VICTOR PERSIEN

AN EXEMPLAR-THEORETIC APPROACH TO CHAIN SHIFTS

AN EXEMPLAR-THEORETIC APPROACH TO CHAIN SHIFTS

VICTOR PERSIEN

In Partial Fulfillment of the Requirements for the Degree of

Master of Arts (M.A.)

Submitted by

Victor Persien

Jun.-Prof. Dr. Ruben van de Vijver (Supervisor)
Jun.-Prof. Dr. Wiebke Petersen (Co-Supervisor)

July 29, 2014

Omnia sponte fluant, absit violentia rebus.

— Iohannes Amos Comenius, *Orbis Sensualium Pictus*.

The most merciful thing in the world, I think, is the inability of the human mind to correlate all its contents.

— Howard Phillips Lovecraft, *The Call of Cthulhu*.

To my father, Dieter Persien.

1953 – 2007

ACKNOWLEDGMENTS

Foremost, I would like to express my sincere gratitude to my supervisors Jun.-Prof. Dr. Ruben van de Vijver and Jun.-Prof. Dr. Wiebke Petersen.

To Ruben: thank you for being interested in supervising my thesis right from the start and for giving me lots of supportive feedback and remarks during the process of researching and writing.

To Wiebke: thank you for supervising my thesis and deepening my interest in and understanding of the application of mathematics to linguistics during my studies and beyond.

Furthermore, I would like to express gratitude towards my late father Dieter without whom I probably would have never become aware of computer programming in the first place.

Additionally, I would like to thank my mother Bettina as well as her significant other Klaus for constantly encouraging me to pursue my studies even if others wouldn't approve so easily.

At last, my gratitude goes out to all of those who have been supportive of me at some point in life or another, especially B., D., E., P, and P.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LISTINGS

INTRODUCTION

When individual sounds shift, they rarely do so in isolation, as a change of their phonetic or phonological properties affects the way in which they contrast with other sounds. Thus, it is often observed that one sound shift invokes another sound shift. For instance, the loss of aspiration of one consonant may be accompanied by the loss of aspiration of other, similar consonants. A special case of such interrelated changes constitute *chain shifts*. These are characterized by one phoneme that takes the position of another phoneme, whereby both contrasts remain preserved.

(1) Chain shift in Franco-Provençal (Martinet 1952: 6):
/ẽ/ → /ɛ̃/ → /ɛ/ → /a/ → [ɔ]

The case depicted in (1) shows a chain shift using the example of a series of sound shifts that have taken place in the Hauteville dialect of Franco-Provençal. Here, /ẽ/ changed to /ɛ̃/, /ɛ̃/ changed to /ɛ/, and so forth. The graphically last shift /a/ to [ɔ] did not result in a merger of two vowel phonemes, because the latter position was previously unoccupied.

Martinet (1952) assumes, that this particular shift was initiated by the raising of /a/ towards [ɔ], whereupon /ɛ/ was lowered towards the gap left behind by /a/, /ɛ̃/ was denasalized, etc. Chain shifts that are the result of consecutive gap-filling are otherwise known as *pull shifts* or *drag shifts* in the literature. But there is, in principle, another scenario that one could propose w.r.t. the above example, namely that it constituted a *push shift*. In such a case then, /ẽ/ would have approximated /ɛ̃/, whereupon the latter would have lost its nasality, and so on.

Crucial with respect to chain shifts is the notion of contrast. In push shifts, contrasts that are on the verge of collapsing are preserved by subsequent sound shifts. In pull shifts, an initial shift allows for the exaggeration of already existing phonetic contrasts, in the above case this is true for several front vowels[1].

Among linguists, reasons for the maintenance and maximization of contrasts are a matter of intense debate. Some argue, that these issues are accounted for by the grammatic system itself (e.g. Flemming 1995, Łubowicz 2012), whereas others assume that repeated instances of language use lead to the gradual selection of variants, that maintain or maximize contrasts within sound inventories, respec-

1 The notation might be misleading in this respect. It is likely not the case, that, for instance, the new phonetic make-up of /ɛ/ equals exactly that of former /a/. Rather, /ɛ/ approached a position circa halfway between novel /ɛ/ and /ɔ/.

tively (e.g. Ohala 1989, Labov 1994, Blevins 2004, Blevins and Wedel 2009, Boersma and Hamann 2008).

This work will argue in favor of the latter position by showing, using a series of computer simulations, that chain shifts can in fact arise from communicative needs, which are negotiated during individual speech acts.

The underlying framework chosen in this work is *exemplar theory* (Nosofsky 1988), a cognitive theory about general concept learning, that has been applied to phonological issues several times during the last two or so decades (e.g. Johnson 1997, Bybee 2001, Pierrehumbert 2001, Blevins 2004, Walsh et al. 2010). It relies on the assumption, that language users do not store mere abstract representations of linguistic units, but have at their disposal detailed memorizations, *exemplars*, of actual instances of language in use. These exemplars consist of various information, like the motor patterns used, a detailed acoustic representation, its meaning, as well as extra-linguistic information, such as the age or gender of the speaker. When producing or perceiving an utterance, language users resort to (parts of) these memorizations in order to en- or decode a piece of information, respectively.

Structure in this context is not accounted for by a set of discrete units and rules. Instead, it is an emergent property, that arises through the similarities and dissimilarities that exemplars share. If, for example, a language user observes that there are two semantically distinct phonetic patterns in use that are basically identical except for one subpattern, which has a different phonetic form in each case, she will assume that these two subpatterns P_1 and P_2 are used for the distinction of meanings. If she then perceives a pattern that contains a subpattern which resembles P_1 or P_2, she will compare the pattern to exemplars in her memory that contain either of them in order to infer its identity in order to make sense of the percept.

This is a very simple example, but if one spins further its underlying thought, it becomes apparent how contrast maintenance or maximization are explainable within exemplar theory. Because of the competition that ensues between exemplars that are representative of concept P_1 or P_2, percepts that are ambiguous w.r.t. to these are less useful in communication and will, as a consequence, be avoided in future utterances or just not stored in the first place. And this leads eventually to the drifting apart of both categories in the sense that more exemplars will be used and stored that are less confusable.

Now, one could be inclined to think that exemplar theory is merely a theory of language acquisition and that language, once it is acquired, behaves like a rule-based system. Although it might be the case that the generalizations which an individual has made about the interaction of linguistic units (which are also generalizations) at a given point in time are describable by means of rule-based grammars, exemplar theory states that these generalizations are not im-

mutable. Exemplars can fade from memory over time and with their withdrawal formerly existing contrasts may become weaker or collapse altogether.

Up to this point, the loss of phonemic contrasts has not been enlarged upon. Let us for an instance assume that example (1) was initiated by a shift of /ẽ/ towards /ɛ̃/. Then it might as well have been the case that both categories merged and the sound shift would have been over. Such cases are well documented in the literature on sound change – but so are push shifts. So, why do categories merge in one case, but chain shift in another? A reason that is often put forward in this regard is that, in the case of a merger, the contrast is simply not important enough in its function of a means to distinguish meanings, and so its disappearance does not negatively affect communication on a larger scale. If a push shift appears, however, the contrast is maintained because of its importance for the distinction of meanings.

Martinet (1952) refers to the importance of a phonemic contrast as its *functional load*. If it is high, the contrast under consideration is important and a chain shift is likely to take place if one of the phonemes drifts towards the other one. If it is low on the other hand, such a situation will probably lead to a merger of both categories. In its initial stage, much criticism was brought up against the functional load as a tool that serves to estimate the likelihood of a merger or push shift, respectively (e.g. King 1967). Nevertheless, in recent times this concept was able to regain popularity due to positive results regarding its validity (Surendran and Niyogi 2006, Wedel et al. 2013a,b).

Against this background, it is easy to be seen that exemplar theory provides us with a large array of assumptions that fit the task of finding a reasonable explanation for the coming about of chain shifts perfectly well. In order to achieve this, two models[2] will be proposed. Both consist of two agents that communicate with one another by exchanging exemplars and it will be explored whether and how the competition among categories can account for the maintenance of contrast and eventually chain shifts and mergers.

The first model is a rather simple one that builds on the one conducted by Pierrehumbert (2001). In this one, communication consists solely of the exchange of phonetic signals, that have to be categorized by a listener. The aim here is twofold. On the one hand it will be investigated if the competition among categories leads to contrast maximization and if this is strong enough to account for pull shifts. On the other hand, the model serves to show that a high difference in phoneme frequency, which affects the density of a category w.r.t. the exemplars it contains, facilitates mergers. We will see that both of these goals can be accomplished but that it largely depends the presence of positive feedback that the listener receives from the speaker.

2 The Java™ source code is available at `github.com/vpersien/csinet`.

In the second model, the agents do not simply exchange speech sounds but embed them in a context in order to transmit meanings, i.e. some sort of extra-linguistic information. The listener will then have to infer the meaning from earlier stored exemplars. The process of inference follows two alternative routes whose selection depends on the availability of suitable exemplars. If the listener chooses the first one, he will attempt to infer the meaning directly from exemplars that have been uttered in the same context and have a similar phonetic make-up. If he picks the other route, he will first assign a phonemic category to the phonetic component and will then use this hypothesis during the inference process. The speaker on the other hand also has two pathways at her disposal. If she has not yet gained enough evidence about the phonetic form of the lexical items that are assigned the meaning she wishes to transmit or if she finds that the item that she initially picked is too easily confusable, she will resort to more prototypical instances of the phonetic components of her utterance. Otherwise, she will immediately prepare the chosen item for production. The factor that the context of the utterances – which can be interpreted as phonetic, syntactic or even extra-linguistic context – aids the inference of a meaning may render the contrast between a given pair of phonemic categories useless. Furthermore, this circumstance allows for a measurement of the usefulness of phonemic contrasts in terms of their functional load. It will be shown that, within this model, once a phonemic category is on its way towards another one, push shifts can ensue due to the joint contribution of competition, homophony avoidance and the lexical distribution of phonemes. Besides, we will see that the functional load of the contrast under consideration can indeed serve as a predictor for mergers or push shifts, but only if one has additional knowledge about the probability distributions that determine the utterance of lexical items.

The thesis is organized into four chapters. In the first one, a theoretical ground is established that is necessary to understand the reasoning behind the simulations. Section 1 will define chain shifts and expand on their relation to homophony avoidance and dispersion maximization. Section 2 serves to equip the reader with a basic understanding of exemplar theory and especially its relation to linguistics. In the subsequent section, it will be explained how sound change is accounted for by exemplar theory. Since this is a huge topic in itself, only two issues will be taken into consideration. Subsection 3.1 will portray how sound change is accounted for without adhering to system-level properties that force sound inventories into a specific shape, e.g. one that maximizes categorical contrasts. The second subsection takes a look at the manner of sound change within exemplar theory, which poses an alternative to earlier accounts such as Neogrammarian sound change or lexical diffusion.

The simulations will be carried out in chapters ii and iii. In both cases, first the model architecture is introduced and afterwards their behavior during several simulation runs is scrutinized and interpreted. The introduction to the second model is a bit more extensive, as it additionally requires a proper definition of a measure for the functional load (7.2) and of the probability distributions that will define the lexical distribution of the phonemic contrasts considered (8.1.2). Eventually, both chapters conclude with a discussion of the observations made during the simulations.

At last, chapter iv offers a discussion of the results and gives a brief outlook concerning possible future work in the same direction as the one in which this thesis is pointing.

Part I

THEORETICAL FOUNDATIONS

1 DIACHRONIC CHAIN SHIFTS

Chain shifts (CSs) are a class of sound shifts A, B where the target position of A is also the source position of B. This process is illustrated by the following example.

(2) Grimm's law:
   a. /b^h/ → /b/ → /p/ → [f]
   b. /d^h/ → /d/ → /t/ → [ð]
   c. /g^h/ → /g/ → /k/ → [x]

Example (2) contains a subset of the sound shifts that are known as Grimm's law, which demarcated the development of Proto-Germanic from Indo-European. In this case the aspirated voiced plosives /b^h,d^h, g^h/ became unaspirated voiced plosives /b,d,g/, the former unaspirated voiced plosives became unaspirated voiceless plosives and /p,t,k/ became their corresponding voiceless fricatives.

However, the established notation does not give us any information about the causal links between the involved sound shifts. With two sound shifts A, B that form a sub-chain of a CS one can distinguish three possible types of CSs:

- Push shifts,
- Pull shifts,
- Regular shifts.

In push shifts, B takes place as a reaction to A. If example (2) constituted a push shift, the loss of aspiration would have triggered the loss of voice in /b,d,g/ which in turn would have caused the fricativization of the voiced plosives.

In pull shifts on the other hand, the moving component of A fills the gap left behind by B. If it were to be the case that Grimm's law was instead a pull shift, the fricativization of /p,t,k/ would have been the first to happen and all the other shifts, from right to left, filled the occurring gaps left by the respective prior shift.

And in regular shifts, a term coined by Łubowicz (2012) in order to characterize synchronic CSs, the observable chain is a superficial result caused by another sound shift or phonological process. Although interesting in its own right, the case of regular shifts will not be pursued any further in this work.

Push and pull shifts are generally associated with to two distinct principles. Push shifts are said to be the result of merger avoidance,

or, more functionally speaking, homophony avoidance. Pull shifts on the other hand are tied to the notion of dispersion maximization. Both concepts and their relation to chain shifts will be discussed in the upcoming two subsections. Note, however, that both principles provide us with little information as to why a sound shift that ultimately lead to a chain shift was initiated in the first place. This is an issue that reduces to the general problem of making out reasons for language change and exceeds the scope of this thesis. In the simulations described in chapters ii and iii, an initiating sound shift will be taken as a given.

1.1 HOMOPHONY AVOIDANCE

If one phoneme P_1 is about to occupy the place held by another phoneme P_2, there are two possible scenarios: P_1 and P_2 merge, or P_2 in turn shifts. The former case will inevitably lead to homophony of the words that formerly contrasted in P_1 and P_2. The latter case then is an alternative strategy to maintain the contrasts that would have otherwise been lost by a merger of P_1 and P_2. This begs the question of what guides the selection for one of the two principles. Labov (1994: 327–9) names five factors that may influence the probability of a merger:

1. The functional load of the opposition.
2. The number of distinctions already made along that phonetic dimension.
3. The number of phonetic features on which the opposition depends.
4. The discriminability of the phonetic features on which the opposition depends.
5. Limitations in the range of movements that would avoid merger.

2.–5. are phonetically based considerations and rather concerned with the question as to why an impending merger is not avoided although it should. The first point is the most important one, both in the linguistic discourse on this topic and as a causing factor. The four latter points will not be discussed any further, but future empirical or computational studies that control for any of these will be certainly of interest.

André Martinet (e.g. Martinet 1952) and linguists of the Prague School proposed that the decision is lead by a quantitative measure called *functional load* (FL) or *functional yield*. The FL of a linguistic opposition is a value that is positively proportional to the number of meanings that are distinguished by this opposition in a given language. However, at the time there was "no complete agreement as

to what this term is meant to cover. In its simplest somewhat unsophisticated acceptation, it refers to the number of lexical pairs which would be complete homonyms if it were not that one word of the pair presents one member A of the opposition where the other shows the other member B" (Martinet 1952: 8). And as Martinet points out, a mere counting of minimal pairs of a given opposition does not yield a satisfying overview of the actual role the opposition plays in the language w.r.t. to their distinguishing meanings, because it does not respect the context the minimal pairs appear in. Let us for an instance assume that in a language there is a phoneme pair P_1, P_2 that seemingly exhibits a high FL (calculated by the minimal pair method). But it turns out that P_1 only appears in nouns whereas P_2 only appears in adjectives. Then, even if both phonemes are randomly confused, the meaning of the vast majority of uttered sentences would still come across unambiguously. So, what is needed instead is a measure that takes into account possibly disambiguating context and, at best, the actual number of occurrences of utterances whose meaning might alter if the distinction between P_1 and P_2 was not made. An early account is given by Hockett (1966) who employed an information-theoretic model that takes into account possible contexts that are able to disambiguate between meanings in the light of a merger of a phonemic contrast. This model was generalized by Surendran and Niyogi (2006) and with it they were able to show among others that in Mandarin tones are as important as vowels, that is a tonal merger would be as grave as a vowel merger, or that consonantal contrasts are more important than vocalic ones (given data from Mandarin, German, Dutch and English).

Even though more sophisticated models than the counting of minimal pairs are available to the researcher, this simple method proves still a decent approximation. Wedel et al. (2013b) were able to show in a corpus study that the number of minimal pairs of a given opposition is negatively correlated with the likelihood of a merger. Additionally, they found a positive correlation of mergers and phoneme frequency if the minimal pair count was low. In a subsequent study, Wedel et al. (2013a) furthermore distinguished between syntactic categories and were able to show that the minimal pair count within categories is a better predictor than across categories. This is congruent with the discussion above, since the loss of contrast across syntactic categories is easier to be compensated for by morphosyntactic means. The authors did not, however, contrast mergers with chain shifts but only calculated the general probability of a merger given an inventory and a list of minimal pairs. Nevertheless, their model predicts a probability of 0.0 for any of the mergers between /i/ and /ɪ/, /ɪ/ and /ɛ/, and /ɛ/ and /e/ in American English, a subset of the vowel space that is often subject to chain shifting. In contrast, their model predicts a positive probability of 0.1 for an /ɑ/-/ɔ/ merger to occur. This merger

is otherwise known as the cot-caught merger and has already taken place in larger parts of Canada and the US.

Despite criticism concerning the validity of the FL as a predictor for mergers (e.g. King 1967), more recent research suggests that the FL is indeed a useful measure. In this thesis it will be explored how phoneme frequencies as well as the incentive of avoiding homophony may affect the coming about of push shifts and how predictive the functional load of the contrasts under observation is.

However, even if it is true that push shifts are the result of contrast maintenance by means of persistent homophony avoidance, at this point it is still not clear how pull shifts might be accounted for. A viable candidate for this matter is the maximization of dispersion between phonemic categories.

1.2 DISPERSION MAXIMIZATION

Speech sound inventories of the world's languages tend to exhibit maximal perceptual distance among their members. For instance, if a vowel system contrasts three vowels, it is more likely to contain [a, i, u] than, say, [a, æ, ɑ]; if it contains 5 vowels, [a, e, o, i, u] is more likely than one that contains only front vowels. In a similar fashion, this pertains to consonant inventories. In this case we say the sound inventory exhibits maximum dispersion. Apart from static observation of existing sound inventories, many sound changes involve the movement of one sound to a position that maximizes the perceptual distance to other surrounding sounds. This is particularly true for pull shifts, where a shifting sound leaves behind a gap that is consequently filled by another sound.

The fact that dispersion maximization seems to be a universal tendency raises the question of whether or not it is an inherent property of the linguistic system. The teleological standpoint is reflected in Dispersion Theory (Flemming 1995) which introduces constraints into OT that explicitly account for the maximization of contrast between surface forms of members of phonological inventories. Another teleologically informed work stems from Liljencrants and Lindblom (1972) who showed computationally that maximizing the distance w.r.t. the first two formant values of vowels results in vowel systems that resemble actual ones.

On the contrary, in recent years a number of authors (e.g. de Boer 1999, Oudeyer 2006, Wedel 2006, Boersma and Hamann 2008, Blevins and Wedel 2009) sought to show that realistically dispersed sound inventories do not have to be accounted for by language-internal mechanisms, but can instead arise as a side-product of the perception-production loop during language evolution or acquisition.

De Boer (1999) studied the evolution of vowel spaces by computational means. In his simulations, agents in a population, each consist-

ing of a three-dimensional vowel space as well as a simplified perception and production apparatus, collectively and distributively build a vowel system from scratch by playing imitation games. In these games, one agent utters a vowel and another, randomly chosen agent seeks to imitate the perceived vowel on the basis of a nearby vowel in his vowel space. The first agent then decides whether the imitated vowel fits into the same category as the initially uttered one and gives accordingly positive or negative non-linguistic feedback. Depending on the feedback, the imitator performs adjustments within his vowel space. A more detailed description of the model will not be given at this juncture. The important thing to note is that through these local interactions a global feedback loop arises that will eventually lead to well dispersed and above that realistically looking vowel systems.

Other than de Boer, Boersma and Hamann (2008) assumed an underlying phonological framework, namely Bidirectional Stochastic OT (Boersma 2007). In their paper they aimed to show that sibilants, represented as spectral mean values on a one-dimensional scale, disperse automatically as a result of the interplay between the involved acquisition algorithm and production constraints. The resultant dispersion, however, is not maximal but rather in an equilibrium state w.r.t. the costs of being misunderstood and production effort. The authors claim this intermediate dispersion to be much closer to what speakers actually produce, since they would not aim for the most extreme values even if these are less likely to be misperceived by the listener. A crucial prediction of the model is that non-optimally dispersed inventories will eventually reach optimality within a few generations purely due to the tension between the goal of becoming both an optimal listener and an optimal speaker.

As a side-effect of every inventory's tendency to exhibit optimal dispersion after a few generations, Boersma & Hamann were able to reenact a push shift that took place in 13th century Polish: After the phonologization of /sʲ/, /s/ and /sʲ/ drifted further apart which in turn lead to progressive backening of /ʃ/ due to dispersion maintenance. However, the model as is will always predict a chain shift in this case and is thus unable to reproduce chain shifts that terminated in a merger.

Without adhering to a phonological framework, Labov (1994) states that dispersion arises naturally as a consequence of the confusability of category members. If solely those percepts that lead to successful communication influenced the representation of the category, those percepts that are less confusable with members of other categories (e.g. the most central ones) would be more often able to contribute to the category representation just because they would be part of more successful acts of communications. A stronger variant of this account is given by Wedel (2006). He assumes that the probability of whether or not an incoming stimulus changes the overall representation of the

category it is categorized in additionally depends on its degree of ambiguity. Or put more concrete, since his is an exemplar-theoretic account (see sec. 2), a percept is less likely to be stored as an exemplar the more ambiguous it is, even if communication was successful. He was able to show in a simulation, that this behavior, as expected, leads to more dispersed categories since more central and peripheral speech tokens of a given category are less likely to be confused with members of other categories and thus even more likely to be stored by the listener. This conjecture could be partially experimentally confirmed (Denby 2013).

Both of the above described phenomena are instances of contrast maintenance. Homophony avoidance serves the goal of maintaining contrast between lexical items when it is threatened by two phonemes that are on the verge of collapsing into one. Via dispersion maximization on the other hand, existing phonemic contrasts, and indirectly also those between lexical items, are improved upon. In view of an ongoing sound shift, these phenomena can come into effect. The former may then result in a push shift, the latter in a pull shift. The present thesis attempts to show how both homophony avoidance and dispersion maximization can be accounted for within the framework of exemplar theory and eventually lead to chain shifts of phonemic categories. But first, a basic understanding of exemplar theory and of the way it explains sound shifts has to be established.

2

EXEMPLAR THEORY

Exemplar theory (ET; Nosofsky 1988) is a theory of general concept learning and categorization that builds on previous theories that commit to family resemblence (Wittgenstein 1998 [1953]), such as prototype theory (Rosch 1975, Rosch and Mervis 1975), rather than tertium-non-datur membership judgements. Just like prototype theory, exemplar theory assumes that categories are built around more prototypical examples, which stem from the experience with members of that category, and membership judgement is accomplished by comparing the given percept to other category members. So, for instance, a salmon would be considered a more typical example of a fish by many people in the Western world, whereas a sea horse only a rather peripheral one, being reflected in response times and error rates. Likewise, instances of pseudo-fishes might be judged more prototypical than sea horses only because of their similarity to more prototypical fishes like salmons.

In order to account for effects caused by similarity, frequency and recency, exemplar theory assumes that mental categories are clouds of actual percepts, exemplars, whose distance from one another depends on their reciprocal similarity. In contrast, rule-based accounts, as well as prototype theory assume the existence of abstract representatives for each category. Categories in exemplar theory on the other hand consist of detailed tokens of remembrance. We will now see how categories are supposed to come about according to this very theory.

2.1 THE EMERGENCE OF STRUCTURE IN EXEMPLAR THEORY

Figure 1: Illustrative example of how exemplar theory accounts for the emergence of structure, i.e. categories. For a detailed explanation, refer to the text.

Structure in exemplar theory is a property that emerges from the repetition of similar percepts. Let us assume for a moment that our mind is a completely blank slate without any pre-wired tendencies. Then, a single percept is in itself meaningless. We can only extract information out of it we compare it to other percepts. Over time we will be able to group similar parts of our history of percepts together and refer to those groups as similar entities. For a useful illustration confer fig. 1. It shows the gradual emergence of an 'ideal' triangular shape from the interference of multiple poorly drawn, semi-opaque triangles. Let us assume that we do not have any conception of what a triangle is. After drawing only four triangular-like shapes we can hardly make any generalizations, metaphorically represented by the darker lines that arise from the iterative overriding, because they are barely visible and do not apply well to novel instances of such shapes. After drawing more and more triangular-like shapes, we can see a common structure, one could say a prototype, emerge from all the overlapping shapes. This is an approximation of the main idea behind many cognitive models or related data structures such as artificial neural networks. The differences lie in the ways how generalizations, i.e. categories, are represented and newly incoming percepts are assigned to a category. Referring to the third panel of fig. 1 as an example, artificial neural networks would store the picture as a whole with more opaque lines being a more probable (or less noisy) triangle[1], and prototype theory would assume that only the thickest lines are stored as a prototype and every newly encountered shape is compared to it. In exemplar theory on the other hand, every single instance of the drawn triangular-like shapes would be stored in memory and incoming shapes would be compared to the most similar ones. This illustrative example is by no means a perfect analogy of neither prototype theory, artificial neural networks or exemplar theory, but hopefully yields a useful idea of what is to be understood when talking of structure in reference to these frameworks.

At this point, the question whether category membership is absolute or fuzzy will be left unadressed. Within the models set up in this work exemplars will be assigned exactly one phoneme label which makes membership absolute. This is, however, only a heuristic in order to be able to account for prototype-related effects and not an elaborate theoretical assumption.

1 This is in fact a simplification. Artificial neural networks generally do not store much more than the weights of the connections between individual neurons. Nevertheless, their probabilistic nature makes them behave in a fashion that is akin to that described in the text.

2.2 EXEMPLAR-THEORETIC LINGUISTICS

Since the end of the 1990s, exemplar theory has been applied to phonetic and phonological phenomena several times (e.g. Labov 1994, Johnson 1997, Bybee 2001, Pierrehumbert 2001, Wedel 2004, 2006). Applications to syntax date back even further, exemplified by the *Data Oriented Parsing* theory (Bod 1992). Johnson (1997) made use of ET to find a way of modeling "speech perception without speaker normalization": Instead of transforming speech signals into a 'neutral' form deprived of the speakers' ideosyncratic characteristics, he chose to store individual signals as exemplars and with them properties he considered by and large responsible for the variation in the signal such as the speakers' sex. After a training phase the model was able to identify synthesized vowels with an accuracy comparable to a human's and, as a side-effect, sex identification with up to 98% accuracy. One could say, Johnson's exemplar model circumvented speaker normalization: Instead of comparing a to-be-classified signal to a single normalized form, it is compared to several forms all of which display some degree of variation. That is, variation is not an undesirable property that has to be leveled, but is deliberately exploited.

The model employed by Johnson focuses on speech perception, but the idea of exemplar theory can be easily extended to speech production as well. Spoken words are motor patterns and within a language these are highly repetitive across lexical items. This high degree of repetition then constitutes what is otherwise known as phonological words, syllables, phonemes, etc. (Bybee 2001, 2006).

In a computational model that combines both perception and production, Pierrehumbert (2001) showed that the gradual spread of lenition from high- to low-frequency lexical items can indeed arise from "exemplar dynamics", that is the interplay between an iterated distorted exchange of speech signals and an exemplar-based lexicon in which they are deposited and from which they are drawn in turn. Furthermore, she gave an early account of how mergers may come about in an exemplar-based situation. Pierrehumbert's work has been seized on by many subsequent modelers and parts of her model will also be adopted and reviewed in a more detailed fashion later in this thesis (see section 4).

In exemplar theory, linguistic structure is accounted for by the same principles in which more general categories come about. Children acquire a language by learning to separate the speech stream into meaningful units and infer from it general patterns, i.e. a grammar, that allow them to construct new sentences on their own or analyze more complicated utterances. But instead of using the input to set parameters or to rerank constraints on the one hand and to build a lexicon of morphemes on the other hand, in exemplar theory children as well as adult language users are assumed to (a) not possess a grammar that

is distinct from a lexicon in the first place and (b) store every single utterance in their exemplar space (or 'lexicon'). Grammatical 'rules' are then inferred directly from the entries in the exemplar space. But exemplars do not reside in memory forever. They fade away gradually over time, formally realized by an activation value.

This treatment allows for two different pathways in perception and production. One the one hand, oft-encountered lexical items, e.g. individual sentences, words or syllables, can be accessed right away. On the other hand, less often or newly encountered items can be parsed or constructed, respectively, from smaller constituents. All knowledge about constituency and principles of combination is inferred from the make-up of exemplars that are present in the exemplar space. Both pathways, though, are not mutually exclusive — in fact, quite the opposite holds true. Exemplar theory would assume that an idiom like "Between a rock and a hard place" will be readily available for a native speaker of English, even more so in perception than in production. Let us assume that all the motor patterns needed to utter this sentence are memorized by heart. Now, the speaker has knowledge about structural properties of the sentence, because she has encountered many similar sentences, words, etc. during her lifetime. If she now wishes to substitute one part of the sentence by another she can do so based on her knowledge. Let us say she wishes to substitute "hard place" by a nonce word "tlaron". She first needs to analyze the sentence sufficiently in order to find a suitable position for the substitution, in this case "hard place". The analysis will then proceed by the construction pathway. "tlaron" is a nonce word, what means that it is most likely not available in her exemplar memory, so the assembly follows the construction pathway. /tla/ is a syllable hardly ever to be stumbled upon in the English language, that is it has little to no activation and has to be composed from its constituent parts, that is /t/, /l/ and /a/ or perhaps /t/ and /la/. /rɔn/ or /rən/ in contrast are more likely to be directly ready for production, but even then the speed with which it can be accessed depends on the frequency and activation of suitable exemplars.

The above example is of course held very simple in order to exemplify how grammar is accessed or sometimes skipped within exemplar theory. For more sophisticated accounts on sentence and structure formation in ET, refer to the literature (e.g. Bod 1992, Bybee 2006, 2013). Nevertheless, the idea that sentences are processed with varying speed depending on the availability and activation of their constituents, on whichever level, is backed by empirical findings on sentence processing (Baayen et al. 1997, Jurafsky 2003). Exemplar theory applied to linguistics provides an integrated approach to explaining several 'performative' phenomena. Among others:

- Within a language user's lifetime, phonemic categories are less prone to change than other constituents, like individual words.

In ET this could be explained by the fact that one perceives and produces a lot more phones than words and hence phonemes are much more densely packed because of which outliers have less impact on the categories as a whole.

- Psycholinguistic experimentation often exploits semantic, phonological or syntactic priming which is but a recency effect in combination with the activation of similar (groups of) exemplars.
- The case of the Al-Sayyid Bedouin Sign Language (ABSL), signed by members of the Al-Sayyid bedouins in Israel, poses a problem for many theories that assume the innateness of phonemes, since this language does not seem to make use of them. All words of ABSL are phonetically holistic and above that largely variable (Sandler et al. 2011). For exemplar theory, though, the absence of a phonemic layer is not problematic at all, since phonemes are categories that are only emergent from repetition found in the speech/sign stream and motor routines. ABSL signs do not exhibit enough repetition (yet?) and hence there are no phonemes in the signers' grammars.
- Children have more difficulties learning novel words that have many lexical neighbors (regarding phonetic similarity) than others which have few. It is suggested, that similar lexical items jointly inhibit the comprehension of novel utterances in their phonetic neighborhood (Swingley and Aslin 2009).

Because in ET categories are abstractions over exemplars and exemplars are stored in the lexicon with every encounter, language change, at least on the level of the individual, happens with every instance of perceived or produced speech. Because of this, in ET linguistic structure is inseparable from its history. To elaborate on this, the next section will focus on approaches to language change and their realization within exemplar theory.

3

SOUND CHANGE IN THE LIGHT OF EXEMPLAR THEORY

Exemplar theory explains linguistic structure as being emergent from multiple instances of language use over time. The synchronic state of a language is thus always tied to the diachronic pathways that lead to it. The current section aims to expand on this theme by giving an explanation of the mechanisms that constitute linguistic change within an exemplar-theoretic context and by contrasting these to former approaches.

As became apparent from the discussion in section 1, the oft-observed phenomenon of contrast maintenance can be explained in several ways. Many authors seek to make the linguistic system itself responsible. In their view, phonetic or phonological units are directly manipulated in order to contrast with other units, i.e. language change is due to contrast maintenance is goal-directed or *teleological*. Exemplar theory on the other hand attempts to step away from the manipulation of abstract units and thus lines itself up with other approaches that see language change as the result of constant variation during language use. Here, evolutionary theory comes in handy as a metaphor: single instances of mutations, i.e. variations in the speech signal, are prone to selection for future use if they adapt to the environment, i.e. lead to communicative success. This view does without a higher-order device that adjusts linguistic units, because the language users themselves take on this task via selection and replication. This will be the topic of section 3.1.

A second point regards the manner in which sound changes spread through the lexicon. In exemplar theory, lexical items do not have a single abstract memory representation, instead particular instances of word exemplars are stored in the lexicon. Phonetic variants of a particular phoneme may thus not only differ context by context, but also word by word. This allows for sound changes that are gradual in two respects: they may first apply to some words and then spread gradually to other words in the lexicon, and the sound change may proceed in minute, unnoticeable steps. This is a type of shift which contrasts with former assumptions about the nature of sound change, such as the Neogrammarian view, which postulates that all words that contain a particular phoneme are affected at once, i.e. abruptly, in the light of a sound shift. Section 3.2 will oppose four kinds of sound shifts which contrast in their gradualness on the lexical and the phonetic level, and outline how exemplar theory explains the one that exhibits gradualness on both levels.

This section concludes with a brief outline of how the findings of this and the preceding section can be combined in order to account for the principles of contrast maintenance that are supposed to underlie the incidence of chain shifts.

3.1 (A)TELEOLOGY AND THE EVOLUTIONARY METAPHOR

Darwinian evolution is a completely ateleological affair: Mutations occur for no particular purpose and if they are by chance helpful (or at least not hindering) in the production of offspring, they will spread. That is, there is no higher agency involved that intentionally selects for individuals. Rather, it is the environment that permits or forbids features to spread. Evolution is not restricted to biology. Instead, the theory of evolution can be applied to any domain whose members are subject to replication, mutation and selection, be it social practices (Dawkins 1976) or algorithms. This generalization of evolution from ecology to other domains is sometimes called Universal Darwinism (Dawkins 1976) or the evolutionary metaphor (e.g. Blevins 2004). The application of evolutionary theory to other domains is by no means new, but began immediately after the publication of Darwin's ideas, most prominently exemplified by the unfortunate example of social darwinism that has served ever since to justify the exploitation of one group of people by another group of people. More fortunate was the application of the theory of evolution to language change by some of Darwin's contemporaries. In his 1850 (pre-"Origins") writing "Die Sprachen Europas in systematischer Übersicht", August Schleicher treats languages as species that may yield daughter languages and can be arranged in genealogical trees. He later (1869) expands on this view with reference to Darwin, stating that languages compete among one another and are selected for on the basis of the territorial expansion of ethno-linguistic groups. Müller (1870) rejects this view, because it relies on extra-linguistic factors and not so much on properties inherent to languages that are subject to extinction. Instead he sees evolution at work at a more fine-grained level:

> How is it that a new word, such as 'to shunt,' or a new pronunciation, such as 'gold' instead of 'goold,' is sometimes accepted, while at other times the last words newly coined or newly revived by our best writers are completely ignored or fall dead? We want an idea that is to exclude caprice as well as necessity [...]. [I]t is the idea of 'Natural Selection' that was wanted [...]. — Müller 1870: 256–7

Thus, in his view, languages relate to ecosystems which are inhabited by linguistic entities, "words and grammatical forms" (Müller 1870: 257), that compete with one another. This stance is already similar to the one taken in this work. His observation, though, that

"the better, the shorter, the easier forms are constantly gaining upper hand" (ibd.) is too simplistic and altogether vague for our purposes: How does one measure goodness or easiness of a linguistic form, and how does, for instance, diphthongization of short vowels come about if it is always the shorter forms that gain the upper hand? In any case, Müller's analogy still proves a useful tool in the quest for the causes of language change.

Evolutionary theory offers an account to variation and change that does without a willful force that performs selection. Where should such a will be located, anyway! Language, however, as a product of cultural evolution, is different from biology in that it is altered and transmitted by agents that do have a will and can hence select for features on their own. Additionally, the human language capacity is most often taken as a mechanism that acts upon linguistic units and could thus in principle account for optimized outputs, whereas an ecosystem is not a mechanism that acts upon its species or individuals.

The principle that sound change comes about via selection from "a pool of synchronic variation" (Ohala 1989) has been taken up several times in more recent times, with (e.g. Blevins 2004, Wedel 2006) or without (e.g. Labov 1994, 2010, Ohala 1981, 1989, 2003) reference to the evolutionary metaphor. The accounts of Ohala and Blevins have in common, that they see the speaker as the main source for variation and the listener in the role of the selector. Selection itself is accounted for by phonological recategorization of the phonetic input. It occurs when the input is misinterpreted by means of confusion with similar sounds, hyper- or hypocorrection (Ohala 1989). There is nothing goal-directed in this view, since phonological change just happens passively as a result of the listener's categorization capacity. Blevins (2004) argues that after recategorization has taken place, the listener, in the role of a speaker, will aim for new prototypes as phonetic targets, which will then spread new variants and influence the categorization of other individuals once again. According to Blevins, the idea of changing prototypes yields a natural explanation for many classes of sound change that oftentimes serve as a justification for teleological accounts, like e.g. dispersion maximization or gap-filling 'in order to' restore symmetry.

Both accounts focus on the listener as the main source of sound change since she is the one that undertakes categorization and is therefore responsible for the make-up of phonological categories. Furthermore, they consider sound change mainly as a cognitive process of individual language users. However, the speakers are the ones to vary their speech signals, and they can do so deliberately. Both authors do not deny a teleological momentum in the choice of variants to be spread. For example, hyperarticulation in infant-directed speech or for pragmatic reasons is such an instance of deliberate preference

in selection of one variant over another. Furthermore, social factors such as low or high variant prestige may constrain the search space of phonetic variants. Language as a system that resides in a community of speakers is thus usually not taken into account in discussions about the teleology of linguistic change. In addition to that, the mind does not treat all variants equally. It filters. As already mentioned, some variants are seen as more representative of the category they belong to than others. Non-teleological accounts of sound change, such as the ones of Ohala and Blevins, are therefore really non-grammatical ones, as they see no forces manipulating categories as a whole. Nevertheless, they do not deny or even do consider teleological forces at work in general cognition (cf. e.g. Ohala 2003: 683).

From an exemplar-theoretic point of view, the distinction between grammar and general cognition vanishes, for grammar is regarded as the "cognitive organization of one's experience with language" (Bybee 2006: 711). In this context, the question whether language change is teleological fades into the background and more important becomes the question as to what extend language change is goal-directed. Do, for example, sound inventories exhibit 'optimal' properties such as symmetry or maximum dispersion because the cognitive apparatus forces their members into these states if they are 'non-optimal'? Or do these properties emerge from more fundamental cognitive processes? The thesis at hand will argue in favor of the latter position.

The evolutionary metaphor is well-suited for the work with exemplar theory. The exemplar space is populated by individual exemplars and the similarity relation between these individuals constitutes 'species', i.e. categories. Both speakers and listeners are responsible for selection and reproduction. The speaker chooses exemplars for speech production according to her communicative needs and may select some or all of them for reproduction via re-storage in her own exemplar space. The listener, then, on the receiving end, selects for some of them and stores them in his exemplar space. Note, that utterance alone is not yet reproduction in the evolutionary sense, since it is not guaranteed that neither the speaker nor the listener (re-)store the signal in their respective exemplar space. Storage depends on several factors such as communicative success, the absence of too similar and highly activated exemplars in the exemplar space, or their degree of ambiguity in categorization. Now, how does variation come about? If all exemplars were transmitted perfectly, within-category variation would hardly ever occur and the coinage of new categories would be the only source of variation. Two principles are relevant for the remainder of this work. First, fluctuation in the speech signal due to factors such as overlap of articulatory gestures, perturbations in the air and other sources of environmental noise, or varying states of the listeners perceptual organs are the norm rather than the ex-

ception. These factors can be said to relate to mutation. Second, it is assumed that exemplars 'mate' in order to construct new exemplars (Rosenbaum et al. 1993). That is, not single exemplars are chosen for reproduction, but clusters of similar exemplars. Pierrehumbert (2001) refers to this principle as *entrenchment*, Wedel (2006) calls it *blending inheritance*.

The current section outlined how sound change can be accounted for without resorting to grammar as the driving force for change, but instead by paralleling speech perception and production with concepts known from evolutionary theory. However, up to this point, we do not yet know in what manner phonetic and phonological changes permeate throughout the lexicon.

3.2 GRADUALNESS OF SOUND CHANGE

According to the Neogrammarians, sound change proceeds in minute, unnoticeable steps and all lexical instances of the changing sound in the inducing context at once. That is, they claim sound change to be phonetically gradual and lexically abrupt. Exceptions from this pattern are then usually explained by analogical change w.r.t. a paradigm different from the original one. This was the predominating view on the nature of sound change for a long time and is still the basis for a lot of work that is done in the field of phonological reconstruction.

Lexical diffusion, introduced by Wang (1969), offers a view on sound change that is complementary to the Neogrammarian one. Here, a given sound change originates abruptly in some words and may later on spread to other words as the changes become established within the individual or the speech community. That is, lexical diffusion is phonetically abrupt and lexically gradual. Under the assumption of the Neogrammarians, sound changes that affect only a subset of all words (guided not necessarily by phonetic or phonological criteria, e.g. morphology, semantics, etc.) in the lexicon would always be the result of analogy to other, already existing forms. Lexical diffusion on the other hand allows for completely novel sound changes to affect only some words. And even more: It claims that sound changes that affect the whole lexicon are the exception rather than the rule and only established at a later point via analogy. Although introduced by Wang as a substitution, lexical diffusion is often regarded as being just an alternative pathway of linguistic change that is different from Neogrammarian sound change (cf. Labov 2010).

The notions of phonetic and lexical gradualness or abruptness, respectively, allow for a typology of a total of four different kinds of sound changes. A third type is constituted by the case of sound changes that are both lexically and phonetically abrupt. These occur when a new (allophonic) variant is introduced in the speech community, for instance via language contact, and its use is not restricted

on groups of lexical items. An example for this kind of sound change was the transition from [r] to [ʁ] in a lot of German dialects (including the standard variant) through the influence of French (Janson 1983). It should be noted in this context, that the existence of two or more variants in the process of a sound change does not make a lexically abrupt change less abrupt as long as the variations are equally likely for every lexical item.

Schuchardt (1885), in his refusal of Neogrammarian sound change, claimed to have observed that sound changes are usually more advanced in lexical items of high frequency and less advanced in those of low frequency. This type of sound shift then constitutes the remaining pattern given by above typology, namely phonetically and lexically gradual sound change. This fourth type fits naturally into exemplar-theoretic models, including this one, because it arises as a by-product of the interplay between bottom-up and top-down processing in speech production. Because high-frequency items receive a lot of activation from similar exemplars with the same meaning, they can be uttered directly. Items of lower frequency on the other hand may have to be constructed from more general constituents, e.g. syllables or phonemes. This proceeds on the basis of stored constituent exemplars which are of course also influenced by high-frequency items, since constituents are only abstractions over similar lexical items. If now high-frequency items are subject to some phonetic change, they 'trail behind' constituents and along with them low frequency lexical items (Bybee 2006, 2012).

3.3 THE RESEARCH PLAN

Now, the insights gained during the last three sections will be combined in order to conduct two computational models and to understand their behavior.

The first one consists of two agents that have at their disposal a crude equivalent of an exemplar space. Both agents communicate with one another by exchanging exemplars. The speaker chooses an exemplar of a phonemic category that he wants to transmit to the listener. But he does not simply pass the exemplar. First, he entrenches it in surrounding exemplars of the same category to form a production target. Seizing the evolutionary metaphor outlined above, the new exemplar is a replication of several other exemplars and inherits from them phonetic traits. After the target is constructed, only the phonetic form is passed to the listener. In perception, the listener compares the percept to stored exemplars with a similar phonetic form in order to estimate its category. At this step, competition between categories ensues for the claiming of the percept and it will be explored under which circumstances this competition leads to the selection, i.e.

storage, of exemplars that are helpful w.r.t. the maintenance or maximization of existing phonemic contrasts.

While in the first model the exemplars, which are representatives of phonemic categories, only exist in isolation, in the second model, they will be further enriched by contextual and semantic information. Now, the new goal of the speaker is to transmit a meaning, and likewise, the listener attempts to decode it by resorting to the contextual and phonetic information consigned by the speaker. Both speaker and listener have to gain access to previously stored exemplars. This proceeds in a way that parallels the interference of top-down and bottom-up processing described in section 2.2. That is, exemplars are readily available for perception and production if they are highly activated. Otherwise, they have to be constructed or decomposed, respectively, by resorting to smaller units, in this case only a context and a phoneme. Additionally, in production, the speaker is enabled to resort to exemplars whose phonetic components are more representative of their phonemic category in order to avoid homophony. Then, the goal of chapter iii is to show that the interplay of high and low activation, bottom-up and top-down, direct and indirect access, successful and unsuccessful communicative acts, more and less frequent words and phonemes, and not least homophony avoidance, can emulate a sense of usefulness with respect to phonemic contrasts and that it is this usefulness that influences their maintenance or loss, resulting in push shifts or mergers.

Part II

THE FIRST SIMULATION

4

INTRODUCTION

In this section, an exemplar-theoretic computational model is conducted that aims to show that chain shifts may arise solely by virtue of competing exemplars during perception and production. That is, a system-level force is not necessary. This first model out of two serves as a proof of concept for the possibility of chain shifting due to exemplar dynamics. It will be expanded in the next chapter in order to explore the possible impact of lexical frequency distributions on the likelihood of chain shifts or mergers, respectively. The first model will explore experimentally the likelihood of a merger in the presence of two colliding phoneme categories as a function of their relative differences in token frequency.

Model 1 is largely based on the one conducted by Pierrehumbert (2001), which lay the foundation for many subsequent exemplar models in linguistics (e.g. Ettlinger 2007, Walsh et al. 2010), albeit with slight changes here and there, that will be pointed out during the course of the text.

The model consists of two agents that communicate with each other by exchanging exemplars. In the beginning of each simulation cycle, one is assigned the role of the speaker and the other one that of the listener. Then, the speaker chooses a label, which represents a phonemic category, that he wants to transmit and picks a vector which, to his understanding, is representative of the category. Vectors in this model are tuples of values between 0 and 1, that serve to represent abstractly some phonetic properties. In the text, they will be treated like the F_1 and F_2 formants of vowels for illustrative reasons, but it is important to note, that the model does not make any claims about specifics of auditory or articulatory features.

After the vector has been prepared for production and transmitted to the listener, the latter attempts to make sense of this signal by assigning a label to it on the basis of similar exemplars that are present in her exemplar inventory. Since there may be exemplars in the perceptual vicinity of the percept, which belong to different categories, competition for the labeling ensues among these. Competition is a crucial factor to the model, for it ultimately determines in which regions of the exemplar space exemplars of a specific category are more likely to be misperceived and which are safe on the other hand. Hence, exemplar models such as the one by Pierrehumbert employ competition in order to account for contrast maintenance or loss. In this respect, the present model is not much different.

Once a label is assigned to the percept, the listener decides whether or not she wants to store it in her exemplar space. This decision depends on the chosen categorization mechanism, of which two will be considered in this work. One relies on feedback given by the speaker regarding the choice of the label, and the other one passes every exemplar. In either case, however, the exemplar is not stored if it is either too ambiguous or already exists in the exemplar space.

The model seems, and is, reduced in many respects: exemplars are restricted to two dimensions, there is no conception of neither articulatory nor auditory factors, communicative acts consist only of the passing around of single phones, etc. But despite its limitations, the simulations will demonstrate that the model can account for chain shifts and mergers in a way that is akin to reality.

4.1 SIMULATION ARCHITECTURE

The model set-up consists of two agents, each of which possess an exemplar space. The exemplars posited in the exemplar space consist of a vector $\vec{x} = \langle x_1, x_2 \rangle \in [0,1]^2$, an activation $\alpha \in [0,1]$ and a label $l \in \mathcal{L}$. Abusing notation, $\vec{x}, \vec{y}, \ldots$ will serve as symbols for both vectors and exemplars, but it will always be clear from the context what they are supposed to represent. Furthermore, there is a production and a perception procedure, each with several sub-procedures, and a distance metric to measure the distance between the vectors of the exemplars.

The activation α is used as a weight that determines to what degree an exemplar contributes to the perception or production of similar exemplars. Its value is dependent on the age t relative to a constant τ that represents the maximum age of any exemplar. The activation is calculated by the monotonically decreasing function

$$\alpha = \exp\left(-\frac{t}{\tau}\right)$$

4.1.1 *Production*

At the beginning of the production step, a label l is chosen with probability p_l. Then, an exemplar $\vec{x}$ with label $\mathcal{L}(\vec{x}) = l$ is chosen at random from the agents exemplar space. This exemplar, however, does not directly serve as a blueprint for production. Instead, the production target $\vec{t}$ entrenched in surrounding exemplars of $\vec{x}$ that carry the same label l. This entrenchment relates to blending inheritance (cf. section 3.1). The influence of the surrounding exemplars decreases with their respective distance to $\vec{x}$. The formula for the entrenchment

of an exemplar $\vec{x}$ inside surrounding exemplars with the same label $\mathcal{L}(\vec{x})$ is given by

$$\mathit{entrench}(\vec{x}) = \frac{\sum_{\vec{y}:\mathcal{L}(\vec{y})=\mathcal{L}(\vec{x})} \vec{y} w_{\vec{y}}(\vec{x})}{\sum_{\vec{y}:\mathcal{L}(\vec{y})=\mathcal{L}(\vec{x})} w_{\vec{y}}(\vec{x})}, \tag{1}$$

where $w_{\vec{y}}(\vec{x})$ is the activation-weighted distance given by

$$w_{\vec{y}}(\vec{x}) = \frac{\alpha_{\vec{y}}}{\mathrm{win}(|\vec{x}-\vec{y}|)+1}. \tag{2}$$

Here, $\alpha_{\vec{y}}$ is the respective activation of $\vec{y}$, $|\vec{x}-\vec{y}|$ is the Euclidean distance between $\vec{x}$ and $\vec{y}$, and win is a window function given by

$$\mathit{win}(x) = \begin{cases} x, & x \leqslant \mu/2 \\ 0, & \text{otherwise} \end{cases}, \tag{3}$$

where μ represents the window size, that is the maximum perimeter inside of which exemplars surrounding $\vec{x}$ contribute to the construction of the production target $\vec{t}$.

After the entrenchment procedure, the resulting vector is dislocated randomly inside a fixed radius δ. This amount of noise represents both the variation in production and perception and added only once for simplicity's sake. At last, x_1 and x_2 are rounded to 0 or 1 if they are smaller than 0 or greater than 1, respectively.

The construction of a target $\vec{t}$ from an exemplar $\vec{x}$ is summarized in the following formula.

$$\vec{t} = \mathit{restrain}(\mathit{entrench}(\vec{x}) + \delta), \tag{4}$$

where $\mathit{restrain}$ is the function that keeps $\vec{t}$ within $[0,1]^2$ and δ is some amount of noise.

The choice of a maximum distance differs from Pierrehumbert's model which uses the k nearest neighbors of $\vec{x}$ instead. The difference is merely due to computational considerations and has no theoretical implications whatsoever. The underlying data structure in which the exemplars are stored is a quadtree and querying for exemplars within a given radius proceeds faster than for the k nearest neighbors of a given exemplar. Even if both variants produce different results they are of so little impact that either choice is equally reasonable.

In order to account for the initial shift, exemplars of one pre-specfied label are additionally dislocated into a certain direction by an amount λ. Which label is chosen depends on the scenario to be modeled. λ is added before $\mathit{restrain}$ applies.

Listing 1 concisely outlines the essential landmarks of the production procedure just described using a pseudocode notation.

```
# receive a random label
l = getRandomLabel()

# fetch a random exemplar marked for this label,...
x⃗ = getRandomExemplarByLabel(l)

# entrench the exemplar in surrounding exemplars
# with the same label
t⃗ = entrench(x⃗, l)

# apply lenition if l is marked for it
if (markedForLenition(l)):
  t⃗ = t⃗ + λ

# add noise
t⃗ = t⃗ + δ

# push the exemplar back into its boundaries
t⃗ = restrain(t⃗)

return t⃗
```

Listing 1: Pseudocode of the production procedure of model 1.

4.1.2 *Perception*

The goal of perception in this model is to assign a label l to an incoming vector $\vec{t}$. In order to achieve this, $\vec{t}$ receives activation from surrounding exemplars that are already stored in the hearer's exemplar space. The activation for each label determines the probability with which the incoming vector is assigned the respective label. This is different from Pierrehumbert's model that uses the majority rule to determine phoneme identity, that is, the label with the highest activation is assigned deterministically rather than probabilistically.

The probability $\phi(\vec{t}, l)$ with which $\vec{t}$ is assigned label l is given by

$$\varphi(\vec{t}, l) = \frac{\sum_{\vec{y}:\mathcal{L}(\vec{y})=l} w_{\vec{y}}(\vec{t})}{\sum_{\vec{y}} w_{\vec{y}}(\vec{t})}, \tag{5}$$

where $w_{\vec{y}}(\vec{t})$ is calculated the same way as shown in formulas 2 and 3 above in the production procedure. Both are repeated in the following for convenience.

$$w_{\vec{y}}(\vec{t}) = \frac{\alpha_{\vec{y}}}{\text{win}\left(\left|\vec{t} - \vec{y}\right|\right) + 1}, \tag{2 rev.}$$

$$win(t) = \begin{cases} t, & t \leqslant \mu/2 \\ 0, & \text{otherwise} \end{cases}. \tag{3 rev.}$$

Different from Pierrehumbert (2001), the same perimeter μ of contributing exemplars is drawn in perception and production. Pierrehumbert, instead, uses a maximum distance in perception and the k nearest neighbors in production. This difference is eliminated in the current model, since there seem to be no apparent theoretical reasons for this decision.

At last, it must be decided whether or not $\vec{t}$ is added to the exemplar space. With regard to this, Tupper (2014) distinguishes three decision protocols, which he refers to as *categorization regimes*.

NO COMPETITION The percept is always assigned the label of the category from which it originated, and is always added to the lexicon.

PURE COMPETITION The percept is assigned the label resulting from the evaluation procedure, and is always added to the lexicon. (citing Blevins and Wedel 2009).

COMPETITION WITH DISCARDS The percept is assigned the label resulting from the evaluation procedure, but only added if the label matches the label of the category from which the percept originated (citing Labov 1994).

Of course, *no competition* is the least plausible one, since it implies that even the most confusable percepts are always added to the inventory, which ultimately leads to the collapsing of all categories. Under *pure competition* on the other hand, the label needs to be evaluated first using a procedure like the one just described. It relies on the assumption that competition among categories is a strong enough factor to keep exemplars of different categories apart. This is the one used in quite a number of the exemplar models dealing with phonological contrast maintenance (e.g. Pierrehumbert 2001, Wedel 2006, Ettlinger 2007, Blevins and Wedel 2009). *Competition with discards* is different from the former two regimes, because it introduces feedback from the speaker as guidance for categorization.

No competition seems pointless at first sight, but it incorporates something that the other two regimes lack, namely top-down reasoning – albeit in an overly exaggerated manner. In the present form of the model, there is no way of inferring phoneme identity from the lexical context of the percepts, because there is none. Competition is entirely based on neighborhood to other phonemic exemplars. No competition constitutes the opposite case, where the lexical context in every single case repairs misperceived or mispronounced items, hence rendering phonology entirely purposeless.

In order to emulate some notion of top-down reasoning in the process of phoneme identification, a success rate σ is introduced. It assures that in $\sigma \times 100\%$ of all cases the phonemic identity of $\vec{t}$ is cor-

rectly recognized. This treatment results in hybrids between no competition and any of the two other regimes, where $\sigma = 1$ is no competition proper, and $\sigma = 0$ yields pure competition or competition with discards, respectively. But although both of the latter can be enriched with a success rate, the simulations below will only employ it in the context of pure competition.

The present model incorporates another criterion for the addition of a percept to the exemplar space, even if $\vec{t}$ passed the tests posed by the chosen categorization regime. Following Wedel (2006), the decision whether or not a percept is added to the exemplar space depends on the certainty with which a label was assigned to it. However, the probability that a percept is *not* added is kept very low and is calculated by

$$\psi(\vec{x}, \mathcal{L}(\vec{x})) = \max(0, 0.4 - \phi(\vec{x}, \mathcal{L}(\vec{x}))),$$

i.e. even when the assignment of the label $\mathcal{L}(\vec{x})$ of $\vec{x}$ was successful in spite of a very low probability, the probability that $\vec{x}$ is not added to the exemplar space never falls short 0.4.

Finally, the percept is added to the exemplar space, but with one further restriction. If an identical exemplar already resides there, $\vec{t}$ is discarded, albeit not entirely, because the age of the other exemplar is reset to $t = 0$. Due to the fact that it is very unlikely that two vectors accidentally share the same values for both x_1 and x_2, identity is determined by a parameter jnd. If two exemplars reciprocally lie within a radius of jnd, they are perceived as identical. For reasons of implementational ease, only the first encounter of an identical exemplar is taken into account in perception, i.e. neither the most proximate one nor all exemplars within this radius.

As before, a summary of the perception procedure is provided in pseudocode-style in listing 2.

```
fetch t⃗
/
# assign a random label considering the joint
# activation of surrounding exemplars
l = getRandomLabel(probability(t⃗,ℒ))

# decide randomly whether the exemplar is to be discarded
# depending on the certainty of its categorization
if (isToBeDiscarded(probability(t⃗,l))):
        discard t⃗
# or else try to add the exemplar based on the chosen
# categorization regime
else:
  if (noCompetition):
     # discard l and choose the original label
     add new Exemplar(t⃗,ℒ(t⃗))
  else if (pureCompetition):
     # add the exemplar with the evaluated label l
     add new Exemplar(t⃗,l)
  else if (competitionWithDiscards and l = ℒ(t⃗)):
     # add the exemplar only if l matches the original
     # label
     add new Exemplar(t⃗,l)
  else:
    discard t⃗
```

Listing 2: Pseudocode of the perception procedure of model 1.

5 THE SIMULATIONS

Now that the basic architecture is established, the simulation will be run and assessed under several parameter settings. The first run in section 5.1 serves the goal of getting to know the basic behavior of the model. Furthermore it will introduce to the reader its multiple parameters and options. After a foundational understanding is achieved, the influence of the model parameters on the occurrence of chain shifts or mergers will be explored in section 5.2. Given the limited space of the medium, the reader is encouraged to fiddle about with the model on their own. The Java™ source code is available under `github.com/vpersien/csinet`. In the final section 6 of this chapter, the results as well as the advantages and disadvantages of the model will be discussed in detail.

5.1 SETTING THE STAGE

The first aim of this section is to find a subset of parameter settings which result in a sufficiently stable behavior of the system. The influence of the remaining parameters on the simulation's performance will then be examined in the further course of the discussion.

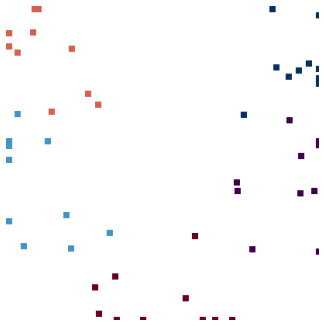

Figure 2: The initial states of all the simulations within this chapter will resemble the basic gestalt of this 5-vowel system.

5.1.1 *Fixing parameters*

Unlike studies that are concerned with the evolution of phonology from scratch (e.g. de Boer 1999), this work studies pathways of change within phonologies that are well-established. It is thus fair to assume that the initial state of the system already resembles existing vowel inventories. Throughout all simulations of model 1, the role model will be the 'classic' five-vowel system with one low vowel, two mid

and two high vowel categories. The respective labels will be called – from left to right, top to bottom – **a**, **b**, . . . , **e**. Each simulation is seeded with ten exemplars per label which deviate randomly from the points $\langle 0.1, 0.1\rangle$,$\langle 0.9, 0.1\rangle$, $\langle 0.1, 0.55\rangle$, $\langle 0.9, 0.55\rangle$ and $\langle 0.5, 0.9\rangle$, each according to its label, where the top-left-most coordinate is $\langle 0, 0\rangle$. This basic set-up is exemplified in fig. 2.

Since the initial situation is relatively stable and it is not the goal of this thesis to evaluate how sound shifts come about in the first place, the movement of the members of one category will be assumed a priori. The simulations are to model the transition from one stable state into another stable state, where the state transitions will be fueled by an initial shift towards a pre-defined target point $\vec{\gamma}$. As already mentioned in 4.1.1, this movement will be achieved by additionally dislocating the exemplars of the label marked for 'lenition' towards $\vec{\gamma}$ by an amount λ.

A requirement for the model parameters is that the system remains stable within both states and state transitions. Stability in this context does not imply the absence of motion but refers mainly to the cohesion of the exemplars within a category. These should neither spread unboundedly nor exhibit too much unprovoked overlap. In a static configuration with $\lambda = 0$, i.e. no initial shift, this should hold particularly true. Additionally, in this case there should be no significant movement of the categories.

$\lvert\mathcal{L}\rvert$	5	Number of labels (categories)
λ	0.03	Amount of lenition of the shifting category
δ	0.5	Amount of noise in production
μ	0.8	Window size for activation and entrenchment
τ	4,000	Maximum exemplar age
jnd	0.06	Just noticeable difference

Figure 3: Parameters that will be held constant throughout the simulations.

Figure 3 gives an overview over the parameters that will be kept constant throughout the simulation runs. This configuration stood the test of stability mentioned above.

λ is set to 0.03 for two reasons. First of all, a greater amount would run the risk of pushing the exemplars of the label which is marked for lenition too prematurely into 'enemy territory' and thus provoke a merger where otherwise a chain shift would have occurred based on frequency alone. Given this observation it seems reasonable to investigate empirically whether the velocity of an initiating sound shift increases the likelihood of a merger to happen. Nevertheless, this is a question for future research, whereas in this work the influence of token frequency, the perception procedure and other parameters

on chain shifts is to the fore. And secondly, λ is not chosen smaller for pragmatic reasons: Such an amount would slow down the process of the simulations unnecessarily.

δ and μ exhibit much interaction. Without noise, there would be no source of variation, but as soon as noise is present, it must be kept in line somehow, e.g. via entrenchment, to prevent exemplars from spreading uncontrollably, as remarked by Pierrehumbert (2001). δ is chosen to be greater than half the window size, for otherwise variation would be constrained to much, resulting in vast uninhabited spaces between categories. But if its value is too large in relation to μ, the effect of entrenchment diminishes.

Apart from its relation to δ during production, the window size μ also affects the range within which exemplars are considered when evaluating the probabilities for the assignment of labels. Too large or too small a value amplifies the effects of the chosen categorization regime regarding contrast maintenance or loss, yielding unsolicited or just unrealistic results.

The maximum age τ of any exemplar has an impact on the immutability of categories. The higher the value the longer categories are prevented from change, the lower it is the less stable categories become during the transition phase. Hence, $\tau = 4,000$ proved a reasonable choice.

The parameter jnd only has a very shallow relationship to its eponym. Its principal purpose is to help define the identity relationship that is used during perception. However, it affects the maximum density of the exemplar cloud, with the effect that a large value results in very sparse categories, which can avert mergers altogether.

Now that the majority of the parameters have been fixed, a first run of the simulation will be carried out.

5.1.2 *A first run*

In the last section some of the model parameters were fixed to increase comparability among the model runs. Now a simulation will be run for the first time to see how the model behaves in a dynamic setting. The parameters that are being used during this run are summarized in fig. 4.

The success rate σ is set to 0.9 for this time. To recapitulate, it determines the ratio of exemplars that are assigned the right label, i.e. the one that matches with the one of the exemplar chosen for production, regardless of the influence of surrounding exemplars in perception. Note that otherwise categorization would solely depend on the distribution of exemplars and their respective label in the vicinity of the incoming signal. But in reality it is also influenced by the context it is uttered in, thus allowing for more phonetic variability. The influence

σ	0.9	Success rate
$\vec{\gamma}$	$\langle 0.1, 0.1 \rangle$	Target position of the initial shift
$p_{\mathbf{a}}$	0.0666...	Utterance probability of **a**
$p_{\mathbf{b}}$	$3 \times p_{\mathbf{a}}$	Utterance probability of **b**
$p_{\mathbf{c}}$	0.2	Utterance probability of **c**
$p_{\mathbf{d}}$	0.2	Utterance probability of **d**
$p_{\mathbf{e}}$	0.2	Utterance probability of **e**

Figure 4: Parameters of the run of model 1.

of this parameter on categorization will be explored more thoroughly in the next section.

The success rate will only play a role in the context of the pure competition regime, the one picked for this run of the simulation. The differences in behavior under the present regime and competition with discards will be examined thoroughly in the next section.

The label whose exemplars undergo the initial shift is **b**. These will be lenited by an amount of $\lambda = 0.03$ in the direction of $\vec{\gamma} = \langle 0.1, 0.1 \rangle$, which is the original mean position of **a**. Furthermore, exemplars of **a** are chosen with only a probability of $p_{\mathbf{a}} = 0.2/3$ whereas the exemplars of the incoming label **b** appear with three times as high a probability. This proportion conforms with Pierrehumbert (2001) where a difference in the utterance probabilities was the main reason for a merger to happen. Below we will see whether this holds true in the present setting, too. The exemplars of the remaining labels are produced with a uniformly distributed probability of $p_{\mathbf{c,d,e}} = 0.2$.

Results

The pictures in fig. 5 show the exemplar space at four different time steps. The plot shows both the x_1 and the x_2 values of the activation-weighted mean of each category. What can be seen is that the exemplars with label **b**, originating in the upper right hand corner, move rapidly towards the space inhabited by the exemplars of label **a**. At time step 30,000 this seems to result in an almost complete overlap. But the third panel shows that category **a** swerves to the bottom of **b**. In fig. 6 it can be seen that **a** and **b** never exhibit a complete overlap in the sense that both share approximately equal activation-weighted mean values of both their x_1 an x_2 components. But eventually, in the fourth picture, category **a** has recovered and now occupies a space at the upper right hand side of the exemplar space. This makes the impression of a push shift, but at closer inspection the results are ambiguous. There was a period were the exemplars of **a** were almost completely covered by other categories, most prominently **b**. This model does not offer a mechanism for recategorization but in reality

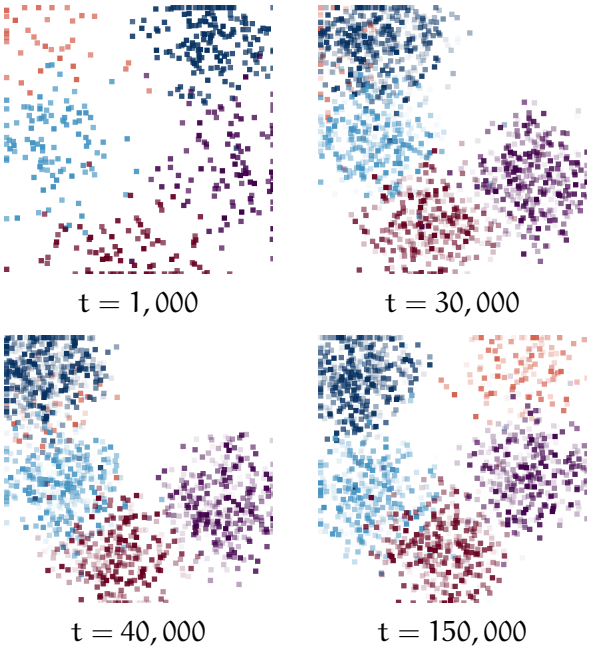

Figure 5: Change of the exemplar space during the first run of the simulation.

a child that would have learned a language with such a distribution during this period would have had little reason to assume a separate category for the exemplars of **a**. Furthermore, the recovering was only possible because there was never a point in time where there were no exemplars with label **a** present. This is due to the success rate which ensures categorization for every label in 90% of all cases. And lastly, this might have happened partly as a consequence of the implementation. The lenition towards the target position never terminates. This causes the distribution of exemplars of label **b** to be very dense and they will never be as widely distributed as the ones of the other labels. In the next section, more evidence for this conjecture will be given and we will likewise encounter more clear-cut cases of push shifts.

Another point of interest regards the situation of the three categories at the bottom of the exemplar space. Category **d** to the right moves downwards immediately. As we have seen, categories with a lower token frequency move away from those with a higher one. This is also the case for **d** as there are more tokens with label **b** present in its vicinity. But as soon as category **b** is sufficiently far away, **d** moves up again. This is signified by an increase in its x_2 value at around time step 30,000 in fig. 6. We do not know at this point whether **d** would have moved up further if **a** had not occupied the upper right corner. Another explanation seems more plausible, though. The categories in the lower half of the exemplar space, all of which exhibit the same utterance probability, do not show much reaction to the movements of their respective neighbors. On the other hand, a much stronger

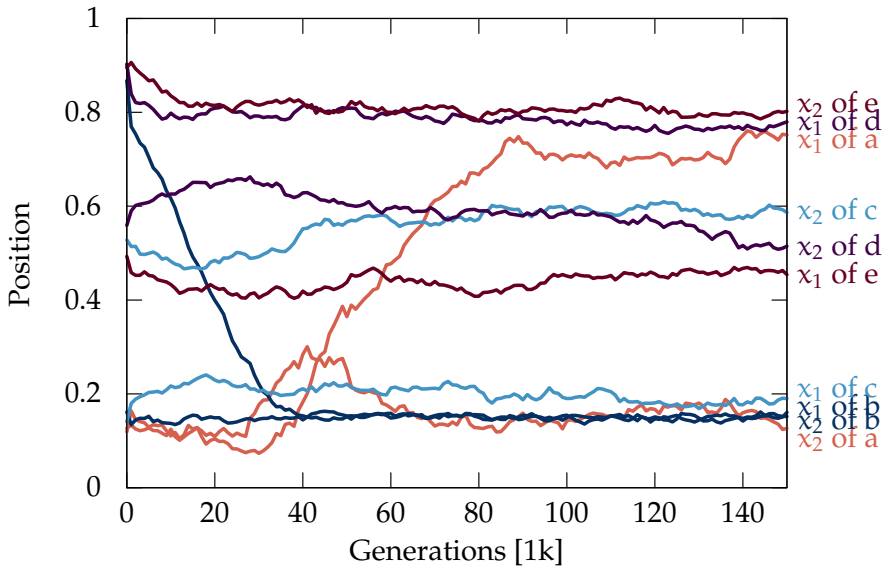

Figure 6: Development of the activation-weighted mean x_1 and x_2 values of all categories over 150,000 generations.

connection can be observed between the movements of **d** and **a**. The increase in **a**'s x_1 value is accompanied by a decrease in the x_2 value of **d**. Again this is signified in fig. 6 from time step 30,000 onwards. This connection becomes clearer at around time step 135,000. It seems that under the present circumstances, larger categories tend to follow smaller ones. In the next section we will see that this is due to the categorization regime, i.e. pure competition, chosen for this run.

Conclusion

The present section provided us with an initial insight in how the model behaves under iteration. It could be seen how the interplay of several settings lead to the coming about of a pull shift as a consequence of dispersion maximization. In the next section, we will expand on this finding by exploring how the choice of the categorization regime influences contrast maintenance between categories. However, given the current conditions it is not possible to unambiguously differentiate between push shifts and mergers. But a comprehensive model of chain shifts in particular and exemplar-theoretic categorization in general should be able to account for pull shifts as well as push shifts and mergers. In the next section it will be shown that the right choice of the categorization regime can help to meet these requirements.

5.2 EXPLORING THE PARAMETER SPACE

The unsatisfying results of the first run of the simulation call for a search for reasons and solutions. In order to achieve both goals, first, the behavior under the pure competition regime is further examined by means of varying the success rate. Afterwards, it is investigated how and if a change of the categorization regime can bring about more fruitful outcomes.

5.2.1 *Altering the success rate*

In this subsection, the effect of the success rate on the emergence of mergers will be explored. The success rate itself allows for a hybrid of the no competition regime, where the hearer automatically chooses the same label as the one chosen by the speaker, and of the pure competition regime, where categorization only depends on the activation of the surrounding exemplars and their respective labels. Blevins and Wedel (2009) made the competition between exemplars responsible for the formation of relatively strict borders between categories. Tupper (2014) on the other hand was not able to replicate their result. In his model, the categories either merged or moved around indefinitely, depending on a model parameter that alters the influence of the density of the categories on selection.

Now it will be evaluated in which way an increased trend towards straight pure competition alters the behavior of the system. We will see that this intensifies the anomalies observed in the last section, hence making pure competition completely unsuitable for the modeling of chain shifts in the context of the present model.

Results

At first, the model is run with the same parameter settings as before, except for the success rate, which is set to $\sigma = 0.5$. Fig. 7 shows a selection of screenshots of the exemplar space between $1,000$ and $400,000$ iterations..

Three of the events shown in fig. 7 are especially noteworthy. First of all, the recovering of category **a** that was also to be observed in the previous section could be replicated (cf. the upper right hand side on the third panel). Second, the hunt of **d** for **a** could be replicated as well (cf. panel four). And lastly, categories **c** and **e** exhibit complete overlap, move in concert from then on, and even jointly chase **a** (cf. panels four through six).

Just as in the run with $\sigma = 0.9$, the occupation of **a**'s initial position by exemplars with label **b** led to a flight of category **a** to the upper right hand corner of the exemplar space, depicted in first three panels of fig. 7. Furthermore, the second panel shows an intermittent trend of **b** towards the refuge of category **a**, i.e. some of the exemplars that

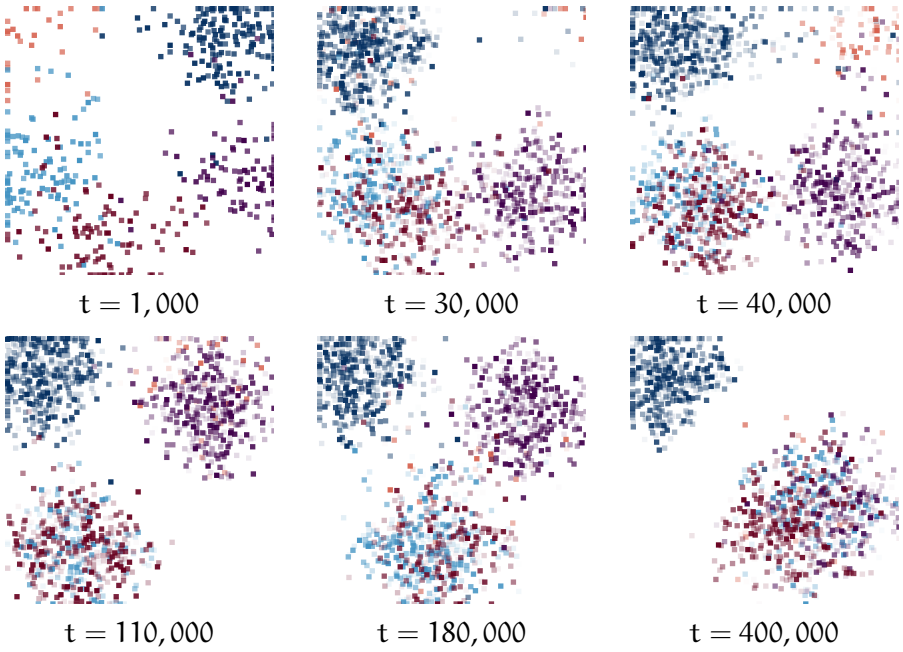

Figure 7: Alteration of the first simulation with $\sigma = 0.5$.

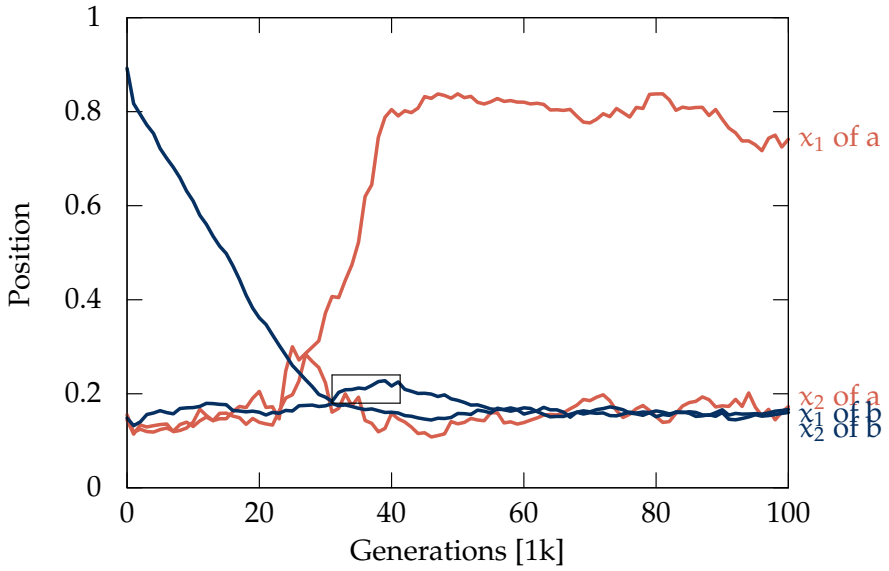

Figure 8: $\sigma = 0.5$. Development of the mean values for x_1 and x_2 of **a** and **b** during 100,000 cycles. The box indicates the time frame within which category **b** drifted along with **a** towards the right of the exemplar space.

were labeled as **a** by the speaker have been claimed by **b** so that more of its labels were able to accumulate in this area for a short period of time. Nevertheless, **b** eventually retreated to its intended target position. The whole chain of events can be retraced in more detail via fig. 8.

To understand this behavior, one has to think again about how the perception-production loop proceeds. During perception, the categories compete with one another for the labeling of the percept. The label is assigned by chance but depends on the activation and amount of the exemplars that carry it already. In 50% of all cases, though, competition is circumvented and the perceived exemplar is assigned the label of the category it originated from. This assured labeling prevents categories from dying out in the sense that there are no exemplars in the exemplar space anymore that are labeled for the respective category. If two categories overlap and their token frequency is *in* proportion, both will stay in place because assignments to either category are equally likely. However, this is not the case for **a** and **b** in the above case, as their token frequency is out of proportion. For each time interval, label **a** has a minimum guarantee of $\sigma \times p_{\mathbf{a}} \times \min(\psi(\vec{x}, \mathbf{a})) \approx 1.33\%$ per tick for claiming an exemplar $\vec{x}$. Even if it completely covered by **b** and the probability of winning competitions drops to near zero, this amount is fix. These guaranteed exemplars will fall inside and outside of the bounds of the larger category. And it is the latter case that leads to an increase of the claiming probability for **a** for two reasons. First of all, for newly perceived exemplars that fall outside of **b** and in the vicinity of the outsiders of **a**, the probability of being assigned to **a** is increased. This is especially true for exemplars that fall to the right of these outsiders. The second and above that major contributing factor is entrenchment. The increase in probability for categories of label **a** to appear further to the right changes the mean value of x_1 of this category and thus the target position of newly uttered exemplars of **a** shifts to the right. But there is more to it. As can be seen in panel 3 of fig. 7, as a consequence of **a**'s shifting to the right, more and more exemplars of **b** are observable on its original outside. This is due to the still very high assignment probability for exemplars that fall into this region. This causes a shift of the mean x_1 value of **b** to the right. This is visible in the marked area of fig. 8 between about step $85,000$ and $120,000$. But this shift proceeds slower than the shift of **a** because more tokens contribute to its mean value, i.e. category **b** exhibits a higher 'inertia'. In this simulation, such a succession does not occur since lenition still applies to exemplars of **b**. In further simulations, though, that are not documented at this point, this conjecture could be confirmed: all else being equal, if both categories share the same initial position and lenition is set to $\lambda = 0.0$, both categories oscillate between the endpoints of the x_1-axis.

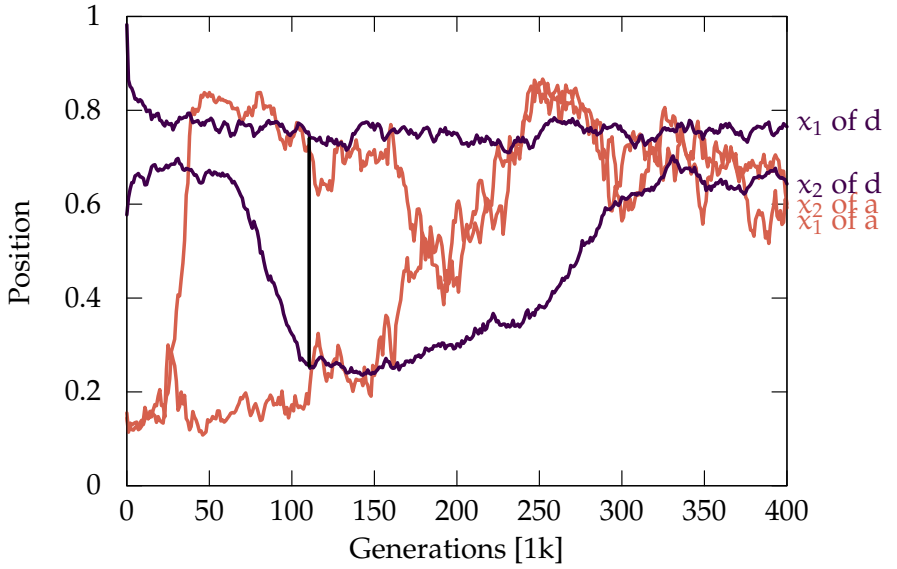

Figure 9: $\sigma = 0.5$. Development of the mean values for x_1 and x_2 of **a** and **b** over a time span of 400,000 cycles. The line on the left indicates the point in time at which **d** covered **a** for the first time.

The above discussion also gives an explanation for the other two phenomena. The superficially observable chase of category **d** after **a** mirrors the behavior of **b** in the third panel of fig. 7: newly uttered exemplars of **a** are assigned label **d** in a lot of cases when they fall in the neighborhood of this category. This causes it to move upwards. But other than exemplars of **b**, those labeled **d** are not restricted by lenition towards a fixed point so that both categories eventually collide. The course of this is depicted in fig. 9. As soon as x_1 of **a** begins to increase, the value x_2 of **d** starts to drop. First rather hesitantly, but shortly after the mean of **a** is situated in the upper right hand corner, said value of **d** decreases rapidly. Then, after a short period of coverage of **a** by **b** between around 110,000 and 160,000 cycles, the game begins anew and more exemplars labeled **a** crop up around the center of the exemplar space. As a reaction to that, the mean of **d** follows behind, but slower since it is more inert than **a**.

The merger of **c** and **e** was initiated by the downward movement of the former which was caused by the arrival of **b** in its vicinity. Further simulation runs have shown that partially overlapping categories tend to merge over time. This is more likely to happen when the success rate is low, that is, this behavior is facilitated by pure competition. Once merged, both categories won't split anymore except possibly through the influence of other categories. When exemplars of label **a** draw nearer, both categories drift towards their position on their own but seen superficially they do so jointly because of their approximately equal inertia.

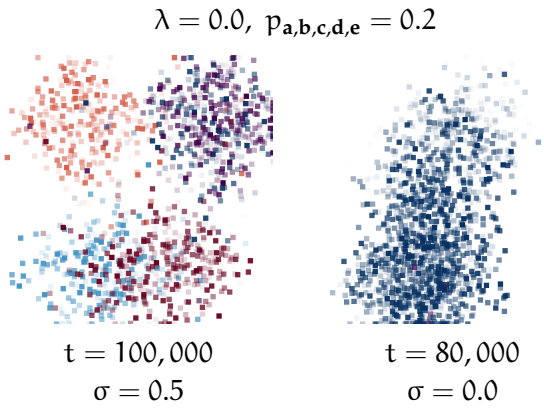

Figure 10: Random category mergers occurring under pure competition.

This undesired behavior of unprovoked category overlapping is not due to chance but constitutes a common theme under pure competition. Categories tend to exhibit partial or even complete overlap if there is no form of top-down processing involved in the assignment of category labels. Fig. 10 gives examples of the system's behavior when there is no initial shift involved and the probabilities of all five labels are equally distributed. In the second case the success rate is additionally set to 0.0. The second panel is especially remarkable, as in this one almost all exemplars are labeled **a**, except for five which are not visible on this picture. This observation justifies the hypothesis that contrast preservation in this setting is fueled by the existence of a success rate.

Conclusion

The above simulations showed almost the exact opposite behavior that one would expect based on the discussion in section 1. Categories with unequal token frequencies moved about in a chain shifting manner, whereas those of equal token frequency exhibited mergers in the sense of a complete overlap. Lowering the success rate allows for longer lasting coverings of the smaller category by the larger one[1] but on the other hand it makes the system as a whole extremely unstable and may eventually lead to the wipe-out of all but one category as demonstrated in the second panel of fig. 10.

To be fair, it is not quite obvious how the parameters of the model relate to real-world phenomena. What time span does a single tick cover in comparison to real time, for instance? Are 5,000 ticks of nearly complete overlap of categories enough to not make a difference during language acquisition? If yes, then at least mergers can

1 A comparison of figs. 6 and 8 yields an instance of this difference. Note how the increase of category **b**'s mean x_1 as a reaction to **b**'s rightward shift between 40,000 ~ 50,000 ticks and 30,000 ~ 55,000 ticks, respectively, does not only cover a greater time span but, as a consequence of this, is also more intense.

be modeled adequately given a sufficiently high success rate, albeit push or pull shifts only by tendency.

The negative results of this section are reproducible under various parameter settings, what makes pure competition less suitable for our purposes. That is not to say that it is to be abandoned altogether. After all, in Pierrehumbert (2001), Blevins and Wedel (2009) and other already cited works, it was responsible for contrast maintenance. A reason for this apparent dissonance could be the chosen evaluation procedure. In Tupper (2014) as well as in this work, categorization is based on chance, whereas e.g. Pierrehumbert implemented the majority rule, which always assigns the most probable label to the percept. At least, the comparison of the differences between pure competition and competition with discards will provide us with insights as to why the performance witnessed above arises in the context of the present model.

5.2.2 *Competition with discards*

Competition with discards is the third of the categorization regimes mentioned in 4.1. Under this one, categories compete for the labeling of exemplars just like under no competition, but instead of unconditionally adding exemplars to the inventory after a label is assigned, competition with discards furthermore demands that the assigned label meets the label of the target exemplar in production.

In this section it will be shown that competition with discards can account for mergers and pull shifts, as well as push shifts under certain conditions. In contrast to the simulations above, mergers in this case are not characterized by mere overlapping, but instead by the extinction of the category label of one of the two partaking categories.

Results

The first run serves the purpose of finding out how the model under this novel regime handles the scenario that we saw in the last section. Again, exemplars of **b** are lenited towards the initial position of **a** and exemplars of the former have three times as high an utterance probability as those of the latter. This time, however, the success rate is set to $\sigma = 0.0$, i.e. there is no top-down reasoning involved but only positive feedback from the speaker. The results of this simulation are given in figs. 11 and 12.

Two major events can be observed in fig. 11. First, after about 50,000 cycles there are no more exemplars labeled **a** existent anymore. And second, the other three categories dispersed towards the corners of the exemplar space. In particular, **d** filled the gap left behind by **b**.

As the exemplars labeled **b** draw closer to their target position at the upper left hand side of the exemplar space, the exemplars of **a** are driven into the corner (first panel of fig. 11), i.e. more newly uttered

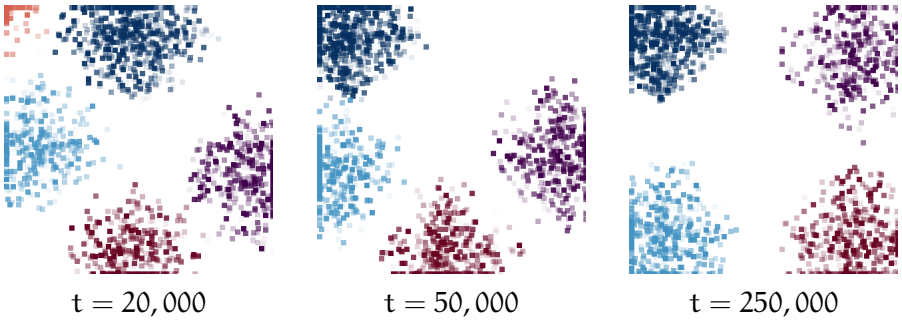

Figure 11: Merger and dispersion occurring under competition with discards.

signals are correctly assigned their original label **a** in this general area. The closer a signal that stems from **a** is located to $\langle 0, 0 \rangle$ the more reliably it is given this label in perception. This point is the safest for members of this category to surface because of the inhibiting influence of the more token-rich categories to its right and bottom. This cornering is amplified by one more property of the model, namely the absence of a success rate that would guarantee the appearance of exemplars of this category in less safe areas. The circumstance that there is no success rate involved also allows for the complete wipe-out of exemplars with label **a** (cf. panel two of fig. 11 and time step 50,000 in fig. 12), since now the categorization probability can drop to zero and hence prevent produced members of **a** to be perceived correctly anymore.

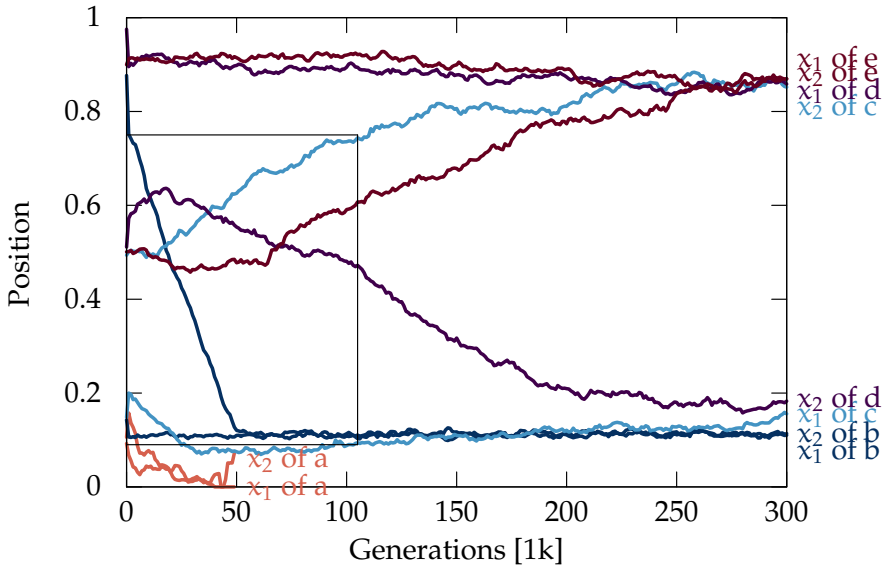

Figure 12: Development of the means of x_1 and x_2 for all categories over a time span of 300,000 cycles under the 'competition with discards' regime. The box on the left marks the time frame that is most characterized by mutually dependent movements of the framed categories.

The second noteworthy phenomenon that occurred during this run is the movement of the other categories, especially the apparent pull shift **d** → **b**. But it turns out that this gap-filling cannot be decidedly called a pull shift. The box in fig. 12 marks the relevant time frame and values that are primarily involved in the sequence of events. As the mean x_1 value of **b** drops, the x_2 value of **c** starts to increase at around 15,000 ticks. At about the same time, **d** starts to displace its mean coordinate upwards. But this alone would never cause it to reach the initial position of **b**. If it were not for the movement of **e**, the combined distance between **c** and **d** in this configuration, i.e. **b** in the upper left and **e** in the bottom center position, would be approximately 1 (conservative rule-of-thumb estimate). This state is reached before the 220th cycle. Instead, **e** does move and it is this movement that causes less exemplars of **d** to appear near its south side and eventually pushes it to the top. Likewise, category **e** is not only pushed towards the right but also pulled into this direction by the departure of **d**.

From a rather abstract point of view, the main mechanism that accounts for sound shifts is contrast maximization. As soon as an equilibrium of the system is disturbed, its elements cooperate to re-establish a stable state. But this observation is very superficial as it seems to imply teleology as being at work. Instead, it is the interplay between the chosen window size, the amount of noise and the positive feedback introduced by competition with discards.

In the last simulation of this chapter the effect of the advancement of **b** towards **a** under equal utterance probabilities for all labels will be examined. If this results in a push shift, this would mean that all of the relevant sound change types under consideration in this work can be accounted for by the current configuration of model 1.

In fig. 13 it can be seen, that a push shift did occur, but it is a quite unorthodox one. The second panel shows that **a** is cornered again by **b** but this time it does not vanish because both categories share the same utterance probability. But it does not move downwards, either, before category **c** has moved away sufficiently far. And at this point in time, all the other categories have already shifted to their final positions. This means that it is not category **a** that drives **c** downwards but the joint work of both **a** and **c**. Finally, the exemplars of **a** reach the center of the exemplar space due to the dispersion maximizing properties that we already encountered in the last run.

The cornering is certainly caused by the chosen shape of the exemplar space. Even the most concrete resemblances of exemplar spaces, e.g. the oral cavity, do not exhibit such properties. Therefore it can be regarded a glitch resulting from the abstractions that needed to be made and as such it is not worth being discussed in any more detail at this point. Especially, because the subsequent shifts have shown that moving categories do push each other around.

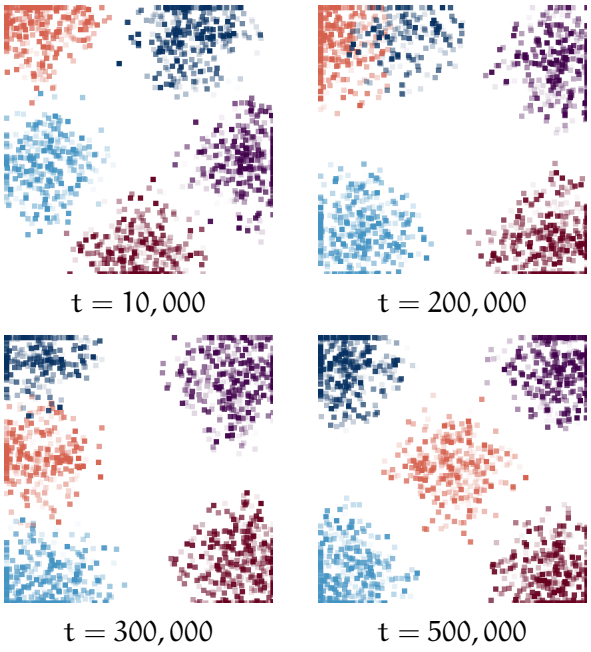

Figure 13: $p_{a,b,c,d,e} = 0.2$. A push shift occurring under competition with discards.

More potential bears the fact that, due to the triangle that is formed by **a**, **b** and **c**, the third one moves as a consequent of the former ones' joint activation. This contrasts with the traditional view on chain shifts as ordered sequences of individual sound changes. If models of exemplar dynamics such as this one are viable conceptions of how language is stored, processed and changes, the above process might shed new light on how chain shifts and sound change in general proceed.

The above results show that exemplar models can signify behaviors that resemble push and pull shifts as well as mergers due to differences in token frequency and that this ability largely depends on the way in which they account for the storage of exemplars. A thorough discussion of the results of this chapter will be provided in the subsequent section.

6

DISCUSSION

The goal of the current chapter was to find out how chain shifts and mergers can be accounted for within a simple model of exemplar dynamics that is based mainly on the exchange of phonemes between agents in a population of two. The behavior of the system was explored under the influence of two different categorization regimes, an altered version of pure competition on the one hand and competition with discards on the other hand. It became apparent that the desired phenomena occurred under competition with discards but could not unambiguously be observed under the other regime.

Pure competition proved difficult mainly because categories become less stable with an increasing amount of 'purity'. Competition in general was introduced to explain contrast preservation within exemplar theory. It is obvious that, if there is no sort of competition involved and percepts are always put into the category they originated from, overlapping categories will retain their overlap forevermore. The addition of competition should in principle be able to find a remedy by favoring exemplars that are further away from the inhibiting area of influence of other categories' exemplars and consequently lead to the dispersion of categories into more safer areas of the exemplar space. But the opposite behavior arises under pure competition: Categories are drawn closer to each other instead – even if they did not exhibit any overlap at the outset.

Co-responsible for this oddity could be another component of the present model, namely the selection procedure for labels. Before a label is assigned, the probability for each label is evaluated based on the activation-weighted distance of the percept to members of other categories. A label is then assigned randomly with respect to the resulting probability distribution. Exemplars that fall somewhere between two category means and have exemplars of both inside their window during evaluation will always have a chance to be mis-categorized and added to the inventory. Consequently each category will contain alien exemplars within their respective clusters. This will draw both categories together because there are more exemplars available for (mis-)categorization between both categories than in the opposite direction.

If on the other hand the assignment is always based on the most probable label, adjacent categories will presumably not merge without any predefined incentive because the area of 50% assignment probability (in the two-category case) serves as a strict border for mis-categorizations. It remains to be ascertained which labeling pro-

cedure is the one at work in human cognition. In any case, however, the combination of pure competition and labeling by chance proves a mischievous choice.

On the contrary, competition with discards yielded more promising results, showing that it is able to account for all of the discussed classes of sound shifts. Furthermore it provided us with more possibilities regarding the causal chains in which sound changes proceed.

On a global scale, the main mechanism that leads to the observed phenomena is dispersion maximization, albeit not in the teleological sense of the word but as an emergent property of the exemplar dynamics at play. Furthermore, this mechanism is sensitive to token frequency in such a way that chain shifts arise among exemplar clouds with a small difference in existent exemplars, and mergers on the other hand occur if this difference is large. This sensitivity partially meets available statistical data on phoneme merger (Wedel et al. 2013b).

If the token frequency of all categories are approximately uniformly distributed, the constant trend towards maximization of contrast results in symmetric categories. If, for some reason or another, a bias affects the members of one category the equilibrium – or at least the current trend if the system is already on a transition from one equilibrium to another – is disturbed and the system starts to re-organize itself until a new stable state is reached.

This behavior can shed new light on the investigation of chain shifts. It is already widely accepted that bigger chain shifts can be made up of a mixture of both push and pull shifts. But in the context of the present model individual shifts could come about as the result of both pushing *and* pulling – a possibility which often remains unregarded in discussions about the causal links within chain shifts. In addition to that, in the simulations push shifts have taken place as a consequence of the joint influence of the members of two distinct categories on the ones of another category. Within the model at hand this is a predictable pattern because an exemplar is less reliably categorized in the face of three categories that compete for its labeling in comparison to just two. This contrasts with the generally held view on chain shifts as neatly ordered sequences of individual shifts, which also shines through in the commonly used arrow notation.

A last point of interest in this discussion regards the occurrence of mergers. In the present model the taking place of mergers is only sensitive to the ratio of token frequencies between the affected categories. The functional load, however, which is put forward by many authors as an explanation for the merging or push shifting of phonemes, respectively, has no clear equivalent in this model. Two nearby categories that exhibit an approximately equal token frequency would always chain shift if one is on the move towards the other. And categories with a high difference in token frequency would always merge

under these circumstances. But token frequency alone does not signify how much importance the phonemic contrast has for the distinction of meanings in a given language. Because functional load is regarded such an important factor in this respect, an extension of the current model that contains the possibility to measure the importance a particular phonemic contrast has in language use will be conducted in the next part of the thesis.

Part III

A MORE SOPHISTICATED APPROACH

7 INTRODUCTION

Given the right combination of settings, the first model was able to account for push shifts, pull shifts and mergers alike. However, the model favored push shifts over mergers only due to frequency differences. The possibility that mergers occur because the contrast exhibits a low functional load is not covered by model 1, as it does not contain lexical information but consists only of phonemic exemplars.

The model presented in the current section aims to fill this gap by introducing a very crude notion of a lexicon. In addition to labels, exemplars will now be enriched by a 'context' and a 'meaning'. The primary goal of the speaker in this altered approach is not to transmit a label, but a meaning, which he encodes in a combination of a context and a vector. The task of the listener then is to decode the message by either attempting to infer the meaning directly from this combination or by a preceding assignment of a label to the vector in order to deduce in a bottom-up fashion which 'word', i.e. combination of a context and a label, the speaker likely intended to transmit.

This novel relationship between the properties of exemplars allows for a shift in perspective w.r.t. the function phonemes have in language use. It is not the case, as model 1 might suggest, that language users speak with one another in order to align their conception of phonemic categories. Rather, they communicate, that is, they intend to exchange extra-linguistic information, called 'meanings' above. Phonemes and other linguistic categories are only means that guide this transmission and they are more useful in their role as guides the more the transmission relies on them, i.e. the less redundant they are.

Model 2 enables the listener to infer a meaning based on context and some phonetic form alone, given that there are not too many similar vectors of other exemplars in her exemplar space, that carry the same context but a different meaning. This would then mean that the phonemic category does not contribute to the distinction of at least two specific meanings, i.e. it does not contrast with other, especially the most proximate, categories. If this is also true for most other meanings that are originally distinguished by the contrast between this category and another one, then it can be said that this contrast is redundant.

This treatment makes for the quantification of the usefulness of a given contrast in the model language in terms of their functional load. The objective of the simulations is to investigate in what way the functional load and the distribution of lexical items that contrast

in the phonemes under observation have an impact on the extinction or maintenance of phonemic contrasts.

To achieve this goal, at first, the model architecture is established and a suitable notion of the functional load is defined in sec. 7.1 and 7.2, respectively. Afterwards, in sec. 8, the model is tested under various probability distributions that define the incentive for the utterance of specific meanings. It will be shown, that (a) chain shifts do contrast with mergers as a function of the usefulness of phonemic oppositions in language use, and (b) the functional load is a reliable measure, but only if one additionally takes into account the lexical distribution of phonemes.

7.1 SIMULATION ARCHITECTURE

As before, model 2 consists of two agents that are communicating with one another whereby each is assigned an exemplar space. Just like in model 1, each exemplar consists of a vector $\vec{x} \in [0, 1]$ and a label l drawn from a set of labels $\mathcal{L}$. But this time exemplars are furthermore enriched by a context $c \in \mathcal{C}$ and a meaning $m \in \mathcal{M}$.

During each speech act, the speaker chooses a meaning m she wants to convey with a probability of p_m and picks a context c to which the meaning is bound. The agent then searches for a suitable vector $\vec{x}$ that defines the general location of the production target $\vec{t}$ via one of two routes. This dual-route approach is inspired by Walsh et al. (2010) who built an exemplar model to process phonological and syntactic information. In their model, as well as in this one, the routes represent top-down and bottom-up processing, respectively. If the top-down route is taken, a lexical exemplar is directly prepared for production, and if the bottom-up route is taken, it has to be built up from lower-level constituencies, in this case 'phonemes' (i.e. exemplars with a specific label). In the present model the decision as to which route is to be taken depends on two conditions:

1. Is the excitatory activation α_m^+ which $\vec{x}$ receives from surrounding exemplars of the same meaning above a threshold θ?

2. Is α_m^+ greater than the inhibitory activation α_m^- received from surrounding exemplars with the same context but a different meaning?

If both are true, $\vec{x}$ is entrenched in exemplars of the same meaning (top-down). If instead the exemplar fails to fulfill at least one of these conditions, it is entrenched in exemplars with the same label (bottom-up). These conditions relate to two different phenomena. The first is based on the assumption, that high-frequency lexical items tend to exhibit more idiosyncratic phonetic forms because they can be uttered directly. Items of lower frequency instead may have to be

constructed from smaller constituents due to their low activation. The second condition is peculiar to the model presented here, as it triggers homophony avoidance which is essentially the inhibitory influence of items with the same phonetic form but a different meaning.

Perception proceeds in a similar fashion. The hearer perceives a vector $\vec{t}$, the production target of the speaker, along with a context c. He then decides whether the activation α_c^+ the exemplar receives from surrounding exemplars marked for the same context c is greater than θ. If this is the case, he attempts to infer the meaning m' directly from nearby exemplars which are marked for c and labels the percept afterwards (top-down). But if the condition is not fulfilled, he first assigns a label l' to the percept based on exemplars of the same context in the neighborhood of $\vec{t}$ and then searches for a meaning based on both c and l'. All these assignments are based on chance just like in the labeling procedure in model 1.

In order to offer an incentive for a category to either move away or stay in place, each cycle ends with feedback from the speaker via comparison of the intended meaning m and the hypothesized meaning m'. If $m = m'$, the exemplar is stored in the exemplar space, otherwise it is discarded.

As in model 1, the activation α of the exemplars decreases as a function of each exemplar's respective age t, calculated by the formula

$$\alpha = \exp\left(-\frac{t}{\tau}\right),$$

where τ is the maximum age an exemplar can reach. If its age goes beyond τ, it is discarded from the exemplar space. An exemplar can 'revive' if a perceived exemplar that is about to be added to the exemplar space is already present in there. In this case the percept is discarded and the age of the stored exemplar is set to $t = 0$. Again, identity is determined by proximity. If an exemplar falls inside a radius jnd of another exemplar and both have the same label, context and meaning, they are perceived as identical in perception. Ties are resolved naïvely by reviving only the first encounter of an identical stored exemplar.

In the following two subsections a more detailed insight in the production and perception procedures is provided.

7.1.1 *Production*

In production[1], a meaning m is chosen from $\mathcal{M}$ with probability p_m. For the calculation of α_m^+ and α_m^- a scoring function is used. This one makes use of a generalized version of the numerator of eq. 5 seen in section 4.1.2. For every combination of $\mathcal{L}$, $\mathcal{C}$ and $\mathcal{M}$, that is every

1 Listing 3 provides a summary of the whole production procedure in pseudocode.

element $\mathcal{S} = \{\mathcal{R}_1, \ldots, \mathcal{R}_n\}$ of $\mathfrak{P}(\{\mathcal{L}, \mathcal{C}, \mathcal{M}\})$, and every combination of a label l, a context c and a meaning m, i.e every element $\{r_1, \ldots, r_n\}$ of $\mathfrak{P}(\{l, r, c\})$, it is defined as

$$score(\vec{x}, \mathcal{S}, r_1, \ldots, r_n) = \sum_{\{\vec{y} \mid \mathcal{R}_i(\vec{y}) = r_i, \forall i : r_i \in \mathcal{R}_i \in \mathcal{S}\}} w_{\vec{y}}(\vec{x}), \tag{6}$$

where $w_{\vec{y}}(\vec{x})$ is the same function given in eq. 2. That one and eq. 3 upon which it is dependent are repeated here for quick reference.

$$w_{\vec{y}}(\vec{x}) = \frac{\alpha_{\vec{y}}}{win(|\vec{x} - \vec{y}|) + 1}, \tag{2 rev.}$$

$$win(x) = \begin{cases} x, & x \leqslant \mu/2 \\ 0, & \text{otherwise} \end{cases}. \tag{3 rev.}$$

Certainly, eq. 6 demands some explanation. The function sums over all distance-weighted activations $w_{\vec{y}}(\vec{x})$, which $\vec{x}$ receives from the $\vec{y}$'s, which are drawn from the set given below the sum sign. The set is restricted to those $\vec{y}$ for which it is true that their property $\mathcal{R}_i$ (e.g. $\mathcal{L}$) matches the value r_i (e.g. l).

In other words: It calculates the joint activation of all exemplars surrounding $\vec{x}$ (weighted by their distance to $\vec{x}$) which have the same label l and the same context c and the same meaning m, as specified in the last three arguments of the function. If at least one of the values is not specified, the remaining values must be equal. If none is specified, all exemplars surrounding $\vec{x}$ are taken into consideration.

Now, and this will help to clarify the definition of score, α^+_m for $\vec{x}$ is calculated by

$$\alpha^+_m = score(\vec{x}, \{\mathcal{M}\}, \mathcal{M}(\vec{x})).$$

This is the summed distance-weighted activation of all exemplars surrounding $\vec{x}$ (including $\vec{x}$ itself) which are marked for the same meaning $\mathcal{M}(\vec{x})$ as $\vec{x}$.

α^-_m on the other hand is calculated by

$$\alpha^-_m = score(\vec{x}, \{\mathcal{C}\}, \mathcal{C}(\vec{x})) - score(\vec{x}, \{\mathcal{C}, \mathcal{M}\}, \mathcal{C}(\vec{x}), \mathcal{M}(\vec{x})),$$

i.e. the joint distance-weighted activation of all exemplars surrounding $\vec{x}$ that share the context $c = \mathcal{C}(\vec{x})$ of $\vec{x}$, but have a meaning different from $\mathcal{M}(\vec{x})$.

After α^+_m and α^-_m are calculated, $\vec{x}$ is entrenched in its surrounding exemplars that are marked for the same meaning or label, depending on whether α^+_m is greater than α^-_m *and* α^+_m is greater than a fixed threshold θ, or not. Entrenchment in this context demands a generalization of the formula given in eq. 1 such that it can handle any $\mathcal{P} \in \{\mathcal{L}, \mathcal{C}, \mathcal{M}\}$, which is given by

$$entrench(\vec{x}, \mathcal{P}) = \frac{\sum_{\vec{y} : \mathcal{P}(\vec{y}) = \mathcal{P}(\vec{x})} \vec{y} w_{\vec{y}}(\vec{x})}{\sum_{\vec{y} : \mathcal{P}(\vec{y}) = \mathcal{P}(\vec{x})} w_{\vec{y}}(\vec{x})}. \tag{7}$$

The decision regarding the identity of $\mathcal{P}$ depends on whether

$$\alpha_m^+ > \alpha_m^- \wedge \alpha^+ > \theta \tag{8}$$

holds. If 8 is true, then one has $\mathcal{P} = \mathcal{M}$, i.e. $\vec{x}$ is entrenched in surrounding exemplars that have the same meaning as itself, otherwise one has $\mathcal{P} = \mathcal{L}$, i.e. it is instead entrenched in exemplars with the same label.

The first subcondition $\alpha_m^+ > \alpha_m^-$ evaluates the probability of confusion of the signal with similar signals. If all else being equal the combination of the chosen context and $\vec{x}$ are more likely to be interpreted as a meaning different from the one intended for transmission, a target $\vec{t}$ is constructed which is more prototypical with respect to the label l of $\vec{x}$. If the probability of confusion is low on the other hand, its construction is based on earlier utterances with the same meaning. But that in turn only happens if c in combination with $\vec{x}$ is regarded an established way of expressing m. This decision depends on the number and activation of earlier perceived exemplars that expressed meaning m, i.e. if $\alpha^+ > \theta$ for some fixed threshold θ.

Furthermore, the above condition determines whether an exemplar whose label is marked for lenition towards a target position $\vec{\gamma}$ is displaced into this direction by some amount λ. This only applies if term 8 holds true, relating to the assumption that sound shifts are initiated by the constant lenition of high-frequency items (see 3.2).

As a last step, some amount of noise δ is added to the entrenched exemplar and the exemplar is forced into the boundaries of $[0, 1]$ if necessary. The ultimate identity of the production target $\vec{t}$ (excluding lenition for brevity's sake) is given by

$$\vec{t} = \mathrm{restrain}\left(\mathrm{entrench}\left(\vec{x}, \mathcal{P}\right) + \delta\right),$$

with $\mathcal{P} = \mathcal{M}$ or $\mathcal{P} = \mathcal{L}$ depending on the term in 8.

7.1.2 *Perception*

The goal of perception[2] lies in the inference of a meaning m' from an incoming combination of a vector $\vec{t}$, the former production target of the speaker, and a context $c = \mathcal{C}(\vec{t})$. There are again two alternative routes that can be taken, this time depending on whether $\vec{t}$ receives sufficient activation α_c^+ by nearby exemplars marked for the same context c. Here, α_c^+ is calculated in a way analogous to the calculation of α_m^+ above, i.e. by the formula

$$\alpha_c^+ = \mathrm{score}(\vec{t}, \{\mathcal{C}\}, \mathcal{C}(\vec{t})).$$

If it turns out that $\alpha_c^+ > \theta$, a probable meaning is inferred directly based on exemplars of the same context. Otherwise, a label hypothesis l' is constructed based on the surrounding exemplars, and only after that a meaning is inferred, based on l'.

2 Again, the perception procedure is summarized in listing 4.

```
# receive a random meaning to be expressed
m = getRandomMeaning()

# fetch a random exemplar marked for this meaning,...
t⃗ = getRandomExemplarByMeaning(m)
# ...its context and its label
c = 𝒞(t⃗)
l = ℒ(t⃗)

# receive excitatory activation of surrounding exemplars
# marked with meaning m
α⁺ₘ = score(t⃗, m)

# receive inhibitory activation of surrounding exemplars
# marked with c but not with m
α⁻ₘ = score(t⃗, c, not m)

# if excitation high and inhibition weak enough...
if (α⁺ₘ > θ and α⁺ₘ > α⁻ₘ):
  # ...entrench the exemplar in exemplars with the same
      meaning...
  t⃗ = entrench(t⃗,m)
  # ...and apply lenition if the label is marked for it
  if (markedForLenition(l)):
    t⃗ = t⃗ + λ
else:
  # ...entrench the exemplar in exemplars with the same
      label
  t⃗ = entrench(t⃗,l)

# add noise
t⃗ = t⃗ + δ

# push the exemplar back into its boundaries
t⃗ = restrain(t⃗)

return t⃗
```

Listing 3: Pseudocode of the production procedure of model 2.

The evaluation of both m' and l' makes use of a generalized version of the probability function given in eq. 5, that is based on the generalized scoring function depicted in eq. 6. Since this altered probability function is very tedious to define, in the following only the special cases which are necessary for the evaluation will be mentioned. The two cases $\alpha_c^+ > \theta$ and $\alpha_c^+ \leqslant \theta$ will now be treated separately.

Case $\alpha_c^+ > \theta$: In this case, the top-down one, first a meaning hypothesis m' is constructed. The probability that $\vec{t}$ is assigned a meaning m^* under a fixed context c is given by

$$\varphi\left(\vec{t}, m^*, c\right) = \frac{\text{score}\left(\vec{t}, \{\mathcal{M}, \mathcal{C}\}, m^*, c\right)}{\sum_{m^{**} \in \mathcal{M}} \text{score}\left(\vec{t}, \{\mathcal{M}, \mathcal{C}\}, m^{**}, c\right)}.$$

After a meaning has been assigned to m', this hypothesis is used to infer a suitable phoneme label l' for $\vec{t}$. The probability for a label l^* under a fixed context c and a fixed meaning m', our meaning hypothesis, is then given by

$$\varphi\left(\vec{t}, l^*, c, m'\right) = \frac{\text{score}\left(\vec{t}, \{\mathcal{L}, \mathcal{M}, \mathcal{C}\}, l^*, c, m'\right)}{\sum_{l^{**} \in \mathcal{L}} \text{score}\left(\vec{t}, \{\mathcal{L}, \mathcal{M}, \mathcal{C}\}, l^{**}, c, m'\right)}.$$

Case $\alpha_c^+ \leqslant \theta$: In the bottom-up case, there has been no sufficient support by exemplars surrounding $\vec{t}$ to infer a meaning directly. In that case, first a label hypothesis l' is constructed and subsequently, with its assistance, a meaning hypothesis m'. The respective probability functions are defined in a way analogous to the ones in the first case.

$$\varphi\left(\vec{t}, l^*, \{c\}\right) = \frac{\text{score}\left(\vec{t}, \{\mathcal{L}, \mathcal{C}\}, l^*, c\right)}{\sum_{l^{**} \in \mathcal{L}} \text{score}\left(\vec{t}, \{\mathcal{L}, \mathcal{C}\}, l^{**}, c\right)},$$

$$\varphi\left(\vec{t}, m^*, c, l'\right) = \frac{\text{score}\left(\vec{t}, \{\mathcal{M}, \mathcal{C}, \mathcal{L}\}, l^*, c, m'\right)}{\sum_{m^{**} \in \mathcal{M}} \text{score}\left(\vec{t}, \{\mathcal{M}, \mathcal{C}, \mathcal{L}\}, m^{**}, c, l'\right)}.$$

After suitable values for m' and l' have been found, the last step consists of comparing m' to the originally intended meaning m. If they are equal, $\vec{t}$, now marked for c, l' and m', is added to the exemplar space. Otherwise it is discarded, because the hypotheses apparently do not lead to successful communication.

7.2 FUNCTIONAL LOAD

The simulations aim to investigate whether the functional load of a phonemic opposition $A \sim B$ has an influence on the appearance of push shifts or mergers. For this reason we need to find a way to measure the functional load $\text{FL}(A \sim B)$ of such an opposition.

```
fetch t⃗

# infer the context from the signal
c = 𝒞(t⃗)

# receive excitatory activation of surrounding exemplars
# marked with context c
α_c^+ = score(t⃗, c)

# if excitation high enough...
if (α^+ > θ):
  # ...assign a random meaning considering the
  # joint activation of surrounding exemplars
  # with the same context and...
  m′ = getRandomMeaning(probability(t⃗,ℳ,c))
  # ...assign a random label based on context and
  # the inferred meaning
  l′ = getRandomLabel(probability(t⃗,ℒ,c,m′))
else:
  # ...assign a random label first and...
  l′ = getRandomLabel(probability(t⃗,ℒ,c))
  # ...assign a random meaning based on this label.
  m′ = getRandomMeaning(probability(t⃗,ℳ,c,l′))

if (m′ = ℳ(t⃗)):
  add new Exemplar(t⃗,c,l,m)
else:
  discard t⃗
```

Listing 4: Pseudocode of the perception procedure of model 2.

The measure of FL used here will be the one originally proposed by Hockett (1966) (see also Surendran and Niyogi 2006). It assigns a *load* H on languages, i.e. a measure of how much information a language transmits. Here, a language is seen as a string of symbols that are in some opposition to one another. The task of finding out how a natural language can be mapped onto this notion of a language is non-trivial, but it nicely fits the present model language. Now, the FL of a given opposition $A \sim B$ is calculated by comparing the load of the language L before a merger of both categories with the load of the derived language L_{AB} after such a merger:

$$\mathrm{FL}(A \sim B) = \mathrm{H}(L) - \mathrm{H}(L_{AB}). \tag{9}$$

The measure used for H is the entropy of a given language, i.e. the predictability of its substrings (Shannon 1949). For an arbitrary language the calculation of its entropy involves the counting of each of its substrings. Since natural languages are infinite, this is an impossible task. Instead, substrings up to a length of n can be counted within a finite subset L' of the language in order to arrive at an n-order approxima-

tion of the entropy. The formula for this n-order approximation H_n for a language L is then given by

$$H_n(L) = -\frac{1}{n} \sum_{w \in W_{L'}} p(w) \log_2 p(w), \tag{10}$$

where $W_{L'}$ is the set of all 1- to n-grams in L' and $p(w)$ is the occurrence probability for each $w \in W_{L'}$. Since the only sources of predictability of a phoneme inside the model language at hand are its possible contexts and a pre-defined utterance probability, it suffices to calculate its entropy with $n = 2$.

All substrings of the language of model 2 consist solely of a context c and a label l. For an example, let us assume that there are three contexts $\{p, t, k\}$ and two labels $\{a, i\}$. This makes for six possible words (i.e. bigram substrings) of which not every one is observed in this example language but only pa, ta ,ti and ki, and each of them appears with a different probability. The probabilities for all observed substrings of this language are summarized in the following table.

$p(pa) = 0.25$	$p(p) = 0.25$
$p(ta) = 0.25$	$p(t) = 0.35$
$p(ti) = 0.1$	$p(k) = 0.4$
$p(ki) = 0.4$	$p(a) = 0.5$
	$p(i) = 0.5$

This makes for an entropy of

$$H_2(L) = -\frac{1}{2} \left(p(pa) \log_2 p(pa) + \ldots + p(i) \log_2 p(i) \right) \approx 2.20992.$$

The entropy of the language which, all else being equal, does not differentiate between a and i is now calculated by substituting both symbols for a single arbitrary symbol X and summing the probability of all arising doublets. Afterwards we arrive at an entropy for L_{ai} of $H_2(L_{ai}) \approx 1.555887$. And consequently the FL of this opposition is $FL(a \sim i) \approx 0.654$.

Eqs. 9 and 10 depict the way in which the FL will be calculated in the following sections. But it is not entirely accurate for two interrelated reasons. (a) There is no previously defined utterance probability for words, only for meanings. Since there are no synonyms, in the beginning of each simulation, meaning probabilities will equal word probabilities, but (b) words can vanish entirely so that their meanings are not expressible anymore. Because of that, the FL of an opposition can actually only be calculated post-hoc, since the disappearance of specific words is not always predictable. Nonetheless, this kind of an a-priori FL gives us a good idea about the strength of an opposition at around the beginning of each run.

8 THE SIMULATIONS

8.1 PRELIMINARIES

8.1.1 *Fixing parameters*

Model 2 is quantitatively much more complex than its predecessor, as there are many more parameters to be considered. What is absent in the present model is a success rate, which means that the addition of exemplars to the lexicon is entirely driven by competition and positive feedback from the speaker. New parameters are the number of contexts, the number of meanings and a threshold θ. Furthermore, one of the main objectives of the simulations is to explore the influence of different utterance probability distributions on the behavior of the system, and there will be three distributions under consideration which will be defined in sec. 8.1.2 below. These will entail two more parameters, namely the slope s of a distribution (only true for two of them) and a multiplier q that defines the ratio of the token frequencies of two categories.

Other than that, the model behaves very sensitive to the alternation of parameter settings. Hence it makes sense to restrict the number of labels, meanings and contexts to the bare necessities. Since our principal interest lies in the interaction of two categories on the verge of merging, the number of labels is restricted to exactly this amount. Furthermore, there will only be five contexts present. This number is just sufficient to display the general 'characters' of the probability distributions. As a consequence of these amounts, the upper bound for the number of meanings is set to ten, and to keep things simple, this will also be the lower bound, i.e. in the beginning of each simulation every possible word that can be made up of the given contexts and labels will be present.

To make use of the newly gained space, category **a** will be initialized at around $\langle 0.5, 0.5\rangle$ and **b** at $\langle 0.9, 0.9\rangle$. Again, the exemplars of **b** will be lenited by an amount of $\lambda = 0.03$ towards $\vec{\gamma} = \langle 0.5, 0.5\rangle$ provided all conditions are met.

This model is much more sensitive to noise than the previous one, so it will be set to $\delta = 0.4$. On the other hand, a window size which is too large can lead to the premature disappearance of low-frequency words because their activation is inhibited too much by a neighboring category. Throughout several test runs, $\mu = 0.5$ turned out to result in a relatively stable and 'natural' behavior.

$\lvert\mathcal{L}\rvert$	2	Number of labels (categories)
$\lvert\mathcal{C}\rvert$	5	Number of contexts
$\lvert\mathcal{M}\rvert$	10	Number of meanings
θ	7.0	Threshold for top-down processing
λ	0.03	Amount of lenition of the shifting category
$\vec{\gamma}$	$\langle 0.5, 0.5\rangle$	Target position of the initial shift
δ	0.4	Amount of noise in production
μ	0.5	Window size for activation and entrenchment
τ	4,000	Maximum exemplar age
jnd	0.06	Just noticeable difference

Figure 14: Parameters that will be held constant throughout the simulations of model 2.

The threshold has been carefully set to $\theta = 7.0$. If it is too high, it may avert lenition. But if it is too low on the other hand, it can prevent mergers altogether. A value of 7.0 yielded the most satisfying results under the current parameter settings.

The remaining parameters are not different from the ones chosen in model 1 and are summarized among the entirety of the constant parameters in fig. 14 for reference.

8.1.2 *Probability distributions*

In order to control for the FL of the phonemic opposition under consideration (note that there is only one), the distribution of the meaning probability p_m will be altered. In this work, there will only be three kinds of distributions, even though the behavior under other distributions is certainly of interest. The first distribution is called the *within-category uniform distribution*, where within a phonemic category the meaning probabilities, and thus those of every word, are equally distributed, but not necessarily across categories. The other two distributions are variants of the Zipfian distribution, where within each phonemic category the meanings are distributed according to Zipf's law, which states, loosely speaking, that within a language some words are very frequent, but most are very infrequent (Zipf 1935). It is plausible to assume that this observation also holds true if one restricts their analysis only to words which contain a specific phoneme. Both variants differ with respect to the cross-category distribution of the context frequencies within each phoneme can be uttered. In the first one, called *dual-Zipfian distribution*, each context exhibits the same utterance probability across categories apart from a constant

factor. In the second one on the other hand, called *cross-Zipfian distribution*, the utterance probabilities for each context are be reversed.

In order to define the distributions in a proper way, some notation needs to be introduced. In what follows, all meanings and contexts will be ordered with respect to some index. Because there are two labels, five contexts and ten meanings, these can be arranged as depicted in table 1.

meanings	m_1	m_2	m_3	m_4	m_5	m_6	m_7	m_8	m_9	m_{10}
contexts	c_1	c_2	c_3	c_4	c_5	c_1	c_2	c_3	c_4	c_5
labels	**a**					**b**				

Table 1: Line-up of meanings, contexts and labels throughout the simulations of model 2.

Furthermore, the set of meanings that can be expressed by means of a label l i denoted as $m(l)$. So, in this case for instance, one has $m(\mathbf{a}) = \{m_1, \ldots, m_5\}$. In addition, the context that is used to express a meaning m is denoted as $c(m)$, e.g. $c(m_7) = c_2$.

Within-category uniform distribution

The within-category uniform distribution simply accomplishes a uniform distribution of utterances within each category, but not necessarily across categories. With it comes a parameter q which determines the frequency ratio between categories in such a way that the probability of **b** to be uttered in a context c is q times as high as for **a**, i.e. $p(c_i\mathbf{b}) = qp(c_i\mathbf{a})$ for all i. This relationship is illustrated in fig. 15. with $q = 3$.

The probability function for this distribution is given by

$$p_q(m) = \begin{cases} \frac{1}{N} & \text{if } l(m) = \mathbf{a} \\ \frac{q}{N} & \text{if } l(m) = \mathbf{b} \end{cases},$$

where $N = |m(\mathbf{a})| + q\,|m(\mathbf{b})|$, i.e. $5 + 5q$ under the current parameter setting.

Dual-Zipfian distribution

Zipf's law is a distribution characterized by the function

$$\mathrm{zipf}(r, s) = \frac{1}{r^s},$$

where s is some real number that influences the slope of the function. In applied contexts, the index r signifies the *rank* of an item in a list decreasingly sorted by frequency.

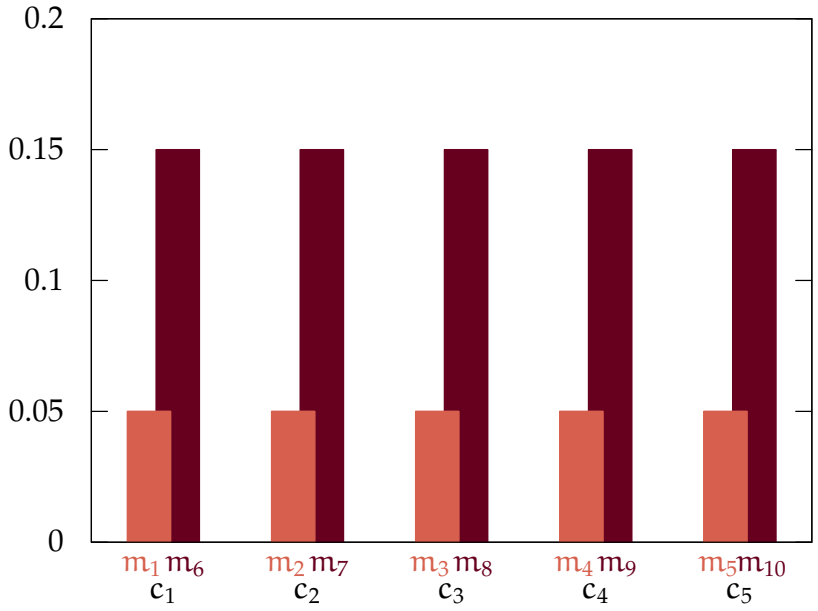

Figure 15: Example of the within-category uniform distribution with $q = 3$.

In the following, the rank of a meaning m will be denoted as $r(m)$. In the case of the dual-Zipfian distribution, the rank of a meaning equals the index of its context, i.e. $r(m) = \mathit{index}(c(m))$, where index is a function that outputs the index of a meaning or context. Analogous to the within-category uniform distribution, there is another parameter q with the same function as before. An example of a dual-Zipfian distribution with $q = 3$ and $s = 2$ is given in fig. 16.

Now, the general distribution (not *probability* distribution) is characterized by the following adapted formula, which will also be used in the definition of the cross-Zipfian probability distribution below.

$$\mathit{zipf}'(m, s, q) = \begin{cases} \frac{1}{r(m)^s} & \text{if } l(m) = \mathbf{a} \\ \frac{q}{r(m)^s} & \text{if } l(m) = \mathbf{b} \end{cases}.$$

Finally, the probability function for the dual-Zipfian distribution is straightforwardly defined as

$$p_{q,s}(m) = \frac{\mathit{zipf}'(m, s, q)}{\sum_{m' \in \mathcal{M}} \mathit{zipf}'(m', s, q)}.$$

Cross-Zipfian distribution

The cross-Zipfian distribution differs from the dual-Zipfian one only with respect to the ranking of the meanings. Let n be the sum of the

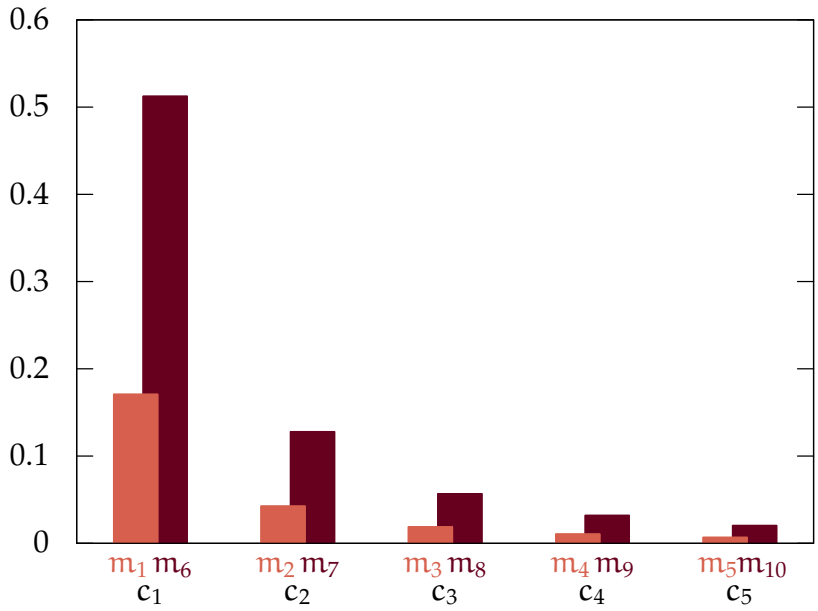

Figure 16: Example of the dual-Zipfian distribution with $s = 2$ and $q = 3$.

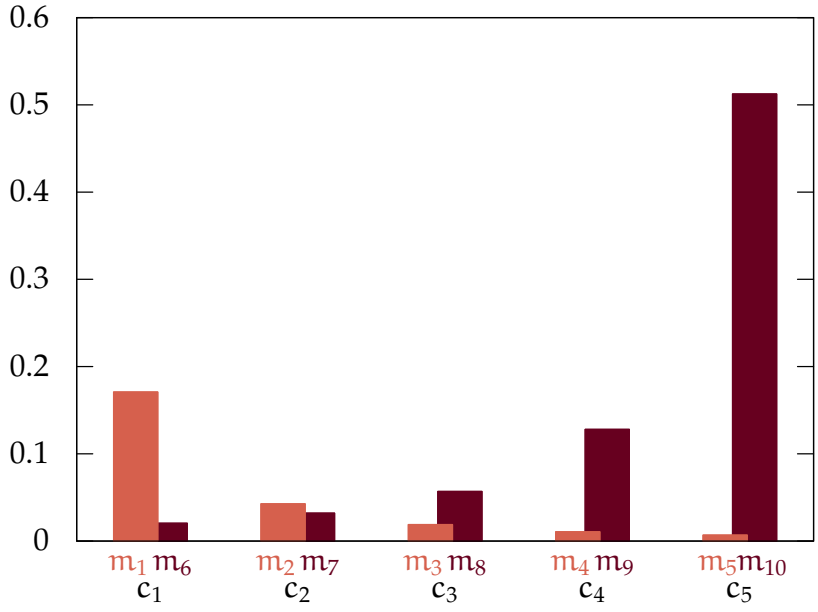

Figure 17: Example of the cross-Zipfian distribution with $s = 2$ and $q = 3$.

number of meanings expressible by **a** or **b**, i.e. $n = |\mathcal{M}|$ in our case, then the appropriate ranking is given by

$$r(m) = \begin{cases} \mathrm{index}(c(m)) & \text{if } l(m) = \mathbf{a} \\ n - \mathrm{index}(c(m)) + 1 & \text{if } l(m) = \mathbf{b} \end{cases}.$$

This ranking results in distributions akin to the one given in fig. 17 for $s = 2$ and $q = 3$.

Now that all necessary parameters have been fixed and explained, the model will be run under the just defined three probability distributions.

8.2 THE SIMULATION

In this section, simulations are run using all of the three utterance probability distributions defined in sec. 8.1.2 and under various parameter settings.

8.2.1 *Within-category uniform distribution*

The very first simulation will play through the scenario of **b** moving towards **a** under a uniform probability distribution for all meanings and consequently also for contexts and labels. This distribution results in a FL for the opposition **a** ~ **b** of 1.0, i.e. the highest possible. Because of that, a push shift is to be expected and its success can moreover be seen as a minimum requirement for the architecture of the model.

The graph in fig. 18 shows the course of the median values of the categories' x_1 and x_2 components throughout the simulation. The error bars signify their first and fifth sextile and can show us by this means whether both categories exhibit any significant overlap or not. From this picture it can be clearly inferred that this was never the case during the simulation. When cluster **b** drew closer to **a**, the latter kept its distance (consider especially the x_1 values in this respect).

It is conspicuous, though, that the x_1 component of **b** never reached a value of 0.5. However, this does not pose a problem since it is merely caused by the breadth and the density of both categories that prevents **a** from being cornered in the upper left. In essence, it can safely be said that if all meanings, and thus all words, appear with equal probability, causing an FL of 1.0, a merger of both categories does not occur.

Now one might ask whether this is the consequence of the chosen probability distribution or the relative token frequencies of both categories. A temporary answer to this question is that at least the equality of their token frequencies can be excluded as a major factor for the coming about of mergers. Even if there are as many as ten

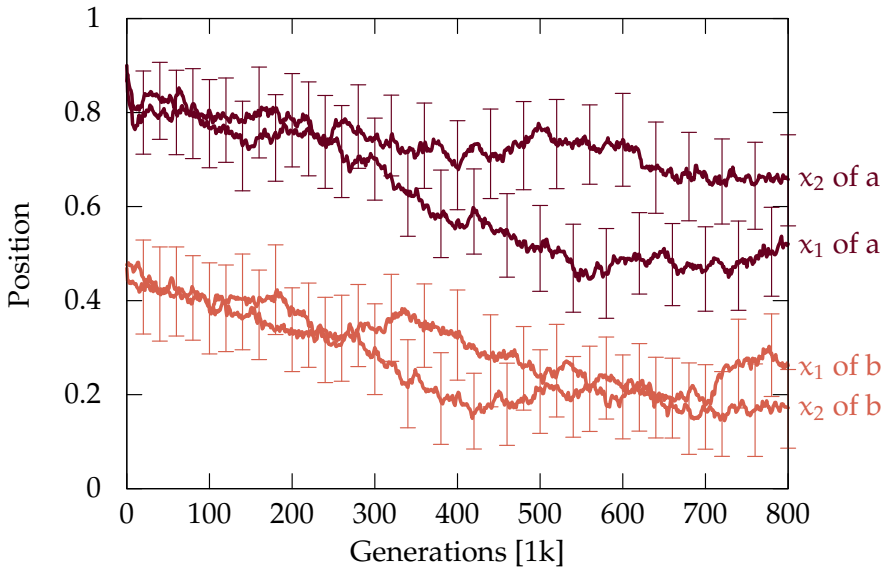

Figure 18: Development of the median values of x_1 and x_2 of both categories over a time span of 800,000 generations, given that all meanings were equally likely to be produced, resulting in $\mathrm{FL}(\mathbf{a} \sim \mathbf{b}) = 1.0$. The error bars depict the first and the fifth sextile of each category.

tokens of **b** for each of **a** (i.e. $q = 10$), yielding an FL of approx. 0.44, chain shifts ensue, as the graph in fig. 19 displays.

This finding contrasts with the run of model 1 summarized in figs. 11 and 12, where a majority of three over one sufficed to wipe out the smaller category. A possible explanation will be given in the next section when the behavior under the cross-Zipfian distribution is evaluated.

8.2.2 *Cross-Zipfian distribution*

Since neither across- nor within-category uniform distributions of the probabilities for meaning utterances are able to account for phoneme mergers, now the cross-Zipfian distribution will be contemplated. As a reminder, under this distribution the meaning probabilities face each other in such a way that the probabilities of two phoneme labels to appear in a specific context are inversely correlated, i.e. given a context c, the probability of **a** to appear in this context is lower the higher the probability for **b** is to appear in this same context. Additionally, within each category, the contexts are ranked and the higher the rank the lower the probability of the respective label to appear in this context – the essence of the Zipfian distribution. The cross-Zipfian distribution can be parametrized further by the steepness of the slope s and a multiplier q which determines the quotient of the token-frequencies across categories.

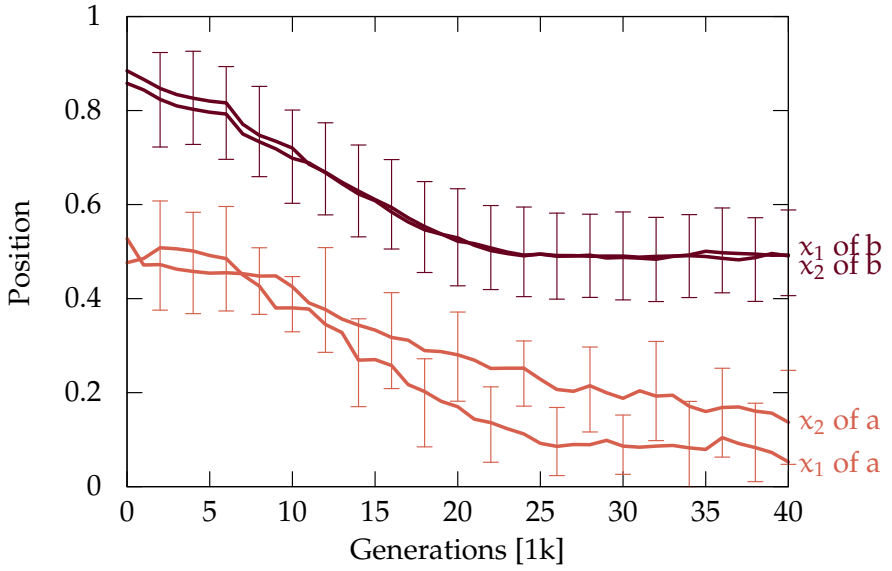

Figure 19: Development of the category medians of their respective x_1 and x_2 values throughout 40,000 cycles using the within-category uniform distribution. The token frequency of **b** is ten times higher than that of **a**, i.e. $q = 10$ and $FL(\mathbf{a} \sim \mathbf{b}) = 0.44$.

Within this section, the main interest lies in the influence of the chosen distribution on the appearance of mergers or chain shifts in general and particularly in the share that s and q have in this respect. The first simulation will give us positive results regarding the appearance of mergers. The subsequent simulations will then serve to examine what impact variations of the parameters s and q have.

Using a slope of $s = 3$ and a multiplier of $q = 4$, yielding an FL of approximately 0.42, a merger of both categories can be observed. This is indicated in fig. 20 by a more or less proper overlap of the sextiles of the x_1 and x_2 values, respectively. Because categories are unlikely to vanish completely, mergers manifest themselves by two characteristics. First, as already mentioned, an approximately complete overlap, and second, by the complementary distribution of contexts within each label can still appear. The graph in fig. 21 shows the success rates for each meaning. For instance, meaning m_5 which is expressed by the word consisting of context c_5 and label **a** reaches incomprehensibility at around step 7,000 with the effect that exemplars which are marked for this combination vanish shortly after. Finally, there are no exemplars left which have the same context but a different label. Such a state figures also the perils of a merger in real life. Expressions may become indistinguishable *in every context* and at least one of them falls out of use, most plausibly the less frequent one. It may then be replaced by another expression, but that case is not covered by the present model.

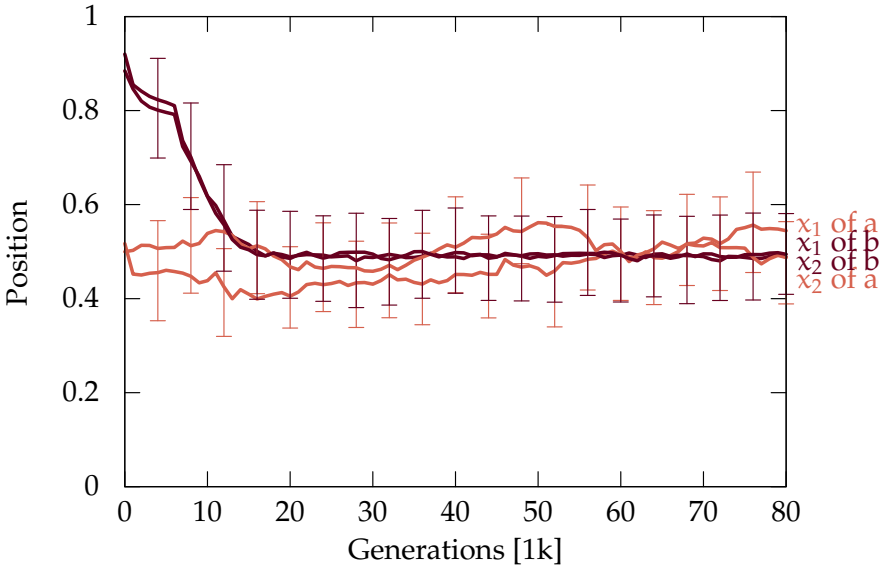

Figure 20: Development of the medians of the categories' respective x_1 and x_2 components using the cross-Zipfian distribution with a slope of $s = 3$ and a token frequency multiplier of $q = 4$, yielding $\mathrm{FL}(\mathbf{a} \sim \mathbf{b}) = 0.42$.

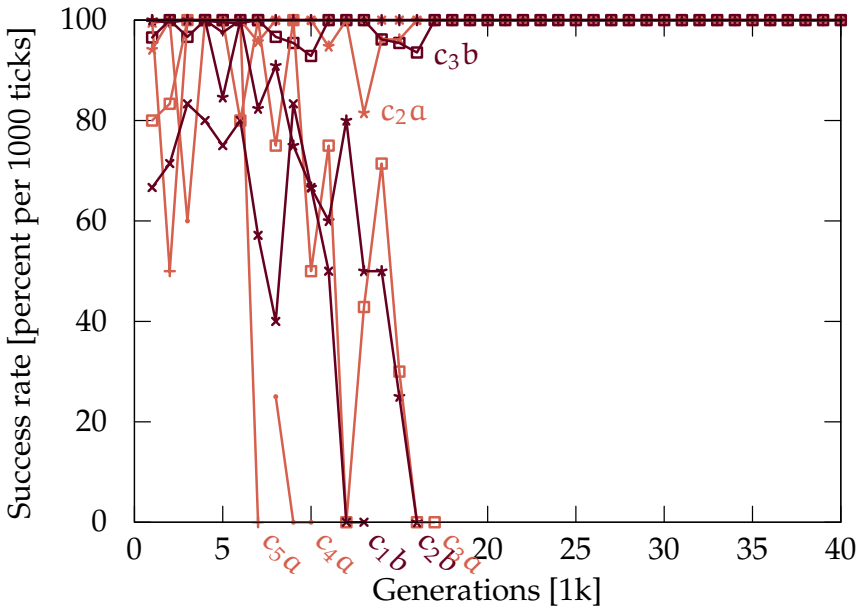

Figure 21: Success rates for the encoding of meanings of all words during the simulation depicted in fig. 20, evaluated with a sample rate of 1,000 cycles.

But why is it that this time category **a** remained largely in place? The reason is, that its high-frequency words are highly successful. To recollect, exemplars with many other exemplars around them that have the same meaning receive much activation from these during production and the production target is constructed on the basis of these alone. If, on the other hand, a meaning is not as well represented, production targets that are supposed to transmit this meaning are constructed with reference to all surrounding exemplars of the same label. This dual-route permits high-frequency words to define the center of their phoneme category. This is true for both categories. Regarding category **b**, lots of exemplars of high-frequency words pop up ever closer to **a**. They can do so almost freely because exemplars of the latter in the same context pose no significant challenge w.r.t. the reliability of the meaning encoding. Low frequency words of **b** are dragged behind because their activation hardly ever exceeds the threshold and as a consequence their target position is always close to regions of exemplars of high-frequency words within **b**. But as more of the exemplars of low-frequency words are uttered near **a**, their meanings become increasingly unencodable which eventually leads to their decline. Regarding category **a**, the situation is basically the same. High-frequency items stay in place, since even the meanings of their exemplars near **b** are successfully communicated, and exemplars of low-frequency words cannot crop up in safer areas as the huge heap of exemplars of high-frequency words defines the position of their production targets.

The same reasoning can explain the chain shifting observed under the within-category uniform distribution. As **b** moves closer to **a**, the area between these two becomes less safe for successful communication. As a consequence, more exemplars appear to the left and the top of **a** because their meaning is more reliably identified. Because frequencies are equally distributed within each category, there are no exemplars of words that avert the movement of **a** away from **b**. So, for the most part, it is the distribution of lexical items within and across categories that has an influence on the probability of a merger and not token frequency alone. Furthermore, the FL proves not quite informative in this respect, either, as it was ~ 0.44 in the uniformly distributed case with $q = 10$, but, with a value of ~ 0.42, only slightly lower in the most recent run.

Nonetheless, the cross-Zipfian distribution per se is not a sufficient predictor for mergers. For example, a parameter setting of $s = 1$ and $q = 1$ results in a chain shift, as fig. 22 shows. This is because the differences in token frequencies of the labels in overlapping contexts are not too large, even for the most extreme cases. For instance, the probability for **a** to appear in c_1 is only five times larger than for **b**. In comparison, in the above case one had $p(c_1\mathbf{a})/p(c_1\mathbf{b}) = 31.25$. As a result, there is more competition, which is reflected in a greater num-

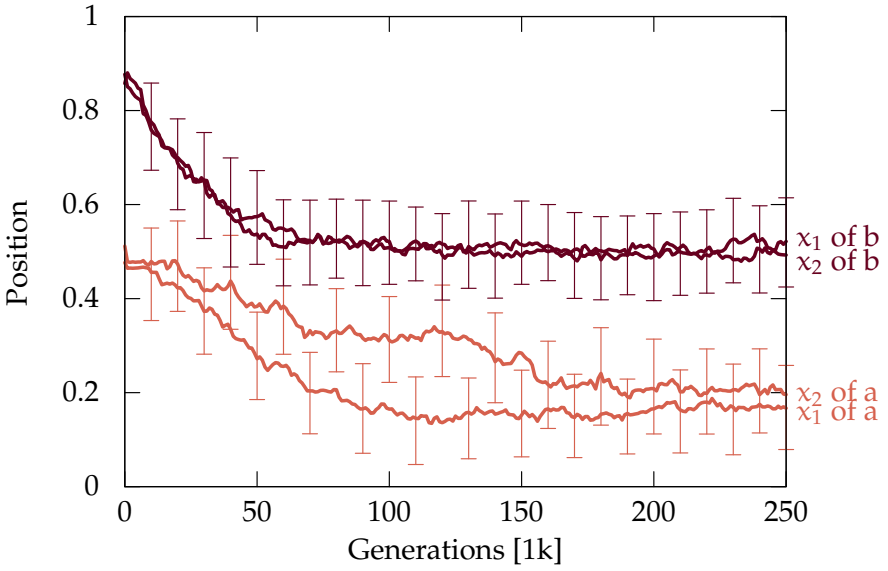

Figure 22: Development of the medians of the categories' respective x_1 and x_2 components using the cross-Zipfian distribution with a slope of $s = 1$ and a token frequency multiplier of $q = 1$, yielding $FL(\mathbf{a} \sim \mathbf{b}) = 0.895$.

ber of simulation cycles that **b** needs to reach its target (ca. 70,000 in this case vs. ca. 15,000 above) and stronger fluctuations of the success rates for all words (see fig. 23).

Both parameters s and q affect the functional load but in a different manner. If q is fixed to 1 and s tends towards ∞, the probabilities of all words approximate 0.0, except for $c_1\mathbf{a}$ and $c_5\mathbf{b}$ which approximate 0.5. Or more generally speaking, one has $p(c_5\mathbf{b}) \to q \times p(c_1\mathbf{a})$ in the limit. The result is an almost complementary distribution. With q towards ∞ on the other hand, the probabilities of all words containing **a** approximate 0.0, hence increasing the likelihood of a wipe-out of all exemplars containing **a**, even without any 'assistance' of **b**.

Of course, it is questionable whether whole phonemic categories ever fall out of use just because every single one of the words containing them fall into oblivion. Nevertheless, the above reflections show that an overall low token frequency of a category, expressed by a high q in the context of this model, may help to facilitate a merger at the sight of an almost complementary distribution of phonemic categories in the contexts they are usually uttered in. Above that, the present discussion also showed that a low functional load might only be a reliable predictor of mergers if the lexical distribution is taken into account as a second variable.

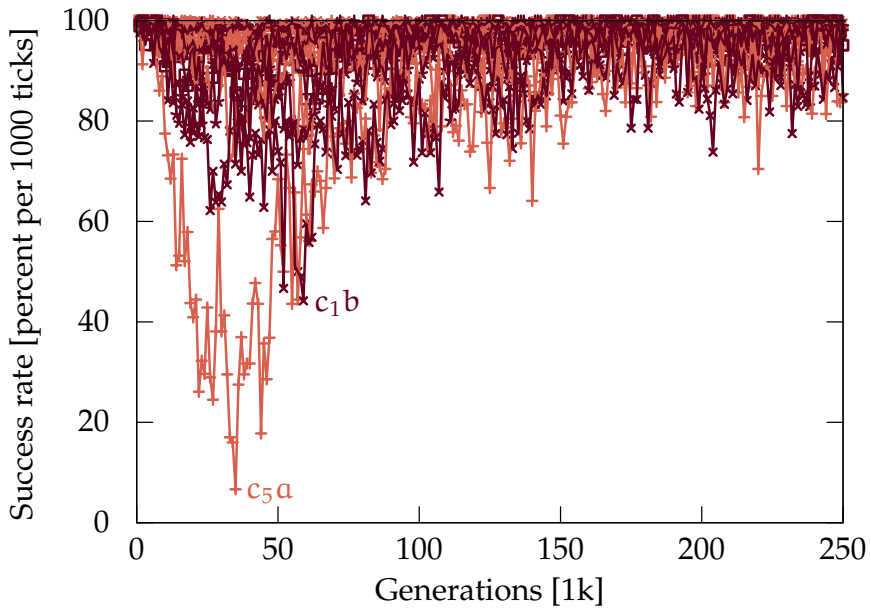

Figure 23: Success rates for the encoding of meanings of all words during the simulation depicted in fig. 22.

8.2.3 *Dual-Zipfian distribution*

As a last point, we want to turn to the remaining distribution, the dual-Zipfian. This section is kept rather concise, because the results follow straightforwardly from the above findings.

Just like under the within-category uniform distribution, the FL is fixed at 1.0 given that $q = 1$. In these cases, mergers never happen, because $p(c_i\mathbf{a}) = p(c_i\mathbf{b})$ for all i. For very high choices of s, e.g. $s = 10$, some meanings may become inexpressible because all exemplars of the words that express them disappear from the exemplar space. But this is solely due to their low utterance probability in the first place and would have happened as well in a stationary setting with $\lambda = 0$, i.e. if there is no initial shift involved.

If on the other hand, the slope is fixed at $s = 1$ and q is set to a high value, the FL assumes a relatively low value. But just like under the first distribution, mergers do not occur. They only would, if it was possible to approximate a complementary distribution, but the cross-Zipfian distribution is the only one of the three that can account for this. Here it shows again, that the FL is not a reliable predictor in its own right but only in the context of the cross-Zipfian distribution.

9 DISCUSSION

The model presented in this section was an alteration of model 1. This time, the exemplars had been enriched by a context and a meaning. This shifted the communicative incentive of the agents from a mere exchange of phonetic units to the transmission of extra-linguistic information, i.e. meanings, via en- and decoding using linguistic utterances consisting of a vector and a context. This allowed for a notion of usefulness concerning a given phonemic contrast. The goal was to show that this usefulness, which is reflected in the lexical, or rather contextual, distribution of phonemic contrasts, has an impact on the decision whether two categories merge or participate in a chain shift.

In order to explore the behavior of the system, three distributions that determine the utterance probability for meanings, expressed by 'words', were taken into account. Furthermore, the functional load of the opposition under consideration was registered before each run. Under two of the distributions, the within-category uniform and the dual-Zipfian distribution, mergers never occurred. Under the cross-Zipfian distribution, however, there appearance was dependent on the slope of the distribution that has a huge impact on the degree of the categories' lexical overlap. The steeper the slope, i.e. the less conflict there is regarding homophones, the more probable mergers became. The functional load of the contrasts was not predictive of mergers in all cases, since it could assume low values under the dual-Zipfian and the within-category uniform distribution even though they caused a resistance to mergers. However, in the case of the cross-Zipfian distribution, its value was able to indicate mergers beforehand.

A surprising result is that a mere majority of tokens on behalf of category **b** could not invoke a merger. This contrasts with the results of model 1. It seems that the incentive to exchange meanings draws stronger boundaries between phoneme categories.

In this model, the FL was not sensitive to the underlying probability distributions. Further investigation is required in order to clarify whether such a sensitivity is reached when analyzing natural language corpora, and if not, whether probability ditributions have to be taken into account as another variable in future research on the likelihood of mergers as a function of communicative usefulness.

Part IV

FINAL DISCUSSION

10 DISCUSSION AND OUTLOOK

10.1 DISCUSSION

The thesis at hand has shown that chain shifts can arise from the dynamics ingrained in exemplar-theoretic models, thus providing further evidence for the validity of these approaches. The two models conducted in this work were able to account for some of the most important phenomena regarding phonological chain shifts, albeit different ones for each model.

In the simulations surrounding the first model, both chain shifts and mergers could be observed. Chain shifts were the result of contrast maintenance between phonemic categories. If the token frequencies of each category were approximately equal, the shift of one category invoked subsequent shifts of other categories, but it was not possible to distinguish properly between push and pull shifts. The reason for this being that the underlying mechanism, which emerged purely from the exemplar dynamics, is the aspiration after maximum dispersion. Once the equilibrium of a system was disturbed by an initial shift, the other categories gradually adjusted to the new situation, ever so slightly maximizing the distance to neighboring categories. Because of this, gap-filling for example could be caused by both a diverging and an approximating category. This indistinguishability of push and pull shifts does not pose a severe problem, however. Many, if not most, chain shifts discussed in the literature are not definitely assignable to either type. Steady dispersion maximization that incorporates elements of both push and pull shifts might thus shed new light on the research on chain shifts. Furthermore, mergers could also be observed during the model runs. They occurred when there was a big difference in token frequency between phonemic categories. This might seem plausible in the context of a model whose notion of communication is a mere exchange of phonemes, but in reality token frequencies of phonemes are hardly ever uniformly distributed, and yet chain shifts ensue.

This last issue was tackled by the second model via emulating the usefulness of phonemic contrasts as a result of communicative needs. Within this model, mergers were more likely to occur the more the lexical distribution of the contrast in question approximated complementarity, otherwise push shifts were the result. The related notion of the functional load proved not a reliable predictor of mergers on its own, as distributions that were unable to approximate complementarity were always associated with push shifts, even if the functional

load was low. However, if it was known beforehand, that the lexical distribution shows signs of complementarity, the functional load was predictive of both mergers and push shifts. It remains to be seen whether this finding can be reproduced empirically.

Another noteworthy point regarding model 2 is that it exhibited sound changes which were both lexically and phonetically gradual. It was observable that the locus within the exemplar space at which the exemplars of a category reside is basically defined by high-frequency lexical items. A shift of an entire category is preceded by the movement of these, and only after their dislocation items with a low frequency can follow. This has to do with the activation which exemplars receive from their environment. Exemplars of high-frequency items receive a lot of activation from surrounding exemplars with the same meaning and can consequently be accessed directly in production. Thus, phonetic variation of other lexical items with a different meaning within the same phonemic category does not affect these exemplars. Exemplars of low-frequency items, however, do not receive sufficient activation from their semantic environment. Thus, their phonetic form has to be constructed on the basis of other members of their phonemic category, most notably the more frequent one. This is in accordance with findings which suggest that there are several types of sound changes which are initiated by high-frequency items. Other lexical items that contribute to the same phonemic category are gradually less progressive in proportion to their frequency (Bybee 2006, 2012).

Both models have in common that they exhibited a behavior during the simulation runs, that is akin to the proposed principles of sound change outlined in sec. 3.1. The noise parameter allows for phonetic variation, but this is uniformly distributed around the centers of the phonemic categories. Nevertheless, competition among categories ensures that signals be misperceived and consequently dismissed. On the other hand, percepts that fall within a safer range are more likely to be perceived correctly and affect future productions. If categories are too close to each other, selection, which is performed by the listener, leads to a directed shift towards another position in the exemplar space. Here, we see two principles at work that have been proposed in the literature. First, the speaker is responsible for variation and the listener for sound change (Ohala 1981, 1989, Blevins 2004). And second, sound change proceeds in a way that is non-goal-directed but instead driven by mutation, selection and reproduction, hence resembling biological evolution (Müller 1870, Blevins 2004, Wedel 2006).

Likewise, both models suffer from the same fundamental shortcomings. The outline of exemplar theory provided in sec. 2 suggested that linguistic structure be an emergent property of repeated exposure to language in use. Nevertheless, the notion of a label is a hard-wired

one. Neither do the agents derive the existence of categories themselves, nor is the set of labels mutable.

Furthermore, it is hard to map the number of cycles that go by during a simulation onto real time. On the one hand, within one run the agents exchange less utterances than an average person in a week, but on the other hand, this relatively small time frame several full-blown sound changes may take place.

To conclude, the computational models presented in this thesis proved able to account for chain shifts as instances of contrast maintenance within the framework of exemplar theory. As such they provided us with further evidence for the capability of explaining cross-linguistic regularities without adhering to teleological forces.

Despite their success, these models are far from comprehensive. Hence, the next section will give suggestions for a handful out of many conceivable extensions.

10.2 OUTLOOK

The models proposed in this work were subject to several trade-offs that seemed necessary in order to give an initial account on chain shifts in exemplar theory. Therefore, this last section will complete the thesis by providing a list with a selection of suggestions that may help inspire future research that is based on the present architectures.

- Assertions about the likelihood of certain aspects of the model behavior were solely based on the observation of multiple runs. No statistical methods were used to validate these conjectures. This could be made up for in future work.

- In the simulations, there were only two agents present which above that were supposed to be representative for a whole community of speakers. It would thus be interesting to see how linguistic change propagates through a given speech community using realistic network models that incorporate the latest findings on social relationships. It would be furthermore interesting to explore how different classes of network structures facilitate specific sound changes.

- Similarly, it could be explored how language changes over the course of multiple generations by including a possibility for agents to pass away, and language learners that have to infer categories from the input that they are exposed to without predefined labels.

- The utterances in both models consist of only a single phonetic component. Even in model 2 the remaining phonetic context is represented by a single abstract context tag. It would hence be

interesting to see how multiple vectors within a single lexical items influence each other, a basic approach of which can be seen in Wedel (2006).

- The only possibility for the agents to avoid homophony is to resort to more prototypical variants. Plausible novel compensation mechanisms could include the coining of new words or contexts.
- Perception and production both use the same kind of phonetic representation. Thus, the model could be enhanced with a unique way of representing articulatory information.

BIBLIOGRAPHY

Baayen, Harald, Ton Dijkstra, and Rob Schreuder. 1997. Singulars and plurals in Dutch: Evidence for a parallel dual route model. *Journal of Memory and Language* 37:94–117.

Blevins, Juliette. 2004. *Evolutionary phonology: the emergence of sound patterns*. Cambridge: Cambridge University Press.

Blevins, Juliette, and Andrew Wedel. 2009. Inhibited sound change: An evolutionary approach to lexical competition. *Diachronica* 26:143–183.

Bod, Rens. 1992. A computational model of language performance: data-oriented parsing. In *COLING-92: proceedings of the 15th International Conference on Computational Linguistics*. Nantes.

de Boer, Bart. 1999. Self-organisation in vowel systems. Doctoral Dissertation, Vrije Universiteit Brussel.

Boersma, Paul. 2007. Some listener-oriented accounts of *h*-aspiré in french. *Lingua* 117:1989–2054.

Boersma, Paul, and Silke Hamann. 2008. The evolution of auditory dispersion in bidirectional constraint grammars. *Phonology* 25:217–270.

Bybee, Joan. 2001. *Phonology and language use*. Number 94 in Cambridge studies in linguistics. Cambridge: Cambridge University Press.

Bybee, Joan. 2006. From usage to grammar: The mind's response to repetition. *Language* 82:711–733.

Bybee, Joan. 2012. Pattern of lexical diffusion and articulatory motivation for sound change. In *The initiation of sound change. Perception, production, and social factors*, ed. Maria-Josep Solé and Daniel Recasens. John Benjamins.

Bybee, Joan. 2013. Usage-based theory and exemplar representations of constructions. In *The oxford handbook of construction grammar*, ed. Thomas Hoffmann and Graeme Trousdale, Oxford Handbooks in Linguistics, chapter 4, 49–69. Oxford: Oxford University Press.

Dawkins, Richard. 1976. *The selfish gene*. Oxford: Oxford University Press.

Denby, Thomas Nathan. 2013. The filtering listener: dispersion in exemplar theory. Master's thesis, University of California, Santa Cruz.

Ettlinger, Marc. 2007. Shifting categories: An exemplar-based computational model of chain shifts. Technical report, University of California, Berkeley Phonology Lab.

Flemming, Edward. 1995. Auditory representations in phonology. Doctoral Dissertation, University of California, Los Angeles.

Hockett, Charles Francis. 1966. The quantification of functional load: a linguistic problem. Technical report, RAND Corporation, Santa Monica, CA.

Janson, Tore. 1983. Sound change in perception and production. *Language* 59:18–34.

Johnson, Keith. 1997. Speech perception without speaker normalization: An exemplar model. In *Talker variability in speech perception*, ed. Keith Johnson and J.W. Mullenix, chapter 8, 145–165. San Diego: Academic Press.

Jurafsky, Daniel. 2003. Probabilistic modeling in psycholinguistics: linguistic comprehension and production. In *Probabilistic linguistics*, ed. Jennifer Hay and Stefanie Jannedy, 39–95. MIT Press.

King, Robert D. 1967. Functional load and sound change. *Language* 43:831–852.

Labov, William. 1994. *Principles of linguistic change, volume I: internal factors*. Number 20 in Language in Society. Oxford: Wiley-Blackwell.

Labov, William. 2010. *Principles of linguistic change, volume III: Cognitive and cultural factors*. Number 39 in Language in Society. Oxford: Wiley-Blackwell.

Liljencrants, Johan, and Björn Lindblom. 1972. Numerical simulation of vowel quality systems: the role of perceptual contrast. *Language* 48:839–862.

Łubowicz, Anna. 2012. *The phonology of contrast*. Advances in Optimality Theory. Sheffield: Equinox.

Martinet, André. 1952. Function, structure, and sound change. *Word* 8:1–32.

Müller, Max. 1870. The science of language. *Nature* 1:256–259.

Nosofsky, Robert M. 1988. Similarity, frequency, and category representations. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 14:54–65.

Ohala, John J. 1981. The listener as a source of sound change. In *Papers from the parasession of language and behavior*, ed. Carrie S. Masek, Robert A. Hendrik, and Mary Frances Miller, 178–203. Chicago: Chicago Linguistic Society.

Ohala, John J. 1989. Sound change is drawn from a pool of synchronic variation. In *Language change: contributions to the study of its causes*, ed. Leiv Egil Breivik and Ernst Hakon Jahr, 173–198. Berlin: Mouton de Gruyter.

Ohala, John J. 2003. Phonetics and historical phonology. In *The handbook of historical linguistics*, ed. Brian D. Joseph and Richard D. Janda, Blackwell Handbooks in Linguistics, chapter 22, 669–686. Oxford: Blackwell.

Oudeyer, Pierre-Yves. 2006. *Self-organisation in the evolution of speech*. Number 6 in Studies in the evolution of language. Oxford: Oxford University Press.

Pierrehumbert, Janet. 2001. Exemplar dynamics: Word frequency, lenition, and contrast. In *Frequency effects and the emergence of lexical structure*, ed. Joan Bybee and Paul Hopper, 137–157. Amsterdam: John Benjamins.

Rosch, Eleanor. 1975. Cognitive reference points. *Cognitive Psychology* 7:532–547.

Rosch, Eleanor, and Carolyn B. Mervis. 1975. Family resemblances: studies in the internal structure of categories. *Cognitive Psychology* 7:573–605.

Rosenbaum, David A., Sascha E. Engelbrecht, Michael M. Bushe, and Loukia D. Loukopoulos. 1993. A model for reaching control. *Acta Psychologica* 82:237–250.

Sandler, Wendy, Mark Aranoff, Irit Meir, and Carol Padden. 2011. The gradual emergence of phonological form in a new language. *Natural Language and Linguistic Theory* 29:503–543.

Schleicher, August. 1850. *Die Sprachen Europas in systematischer Übersicht*. Bonn: H. B. König.

Schleicher, August. 1869. *Darwinism tested by the science of language*. London: John Camden Hotten.

Schuchardt, Hugo. 1885. *Ueber die Lautgesetze. Gegen die Junggrammatiker*. Berlin: Robert Oppenheim.

Shannon, Claude Elwood. 1949. The mathematical theory of communication. In *The mathematical theory of communication*, ed. Claude Elwood Shannon and Warren Weaver. Urbana, IL: University of Illinois Press.

Surendran, Dinoj, and Partha Niyogi. 2006. Quantifying the functional load of phonemic oppositions, distinctive features, and suprasegmentals. In *Competing models of linguistic change: evolution and beyond. In commemoration of Eugenio Coseriu (1921-2002)*, ed. Ole Nedergaard Thomsen. Amsterdam & Philadelphia: Benjamins.

Swingley, Daniel, and Richard N. Aslin. 2009. Lexical competition in younger children's word learning. *Cognitive Psychology* 54:99–132.

Tupper, Paul. 2014. Exemplar dynamics model of the stability of phonological categories. URL http://arxiv.org/abs/1405.0049.

Walsh, Michael, Bernd Möbius, Travis Wade, and Hinrich Schütze. 2010. Multi-level exemplar theory. *Cognitive Science* 34:537–582.

Wang, William S.-Y. 1969. Competing changes as a cause of residue. *Language* 45:9–25.

Wedel, Andrew. 2004. Self-organisation and categorical behavior in phonology. Doctoral Dissertation, University of California, Santa Cruz.

Wedel, Andrew. 2006. Exemplar models, evolution and language change. *The Linguistic Review* 23:247–274.

Wedel, Andrew, Scott Jackson, and Abby Kaplan. 2013a. Functional load and the lexicon: Evidence that syntactic category and frequency relationships in minimal lemma pairs predict the loss of phoneme contrasts in language change. *Language and Speech* 56:395–417.

Wedel, Andrew, Abby Kaplan, and Scott Jackson. 2013b. High functional load inhibits phonological contrast loss: A corpus study. *Cognition* 128:179–186.

Wittgenstein, Ludwig. 1998 [1953]. *Philosophische Untersuchungen*. Berlin: Akademie-Verlag.

Zipf, George Kingsley. 1935. *The psychobiology of language*. Boston: Houghton-Mifflin.

COLOPHON

This document was typeset using LaTeX and BibTeX with help of TeXstudio (texstudio.sourceforge.net) for editing and Jabref (jabref.sourceforge.net) for references.

The layout of this thesis is `classicthesis`, developed and arranged by André Miede, available under code.google.com/p/classicthesis.

The fonts used in this thesis are Palatino for plain text, AMS Euler for formulas, and `Bera Mono` for listings.

The simulation stills in chapter ii are taken from the custom-made graphical output of the simulation program. The graphs in chapters ii and iii were plotted using the freely available gnuplot (gnuplot.info).

The color scheme used in this thesis is taken from colorbrewer2.org and intended to be convenient for most kinds of color vision deficiency.

The models were implemented using the Java™ programming language. Their source code, licensed under GNU GPL, is available under github.com/vpersien/csinet.

Final Version as of December 5, 2014 (version 1.1).