate, then the constraint that stops the correct candidate from appearing is demoted below the constraint that would stop the incorrect but chosen candidate from appearing. As an example, consider tableau 2.

A	*B	*A	*C	*D
✓ A		*!		
B	*!			
C			*!	
☞ D				*

Tableau 2

In this tableau, the correct output candidate for the input form *A* is A. However, the current constraint ranking causes the incorrect candidate D to be chosen as the optimal candidate. To correct this mistake, constraint *A is demoted to stand below *D. This way, the learner has updated his constraint ranking such that in the future, output candidate A will be chosen for input form *A*. This new state is shown in tableau 3.

A	*B	*C	*D	*A
☞ A				*
B	*!			
C		*!		
D			*!	

Tableau 3

Tesar and Smolensky argue that this demotion principle is used until all constraints have reached their correct position (i.e., the position in which the correct output candidate is always chosen). At that point, we might say that learning has stopped.

There are at least two assumptions in this model that are problematic to our model of early acquisition. First, models like these can only work if one assumes that the learning process is supervised by some 'error detection mechanism'. In this particular model, it is assumed that the child's knowledge of underlying phonemic structures can tell the child when the chosen output candidate is incompatible with the input form. In our example, it means that the child has the knowledge that the surface form *A* has an underlying form A. However, it is unclear and often left unspecified how the child should have acquired this knowledge. As this is a problem not only for the Constraint Demotion model but for many accounts of phonological acquisition, I will further discuss this problem in section 2.3.4.

Boersma & Hayes (2001, p. 64-66) point out a second problem of the Constraint Demotion model, as they state that the model cannot deal with free variation. An example of this problem is given below.

Sometimes, there is more than one generally accepted way to pronounce a word (e.g. the pronunciation of *lawyer* as either ['lɔ jər] or ['lɔɪ ər]). In these cases, the child may hear both of the pronunciations from time to time. Now, suppose the child hears an instance of ['lɔ jər], as can be seen in tableau 4. At the current state of the tableau, the ranking of the constraints predicts an incorrect output candidate. As we know, the way to solve this is by demoting the high ranked constraint below the low ranked constraint.

[ˈlɔ jər]	*/ˈlɔ jər/	*/ˈlɔɪ ər/
✓ /ˈlɔ jər/	*!	
☞ /ˈlɔɪ ər/		*

Tableau 4

But then, suppose the child hears an instance of [ˈlɔɪ ər]. Since we just demoted the first constraint such that the new ranking also predicts an incorrect outcome (see tableau 5), again we must solve this by demoting the constraint that is now ranked too high. This process keeps going on whenever the child hears one of the two variations that is at that point wrong according to the tableau. In other words, free variation prevents the tableau from ever reaching a stable state.

[ˈlɔɪ ər]	*/ˈlɔɪ ər/	*/ˈlɔ jər/
☞ /ˈlɔ jər/		*
✓ /ˈlɔɪ ər/	*!	

Tableau 5

As we have seen, the Constraint Demotion principle cannot adequately handle free variation. A reason for this can be that the learning steps in this model are too severe: the child only needs to receive one input sound that is incorrect according to its current knowledge before it decides to (often quite severely) alter its constraint ranking. This way, in the case of free variation or even when the child makes a mistake, the constraints in the tableau change position even when they are not supposed to. To solve this, Boersma (1997) proposes to use a model that changes the constraints' ranking values *gradually*, such that constraints only get re-ranked in an absolute sense after the child has heard sufficient input data that would warrant this re-ranking. This is called the Gradual Learning Algorithm. Boersma (1997) implements this algorithm through assuming that each constraint comes with an inherent *ranking value*. With each learning step, only the ranking values of certain constraints are changed. Absolute re-ranking of constraints can still happen, but only after the ranking value of one constraint falls below the ranking value of another constraint. By doing this, constraint rankings can remain stable even in the case of free variation or errors made by the listener.[7] This advantage of the Gradual Learning Algorithm over discrete re-ranking models such as the Constraint Demotion Model is the reason that we will use the Gradual Learning Algorithm in our model of early acquisition. Also, in line with the Gradual Learning Algorithm we assume that learning does not necessarily imply strict demotion of constraints, but that it can also imply the promotion or simultaneous demotion and promotion of constraints.

This section showed how the problems faced with Tesar and Smolensky's (1998) Constraint Demotion principle can largely be solved by assuming a gradual learning model. The robustness of such a model makes it a very attractive option when modelling early phonological acquisition. In chapter 3, we will show the implementation of the assumptions into our model.

2.3.4 Unsupervised learning

Probably the biggest problem that comes with most OT accounts of phonological acquisition is that they all rely on some sort of supervised learning, i.e., they all assume that the child has access to a lexicon that functions as an error detection mechanism (e.g. Boersma

[7] To make this process work completely, one also needs to build in some kind of algorithm that ensures that the ranking values of constraints can diverge to a certain extent, even after re-ranking of these constraints. A description of how this can be achieved (namely through assuming a small amount of stochastic noise) is given in section 4.3.

& Hamann, 2008; Tesar & Smolensky, 1998; Boersma & Hayes, 2001). For example, let us assume that the child hears a speaker who happens to pronounce words such as *food* /fuːd/ with a more fronted [ʉ] (a phenomenon that can be found in various British dialects). If we assume that our child has not yet learned the word *food*, then chances are it might perceive this [ʉ] as an instantiation of /i/, resulting in the wrong conclusion that [ʉ] is the phonetic realisation of the phonological form /i/. However, in all accounts of supervised learning children are assumed to have access to an inventory of lexical entries (or in some other way to the sounds' underlying representation) to which the incoming sound could be checked. In other words, it is assumed that our child has already acquired the word *food* (including its underlying form /fuːd/) such that, should they hear the pronunciation [fʉːd], they know from their lexical knowledge that [ʉ] should be perceived as an instantiation of /u/. With this knowledge, a child can re-rank its constraints in such a way that relatively fronted pronunciations of [u] will still be perceived as tokens of the phoneme /u/. The lexicon under this view acts as a mechanism that can be used to check whether incoming sounds are mapped onto the right phonological entry.

This view on phonological acquisition may be empirically sound at times when a child has already acquired some words, but it cannot be the whole story. For one, such representations suffer from a problem commonly called the *bootstrapping problem* (Ramus, 2002). Applied to phonological acquisition, the problem can be defined as follows: the lexicon can help the child in acquiring phonological knowledge. But in order to acquire words in the first place, the child needs to have some phonological knowledge. This logically means that both language modules need each other in order to initiate any learning. It is unclear, then, how language acquisition can start in the first place. If we assume that phonological learning is at all times guided by supervised learning mechanisms, actual learning can never happen. That is why we simply need to assume that children can also acquire phonological knowledge without the help of the lexicon, at least in their earliest stages.

There is also much empirical evidence that suggests that infants already start distinguishing different phonemes before they start learning words, as was shown in the introduction. If this is the case, the infant cannot use the lexicon as a checking mechanism because the infant simply does not have a lexicon yet. Still, infants seem to develop some insight into what the phoneme inventory of their language looks like before they know any words. This again suggests that it is not necessary to have a lexicon before infants can start learning the phonemes of a language. In this thesis, I therefore want to develop a model that can explain exactly how this kind of unsupervised learning works. What such a model should look like and what kind of constraints should be used in these cases is something I will develop further in the next sections.

2.4 Detailed assumptions

In section 2.3, I considered four general assumptions I will make in our OT model that are usually not made within standard OT. To summarize, the following assumptions will from now on be made:

1. In our model, every input form will by definition be compared to a finite and reasonably small number of output candidates.

2. In our model, the set of phonological constraints an infant uses either emerges from language input, or is based on general cognitive abilities or physiological properties. These constraints are then by definition not formed from innate, language-specific knowledge.

3. In our model, learning will not imply the immediate demotion of a constraint with each learning step. Instead, learning will be represented as proceeding gradually through changing the constraint's inherent ranking value with each learning step. Learning can furthermore also be achieved through promotion or simultaneous demotion and promotion of constraints.

4. In the first stage of our model, learning will proceed without the supervision of the lexicon.

These four assumptions define the general framework of our model. However, we also need to specify the more detailed properties of the model, which is what we will do in the next sections. These more detailed assumptions are mostly concerned with the type of constraints we use in our model. More specifically, the last two sections will focus around the choice to use a group of so-called *warp constraints* instead of e.g. cue constraints. But first, in the next section we will define the general properties of the constraints used in the model and argue for the choice to use a match-driven mechanism instead of an error-driven mechanism.

2.4.1 Error-driven vs. match-driven mechanisms

As of yet, relatively little attention has been given to the issue of unsupervised learning within the OT acquisition literature. One of the few OT accounts of unsupervised learning is given in the paper by Boersma et al. (2003). In their view, unsupervised learning is achieved through the interaction between three groups of constraints, called PERCEIVE constraints, WARP constraints and CATEGORY constraints. Perceive constraints are a group of constraints of the kind PERCEIVE([340]). These constraints basically tell the infant that it should perceive an incoming acoustic value as a meaningful sound that they should process in some way. Whether or not a particular value is cognitively processed depends on the position of the corresponding perceive constraint in the tableau. Taking the example given above, when an infant hears a sound with an F_1 value of 340 Hz, then the choice whether or not to process this sound depends on the position of the constraint PERCEIVE([340]). More specifically, in a state where perceive constraints are ranked lower than the other groups of constraints, infants do not perceive incoming sounds at all. After a certain age (or maybe immediately after birth), due to certain innate abilities further left unspecified by the authors, infants start to re-rank their perceive constraints until they come to stand above the other two groups of constraints. At that point, infants start to perceive all speech sounds as meaningful units.

Category constraints are constraints that punish the mapping of a particular sound onto a particular category. For example, the constraint *CATEG /320/ tells the infant not to map any incoming sound onto the category /320 Hz/. Note here that the notion 'category' has a different meaning than how it is used in regular phonological theories: here, the categories do not refer to abstract sound representations, but rather to the (slightly more abstract) categorization of purely acoustic properties such as frequency or duration. Initially, whenever infants hear a sound with a certain acoustic value, this value is mapped onto the 'category' that is acoustically closest to this value. In Boersma et al. (2003), should the corresponding constraint be ranked higher than another, less optimal constraint, then the corresponding constraint is lowered and the less optimal constraint is raised in order to correct the mistake. As an example, consider that the infant hears a sound with the F_1 value [340 Hz]. Initially, the infant was 'told' not to perceive this sound at all. This is formalized in the fact that the constraint *CATEG /340/ is ranked above the constraint PERCEIVE([340]), as seen in tableau 6. At one point, however, the infant's innate learning device tells the infant that this is wrong and that it actually should perceive the sounds it hears. The infant reacts to this by lowering (by an unspecified amount) the ranking value of the category constraint and raising (again, by an unspecified amount) the ranking value of the perceive constraint.[8]

[340]	*Categ (/320/)	*Categ (/340/) →	Perceive ([340]) ←
/320/	*!		
✓ /340/		*!	
☞ /-/			*

Tableau 6 (adapted from Boersma et al., 2003, p. 1014)

We can see that in this sense, learning under Boersma et al.'s (2003) model is still error-driven, as infants only re-rank their constraints in case of a mismatch between what they perceive and what they (think they) should have perceived.

The last group of constraints, the so-called warp constraints, requires that every incoming sound should be mapped onto the most similar available category. For example, the constraint *Warp (40) states that any incoming sound should not be mapped onto a category that is more than 40 Hz off that sound. So when this constraint is ranked highly, it is unlikely that an infant will map an incoming sound of 320 Hz onto the category /360/. Similarly to the other types of constraints, these constraints can be re-ranked in case of an encountered error.

According to Boersma et al. (2003), the first stage of phonological acquisition does not yet involve supervised learning in that the lexicon does not yet help in the re-ranking of constraints. However, in another way their account of learning is still rather traditional in that learning (through the re-ranking of constraints) only occurs when the infant 'knows' that a certain input has been mapped onto a non-optimal candidate, making learning in this model to be error-driven. In my opinion, this way of modelling early acquisition assumes that infants have a kind of (innate?) awareness such that they know that some mappings of input values onto output candidates are more preferred than others. This is in a way rather unexpected when we also assume that infants start learning without any prior knowledge of what combinations are more or less preferred. I therefore think that, when we want to model the earliest stages of phonological acquisition, a model that does not in any way rely on an error-driven mechanism is more elegant than a model that does. That is why, in our model of phonetic categorization, learning will always be what we will call *match-driven*, i.e., we will assume that infants only change their constraint ranking when they encounter a match between input form and output candidate. How this assumption is implemented into our model is shown in chapter 3.

2.4.2 Unidimensional mapping

We have seen in the previous section that Boersma et al.'s (2003) model uses warp constraints in order to be able to map certain values on different but acoustically close phonetic categories. However, this routine was later abandoned, especially due to the fact that warp constraints cannot handle the mapping from more than one acoustic dimension onto one abstract category (Silke Hamann, personal communication). For example, it is unclear how warp constraints can map both the acoustic values 340 Hz (F_1) and 120 ms (duration) onto one phonetic category.

However, the question whether this really poses a problem for unsupervised learning relies on the assumptions one makes about this kind of learning. Warp constraints indeed

[8] Note that this does not automatically imply that the constraints are also absolutely re-ranked, as the model by Boersma et al. (2003) is a gradual one.

seem to be less useful as a tool to model the stage in which children start to map multiple acoustic cues onto one phonological category. For example, it is not unthinkable that, at a certain stage, children will use multiple cues to decide to which phoneme a certain sound should be mapped. This is in particular relevant in cases where certain phonemes partly use identical phonological feature settings. A simple example can be given for Dutch, that makes use of phonemes that only differ on the voicing dimension, e.g. /t/ and /d/. When a child then hears [d], not only does it need to make use of the acoustic cues that can help them to distinguish the sound from voiceless /t/ (such as Voice Onset Time, duration, amplitude in the closure phase, etc.), it also needs to make use of additional acoustic cues to distinguish the sound from voiced but otherwise different sounds, such as /n/, /z/, /g/, /u/ etc. Since each warp constraint can only work on one acoustic dimension, this would imply two things: one, that there are separate warp constraints for each acoustic dimension, and two, that children cannot use several cues at once in phonological mapping. Whereas the first implication does not really pose a problem (to me at least, it is not unthinkable that children could have several groups of warp constraints for each acoustic dimension they use as a cue), the second implication is more problematic: it would mean that children (and adults, for that matter) always separate each acoustic cue from an incoming sound, find the optimal candidate for them and then link these candidates together to map them onto their desired phonological category. Not only is this probably very time-consuming, it also gives us additional problems when, for example, different cues lead to different optimal candidates. In these cases, how should a child know which candidate to choose? So in sum, warp constraints become problematic as soon as children start to map incoming sounds onto actual phoneme categories.

However, the aforementioned problems do not necessarily mean that warp constraints are not useful at all. Recall that we know from empirical evidence that infants already start to form distributional categories before they start learning phonemes. In this stage of language development, we might still want to assume the existence of warp constraints based on the nature of this kind of learning. In particular, literature on visual category learning (Feldman, 2000; Ashby et al., 1999), general category learning (Love, 2002) as well as literature on auditory categorization (Goudbeek et al., 2009; Maye et al., 2008; Holt & Lotto, 2006) seems to suggest that, in early stages of acquisition, infants make use of only one acoustic cue at the time to distinguish between categories. It is only with the help of feedback (i.e. the supervised learning mechanisms mentioned earlier for auditory category learning) that infants make use of multiple cues at the same time.[9] In other words, we do not expect that pre-lexical infants treat the different acoustic cues in speech as if they are inseparable. Rather, I will follow Boersma et al. (2003) who assume that in these early stages of phonological (or rather, phonetic) acquisition, infants will treat these cues separately, i.e., that they do not make a connection between the different cues. Under this view, if an infant for example hears an /i/ with an F_1 of 300 Hz and an F_2 of 2200 Hz, they will store both cues separately without making a connection between the two, thereby making two different categories /300/ and /2200/ without storing the information that these two categories came from only one speech sound. Using this assumption, we eliminated the problem that warp constraints cannot handle mapping from multiple cues onto one category, simply because infants do not yet make the abstraction that multiple cues should be mapped onto only one category. In line with Boersma et al. (2003), this abstraction will be made later, when infants learn to map multiple acoustic cues onto single phoneme categories. At this stage of language acquisition, the warp constraints will simply become obsolete, just like the perceive constraints become obsolete as soon as

[9] But take note of Benders (2013, ch. 5), who suggests that infants do actually use multiple acoustic cues at the same time, even in early acquisition.

infants begin differentiating incoming sounds. We will come back to this when we discuss the modelling of supervised learning in chapter 4.

2.4.3 The role of physiological perception

In the previous section, I explained why the fact that warp constraints cannot map multiple acoustic cues on one phonetic category is not necessarily a problem. Another argument against using warp constraints is the one given by Van Leussen (2013), who states that the specialized role of the warp constraints makes them teleological in nature. In this I disagree with him, most importantly because the warp constraints in my view were not created in some way only to make perceptual warping possible. Indeed, as Boersma et al. (2003) do not specify why and on what grounds the infant is born with a group of warp constraints, it seems as if these warp constraints are only there because it would otherwise be impossible for the infant to create phonetic categories. This way, the warp constraints do seem to be teleological in nature. However, this problem can be solved by explicitly linking these warp constraints to the properties of our auditory perception. In this thesis I therefore propose that the warp constraints are simply the formalization of the physiological nature of our hearing and that they are in that way completely unrelated to any teleological 'goal'. How exactly the warp constraints should be related to the auditory system of human beings will be discussed below, but let us first shortly consider the alternative for the warp constraints given by Van Leussen (2013). His alternative seems to be to adopt a combination of cue constraints and structural constraints (roughly equivalent to our category constraints). Cue constraints are constraints of the type *[340]/360/ that regulate the mapping between the input form and the output candidate; an early appearance of this group of constraints is found in e.g. Escudero & Boersma (2003). In particular, these constraints militate against the mapping of e.g. the value [340] onto the phonetic category /360/. The main downside to these constraints is that in order to make them work properly, one should make a constraint for each input-output combination. For example, even if we assume that infants are only able to distinguish 20 different acoustic values on one dimension (which is a fairly conservative estimate), then we would still need to assume a number of (20x20=)400 different cue constraints. This number should then be multiplied by the number of acoustic dimensions we assume, thereby quickly reaching more than 1000 constraints. To me, such large constraint inventories would strain the child's cognitive capacities too heavily to be a likely representation of what happens in early acquisition. With warp constraints, we do not need to make input-output pairs, which means that the number of constraints can be kept much lower. It is this property of warp constraints that makes them in my opinion preferable over a model that uses cue constraints in this stage of acquisition.

Let us now return to the link between the humans' physiology and the warp constraints. To clarify this link, it is necessary to first summarize the relevant properties of the humans' auditory system.[10]

The ear is normally divided into three components: the outer ear, which consists of the pinna and the ear canal; the middle ear, which consists of the ear drum and the ossicles (ear bones) and the inner ear, which consists of the cochlea and the balancing organ. To us, the most important part of the ear is the inner ear, more specifically the cochlea. The cochlea itself is a tube rolled up into a coil with a length of approximately 32 mm. It is filled with a fluid called perilymph. The cochlea is further divided into two compartments by the basilar membrane, called the scala vestibuli and the scala tympani.

[10] The information in this section is taken from Rietveld & Van Heuven (2009, ch. 10) and Hayward (2000, ch. 5).

The helicotrema forms a passage between these two compartments. The basilar membrane itself is a membrane that more or less resembles a flipper: small and stiff near the base, but wider and more flexible near the periphery (or apex). Figure 2.1 shows the cochlea as if it were unrolled.

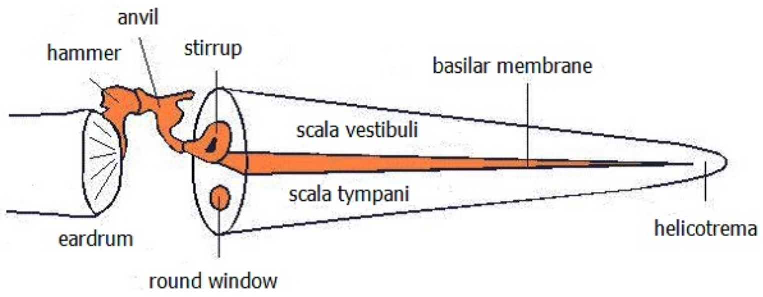

Figure 2.1: The cochlea in unrolled position (figure adapted from Rietveld & Van Heuven, 2009, p. 213).

When a sound signal reaches the cochlea, the sound waves cause the fluid within the cochlea to vibrate. This vibration of fluid consequently causes the basilar membrane to vibrate as well. The motion of the basilar membrane resembles that of a wave, with the motion moving from the base of the membrane (near the stirrup) to a set point away from the base of the membrane. The motion gradually increases until it reaches its maximum at a certain point on the membrane, after which the motion quickly decreases. The point where the intensity of the wave is strongest depends on the frequency of the incoming sound: if the sound has a low frequency, the wave of the basilar membrane will peak near the apex. If the sound has a high frequency, the peak of the wave will lie more towards the base. This is caused by the fact that the base of the basilar membrane is more stiff than the apex of the membrane, such that only high frequency sounds can really cause this part of the membrane to vibrate. The motion of the basilar membrane for sounds with different frequencies is shown in figure 2.2.

On the basilar membrane lies a large number of hair cells. Whenever parts of the basilar membrane vibrate, the hair cells on these parts react to this by firing signals to other nerve fibres. These nerve fibres will eventually lead the signals to the brain, where they are processed into actual sound perceptions. Each hair cell has a frequency value to which they are most sensitive. This means that hair cells will react most to frequency values that are equal or similar to their preferred frequency value.

So how can we relate the actions on the basilar membrane to the properties of our warp constraints? The fact that, under the proposed view, warp constraints are immediately linked to the physiological properties of the basilar membrane has important consequences for the formalisation of our learning model. In principle, the physiological make up of the basilar membrane is roughly the same for all human beings in that, at birth, the sensitivity of the hair cells is the same for all infants (or at least those who do not have any form of hearing deficiencies). Also, even though the sensitivity of the hair cells often decreases as people get older, this sensitivity usually does not decrease significantly within children. In other words, the sensitivity of the hearing organs does not change significantly during the time that children acquire their native language. This also means that the ability to perceive a certain sound value as a similar but different sound value is constrained by the physiological properties of the basilar membrane and its hair cells and that these constraints cannot change during the process of learning a language. These properties of

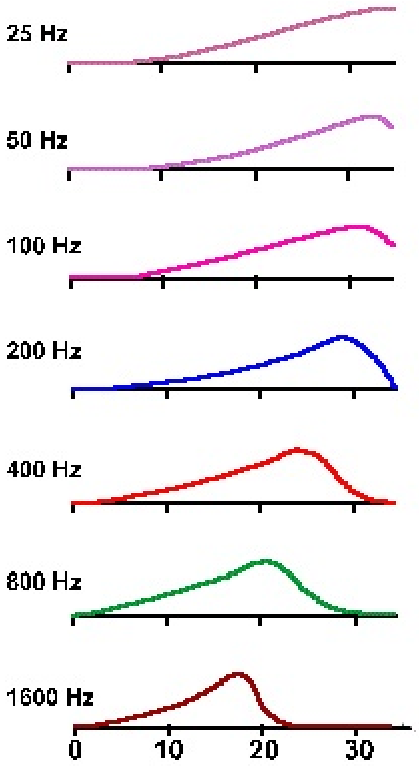

Figure 2.2: The motion of the basilar membrane's waves at different sound frequencies (figure taken from http://michaeldmann.net/mann8.html). The numbers on the horizontal axis indicate the distance in mm from the base of the membrane.

hearing are now formalized in the fact that, within our model, the ranking values of our warp constraints are fixed. Consequently, this means that the warp constraints are unable to move according to input, a property that our category constraints do actually have. We explicitly depart here from the model used by Boersma et al. (2003), who assume that warp constraints can actually move around the tableau.

Another assumption about the warp constraints that springs from the rather direct connection we make between these group of constraints and the properties of the basilar membrane is the assumption that the number of warp constraints is not only finite, but in fact rather small. Recall that each frequency value of a sound is linked to the position on the basilar membrane where the wave motion reaches its peak. However, to arrive at this point the sound has to 'travel' along the basilar membrane, with the movement of the membrane gradually increasing until it reaches its peak. This means that not only the hair cells at peak position are stimulated to fire, but also those hair cells that lie between the base of the basilar membrane and the position where the motion is largest, as well as the hair cells that lie a short distance after the peak. It is through exactly this property of the basilar membrane that the frequency of a sound can also be perceived as another, similar frequency value. At the same time, it is also important to realize that the firing of a hair cell is always a matter of all-or-nothing; hair cells cannot fire with different strengths. The only relevant property is the so-called *firing rate*, i.e., the number of fires of a hair cell in a fixed period of time. Hair cells occasionally fire when there is no stimulation of the basilar membrane; this is also called the *spontaneous rate* of a hair cell. The firing rate increases when the hair cell is stimulated by movement of the basilar membrane.

In this thesis I will make the assumption that the firing rate of a hair cell needs to reach a certain threshold value before it can be perceived by the brain. This can perhaps be explained better by using an example. Consider that the child hears a frequency value of 1000 Hz. Obviously, the firing rate will be highest for the hair cells whose preferred frequency is 1000 Hz, but the surrounding hair cells will also have an increased firing rate. These firing rates will be lower than the firing rate of the hair cells with the preferred value of 1000, and the further away the hair cells are from the 'optimal' hair cells, the more their firing rate will decrease. If we arbitrarily assume that the firing rate decreases with 50% for hair cells whose preferred frequencies differ with 50 Hz (in this case, the hair cells whose preferred values are 950 and 1050 Hz), then the firing rates of those hair cells with preferred values of 800 and 1200 Hz is only 6.25% of the firing rate of the optimal hair cells. If we then assume that a certain frequency value can only be perceived if the firing rate of the corresponding hair cell is at least 10% of the firing rate of the optimal hair cells, then we can say that frequency values of 1000 Hz can only be perceived by the brain as a frequency value between approximately 840 Hz and 1160 Hz. If we say that this is a property of the way human hearing works, then that would make it simply impossible for humans to perceive sounds of a certain frequency as a sound of another frequency if that frequency is too far away from the actual frequency.

Even though this assumption does not seem to be empirically attested for in the phonetic literature, we may still find indirect evidence for this mechanism in findings on masking effects. In general, we find that when we hear two sounds of a similar frequency, these frequencies can influence how well we perceive the separate sounds. In particular, the effect of masking is that it becomes more difficult to perceive these separate sounds. The reason for this is that when two sounds with similar frequency values are heard simultaneously, the areas on the basilar membrane that are influenced by these sounds overlap. The critical band defines the area within which two sounds mutually influence the perception of both sounds. Since masking effects are caused by overlap on the basilar membrane, defining the general width of the critical band also indirectly defines the area on the basilar membrane for which the hair cells meet the minimum firing rate value, i.e., it defines the width of frequency values that are perceived as actual sounds by the brain. Rietveld & Van Heuven (2009, p. 224) define the width of the critical band at 90 Hz for frequencies below 650 Hz and at approximately 15% for frequencies above 650 Hz. We can now use these values to limit the number of warp constraints we use in our simulations: if we for example assume that the Just Noticeable Difference lies at 20 Hz[11] and that the width of the critical band cannot be larger than approximately 360 Hz (assuming that the maximum F_2 frequency lies at around 2400 Hz), then for the first two formant dimensions we would in principle never need to have more than 18 warp constraints. Of course, the number of constraints may often be lower, as e.g. F_1 values rarely exceed 900 Hz.

In this section, we have seen that the properties of our hearing have important consequences for the properties of the warp constraints in our model. Furthermore, this chapter in general defined and discussed the various assumptions we make for our model of early acquisition. In the next chapter, I will show how these assumptions are implemented in an actual learning model.

[11] In reality, the value of the Just Noticeable Difference depends on the frequency of the sound. In particular, JND values seem to lie around 14.5 Hz for frequency values below 800 Hz and around 1.5% of the perceived frequency value for values above 800 Hz (Kewley-Port & Watson, 1994). This also shows that our auditory perception is scaled logarithmically rather than linearly at least for values in the higher frequency regions. This property of auditory perception can be represented by using frequency scales that work with logarithmic units, such as the ERB scale and the Bark scale.

Chapter 3

Modelling unsupervised learning

3.1 Introduction

In this chapter I will define the exact characteristics of the learning model. I will do this with the help of a language example. To make this example as clear as possible without losing out on validity, we will use a constructed ‘fantasy language’, of which the properties will now be outlined.

For demonstration purposes, let us see how the model can predict the way in which children learn to distinguish different vowels in a language. Our fantasy language makes use of two different vowels, namely /ɔ/ and /ɛ/. In this language, the typical F_1 value for /ɔ/ lies at around 520 Hz, while the typical F_2 value for this phoneme lies at around 800 Hz. For /ɛ/, typical F_1 and F_2 values lie at around 600 Hz and 1800 Hz, respectively.[12] Of course, a situation like the one mentioned for our fantasy language is probably unlikely to occur in a real language: first, most languages have more than two vowels in their phoneme inventory – Maddieson (2011) finds only four languages with a two-vowel inventory in his sample of 564 languages – and second, it is shown that languages tend to make use of phoneme inventories where all phonemes show maximal auditory dispersion (Flemming, 1995; Boersma & Hamann, 2008). Therefore, it would be unlikely that in a situation where a language uses only two vowels, these vowels would have F_1 values that are so close together. But since the situation sketched for our pseudo-language is best at demonstrating all aspects of our learning model without becoming too complex, let us for now assume that this situation could in principle be found in (an admittedly idealized) language.

3.2 Learning through promotion

In the first stage of phonological acquisition, an infant learns to distinguish the various phonetic categories of a language. In our example, let us for now focus only on the acquisition of the F_1 category of the phoneme /ɔ/. In our view, at the beginning of this stage an infant does not yet know anything about the phonological properties of its language. The only thing we may say it has learned by now is that the sounds it hears are meaningful and that it ‘should do something’ with those sounds. In our model, the fact that infants now start to process the incoming sounds they hear is formalized in that the perceive constraints introduced by Boersma et al. (2003) are now all ranked above all other constraints, such that they do not play a role any more in the learning process.[13]

[12] The frequency values are roughly based on the values found in the vowel chart constructed by Bruce Hayes, retrieved at http://www.linguistics.ucla.edu/people/hayes/103/Charts/VChart/FormantPlot.jpg

[13] Because of this fact that the perceive constraints are no longer useful in this stage of acquisition, in all further tableaux these constraints will be left out.

Now that the perceive constraints have become obsolete, learning at this stage can be defined as the interaction between category constraints and warp constraints. Recall from before our assumption that all warp constraints take a fixed position within the tableau. Additionally, the least severe of these warp constraints (i.e., the one that can only warp acoustic values on those values that are acoustically close to the actual value), is ranked lowest, and the most severe of the warp constraints (the one that can warp an acoustic value on another value that is acoustically very far apart) is ranked highest. Furthermore, at this initial stage all category constraints are ranked at the same height below the warp constraints. This state-of-affairs is formalized in tableau 7. Our learning model is different from the model used by Boersma et al. (2003): whereas they represent learning as the simultaneous promotion and demotion of various constraints, in this model learning is represented as the promotion of non-matching category constraints. Furthermore, because we assume with Boersma et al. (2003) that learning in our model is gradual, all constraints are also given an inherent initial ranking value (in the tableaux, these values can be found immediately above the constraints). For the category constraints, this value is randomly set at 100 and for the warp constraints, these value are randomly set at 200, 300 and 400. The list of output candidates is in principle made up of all perceivable frequency values (i.e., roughly all frequency values between 20 and 20.000, with a Just Noticeable Difference of 20 Hz taken as the distance between two adjoining output candidates), but for practical reasons, the tableaux in this thesis will only show those output candidates that are relevant in each specific situation.

	400	300	200	100	100	100	100
	*W(60)	*W(40)	*W(20)	*/500/	*/520/	*/540/	*/560/
/500/							
/520/							
/540/							
/560/							

Tableau 7

Now, suppose the infant hears a speech sound with a frequency of 520 Hz. Since all category constraints are still ranked below the warp constraints, the infant will automatically map the incoming sound onto the category /520/, as is shown in tableau 8. Also, learning now starts in that the value of the category constraints of all non-matching candidates (in this case */500/, */540/ and */560/) is increased by 1 (again, one might also choose to give a different value to each learning step, depending on how fast or slow you want learning to be. For example, setting the learning step at 0.1 means that the infant needs to hear more input before it can re-rank its constraints). The result of this is that all category constraints except for */520/ now have a ranking value of 101. The ranking value of */520/ stays at 100.

	400	300	200	← 101	100	← 101	← 101
[520]	*W(60)	*W(40)	*W(20)	*/500/	*/520/	*/540/	*/560/
/500/			*!	*			
☞ /520/					*		
/540/			*!			*	
/560/		*!					*

Tableau 8

Since all category constraints were initially ranked at an equal height, the first learning step causes that the category constraints */500/, */540/ and */560/ are now also absolutely ranked above */520/. This is also shown in tableau 9, where the dashed line between */520/ and the other category constraints changed into a straight line.

	400	300	200	101	101	101	100
	*W(60)	*W(40)	*W(20)	*/500/	*/540/	*/560/	*/520/
/500/							
/520/							
/540/							
/560/							

Tableau 9

This process of promoting all non-matching category constraints continues whenever an infant hears a speech sound. Suppose now that the infant has heard a lot of sounds with a frequency of 520 Hz. In fact, after the infant has heard 101 instances of [520], what happened is that the ranking values of the constraints */500/, */540/ and */560/ are now higher than the ranking value of the constraint *Warp(20). As a consequence, at this point these category constraints will have come to stand above the *Warp(20) constraint. Then, when an infant suddenly hears a sound with a frequency of 500 Hz, what now happens is that it will categorize this sound as 520 Hz. This process is also formalized in tableau 10.

	400	300	201	201	201	200	100
[500]	*W(60)	*W(40)	*/500/	*/540/	*/560/	*W(20)	*/520/
/500/			*!				
☞ /520/						*	*
/540/		*!		*			
/560/	*!				*		

Tableau 10

But note that when our infant hears a sound of 560 Hz, it will still categorize it as 560 Hz and not as 520 Hz. This is because the category constraint */560/ is not yet positioned above the *Warp(40) constraint. This is also demonstrated in tableau 11.[14]

	400	300	201	201	201	200	100
[560]	*W(60)	*W(40)	*/500/	*/540/	*/560/	*W(20)	*/520/
/500/	*!		*				
/520/		*!					*
/540/				*		*!	
☞ /560/					*		

Tableau 11

In cases where an input form is mapped onto an output candidate with a different acoustic value, I propose that none of the ranking values is changed and that, consequently, there is no learning step. Why we choose to do this is explained in section 3.3.

This way of learning continues throughout the first stage of phonological acquisition: whenever the infant hears a sound, if possible this sound is mapped on the output candidate with an identical acoustic value and the constraints with different values are promoted. If the sound is mapped onto an output candidate with a different acoustic value, no learning takes place. The category constraints stop moving once they have become promoted above the highest ranked warp constraint for that acoustic dimension (correlating with

[14] In this tableau, please note that the constraints */540/ and */560/ are ranked equally high. In this case, the infant makes the decision to map [560] on /560/ based on the fact that the candidate /540/ violates an additional lower-ranked constraint, namely *Warp(20).

the threshold value of the hair cells' firing rate).[15] This fact also ensures that there is a point at which learning stops, namely at the moment that all category constraints are either ranked above the highest-ranked warp constraint or below the lowest-ranked warp constraint. This is different from the model by Boersma et al. (2003), where learning does not stop automatically (Paul Boersma, personal communication). Eventually, this learning mechanism causes a number of category constraints to become ranked above one or more of the warp constraints, while some other category constraints are still ranked below any of the warp constraints. Then, at the end of the first stage, the infant evaluates the place of all category constraints in the tableau. More specifically, the infant looks at those category constraints that are not ranked above any of the warp constraints; these form the basis of the phonetic categories found in the infant's language. Then, it looks at the position of the other category constraints in the tableau. In principle, many of the input values will have already been mapped onto other output values than the identical value. In the end, this tendency will be secured by the infant in that it will explicitly consider that certain acoustic values are actually tokens of a phonetic category of another value. In particular, all values that correspond to a category constraint that *at least* stands above the warp constraint that would warp these values onto a category constraint that is still below any of the warp constraints is from then on taken to be part of that phonetic category. For example, if at the end of the first stage the constraint */560/ is either ranked above the constraint *Warp(40) or above one of the even more severe warp constraints while the constraint */520/ is still below all warp constraint, then the input value [560] will from then on always be considered to be a token of the phonetic category /520/.

In cases where certain values could in principle be part of more than one phonetic category, the category that has an acoustic value closest to the value of the input sound is taken to represent that input value from then on. For example, suppose that both the constraints */520/ and */600/ are eventually positioned below all warp constraints (implying the existence of two phonetic categories /520/ and /600/). If we also have a constraint */540/ that is positioned above *Warp(60) (which means that this value could in principle be an instance of both phonetic categories), then the choice is made to consider values of [540] to be a token of the acoustically closer phonetic category /520/, as is predicted by the ranking of the constraints, as we can see in tableau 12.

	423	400	300	200	140	128
[540]	*/540/	*W(60)	*W(40)	*W(20)	*/520/	*/600/
☞ /520/				*	*	
/540/	*!					
/600/		*!				*

Tableau 12

In the unfortunate case where an acoustic value falls precisely between two phonetic categories (e.g. the value [560] when we assume the existence of the phonetic categories /520/ and /600/), then the choice to which category this value should be mapped will be based on which of the two constraints correlating to the phonetic categories has the lowest ranking value. This is shown in tableau 13.

[15] The learning mechanism might even be more sophisticated in that, for every input value, the number of warp constraints that is chosen corresponds with the threshold value of that particular input. For example, this would mean that when an infant hears an F_1 value of 200 Hz, fewer warp constraints would be present in the tableau than when an infant hears an F_1 value of 700 Hz. However, since such a model would be next to impossible to virtually simulate, in this thesis I will assume that the number of warp constraints is fixed for every acoustic dimension.

	478	400	300	200	140	128
[560]	*/560/	*W(60)	*W(40)	*W(20)	*/520/	*/600/
/520/			*		*!	
/560/	*!					
☞ /600/			*			*

Tableau 13

Because we predict that the ranking values of two phonetic categories are often relatively close together, it is quite probable that the infant initially chooses to map the value on an incorrect category (in this case, that it maps instances of [560] on the category /600/, even though [560] should be part of the phonetic category /520/). In these cases, we assume that in the next stage the child will correct this mistake using the extra information given by its acquired lexical knowledge.

Another problem can arise when a particular value cannot be taken to be a member of any of the phonetic categories, because the corresponding category constraint is not yet ranked high enough. An example of such a situation is given in tableau 14.

	500	489	419	400	300	200	140
[440]	*W(80)	*/440/	*/500/	*W(60)	*W(40)	*W(20)	*/520/
/440/		*!					
☞ /500/			*	*			
/520/	*!						*

Tableau 14

In tableau 14, note that even though the constraint */500/ is ranked higher than *W(60), the ranking value of this constraint is lower than the ranking value of */440/. This is actually expected; the fact that the value [500] is acoustically closer to the value of the phonetic category /520/ than the value [440] implies that the infant will hear [500] more often than [440]. Using the mechanisms of the model, this would also imply that when an infant cannot map a certain acoustic value onto an existing phonetic category, it will map the value onto the most similar acoustic value possible. So even when the infant would not be able to map [440] on /500/ (e.g. because the constraint */500/ was ranked below *W(60)), then the infant would still map this value onto /480/ (provided that the constraint */480/ was ranked above *W(40)). Additionally, the infant will then create extra phonetic categories for all values that cannot be mapped onto any of the 'real' phonetic categories. Of course, this means that the child has more phonetic categories on one dimension than there are actually present in its language. In the next stage of acquisition, then, we assume that the child will use its lexical knowledge to find out to which of the real phonetic categories these pseudo-categories (i.e., the categories to which only one or very few acoustic values are mapped) should belong. How this is done exactly will be further specified in section 4.2.

The model described in this section shows us one possible way in which early acquisition can be explained. Another model that involves demotion of the matching constraints is given in section 3.5. The next two sections further explore what should happen in cases of mismatches between the input form and the chosen output candidate, and give a short example of how more realistic learning can be represented.

3.3 Learning in the case of mismatches

An important question to ask is: what happens when a certain acoustic value is mapped onto the category of a different acoustic value? In these cases, does the infant then still learn by promoting certain constraints? In my view, this does not happen, and it does not do so for a very important reason. Consider the possibility that, even when values are mapped onto the 'wrong' candidates (even though errors are not really possible in this model), an infant will still promote all category constraints that are linked to the non-matching candidates. This would mean that, once a category constraint has come to stand above a certain warp constraint while the category constraint that differs from the promoted candidate by exactly the value that is given by the mentioned warp constraint, the promoted category constraint will from then on always be promoted, even when the infant hears input values that are identical to the value of the promoted constraint. To make this situation a little less abstract, let us take another look at tableau 10.

	400	300	201	201	201	200	100
[500]	*W(60)	*W(40)	*/500/	*/540/	*/560/	*W(20)	*/520/
/500/			*!				
☞ /520/						*	*
/540/		*!		*			
/560/	*!				*		

Tableau 10

In this tableau, the category constraint */500/ is ranked above the *Warp(20) constraint, while the category constraint */520/ is below this warp constraint. Would we still choose to promote category constraints even when the chosen output candidate differs from the perceived sound value, then the constraint */500/ would always be promoted when the infant hears a sound with a value of 500 Hz. This leads to a situation where */500/ will always get promoted, regardless of the input the infant receives. Since this is obviously an undesirable state, I will refrain from assuming that any promotion will occur in cases where incoming sounds with a particular acoustic value will be mapped onto an output candidate with a different value. Instead, I will assume that in these cases none of the category constraints will be promoted, such that there is no learning step for the infant.[16]

3.4 Realistic learning: promotion

In principle, we could assume that the infant practically never hears any speech sound with another F_1 value than 520 Hz, such that in the end all category constraints except for */520/ will be ranked above the highest warp constraint. This would mean that, eventually, any sound would be mapped onto the phonetic category /520/. However, this is quite an idealization in that in reality, speakers do not simply pronounce a vowel with exactly the same formant frequencies every single time. A more realistic approach might be that the frequency values of a particular phoneme are taken from a more or less Normally distributed pool of values, such that the further away a particular value is from the prototypical value, the less likely it is to be pronounced. In our example, our prototypical (or mean) value is obviously 520 Hz. Now, let us assume that the standard deviation in this distribution lies at 20 Hz and that any differences below 20 Hz will not be perceived as a different sound. Following typical values in the case of a Normal distribution, we can then say that there is a chance of approximately 68% that a speaker will produce a sound that the infant will perceive as 520 Hz. Since both 540 Hz and 500

[16] One might also consider the possibility that, in the case of mismatches, the category constraint that is linked to the actual input value will be demoted. Another possibility in these cases could be that the constraint that is linked to the winning candidate will be promoted. To me, this might theoretically be possible, but since the implementation of these kinds of mechanisms would greatly increase the complexity of the model, I will not further investigate these options.

Hz are 1 standard deviation away from 520 Hz, the chance that a speaker will produce a sound that the infant will perceive as either 540 Hz or 500 Hz is approximately 27%. For 560 Hz and 480 Hz, lying 2 standard deviations away from the mean, the percentage is approximately 5% (the occurrence of F_1 values that are more 3 or more standard deviations from the mean is so low that we can consider their occurrences to be negligible). Now, let us assume that an infant hears 100 tokens of the phoneme /ɔ/. With the assumption that the F_1 frequency values are Normally distributed, we predict that 68 of these tokens will have the input form [520], 13.5 the input form [500], 13.5 the input form [540], 2.5 the input form [480] and 2.5 the input form [560]. We can simulate this situation by creating a list that contains these exact numbers of tokens and then randomizing the order of these tokens using an on-line list randomizer.[17] The randomized list used for the simulation can be found in Appendix A.

Tableau 15 shows the initial ranking of the constraints, before the infant has heard any input values.

	400	300	200	100	100	100	100	100
	*W(60)	*W(40)	*W(20)	*/480/	*/500/	*/520/	*/540/	*/560/
/480/								
/500/								
/520/								
/540/								
/560/								

Tableau 15

The value of each learning step is set at 10. Now, according to the list our virtual infant first hears an input value of [500]. According to our model, the infant processes this value by increasing the ranking value of all category constraints except for */500/ to 110. Then, the infant hears 9 tokens of [520], followed by one token of [540], followed by two more values of [520]. After promoting all non-matching category constraints according to input, tableau 16 shows the constraint ranking after the infant has heard 13 input tokens.

	400	300	230	230	220	220	200	120
	*W(60)	*W(40)	*/480/	*/560/	*/500/	*/540/	*W(20)	*/520/
/480/								
/500/								
/520/								
/540/								
/560/								

Tableau 16

We can see that all category constraints except for */520/ are now ranked above *W(20). Since the values [500] and [480] are only 20 Hz away from /520/, from now on whenever the infant hears one of these two values, they will be mapped onto the phonetic category /520/. In accordance to what was said in section 3.3, in these cases there will be no learning step. Therefore, the next value [540] the infant hears will cause no change in any of the ranking values.

Tableau 17 shows the constraint ranking after 36 perceived input tokens.

[17] As we can see, some of the values have a non-integer amount of tokens. To solve this, we will assume that one of the values is heard one time more often than the other value (for example, that there are 13 tokens of [500] and 14 tokens of [540]). Which value is assigned which amount of tokens is chosen at random.

	420	410	410	410	400	300	200	130
	*/560/	*/480/	*/500/	*/540/	*W(60)	*W(40)	*W(20)	*/520/
/480/								
/500/								
/520/								
/540/								
/560/								

Tableau 17

At this point, all category constraints except for */520/ have already come to stand above the highest-ranked warp constraint. As we assumed that learning stops after all category constraints are ranked either higher than the highest-ranked warp constraint or lower than the lowest-ranked warp constraint, we can say that learning is now finished. All further incoming sounds are now automatically mapped onto the phonetic category /520/, as is shown in tableau 18.

	420	410	410	410	400	300	200	130
[500]	*/560/	*/480/	*/500/	*/540/	*W(60)	*W(40)	*W(20)	*/520/
/480/		*!					*	
/500/			*!					
☞ /520/							*	*
/540/				*!		*		
/560/	*!				*			

Tableau 18

This section thus showed that, even with more realistic input distributions, our promotion-based model still yields the correct outcomes.

3.5 Learning through demotion

In section 3.2, I described a model that defines learning as the promotion of non-matching category constraints. However, another possibility might be to model learning the other way around, namely as the demotion of the matching category constraint. In this section, I will explore this possibility.

The great advantage of a demotion model when compared to a promotion model is that the learning steps of a demotion model feel more intuitive and straightforward than those of a promotion model. When modelling early phonological acquisition, the preferred model is a model that yields the correct outcomes with the least complicated learning mechanism. A demotion model clearly has a less complicated mechanism than a promotion model: whereas in a demotion model each learning step simply involves the demotion of the category constraint corresponding to the heard value, in a promotion model each learning step involves the promotion of *all* category constraints that do not correspond to the heard value. Obviously, the latter step involves more cognitive effort than the former step, which is why a demotion model in general would be preferred over a promotion model.

We already stated that we only want to use models that are explicitly match-driven. However, at first glance it looks as if match-driven mechanisms always produce incorrect outcomes within a demotion-based model. To demonstrate this, let us consider the OT tableau of a mechanism that always demotes the category constraint that corresponds to the correct output candidate. In the initial state of this set up, we will need to rank our category constraints at the same height above the warp constraints. This is shown in tableau 19, again using random values for all constraints.

	500	500	500	500	400	300	200
	*/500/	*/520/	*/540/	*/560/	*W(60)	*W(40)	*W(20)
/500/							
/520/							
/540/							
/560/							

Tableau 19

Now, let us assume again that an infant hears a sound with an F_1 frequency of 520 Hz. Hearing this, the infant will proceed to lower the ranking value of the category constraint */520/. The consequence of this is that this constraint will immediately fall below the other category constraints (as it now has a ranking value of 499 whereas all other category constraints still have a ranking value of 500). Then, consider the possibility that an infant now hears an F_1 frequency of 540 Hz (as shown in tableau 20).

	500	500	500	499	400	300	200
[540]	*/500/	*/540/	*/560/	*/520/	*W(60)	*W(40)	*W(20)
/500/	*!					*	
☞ /520/				*			*
✓ /540/		*!					
/560/			*!				*

Tableau 20

In the current state of the tableau, all category constraints are ranked above the category constraint */520/. Therefore, what happens in this situation is that it will always be worse to map an incoming F_1 value onto any other category than the category /520/. What this in other words leads to is that the infant will always map sounds with a frequency of 540 Hz, or sounds with any other frequency for that matter, onto the category /520/. This shows that once one category constraint has been demoted, it becomes logically impossible to map new incoming sounds on anything other than the first demoted category. In our example, the first demoted constraint by coincidence belonged to the category with the correct category value /520/. But in cases where the infant happens to hear another value before [520], this would lead to the creation of an incorrect phonetic category in the end. Furthermore, in this situation an infant can always only create one phonetic category for each acoustic dimension. It thus seems that a match-driven demotion-based model is way too error-prone, as the infant would base the value of its phonetic category only on the first sound it ever heard. We might however still be able to create a demotion-based model by changing the behaviour of the category constraints.

One of the most important notions within OT is the notion of *constraint interaction*, i.e., the idea that which output candidate is chosen depends on the relative ranking of the constraints. It seems that exactly this assumption causes our demotion-based model to currently predict incorrect outcomes. But since we also stated that the ranking position of the warp constraints correlates with the properties of the basilar membrane, it would be odd to suddenly change the behaviour of the warp constraints. However, obviously it is also strange to assume that just one token of an acoustic value is so important that it immediately pulls all other acoustic values towards it. To solve this, I propose the option that each output candidate is no longer compared to any of the category constraints, but only to the warp constraints. Tableau 21 further illustrates this mechanism.

	499	480	477	432	400	300	200
[560]	*/560/	*/500/	*/540/	*/520/	*W(60)	*W(40)	*W(20)
/500/					*!		
/520/						*!	
/540/							*!
☞ /560/							

Tableau 21

In tableau 21, we assume that the infant has already heard a number of input values, of which [520] was heard most frequently. In traditional OT models, the fact that */520/ is ranked lower than all other category constraints would create a situation where all further incoming values are always mapped onto the output candidate /520/. But note in this tableau that the choice of output candidate is not influenced by the position of the category constraints, such that there are no violation marks in any of the grids belonging to the category constraints. The only constraints that the output candidates are influenced by are the warp constraints. Since warp constraints by definition punish the mapping of an input value onto an output candidate that is not acoustically identical to the input value, in this model all input values are by definition mapped onto the output candidate that has the same acoustic value. Here, this means that even though [560] is heard least frequently, a token of [560] is still mapped onto the output candidate /560/. The learning steps then stay the same as in other demotion-based models in that the ranking value of */560/ is now lowered from 499 to 498.

When an infant continues this process, eventually it will arrive at the situation in tableau 22. Here, the fact that the infant hears the F_1 value [520] causes the ranking value of the constraint */520/ to fall below the ranking value of the constraint *W(60), resulting in the demotion of */520/ below *W(60). Formally, this does not change anything; at this time, all input values are still automatically mapped onto the closest output candidate because of the fact that there is no interaction between the category constraints.

	498	475	474	→ 399	← 400	300	200
[520]	*/560/	*/500/	*/540/	*/520/	*W(60)	*W(40)	*W(20)
/500/							*!
☞ /520/							
/540/							*!
/560/						*!	

Tableau 22

The process of automatically mapping all incoming acoustic values on the closest output candidate pretty much dictates this whole first stage of phonological acquisition. Every time an infant perceives an input value, the ranking value of the category constraint corresponding with the chosen output candidate is lowered, and this process does not stop until the end of the first stage. Similarly to Boersma et al. (2003), in this model the tableaux do not reach a stable state, and as such learning does not stop in this model. Rather, at the end of this first stage (which is presumably reached automatically at the moment an infant starts to acquire lexical knowledge) the infant evaluates the final state of the tableau to see which acoustic values should be mapped onto which acoustic categories. More specifically, at this point the infant considers which of the category constraints are ranked below the lowest-ranked warp constraint. These values are then considered to be the 'core members' of a phonetic category to which all surrounding values are from then on mapped. This does not mean that all values an infant ever heard are retrospectively re-labelled,[18] but rather that the infant uses this new knowledge about the phonetic categories of its language when it constructs the new constraints that are needed in the next stage of phonological acquisition (this process will be explained in detail in chapter 4). For now, we assume that the infant gains the passive knowledge that, from that moment on, any incoming sound should be considered to be an instance of the acoustically closest phonetic category.

To shortly exemplify this final stage, consider again our fantasy language. We already stated that this language's vowel inventory consists of two vowels with typical F_1 values of 520 Hz and 600 Hz. When an infant successfully reached the final state of the first stage of phonological acquisition, we would expect that in the end both of the category constraints */520/ and */600/ will be positioned below *Warp(20). Using this information, the infant then infers that there are two phonetic F_1 categories /520/ and /600/ and that all other F_1 values are instances of one of these two categories. We additionally assume that the infant will always categorize a value as the category that is acoustically closest. When an input value is equally close to two phonetic categories, then the same assumption is made as in the promotion-based model; namely, that the value is then added to the category with the lowest ranking value and that the infant in the next stage decides whether the value really belongs to that category. In section 3.6 we will apply the model's properties on a more realistic distribution.

Using a demotion-based model where the category constraints do not influence which output candidate is chosen, unsupervised learning in its earliest stages is basically reduced to counting the occurrences of each input value. Some might say, then, that this model really is not an OT-based model any more, but rather a 'counting model' wearing an OT jacket. We might consider this to be a disadvantage of demotion-based models. We will further discuss the advantages and disadvantages of both promotion- and demotion-based models in section 5.2.

3.6 Realistic learning: demotion

The application of a more realistic input situation first described in section 3.4 can also be used for the demotion-based model. That way, we can more easily see how the model works with more and more realistic input data.

For the mini-simulation, let us assume the same input values and distributional properties of the values that were defined in section 3.4. This would mean that, again taking 100 tokens from the distribution, we feed the model with 68 tokens of [520], 13.5 tokens of both [500] and [540] and 2.5 tokens of both [480] and [560]. The (randomized) order in which we feed the virtual infant with the input tokens is the same as the order used in section 3.4. This order can be found in Appendix A. We set the initial ranking value of the category constraints at 500 and those of the warp constraints at 400, 300 and 200, respectively. Doing this, we begin at the initial state given in tableau 23.

[18] Assuming that an infant re-labels all values after it gets new information would imply that the infant stores every value it ever hears in an exemplar theory-like style, as e.g. happens in Pierrehumbert (2003). In general, I think that storing every input token is unnecessary and only causes heavy memory capacity issues.

	500	500	500	500	500	400	300	200
	*/480/	*/500/	*/520/	*/540/	*/560/	*W(60)	*W(40)	*W(20)
/480/								
/500/								
/520/								
/540/								
/560/								

Tableau 23

With each learning step, the ranking value of the category constraint corresponding to the input value is lowered by 10. Since learning in the demotion-based model does not stop until the end of the first stage, we will carry out the simulation by only looking at the constraint ranking after the infant has heard all 100 input values. Tableau 24 shows the final ranking of the model.

	480	470	400	370	360	300	200	-180
	*/480/	*/560/	*W(60)	*/540/	*/500/	*W(40)	*W(20)	*/520/
/480/								
/500/								
/520/								
/540/								
/560/								

Tableau 24

We can see that the model nicely did its job, as we arrive at a state where only the category constraint corresponding to our future phonetic category /520/ has fallen below the last warp constraint. However, we can also see that the constraints */500/ and */540/ have fallen below the first warp constraint. This immediately shows an important complicating factor of this model: as learning does not stop, it becomes possible that too many constraints fall below the last warp constraint, thereby creating too many phonetic categories. In our example, this would already happen after the infant hears approximately 300 more input tokens. In other words, in this model the first stage of acquisition should finish in time. If it goes on for too long, eventually the infant will always end up with too many phonetic categories. It is not clear to me how this could be resolved. In section 5.2, we will further compare the two models.

In this chapter I described two models of early acquisition. Both models showed that, in principle, it is possible to represent early acquisition as the process of unsupervised, match-driven learning that leads to the emergence of discrete phonetic categories. In the next chapter, I will describe how learning proceeds from here: in particular, I will show how to go from discrete phonetic categories to abstract phonological features and eventually to a number of different phonemes.

Chapter 4

Modelling supervised learning

4.1 Introduction

By now, the infant has arrived at a stage where it has learned different discrete phonetic categories solely based on raw input data. Now, at a certain point, the child begins to acquire lexical knowledge. At this point, the child is ready to proceed to the next stage of phonological acquisition. In this stage I will follow Boersma et al. (2003), who propose that the child will start to map acoustic values onto phonological features, such as [mid] and [back]. How this is done exactly and how the acquisition of phonetic categories can help with this will be described in section 4.2.

A child not only needs to learn to map acoustic input onto phonological features; it needs to learn how to map values from different acoustic dimensions onto one phoneme category as well. Recall from section 2.4.2 that, for the first stage of phonological acquisition, we assumed that infants map sounds onto phonetic categories on a one-to-one basis, i.e., that infants do not yet map several cues onto one single category. But once a child has learned the inventory of phonological features a language uses and which acoustic values should be mapped onto which of these features, it should also learn how different features can be combined into representing one phoneme. In section 4.3, I will propose an OT-based formalization that might explain the processes involved in this kind of phonological learning.

4.2 From phonetic categories to phonological features

In chapter 3, I described two models that both might account for the way children acquire phonetic categories. Regardless of which of these two models is the best predictor, both models need to be altered in order to accommodate to the new stage of phonological acquisition. This process will roughly be the same for both models and will be further elaborated on in this section.

Let us first look at the initial state of the model in this phase. By now, the infant should have arrived at a point where it has a number of phonetic categories onto which all incoming sounds are mapped. These phonetic categories an infant has at this stage are, in a way, rather meaningless. What I mean by this is that the categories are just composed of some values that are only there because the infant heard these values most often in the earlier stage. To make the categories more meaningful, infants should at this stage begin to map the perceived values onto other entities than just phonetic categories. The first question then is: what are the properties of these new entities?

In this thesis, in line with Boersma et al. (2003) I will assume that the infant does not immediately go from phonetic categories to phoneme categories, but first uses an intermediate step where it goes from phonetic categories to phonological features.[19] There are some empirical tendencies that might argue for the existence of phonological features as real mental constructs. For example, Boersma & Chládková (2011) show that adult speakers seem to make use of the information given by separate phonological features in the perception of vowels. Furthermore, the phenomenon of *canonical babbling* might also argue in favour of this existence: empirical studies (e.g. Oller, 1980; Vihman et al., 1985) have shown that there is clear continuity between the sound forms in babbling and the phonetic forms of early speech. This way, we might assume that the function of canonical babbling is that the infant practices the sounds it perceives from other speakers. One result of this practice then possibly is that the infant learns to link the perceived sounds to actual articulatory positions and movements. For example, by uttering [ba] the child might learn that the sound [a] (with its corresponding acoustic values) can be articulated by opening the mouth quite far and not doing anything with the tongue, i.e., letting the tongue rest in middle position. Assuming this, we then also state that one function of canonical babbling is that the child through practice learns to distinguish the different phonological features that appear in its native language. We can then formalize this by positing an intermediate stage where all discrete phonetic categories become mapped onto discrete phonological features before they will become mapped onto phoneme categories.[20]

We now need to specify the formal steps that will alter the constraints of the previous stage into constraints that are fit for this particular stage. First, let us look at a possible final stage of the phonetic learning process. Taking again our example tableau, let us assume a final stage in which all sounds with an F_1 value between 480-560 Hz will be mapped onto the phonetic category /520/. Tableau 25 shows a possible final state of the promotion-based model; tableau 26 shows this final state when assuming a demotion-based model.

	400	301	301	300	201	200	124
[560]	*W(60)	*/560/	*/540/	*W(40)	*/500/	*W(20)	*/520/
/500/	*!				*		
☞ /520/				*			*
/540/			*!			*	
/560/		*!					

Tableau 25

	490	400	359	351	300	200	68
[560]	*/560/	*W(60)	*/500/	*/540/	*W(40)	*W(20)	*/520/
/500/		*!	*				
☞ /520/					*		*
/540/				*!		*	
/560/	*!						

Tableau 26

As we can see, all values between 480 Hz and 560 Hz will now be mapped as an instance of the category /520/, regardless of which model we use. Now, for the next stage, consider that a child has also acquired a phonetic category /600/ to which all sounds with a

[19] In line with Chomsky & Halle (1968), we will define the phonological features as articulatory commands, but note that this is not the only option. Another option could be to define phonological features on a strictly phonological level, an option advocated by e.g. Clements (1985).

[20] By assuming that the phonological feature labels are linked to real articulatory commands, I depart from the assumption made by Boersma et al. (2003) that the infant gives its phonetic categories *arbitrary* feature labels.

frequency between 560 and 640 Hz will be mapped. We can now show how a child deals with the mapping of F_1 frequencies belonging to two different phonetic categories onto two different phonological features.

In this stage of learning, the old warp and category constraints have become obsolete, as the child does not label incoming values as what we called 'phonetic categories' any more. In order to accommodate to the new situation, Boersma et al. (2003) use a model where the child now uses the information gathered in the first stage to form a new group of so-called CUE constraints. I will adopt this model (with some minor alterations). We assume that the child transforms the information from the previous stage by giving all values that were previously mapped onto the same category constraint a feature label, while at the same time making sure that all these values are initially ranked equally high. To make this process more clear, consider our earlier example where all values between 480 Hz and 560 Hz were mapped onto the category /520/. The final state of the tableau (whether demotion- or promotion-based) tells the child that its language has a phonetic category of which 520 Hz is the prototypical member, and that all frequency values between 480 and 560 Hz should therefore be considered to be members of this phonetic category. Furthermore, due to some knowledge the child has about the correspondence between acoustic properties and the anatomy of the speech organs (knowledge that may be learned through babbling), the child knows it should relate /520/ to the phonological feature [mid]. In the new stage, the child uses this information to form a new set of cue constraints of the type *[520]/[mid], which should be read as "a value of 520 Hz should not be mapped onto the feature [mid]". The child also makes similar constraints for the other values that were previously mapped onto the category /520/, in this case all values between 480-560 Hz. These values will initially be ranked at the same height.

In our example, the child also heard F_1 values between 560 Hz and 640 Hz which were mapped onto the category /600/. Assume now that the child knows that 600 Hz belongs to a feature [mid-low] in its language. In a similar procedure as the one described in the last paragraph, the child will now proceed by making new cue constraints for all values between 560-640 of the type *[600]/[mid-low]. Initially, these cue constraints will all be ranked at the same height, as well as being at the same height as the other group of cue constraints. Furthermore, the child will also make cue constraints in which all F_1 values will be paired with what the child considers to be the 'wrong' feature, e.g. *[520]/[mid-low]. These cue constraints will initially be ranked higher than the cue constraints with 'correct' pairings of F_1 values and features, since the child assumes from the information it acquired in the previous stage that it is for example more unlikely that a value of 520 Hz should be mapped onto the feature [mid-low] than onto the feature [mid]. In the end, the child has thus transformed the information given to him by the final ranking of the category constraints relative to the warp constraints into a set of 40 cue constraints that specify to which phonological feature various F_1 values should be mapped. These cue constraints are initially ranked as follows:

{*[480-560]/[mid-low], *[560-640]/[mid]} >> {*[480-560]/[mid], *[560-640]/[mid-low]}

We can see from this ranking that, initially, all cue constraints with 'wrong' pairings of F_1 value and phonological feature are ranked at the same height above all cue constraints with 'correct' pairings of F_1 value and phonological feature.

Now how does learning proceed in this stage of phonological development? First, contrary to the first stage of phonological acquisition, a child now starts to acquire lexical knowledge, i.e., it starts to learn words. Children usually start to produce their first words around 10-11 months of age (Kit, 2003, p. 4), although they probably start to perceive and learn these words a little earlier. This lexical knowledge can in turn guide the child in phonological acquisition by telling the child when it incorrectly mapped an incoming sound onto a particular phonological feature. To make this more clear, consider the following example in which a child hears a word that contains a vowel with an F_1 value of 540 Hz. According to its acquired knowledge, the child will map this sound onto the feature [mid]. But let us assume now that in this particular case the value of 540 Hz is supposed to be mapped onto the feature [mid-low]. The child knows that it has made a mistake in mapping [540] onto [mid], for example because the lexical context tells the child that it should have perceived *less* instead of *loss*. The child then proceeds to correct this mistake by re-ranking some of its constraints. In particular, in this case the child re-ranks its constraints by degrading the constraint *[540]/[mid-low] (since the child found out that it is ranked too high) and by upgrading the constraint *[540]/[mid] (since, according to the input, this constraint is ranked too low). This procedure, being error-driven, more closely resembles the way in which traditional re-ranking of constraints is assumed to proceed. However, there are some differences as well: recall that traditional OT learning accounts are often assumed to be discrete (i.e., re-ranking of constraints always means that one constraint is placed above or below another constraint with each learning step). In this thesis, I will assume that learning not only involves the gradual demotion or promotion of constraints in the first stage of acquisition, but all the way throughout the acquisition process. In other words, we will keep representing each learning step with a change of the ranking values of certain constraints. As an example, again consider that the child wants to re-rank the constraints associated with the F_1 value [540] because its grammar at this point depicts an incorrect winning output candidate. In our initial state, let us assume that all constraints that depict incorrect pairings of value and feature are randomly given a ranking value of 100, while all constraints that depict correct pairings of value and feature are given a ranking value of 0. Thus, in the initial state the constraint *[540]/[mid-low] has a ranking value of 100 while the constraint *[540]/[mid] has a ranking value of 0. Having set the values like this ensures that all values of 540 Hz are normally mapped onto the feature [mid], as we can see in tableau 27 (leaving out all constraints that are unaffected for now).

		100	0
	[540]	*[540]/[mid-low]	*[540]/[mid]
✓	[mid-low]	*!	▨
☞	[mid]		▨ *

Tableau 27

However, since the child knows it made a mistake, it wants to correct this mistake in some way. In our model, the child does so by degrading the highest-ranked constraint by a set value of 1, while at the same time upgrading the lowest-ranked constraint by this same value of 1. This is the process as described by the Gradual Learning Mechanism (Boersma, 1997). Doing this, after correction the values of the constraints have been updated to 99 and 1, as can be seen in tableau 28.

	99	1
	*[540]/[mid-low]	*[540]/[mid]
[mid-low]		
[mid]		

Tableau 28

As each learning step only involves changing the constraint's inherent ranking values by small amounts, constraints will only change positions after the child has had sufficient input to reasonably assume that the affected constraints really point to an error in the relative ranking of these constraints. Put differently, in traditional OT learning models it was very easy for the child to re-rank certain constraints that should not be re-ranked, only because it heard one instance of a sound that violated the current constraint ranking. With a gradual learning algorithm, constraint re-ranking is less error-prone since it will only happen after the child has found enough evidence to justify such a re-ranking. In our case, for example, a child will only re-rank the two constraints after it has heard 51 instances where the current constraint ranking gives it an incorrect outcome, as we can see in tableau 29 (but of course the initial values or ranking steps can be altered to increase or decrease the amount of input a child needs to hear before it can actually re-rank its constraints). Thus, using a gradual learning model instead of a discrete learning model in general improves the robustness of the constraint ranking.[21]

	51	49
[540]	*[540]/[mid]	*[540]/[mid-low]
☞ [mid-low]	*!	
[mid]		*

Tableau 29

In chapter 3, we also wondered what would happen if an infant in its first stage heard input values that could not be mapped onto any phonetic category. We then posited the possibility that the infant would map these values onto another value that was acoustically closest to a phonetic category as possible, thereby making a new phonetic pseudo-category. The lexical information in this stage would then further tell the child how to process this category. Now, let us assume that the child created an extra phonetic category /480/. In principle, this category can equally likely be mapped onto either the feature [mid] or the feature [mid-low]. We formalize this in that, initially, the two cue constraints *[480]/[mid] and *[480]/[mid-low] are ranked equally high, e.g. both begin with a ranking value of 50. Even though in the beginning the child may make mistakes in the choice of output candidate, eventually the child's lexical knowledge will tell it which phonological feature is most likely the right feature for that input value. The child can then easily alter the ranking values of both constraints, as is shown in tableau 30.

	51	49
[480]	*[480]/[mid]	*[480]/[mid-low]
✓ [mid]	*!	
☞ [mid-low]		*

Tableau 30

In this tableau, the constraint militating against the mapping of [440] onto [mid] is still ranked highest, but the lexicon told the child that this is wrong. Therefore, the ranking values will be changed and eventually, the constraints will change positions such that this value from then on will correctly be mapped onto the feature [mid].

To sum up, in this stage of learning children start to map acoustic values onto phonological features, using the information they gathered in the first stage of phonological acquisition. In this stage learning no longer happens more or less automatically. Instead, the child does not alter its constraint ranking or the constraints' ranking values when it hears input that is compatible with the current ranking of its constraints, but only when the child hears a sound that is incompatible with what the child should have heard, with the child's lexical knowledge telling it when a certain sound is perceived incorrectly.

[21] In this set up, though, the ranking of the constraints is immediately altered again after the child hears another token of [540] that should be mapped onto the feature [mid]. How we prevent this kind of instability to happen will be explained in section 4.3.

When a child learns, it does so by changing the ranking values of the affected constraints by small steps at a time. Once the ranking value of a previously higher-placed constraint falls below the ranking value of a previously lower-placed constraint, these constraints will switch places. Once this happens, the child will further assume that it should perceive sounds according to the new constraint ranking. This stage thereby ensures that all input values are mapped onto the right phonological features. This is especially important in cases where the information gathered in the previous stage is vulnerable to mistakes, for example because two phonological features belonging to different phonemes lie acoustically close together. In general, we can maybe state it such that the first stage of acquisition involves the creation of broad and roughly-defined phonetic categories. These categories are fine-tuned in the second stage and additionally given new feature labels. In the third stage, described in the next section, phonological features from different acoustic dimensions are then combined into different phoneme categories.

4.3 From phonological features to phoneme categories

In the previous section, I have shown how the mapping of acoustic values onto phonological features can be modelled using an OT-based Gradual Learning Algorithm. But a child also needs to learn how several phonological features should be mapped onto one phoneme category. A description of a model that can explain such a mechanism is given in the current section. This model is based on the model used in Boersma & Hamann (2008), with the only difference being that they use acoustic values as input forms instead of phonological features.

From the previous stages, our child has learned that there are two F_1 categories of which 520 Hz and 600 Hz are prototypical members, and that these two categories relate to the phonological features [mid] and [mid-low]. Furthermore, the child has fine-tuned exactly which F_1 values belong to which feature. Now, for the next stage, let us assume that a child at the same time learned that there are two F_2 categories of which 800 Hz and 1800 Hz are prototypical members. The child also learned that all sounds with an F_2 between 1700 and 1900 Hz are values that belong to the category /1800/ and that all sounds with an F_2 between 740 Hz and 860 Hz are values that belong to the category /800/. Later on, the child learned to associate the category /1800/ with the phonological feature [front] and the category /800/ with the feature [back]. Our last assumption is that the child through its emerging lexical knowledge now knows that its language has two vowels and that they are arbitrarily labelled /ɔ/ and /ɛ/.

In our language, the phoneme /ɔ/ is most commonly pronounced by adult speakers as a phoneme with the features [mid, back]. The phoneme /ɛ/, in contrast, is most commonly pronounced with the features [mid-low, front]. The combinations [mid, front] and [mid-low, back] do not normally correspond to a particular phoneme label, but both of these combinations may still occasionally be pronounced by a speaker. How such combinations are perceived depends on the positions of the constraints in the child's tableau, as we will see later.

At this stage of acquisition, we assume in line with Boersma & Hamann (2008, p. 235) that the child already has correct lexical representations (i.e., it can already correctly distinguish the different phonemes that a word is built up from), but that it does not yet know how exactly these phonemes are linked to the child's acquired phonological features. Also, at this point the child still handles the incoming cues from different acoustic dimensions separately; that is, the child has not yet learned to combine values from multiple acoustic dimensions into one abstract phoneme representation. The child's task now is to link the several acoustic features to the correct phoneme labels. To make this task possible, the child first constructs a new inventory of cue constraints. In this particular case, the child combines all feature combinations with one of the two possible phoneme labels. The result is the following set of 8 cue constraints:

*[mid-low, front]/ /ɔ/, *[mid-low, front]/ /ɛ/
*[mid, front]/ /ɔ/, *[mid, front]/ /ɛ/
*[mid-low, back]/ /ɔ/, *[mid-low, back]/ /ɛ/
*[mid, back]/ /ɔ/, *[mid, back]/ /ɛ/

Additionally, we assume that each incoming sound is first transformed to the corresponding phonological features (in line with the final state of the tableau from the previous stage) before it is used as the input value in the new tableau. In other words, each input value no longer represents purely acoustic values, but rather abstract phonological features.

The phoneme labels phonologists use are defined only by convention, which means that the only reason that e.g. /ɔ/ is given that label and not e.g. the label /ɛ/ is because we made the agreement to label /ɔ/ as /ɔ/ and not as /ɛ/. This also means that the link between acoustic features and phoneme labels is completely arbitrary. This is formalized in that in the initial state of this stage, the various constraints are all ranked at the same height, as we can see below. This also automatically implies that all constraints start with the same initial ranking value, here randomly set at 100.

{*[mid-low, front]/ /ɔ/, *[mid-low, front]/ /ɛ/, *[mid, front]/ /ɔ/, *[mid, front]/ /ɛ/, *[mid-low, back]/ /ɔ/, *[mid-low, back]/ /ɛ/, *[mid, back]/ /ɔ/, *[mid, back]/ /ɛ/}

This state is different from the initial state in the previous stage, where the child's acquired knowledge about the phonological properties of its language was formalized in an initial state where some of the constraints were ranked higher than other constraints.

Now the child can start learning again. Similarly to the previous stage, learning in this stage is supervised by the child's lexical knowledge.[22] For the first learning steps, consider tableau 31, where the child hears the feature values [mid-low] and [front].

	100	100
[mid-low, front]	*[mid-low, front]/ /ɔ/	*[mid-low, front]/ /ɛ/
/ɛ/		*
/ɔ/	*	

Tableau 31

Because all constraints are still ranked at the same height, the tableau cannot tell the child which output candidate it should choose. To still make a decision, there are basically two routes we can follow. The first option is to assume that every input value is perceived with a small amount of stochastic 'noise' as described in Stochastic OT (Boersma, 1997), such that even with identical ranking values, one constraint is still always considered most optimal. This constraint then decides the chosen output candidate. Because this output candidate is chosen at random, in our case there is a 50 % chance that the incorrect output candidate is picked. It is then assumed that this potential error is soon corrected when the child notices that the position of the cue constraints in the tableau does not correspond with what the lexicon tells the child is the correct output candidate. Another option is to assume that the lexicon also guides the decision of which output candidate to choose whenever the ranking of the constraints prevents the child from choosing. In this case,

[22] Note that the assumption that the learning process is supervised by the lexicon in both the second and the third stage implies that each lexical representation contains both phonemic and featural information.

the fact that the child e.g. heard the word *shop* forces the choice of output candidate /ɔ/ over output candidate /ɛ/.

Regardless of which path is chosen, the choice of output candidate forces the re-ranking of the constraints. If we assume that /ɔ/ is chosen as the output candidate, the ranking value of the constraint *[mid-low, front]/ /ɔ/ is lowered to 99 and the ranking value of *[mid-low, front]/ /ɛ/ is increased to 101. Consequently, the latter constraint is now also absolutely ranked above the former constraint. Now, we would normally assume that this ranking state stays the same as long as the child keeps hearing instances of [mid-low] and [front] that correctly correspond to the phoneme /ɔ/. However, the downside then is that only one occurrence of [mid-low, front] that should be linked to the phoneme /ɛ/ immediately causes the ranking values of the two constraints to become equal again, regardless of how many instances of [mid-low, front]-/ɔ/ the child heard before. In other words, the frequency of occurrence of a particular feature-phoneme combination cannot influence the stability of the corresponding constraint, making that the position of all constraints remains unstable indefinitely. This is not what we want, as the whole reason we use a gradual learning model instead of a discrete one is to improve the robustness of the constraints.

In this respect, the advantages of Stochastic OT quickly become apparent. In Stochastic OT, the fact that each sound is perceived with some evaluation noise ensures that e.g. an instance of [mid-low, front] may still be perceived as /ɛ/, even when the constraint disfavouring /ɛ/ is ranked slightly higher than the constraint disfavouring /ɔ/. In these cases, the child reacts to this by further increasing the ranking value of *[mid-low, front]/ /ɛ/ and lowering the ranking value of *[mid-low, front]/ /ɔ/. This is also shown in tableau 32, where the evaluation noise causes /ɛ/ to be chosen over /ɔ/, even though the ranking value of the constraint favouring /ɔ/ is lower than the ranking value of the constraint favouring /ɛ/. The child then reacts to this by changing the ranking values of the constraints to 98 and 102, respectively.

	99	101
[mid-low, front]	*[mid-low, front]/ /ɔ/	*[mid-low, front]/ /ɛ/
☞ /ɛ/		*
/ɔ/	*!	

Tableau 32

A stochastic model greatly increases the robustness of the constraints, as it makes it possible for the constraints' ranking values to diverge beyond their smallest possible distance. In general, the constraints now reach a stable state as soon as the distance of their respective ranking values exceeds the point where the influence of the stochastic noise becomes so small that it can be considered negligible. For example, if we assume that the influence of stochastic noise can be defined with a standard deviation of 5.0 in a Normal distribution (meaning that the chance that the value of a constraint at one point in time is 5 or fewer points away from the actual ranking value of that constraint is approximately 68%), then this means that when the ranking value of one constraint differs around 15 points (or 3 SD's) from the ranking value of another constraint, the chance that the first constraint will become lower ranked than the second constraint due to evaluation noise is so small that this chance becomes negligible. At that point, one can say that these two constraints have become stable with respect to one another. Changes then only happen when the lexicon tells the child that the chosen output candidate is incorrect. But even then, a single incorrect output candidate does not immediately change the absolute ranking of the constraints.

We also considered what would happen when a child heard feature combinations that normally do not correspond to an actual phoneme. As the child does not yet know which feature combinations normally belong to which phoneme categories, these instances will not be treated differently from the normal situation (we also need to assume that the child does not mind having multiple feature combinations for just one phoneme). One thing that makes this model quite elegant is that with each input value, only the constraints that directly correspond to this value can be altered. So when we assume that the child hears the uncommon combination of [mid, front], only the constraints *[mid, front]/ /ɔ/ and *[mid, front]/ /ɛ/ are possibly affected, while the other constraints by default remain unaltered. This makes way for a situation where the child can treat uncommon input combinations equally to the standard combinations without labelling some of the combinations as more or less prototypical. The only possible difference is that, due to the likely low frequency of occurrence of the uncommon combinations, the state of the corresponding constraints is less stable. This is not at all a problem, as it actually quite nicely reflects the uncertain phonemic properties of such utterances.

So how does learning normally develop in this stage? With every input a child hears, it checks whether the current constraint ranking is still consistent with the chosen output candidate, as 'told' to the child by the lexicon. In the early stages, when the child has heard relatively few input tokens, it is predicted that constraint re-ranking happens relatively often. With more input tokens, the position of the constraints generally becomes more and more stable. This way, the link between phonological features and phonemes becomes stronger with more input, eventually reaching a point where the link is so strong that the constraints stop moving. This does not necessarily happen to all constraints, as it is perfectly possible for a language to have separate phonemes whose acoustic values occasionally overlap. In these cases, the constraints might still show occasional value re-ranking or even switch positions.

Even though the acquisition process is more or less finished as soon as most constraints reach a stable state, complete acquisition might never be reached. Even though older children (and adults, for that matter) can probably automatically map incoming values onto the right phoneme category, it might still occasionally be necessary to change the ranking values of certain constraints. Situations where this might happen are for example when listeners hear speakers from other dialects or sociolects, who happen to pronounce certain phonemes slightly differently than what the listener is used to. Actual constraint re-ranking might even happen as a result of phonological change (e.g. with phonological mergers or push chains), although such situations might be quite rare to happen completely during just one generation. We will come back to this in the discussion.

Chapter 5

Unsupervised learning: possible simulations

5.1 Introduction

Despite all assumptions and examples in our model of unsupervised learning, the only real way to see if our model actually works is to virtually simulate actual phonological acquisition using preferably real (or plausible) language data. But as a description of the implementation and testing of such simulations would probably require me to write an additional thesis, this task falls outside the scope of this thesis. Instead, I will use this chapter to shortly outline the considerations one needs to make before they can actually start running computer simulations. The discussion will only focus on the simulation of the unsupervised models, as the simulation of supervised models has already been extensively discussed in the literature (see e.g. Tesar & Smolensky, 1998; Boersma & Hayes, 2001; Boersma & Hamann, 2008). A large part of the discussion will in particular focus on the question which of the two models discussed in the section on unsupervised learning - the promotion-based model or the demotion-based model - is more fit to represent actual learning. This discussion will be described in section 5.2.

If one wants to simulate models using real language data, there are still a number of parameters that need to be defined. In section 5.3, I will shortly discuss these parameters.

5.2 Promotion vs. demotion

In chapter 3 I showed that there are (at least) two ways to model unsupervised learning: a model that represents learning through promotion of constraints and one that represents learning through demotion of constraints. Still, even if we can make two models that in theory should work, at one point we need to commit ourselves to one of these models as the model that has the best balance between being likely to represent real phonological acquisition and being easy to virtually implement. In this section, I will discuss the (dis)advantages of both models to come to the conclusion of which model is eventually preferred.

A demotion-based model is preferred over a promotion-based model in that the learning steps are more straightforward and therefore easier to implement. In a promotion model, each learning step requires the increase of the ranking value of a large number of constraints. In a demotion model, in contrast, each learning step only requires that the infant lowers the ranking value of the matching constraint. In other words, a model that is based on demotion requires considerably less cognitive effort, is easier to virtually simulate and is in those respects preferred over a promotion-based model. This way of representing learning is also more like the way learning has been represented in traditional OT learning accounts. However, in another way our demotion model is much further away from standard OT models and their inherent assumptions than the promotion model: there is no interaction between the various category constraints in the demotion model, which basically reduces the learning process to simple counting. As a consequence, at the end of the first stage the infant suddenly needs to re-interpret all of its phonological knowledge in order to accommodate to the new situation where all values are instances of a certain phonetic category. In the promotion model, the infant gradually builds these categories, and as such the emergence of categories in this model is not as sudden as in the demotion model. A last disadvantage of the demotion model over the promotion model is that the constraints in a demotion model never reach a solid, final state. In other words, in a demotion model learning never stops. A consequence of this is that the first learning stage should stop in time, because it would otherwise be possible that too many constraints fall below the lowest warp constraint, with the result that the infant would assume more phonetic categories than there are present in the language. In the promotion-based model, the emergence of too many phonetic categories would only be possible with too little input, not with too much input; this has to do with the fact that learning only involves the movement of non-matching category constraints, and that there is no movement of matching category constraints. Also, the fact that there are no learning steps for a particular input form after the corresponding category constraint has been moved behind the last warp constraint makes the chance of ending up with too few phonetic categories much smaller.

In sum, we might say that a demotion-based model is both easier to learn and implement than a promotion-based model, which would make demotion preferred over promotion. However, the demotion-based model comes with a number of disadvantages that the promotion-based model does not have. These disadvantages include the fact that phonetic categories in a demotion-based model cannot emerge gradually, the fact that learning never stops and the fact that the model only predicts correct outcomes with exactly the right amount of input data. These disadvantages make the demotion-based model so fragile that I think that the promotion-based model in the end is the best choice for modelling early acquisition. But obviously, the best way to test the two models is by simulating them with real language data. Even though we cannot do that here, in the next section I will still give a couple of things to keep in mind for whoever wishes to carry out the simulations.

5.3 Defining the parameters

Even after all the assumptions made in the description of the models, we are still required to make some additional assumptions when we want to simulate the learning process with real language data. First, we need to have language data we can use for the simulations. This language data preferably comes from a corpus of infant directed speech (IDS), data which is of course as empirically sound as possible. Benders (2013, ch. 5) for example did this with a corpus of Dutch IDS from which different tokens of /ɑ/ and /aː/ were randomly taken. Another option is to base the simulations on a phonemic situation that is found in one or more languages, but of which the exact acoustic properties are invented. This is for example done by Boersma & Hamann (2008), who simulate their model with sibilant inventories that are based on actual sibilant inventories found in various languages, but of which the exact acoustic properties are invented by them. Whichever route one wants to take, to simulate unsupervised learning we still need to specify the amount of input the infant approximately receives during the time that unsupervised learning is active.

First, let us assume that unsupervised learning begins as soon as an infant is born and is finished at six months of age. Based on corpus research, Swingley (2007, p. 461) estimates that children hear around 48800 word tokens in a period of 3 weeks. For a period of 6 months, this would mean that an infant hears almost 4 million word tokens. Of course, this does not mean that the infant also hears 4 million tokens of each phoneme. But if we for example want to simulate a language with 5 different vowel phonemes that all occur equally often, and if we assume that a word token on average contains 1.5 vowels, then calculations show that each vowel will on average be heard 1.2 million times within the six-month period. One can then use this number for the simulations by specifying that for every vowel, 1.2 million tokens must be randomly taken from the vowel's acoustic distribution and then 'fed' to the virtual infant. This way, the virtual simulations come as close to what actually happens within an infant's first 6 months as they can get.

Alongside the number of tokens fed to the virtual infant, there are a number of additional important parameters that we need to define.[23] These parameters include the initial ranking values of both the category and the warp constraints and the value of each learning step (also called *plasticity*). Because the specification of these values depends on how fast one wants learning to be, it is perhaps best to simply try out through trial-and-error which settings work best. Regarding the warp values, the distance between the ranking values of two adjoining warp constraints should ideally correspond to the difference in percentage between two fixed places on the basilar membrane.

Both models described in this thesis are best simulated with Stochastic OT (Boersma, 1997), where each input form comes with a small amount of stochastic (or *evaluation*) noise. In simulations, this is represented by the fact that for each evaluation, the ranking values of the constraint are not exactly identical to the value that was assigned to them. In fact, the values are a little bit higher or lower thanks to the evaluation noise. The range of this variation from the absolute ranking value depends on the amount of evaluation noise one chooses. The evaluation noise's number corresponds to the standard deviation within a Gaussian function, such that an evaluation noise of 2.0 can also be read as a Gaussian distribution with a standard deviation of 2.0. The mean of this Gaussian distribution is constant at 0 (Boersma, 1998, p. 331). If we choose higher values for the evaluation noise, then the margins of the possible ranking values are wider, making it easier for constraints to switch places and thus for an input form to be assigned the wrong output candidate. What number one should choose depends on how far one wants the margins between two adjoining constraints to be. In general, with higher evaluation noise numbers the virtual infant needs to hear more input tokens before the model reaches a stable state, but the outcomes are also more robust.

In this section, I have listed all properties for which one needs to define the values when running virtual simulations. Of course I hope that, at some point in time, this description will actually be used (whether by myself or someone else) to test the reliability of the two unsupervised models.

[23] Note that this description is based on the assumption that the simulations are carried out in the phonetics software program Praat, where all of the described parameters can easily be defined. Praat can be downloaded for free from http://www.fon.hum.uva.nl/praat/.

Chapter 6

Conclusions and discussion

The aim of this thesis was to see if it is possible to model unsupervised phonological acquisition. Earlier non-OT models that tried to account for unsupervised learning are insufficient either because they cannot make precise when exactly phonetic categories have emerged or because they rely on difficult computational learning mechanisms that are unlikely to represent real learning processes. In this thesis, I therefore developed a more extensive and slightly altered version of the model by Boersma et al. (2003) that aims to be both empirically adequate and sufficiently precise. In my opinion, the altered model in this thesis has three advantages over the model described by Boersma et al. (2003):

1. Whereas Boersma et al.'s model used a group of teleological warp constraints invented only to make perceptual warping possible, our model removed this teleological nature of the warp constraints by directly linking the properties of this group of constraints to the physiological properties of the humans' auditory system. In other words, by assuming that these constraints are a formalization of innate physiological properties we removed the need to assume any sort of innate 'warping device'.

2. In Boersma et al.'s model, the infant only learns when it encounters a mismatch between the sound it perceived and the category to which this sound is mapped. This kind of error-driven learning implies that the infant has some innate device that tells it which input-output combinations are preferred. In our model, learning is instead match-driven: with such a device, learning happens as the result of an automatic mapping between the perceived input form and a phonetic category with the same value. This way, we do not need to assume that the infant knows which input-output combinations are more or less preferred.

3. In Boersma et al.'s model, there is no mechanism that causes learning to stop. As a consequence, if the infant receives too many input values, this causes an incorrect number of phonetic categories. In the promotion version of our model (but not in the demotion version), learning stops as soon as all category constraints are either ranked above the most severe warp constraint or below the least severe warp constraint. In other words, the fact that learning in this model stops automatically makes the promotion-based model more robust than Boersma et al.'s model.

I furthermore showed how this model of unsupervised learning can be extended to later stages, where learning becomes guided by the lexicon. The combination of these models yields a complete model of phonological acquisition all the way from the start until phonological acquisition has finished. This model can be represented as in figure 6.1, with different types of constraints working on different levels of abstraction.

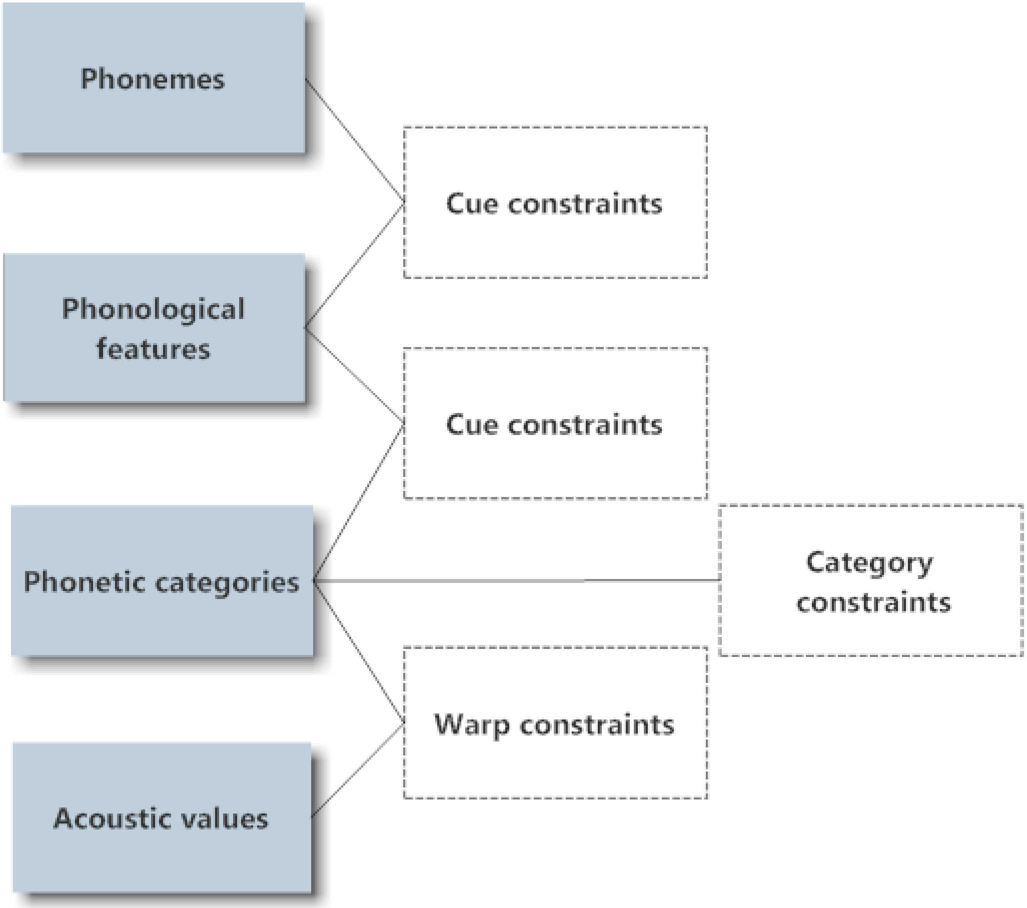

Figure 6.1: An integrated model of phonological acquisition.

In the first stage of phonological acquisition, learning involves the mapping of acoustic values onto phonetic categories. A group of warp constraints guides this mapping by making it possible for acoustic values to be mapped onto phonetic categories with different values, while the group of category constraints limits the number of phonetic categories that can be formed. In the second stage, the created categories form the basis for a new group of cue constraints that guide the mapping from the phonetic categories onto a number of articulatory-based phonological features. From this point on, learning becomes guided by the child's lexical knowledge. In the third stage, the cue constraints guide the mapping from phonological features belonging to different acoustic dimensions onto single phoneme categories.

Two variants of the unsupervised model were developed: one variant that is based on the demotion of matching category constraints and one that is based on the promotion of non-matching category constraints. It was shown that both models in principle seem to yield the right outcomes when applied to invented language data. Initially, our aim was to also virtually simulate both of these models with more empirically adequate language data. It however turned out that this was a task too big to carry out in this thesis. I hope that, in the future, such simulations can give definite conclusions about the reliability of both models, and that they can show which of the two variants works best. Furthermore, the model used by Boersma et al. (2003) involves the simultaneous promotion and demotion of constraints. It might be interesting to research whether this procedure could also be implemented in a model that uses fixed warp constraints, and if so, what the potential (dis)advantages of such a procedure would be over using only promotion or demotion within a model.

In section 4.3, we stated that phonological learning does not necessarily stop after childhood. In fact, the great diversity in which different speakers of the same language can pronounce various sounds and the constant evolution of spoken language might force listeners to keep updating their constraint rankings, even as adults. The fact that we assumed that, after the second stage, listeners perceive incoming sounds no longer as raw acoustic values but as phonological features might in this respect be slightly problematic. Especially the more subtle, purely phonetic changes might be hard to implement with this assumption. If we assume that, through adulthood, listeners can still slightly shift their feature boundaries, then we must somehow be able to account for that within our model. It is therefore worth investigating both if and to what extent listeners actually do shift their feature boundaries and, if they do, how we can account for this within our model.

The language example we used to illustrate our models was admittedly simplified. In further research, it is worth investigating if the models can also handle more complicated phoneme inventories. Some options to consider are cases where many different phonemes are found on a particular acoustic dimension or languages where single phonological features are used for more than one phoneme category, e.g. a language where two vowels make use of the feature [mid].

Even though our model used the Gradual Learning Algorithm (Boersma, 1997) combined with the framework of Stochastic OT (Boersma, 1997) in the last two stages, it might also be interesting to see how the model works within other variants of OT. In particular, it can be interesting to see how our model fares within Harmonic Grammar (Pater, 2009), a version of OT that uses *weighted* instead of ranked constraints. An important consequence of weighted constraints is that the combination of several low-weighted constraints have the possibility to outweigh a single heavier-weighted constraint. It may be worthwhile to see what the effects of this different approach are on our learning model.

References

Adriaans, F., & Swingley, D. (2012). Distributional Learning of Vowel Categories is Supported by Prosody in Infant-Directed Speech. In N. Miyake, D. Peebles, & R. Cooper (Eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 72–77). Cognitive Science Society.

Ashby, F. G., Queller, S., & Berretty, P. M. (1999). On the dominance of unidimensional rules in unsupervised categorization. *Perception & Psychophysics*, *61*(6), 1178–1199.

Ashley, K., Disch, L., Ford, D., MacSaveny, E., Parker, S., Unseth, C., ... Yoder, B. (2010). How many constraints are there? A preliminary inventory of OT phonological constraints. *Occasional Papers in Applied Linguistics*(9). Retrieved from http://www.academia.edu/1010234/How_many_constraints_are_there_A_preliminary_inventory_of_OT_phonological_constraints

Benders, T. (2013). *Nature's distributional-learning experiment: Infants' input, infants' perception, and computational modeling.* (Unpubished doctoral dissertation, University of Amsterdam)

Best, C. T. (1995). Learning to Perceive the Sound Pattern of English. In C. Rovee-Collier & L. Lipsitt (Eds.), *Advances in Infancy Research* (Vol. 9, pp. 217–304). Ablex.

Boersma, P. (1997). How we learn variation, optionality, and probability. In *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam* (Vol. 21, pp. 43–58).

Boersma, P. (1998). *Functional Phonology: Formalizing the interactions between articulatory and perceptual drives.* Holland Academic Graphics/IFOTT. (Doctoral dissertation, University of Amsterdam)

Boersma, P., & Chládková, K. (2011). Asymmetries between speech perception and production reveal phonological structure. In *Proceedings of the 17th International Congress of Phonetic Sciences* (pp. 328–331).

Boersma, P., Escudero, P., & Hayes, R. (2003). Learning Abstract Phonological from Auditory Phonetic Categories: An Integrated Model for the Acquisition of Language-Specific Sound Categories. In M. J. Solé, D. Recasens, & J. Romero (Eds.), *Proceedings of the 15th International Congress of Phonetic Sciences* (pp. 1013–1016). Barcelona: Casual Productions.

Boersma, P., & Hamann, S. (2008). The evolution of auditory dispersion in bidirectional constraint grammars. *Phonology*, *25*, 217–270.

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

Chomsky, N., & Halle, M. (1968). *The Sound Pattern of English.* ERIC.

Clements, G. N. (1985). The geometry of phonological features. *Phonology yearbook*, *2*, 225–252.

Escudero, P., & Boersma, P. (2003). Modelling the perceptual development of phonological contrasts with Optimality Theory and the Gradual Learning Algorithm. In S. Arunachalam, E. Kaiser, & A. Williams (Eds.), *Proceedings of the 25th Annual Penn Linguistics Colloquium. Penn Working Papers in Linguistics* (Vol. 8.1, pp. 71–85).

Feldman, J. (2000). Minimization of Boolean complexity in human concept learning. *Nature*, *407*(6804), 630–633.

Flemming, E. S. (1995). *Auditory Representations in Phonology.* (Doctoral dissertation, University of California. Published 2002, Routledge.)

Goudbeek, M., Swingley, D., & Smits, R. (2009). Supervised and Unsupervised Learning of Multidimensional Acoustic Categories. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(6), 1913–1933.

Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The Faculty of Language: What Is It, Who Has It, and How Did It Evolve? *Science*, *298*(5598), 1569–1579.

Hayward, K. (2000). *Experimental Phonetics.* Longman London.

Holt, L. L., & Lotto, A. J. (2006). Cue weighting in auditory categorization: Implications for first and second language acquisition. *The Journal of the Acoustical Society of America*, *119*(5), 3059–3071.

Hume, E. (2011). Markedness. In M. Van Oostendorp, C. J. Ewen, E. Hume, & K. Rice (Eds.), *The Blackwell Companion to Phonology* (Vol. 1, pp. 79–106). John Wiley & Sons.

Idsardi, W. J. (2006). A Simple Proof That Optimality Theory Is Computationally Intractable. *Linguistic Inquiry*, *37*(2), 271–275.

Kager, R. (1999). *Optimality Theory.* MIT Press.

Kewley-Port, D., & Watson, C. S. (1994). Formant-frequency discrimination for isolated English vowels. *The Journal of the Acoustical Society of America*, *95*, 485–496.

Kiparsky, P. (1982). Lexical Phonology and Morphology. In I.-S. Yang (Ed.), *Linguistics in the Morning Calm* (pp. 1–91). Hanshin.

Kit, C. (2003). How Does Lexical Acquisition Begin? A Cognitive Perspective. *Cognitive Science*, *1*(1), 1–50.

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., & Lindblom, B. (1992). Linguistic Experience Alters Phonetic Perception in Infants by 6 Months of Age. *Science*, *255*(5044), 606–608.

Love, B. C. (2002). Comparing supervised and unsupervised category learning. *Psychonomic Bulletin & Review*, *9*(4), 829–835.

Maddieson, I. (2011). Vowel Quality Inventories. In M. S. Dryer & M. Haspelmath (Eds.), *The World Atlas of Language Structures Online.* Munich: Max Planck Digital Library. Retrieved from http://wals.info/chapter/2

Maye, J., Weiss, D. J., & Aslin, R. N. (2008). Statistical phonetic learning in infants: Facilitation and feature generalization. *Developmental Science*, *11*(1), 122–134.

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*(3), B101–B111.

Mehler, J., Jusczyk, P., Lambertz, G., Halsted, N., Bertoncini, J., & Amiel-Tison, C. (1988). A precursor of language acquisition in young infants. *Cognition*, *29*(2), 143–178.

Moon, C., Cooper, R. P., & Fifer, W. P. (1993). Two-day-olds prefer their native language. *Infant behavior and development*, *16*(4), 495–500.

Oller, D. K. (1980). The emergence of the sounds of speech in infancy. *Child phonology*, *1*, 93–112.

Pater, J. (2009). Weighted Constraints in Generative Linguistics. *Cognitive Science*, *33*(6), 999–1035.

Pierrehumbert, J. B. (2003). Phonetic Diversity, Statistical Learning, and Acquisition of Phonology. *Language and speech*, *46*(2-3), 115–154.

Polka, L., & Werker, J. F. (1994). Developmental Changes in Perception of Nonnative Vowel Contrasts. *Journal of Experimental Psychology: Human Perception and Performance*, *20*(2), 421–435.

Prince, A., & Smolensky, P. (1993). *Optimality Theory: Constraint Interaction in Generative Grammar* [Tech Report No. 2]. (Rutgers Center for Cognitive Science, Rutgers University)

Ramus, F. (2002). Language discrimination by newborns: Teasing apart phonotactic, rhythmic, and intonational cues. *Annual Review of Language Acquisition*, *2*(1), 85–115.

Rietveld, A. J. M., & Van Heuven, V. J. (2009). *Algemene Fonetiek* (3rd ed.). Bussum: Coutinho.

Swingley, D. (2007). Lexical Exposure and Word-Form Encoding in 1.5-Year-Olds. *Developmental psychology*, *43*(2), 454–464.

Tesar, B., & Smolensky, P. (1998). Learnability in Optimality Theory. *Linguistic Inquiry*, *29*(2), 229–268.

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*, *104*(33), 13273–13278.

Van Leussen, J. W. (2012). *(Un)supervised Category Formation in Multi-Level Phonological Grammars.* (Poster presented at LabPhon 13, July 27–29, University of Stuttgart, Germany)

Van Leussen, J. W. (2013). *Simulating unsupervised category formation using cue and structural constraints.* (Unpublished manuscript)

Vihman, M. M., Macken, M. A., Miller, R., Simmons, H., & Miller, J. (1985). From babbling to speech: A re-assessment of the continuity issue. *Language*, *61*(2), 397–445.

Werker, J. F., & Lalonde, C. E. (1988). Cross-Language Speech Perception: Initial Capabilities and Developmental Change. *Developmental Psychology*, *24*(5), 672–683.

Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development*, *7*(1), 49–63.

Appendix A

Order of tokens

1. [500]
2. [520]
3. [520]
4. [520]
5. [520]
6. [520]
7. [520]
8. [520]
9. [520]
10. [520]
11. [540]
12. [520]
13. [520]
14. [540]
15. [520]
16. [520]
17. [480]
18. [500]
19. [520]
20. [520]
21. [520]
22. [520]
23. [520]
24. [500]
25. [520]
26. [500]
27. [500]
28. [500]
29. [520]
30. [520]
31. [520]
32. [520]
33. [500]
34. [520]
35. [520]
36. [520]
37. [500]
38. [520]
39. [540]
40. [500]
41. [540]
42. [520]
43. [520]
44. [520]
45. [540]
46. [520]
47. [520]
48. [520]
49. [520]
50. [520]
51. [520]
52. [520]
53. [520]
54. [520]
55. [500]
56. [540]
57. [520]
58. [520]
59. [540]
60. [560]
61. [520]
62. [540]
63. [540]
64. [520]
65. [540]
66. [480]
67. [500]
68. [520]
69. [520]
70. [520]
71. [540]
72. [520]
73. [520]
74. [520]
75. [520]
76. [520]
77. [520]
78. [560]
79. [540]
80. [500]
81. [520]
82. [520]
83. [520]
84. [520]
85. [520]
86. [520]
87. [560]
88. [520]
89. [520]
90. [540]
91. [500]
92. [520]
93. [520]
94. [520]
95. [520]
96. [520]
97. [520]
98. [500]
99. [520]
100. [520]