Chapter 1

Towards hybrid modelling:
Optimality Theory & Exemplar Theory

In phonology, it is not common to investigate grammar and frequency together. Grammar is usually investigated by generative phonologists, and lexical frequency is usually examined by usage-based phonologists. These are traditionally distinct approaches. During the last few years, it has become increasingly clear that linguistic data must be accounted for in a model combining aspects of both approaches. The theoretical frameworks of the mainstream phonological models have therefore been further developed. From a generative perspective, the grammar has been augmented with the possibility to account for lexical frequency effects, like in Stochastic OT and Noisy Harmonic Grammar. From a usage-based perspective, the lexical model has been expanded with a production module and the notion of a “usage-based grammar”: a grammar that emerges from the lexicon. Whereas in Stochastic and Noisy Harmonic Grammar, the grammatical model is strongly developed, the role of the lexicon remains rudimentary. Contrastively, whereas in usage-based grammar the lexicon is fully developed, the model lacks a comprehensive grammar. Another important reason for the underdevelopment of these two components of language, grammar, and lexicon, is the lack of data that reveals the nature of the relation between the phonological grammar and the lexicon, that is, data that show the interaction between phonological grammar and other aspects such as frequency. This thesis provides such relevant data, to show possible interactions between the phonological grammar and the lexicon, and proposes a hybrid model with a fully fledged grammatical model as well as lexicon, and develops the connection between the two.
Linguistics during the twentieth century is characterized by a schism between two approaches. Competence vs. performance, langue vs. parole, I-language vs. E-language: these pairs all refer to two properties of language which were—and still are—usually investigated separately. De Saussure (De Saussure & Baskin 2011) was the first to make a distinction between langue and parole. These notions refer to the system of language and the physical realization of language, respectively. De Saussure considered langue and parole as two separate aspects of language, which are contrastive and form two different fields of study.

“S’il est vrai que langue et parole se suppose l’une l’autre, en revanche, ils sont si dissemblables qu’ils sont chacun leur théorie séparée.”

“Although langue and parole are complementary to each other, they are so dissimilar that they form each their own, separated theory”

(De Saussure, Cours de linguistiques générale S2.17, in De Saussure and Baskin 2011).

For de Saussure, langue is the essential part of language and the only part that is relevant for linguistic investigation.

“L’étude de language comporte donc deux parties; l’une, essentielle, a pour objet du langue, qui est sociale dans son essence et indépendante de l’individu.”

“The study of language consist therefore of two parts, one, essential, is ‘langue’, which is social in nature and independent of the individual”

(de Saussure, Cours de linguistiques générale ntr IV al.5, in De Saussure and Baskin 2011).

Similarly, half a century later, Chomsky distinguished between competence and performance:

“We thus make a fundamental distinction between competence (the speaker-hearer’s knowledge of his language) and performance (the actual use of language in concrete situations)”

(Chomsky 1965).

Although the separation is similar, Chomsky deviates from the Saussurian concept of langue, which he regards to be “a merely systematic inventory of items” (Chomsky 1965: 4). This dualistic approach resulted in linguistic research along two distinct lines during the twentieth century: on the one hand, the structuralist and generative investigation of the system of language and the innate knowledge of language, and on the other hand, the
sociolinguistic, psycholinguistic, and usage-based studies of language use, variation and change.

One of the reasons for the opposition between usage-based and generative approaches in phonology lies in the nature of the data that are investigated in phonology: speech sounds and their alternations. Speech sounds may differ from each other either gradually or categorically: that is, sounds can behave as either minimally differing on a continuous scale or as discrete units (see Ernestus (2011) for an overview). Continuous variation, characterized by subtle differences on a gradient scale, is grounded in the wealth of pronunciation differences of the same words, e.g. in many lenition processes. Some textbook examples are Spanish plosive lenition (e.g. Bybee (2001) and references cited there) or vowel reduction in many languages (e.g. van Bergem (1995), Crosswhite (2001), Jurafsky et al. (2001), among many others). Usage-based phonology focuses on such continuous processes and models them in the lexicon, i.e. all memorized words or morphemes, which, strictly speaking, often belong to the field of performance, like frequency effects, recency effects, or speech rate. Frequency effects are differences in behaviour between high-frequency (HF) linguistic items and low-frequency (LF) linguistic items (such as words, roots, or suffixes) that cannot be explained otherwise (see below in this section). On the other hand, sounds may also behave in a categorical way. Categorical alternation between sounds is grounded in categorical perception, the general tendency to perceive items on a continuous scale as discrete units (e.g. Harnad (1990) for a overview). Categorical perception in language leads to the classic phonemic representations, vz. sounds that contribute to a difference in meaning. Allophones, different alternants of a phoneme, are also categorical: they are supposed to have a particular feature, either monovalent or binary (Archangeli (1988), Dresher et al. (1994), among others) that other allophones do not have. An example of this allophonic alternation is English nasal place assimilation, where the phoneme /n/, underlyingly coronal, surfaces as labial [m] or velar [ŋ] if the following consonant is labial or velar respectively [m]plement or [ŋ]capable. Phonemic, but also allophonic, representations abstract away from phonetic detail. Alternations of categorically distinct speech sounds, that occur in well-defined phonological contexts, form the backbones of phonological competence or phonological grammar. Phonological grammar is usually assumed to consist of a set of rules or constraints that takes a lexical form as input and generates a phonetic output. Nasal place assimilation is a well-investigated categorical process in many languages. Other examples are final -t/-d deletion in English (see Coetzee (2004) for an overview) and rendaku in Japanese (see also chapter 5 and references cited there). Whereas usage-based phonology focuses mainly on continuous variation data, generative phonology largely focuses on such systematic categorical alternations. Thus the dual behaviour of speech sounds, i.e. continuous and categorical, lies at the heart of the dichotomy between usage-based and generative phonology.

Although the lexicon and the grammar are two distinct parts of language, which will be shown in this chapter, it remains to be seen what exactly the division of labour is between
the two. Therefore, a growing number of linguists currently seek a way to combine lexical and grammatical aspects in a single model. One of the challenges formulated by Jackendoff (2007) for linguistics in the future is to bridge the gap between the lexicon and the grammar. Such a gap, Jackendoff (2007) argues, does not really exist: in fact the lexicon and the grammar form a continuum. Similarly, Smolensky & Legendre (2006) claim that there is enough evidence which shows that both usage-based as well as generative approaches are on the right track; but whereas the lexical level stands for storage of concrete and detailed concepts and neural activation of these concepts, the grammar stands for the abstract concepts and higher mental representations of language. Both are needed in the architecture of a linguistic model (Smolensky & Legendre 2006). More specifically, for phonology, van de Weijer (2009, 2012) argues that a combination of usage-based and generative phonology has the advantage of being more psycholinguistically realistic, more balanced regarding storage, perception, and production, and able to account both for variation in the individual language user and the systematic typological differences between languages.

There exist, of course, different approaches to meet the call for hybrid modelling. First, one could extend the usage-based model with a grammar. Conversely, it would be possible to extend the generative model with lexical information. Thirdly, an entirely new model could be developed. Finally, one could try to combine the best of both ways: the mainstream usage-based approach and the mainstream generative approach. Section 1.1 and 1.2 will show that the possibilities to extend usage-based lexical models with a grammar and to extend the grammar with lexical information are either too limited, or the concepts are not fully articulated. New models are always a possibility, and a number of computational phonological models have been developed, although, according to Ernestus & Baayen (2011) they are also not yet fully operative. Both usage-based as well as generative phonology have developed models of either the lexicon or the grammar. The most widely used model in usage-based phonology is Exemplar Theory (Goldinger (1996), Johnson (2007)) and the leading theory in generative phonology Optimality Theory (Prince & Smolensky 2004).

Advantages and shortcomings of both models will be described in §1.1 and §1.2, where I will also show that Exemplar Theory (ET) lacks a clear notion of grammar, although reference to grammar is sometimes made. On the other hand, I will show that Optimality Theory (OT) is not adequate in accounting for frequency effects, although several models claim that they can do so. In other words, neither ET, nor OT provides a full account of language. Given the strength of the two models in their respective fields, however, it seems worthwhile to investigate the possibilities to combine them. A first step towards this approach has been taken by van de Weijer (2009, 2012), discussed in §1.3.3. In order to work out this concept further, however, empirical data are needed in which clear interactions between lexical and grammatical processes are investigated. The first aim of this thesis is to provide data which show such interactions. The second aim is to model these data in a hybrid model, based on ET and OT. But let us first look at well-established facts about frequency effects.
During the past two decades, new usage-based data have rapidly become available. The possibility to store large databases stimulated corpus research, which made linguists more aware of the wealth of variation in spoken data and the role of frequency of occurrence in this variation. It also became clear that this variation is neither random, nor can it be referred to as 'noise'; systematic patterns are found in variation that are related to reduction, frequency of occurrence, and lexical diffusion. In addition, many patterns of change which were earlier considered as Neo-grammarian, i.e. discrete, sound change turned out to be lexically diffused and continuous (see Labov (1994), Phillips (2006) for an overview). Moreover, frequency effects appeared to play an important and a systematic role in lexical diffusion. Numerous studies have convincingly shown that there are two lexical frequency effects that play a role in lexical diffusion: depending on the process, HF words can either be the last words to change, or HF words are the first words to change. In analogical change, LF words are more susceptible to change to fit into the analogical pattern than HF words. A textbook example comes from the strong verbs in Germanic languages: whereas Germanic verbs historically may have had two different stems, this became regularized over time, except for HF words, which have maintained their irregular stems (e.g. write-wrote-written and read-read-read), for a thousand years or more (Bybee (2001), van de Weijer (2012: 39-42)). Let us call this frequency effect, which occurs in analogical change, a Type I frequency effect.

(1) **Type I frequency effects (analogue change)**

Frequency of occurrence correlates with analogical change such that HF words are less likely to undergo analogical change and LF words are more likely to undergo analogical change.

This Type I frequency effect is explained by the relatively strong mental representation of HF words, which is difficult to change. The result is that, in language change, HF words appear to be conservative, unwilling to change, and may eventually retain autonomous behaviour, acting as lexical exceptions. On the other hand, HF words are the first to change if they are subject to lenition and reduction (Hooper 1976), or assimilation, which is also a kind of reduction (Phillips 2006). Let us refer to this reduction-related frequency effect as a Type II frequency effect.

(2) **Type II frequency effects (reduction)**

Frequency of occurrence correlates with reduction such that HF words are more likely to undergo reduction and LF words are less likely to undergo reduction.

Type II frequency effects occur because, by their very nature of being often used, HF words are subject to more automation processes on the articulatory level and consequently undergo more reduction. Reduction processes may lead to change, however, they may also occur synchronically. One could pose the question what occurs in other situations, when neither
analogue change, nor reduction occurs. For instance in stable variation without reduction or loanword integration. Such examples will be addressed in chapters 4, 5, and 6. Further, it should be noted that, of course, no sharp division line between HF and LF items exists; rather, LF and HF words form a continuum and theoretically speaking, an HF word in one process may behave as an LF word in another process. For instance, a word is subject to two patterns of change in a particular language, e.g. grammaticalization of irregular verbs and sound change. Compared to all other verbs, the word may be relatively HF. Compared to all nouns that are subject to change, the word may be relatively LF. We will see examples of this in chapter 5 and 6. We will see examples of relative frequency in chapter 5 and 6.

In this section, I discussed the fact that phonological research in the twentieth century is characterized by two approaches, generative phonology and usage-based phonology, which are usually considered to be incompatible. Subsequently, I pointed out that an increasing call for hybrid modelling exists. I will investigate the possibility to combine Exemplar Theory with Optimality Theory, on the basis of data which show an interaction between frequency effects and grammar. In the remainder of this chapter, I will introduce Optimality Theory and the way in which OT accounts for frequency effects in §1.1. Subsequently, I will introduce Exemplar Theory including the concept of usage-based grammar in §1.2. Section 1.3 provides the research questions and hypotheses. The final section (§1.4) provides an overview of this thesis.

1.1 From grammar to usage

In this section, I will outline the basic principles of Optimality Theory (§1.1.1), including two observations that lie at the heart of Optimality Theory, namely phonological typology (§1.1.2) and The Emergence of The Unmarked (§1.1.3). Further, we will look at the fashion in which frequency effects can be accounted for in the optimality-theoretical framework (§1.1.4).

1.1.1 Optimality Theory

The current leading generative model in phonology is Optimality Theory (OT) (Prince & Smolensky 2004), a constraint based grammar. OT fits into the feed-forward model of generative phonology, in which the lexicon provides an input for the grammar, which is an underlying, phonemic, form that is stored in the lexicon, and the grammar generates a phonetic output. The OT grammar consists of a small number of fixed components: the generator of input, candidates that are possible outputs, a set of ranked constraints, and an evaluation system EVAL. The generator GEN is a mechanism that generates candidates, possible phonetic output forms that can basically have any form, but for convenience, usually only the outputs that are the most relevant, and that differ minimally from the input, are presented. The set of constraints CON forms the grammar and works as a filter for the candidates. A candidate can either violate or satisfy a constraint. The best candidate, the winner, which represents the actual output, is the candidate that violates only the most lowly ranked constraints, in comparison with other candidates (McCarthy 2008). Basically, there are two kinds of constraints: markedness constraints and faithfulness constraints. Ideally,
markedness constraints are grounded in phonetics and prohibit sequences that are difficult to articulate or perceive. Faithfulness constraints preserve lexical contrast, and require that input and output are identical (Kager 1999: 11). In principle, since markedness occurs at the cost of faithfulness, and vice versa, markedness and faithfulness constraints are in conflict with each other (Kager 1999).

As an illustration, let us consider word-final -n deletion in Dutch. In Dutch, the suffix -ən is often reduced to -ə: for instance, in the infinitive and plural marker in verbs: /lezən/ → [lezə] ‘to read’ (Booij (1995), van Oostendorp (2005), van de Velde & van Hout (1997)). Under the assumption that the input contains a final -n, the most relevant candidates to consider are the faithful candidate [lezən] and the winning candidate [lezə]. One constraint that is needed for word-final -n deletion in Dutch is a constraint that prohibits [ən] at the end of the word.¹

(3) *ən]

Assign a violation mark to the word-final sequence [ən].

Furthermore, a faithfulness constraint is needed that compares the output to the input: a constraint that requires all segments to be realized.

(4) MAX-IO (McCarthy & Prince 1995)

Every segment of the input has a correspondent in the output.

(“No phonological deletion”).

Since [lezə] is favoured over [lezən], *ən] >> (dominates) MAX-IO. A candidate that performs equally well in this mini-grammar is [lezən], which is ungrammatical in Standard Dutch. So, we also need a constraint that prevents [lezən] to be the correct output, that is, a constraint that prohibits syllabic nasals.

(5) *NUC/nasal (De Lacy 2001)

Assign a violation mark to any syllabic nasal.

Since [lezən] is not a possible output in Standard Dutch, we derive the ranking *NUC/nasal >> MAX-IO. The ranking between *ən] and *NUC/nasal cannot be decided, in other words, these constraints are not crucially ranked. This grammar is illustrated in tableau (6). Constraint violations are indicated by an asterisk * and if this violation is crucial, which means that there is at least one candidate that is evaluated as more ‘harmonic’, or more grammatical, an exclamation mark ! is added. Crucial constraint ranking is indicated by solid lines and non-crucial constraint ranking is indicated by dashed lines. Cells that are irrelevant are shaded. Finally, the winning candidate, the actual output, is pointed with 🅰️(6).

¹ Final -n deletion does not occur in stems (van Oostendorp (2005)), but for reasons of clarity this is omitted from the analysis.
1.1.2 **Typology**

Constraints in OT are supposed to be universal and different languages vary in their constraint ranking. OT as such makes a typological prediction (which is absent from rule-based approaches). If the constraints that we used above are re-ranked, we arrive at other possible grammars. Typology is a core property of OT, which is formulated in the principle of Richness of the Base:

"Richness of the Base. The source of all cross-linguistic variation is constraint re-ranking. In particular, the set of inputs to the grammars of all languages is the same. The grammatical inventories of a language are the outputs which emerge from the grammar when it is fed the universal set of all possible inputs"

(Prince & Smolensky 1993: 191).

A comparison between word-final -en reduction in Dutch and German illustrates how typology is accounted for in OT. Many Dutch-German cognates exist in which reduction of the suffix <en> occurs, but the two languages differ in the way reduction is applied. Let us take the cognate Dutch <lezen> /lezən/ and German <lesen> /lezən/ ‘to read’. Whereas in Dutch, the final -n is deleted (see §1.1.1), in German, the schwa is deleted (Kohler & Rodgers 2001), resulting in [lezn] ‘to read’. The same constraints as in (6) are used. Additionally, in order for [lezn] to be the winning candidate, the final consonant should be realized and therefore we invoke the constraint \{RIGHT\}ANCHOR-IO.²

\[
\text{(7) } \{\text{RIGHT}\}\text{ANCHOR-IO (McCarthy & Prince 1995)}
\]

Any element at the designated periphery of the input has a correspondent at the designated periphery of the output.

² Anchor-IO belongs to the family of alignment constraints, which requires that certain prosodic boundaries coincide with morphological boundaries. Here the initial segment of the phonological word coincides with the right edge of the morphological word.
Since [lezn] is favoured over [lezə] in German, the partial ranking must be {RIGHT}ANCHOR-IO >> MAX-IO, *NUC/vowel. The ranking between *ən and {RIGHT}ANCHOR-IO cannot be decided. The tableau for German [lezn] is provided in (8).

(8) Tableau for schwa deletion in German.

<table>
<thead>
<tr>
<th>/lezn/</th>
<th>*ən]w</th>
<th>{RIGHT}ANCHOR-IO</th>
<th>MAX-IO</th>
<th>*NUC/vowel</th>
</tr>
</thead>
<tbody>
<tr>
<td>[lezn]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>*!</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Since constraints are universal, {RIGHT}ANCHOR-IO is also supposed to be present in Dutch, although it must be low-ranked (9).

(9) Tableau of word-final -n deletion in Dutch.

<table>
<thead>
<tr>
<th>/lezn/</th>
<th>*ən]#</th>
<th>*NUC/vowel</th>
<th>MAX-IO</th>
</tr>
</thead>
<tbody>
<tr>
<td>[lezn]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[lezə]</td>
<td></td>
<td>*!</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>*</td>
</tr>
</tbody>
</table>

Tableaux (8) and (9) show the typological difference in final -en deletion between Dutch and German, just by re-ranking the constraints. The fact that typology naturally emerges from the universal character of the constraints and the ranking mechanism makes OT superior to rule-based phonology.

1.1.3 The Emergence of the Unmarked

Constraints may be so low-ranked that they do not contribute to usual output. But in some cases, i.e. when there is no lexical input, one can observe that these low-ranked constraints do exist in the language. This is called the Emergence of the Unmarked (TETU) and was first described by McCarthy & Prince (1994). TETU effects typically occur in epenthesis and reduplication. The reason is that if there is no lexical input, faithfulness constraints, which compare lexical input to the output, are all vacuously satisfied and therefore candidates can only violate markedness constraints. As a result, the most unmarked candidate will win. For instance, in epenthesis, if a vowel is inserted, it is usually the most unmarked vowel (which is a central vowel, or, if there is no central vowel, a front vowel (Rice 1995). If a consonant is
epenthesized, it is generally $h$, which has no supralaryngeal place features (Lombardi 2002). As for reduplication, the reduplicant is usually formed in relation to the base, the output rather than in relation to the input. Thus the reduplicant has no lexical input, and will vacuously satisfy all IO-correspondence constraints. The most unmarked candidate will thus be the winner. This can be illustrated by reduplication in Leti.

In Leti, an Austronesian language spoken on the island Leti, reduplication occurs at the left-hand side of the base. Any marked structure of the base, however, is not copied in the reduplicant. Whereas lexical forms may have geminates, the reduplicant does not. Whereas lexical forms may have long vowels, the reduplicant does not. And whereas the lexical forms may contain a consonant cluster, the reduplicant does not (Sloos & van Engelenhoven 2011).

<table>
<thead>
<tr>
<th>(10)</th>
<th>Lexical form</th>
<th>Gloss</th>
<th>Reduplication</th>
<th>Gloss</th>
</tr>
</thead>
<tbody>
<tr>
<td>mmerna</td>
<td>to become swift</td>
<td>me-mmerna</td>
<td>swift(ly)</td>
<td></td>
</tr>
<tr>
<td>luuβu</td>
<td>rest, remnant</td>
<td>lu-luuβu</td>
<td>remaining, rest. ADV</td>
<td></td>
</tr>
<tr>
<td>mili</td>
<td>to be sour</td>
<td>m-li-lili</td>
<td>sour</td>
<td></td>
</tr>
</tbody>
</table>

The markedness constraints that are responsible for the restrictions on the reduplicant are $^*C$, $^*V$, and $^*CC$ respectively.

(11) $^*C$:
Assign a violation mark to any long consonant.

(12) $^*V$:
Assign a violation mark to any long vowel.

(13) $^*CC$
Assign a violation mark to any consonant cluster.

Crucially, as shown in (10), the lexical forms do allow for geminates, long vowels, and complex onset clusters, so faithfulness constraints like MAX-IO are higher ranked and therefore, without reduplication, the constraints in (11)-(13) are “invisible”. The output candidate violates a maximality constraint which compares the base and the reduplicant (14). Tableau (15) shows an example of reduplication in Leti.

(14) MAX-BR (McCarthy & Prince 1995)
Every segment of the base has a correspondent in the reduplicant.

---

3 The analysis is simplified here; in Sloos & van Engelenhoven (2011) it is proposed that the reduplicant corresponds to the input rather than to the base.
1.1.4 Frequency effects in Optimality Theory

In the light of the knowledge about frequency effects, some phonologists undertook the challenge to account for such frequency effects in OT. Several OT models in which frequency effects are modelled have been proposed: Stratal OT (Anttila (1997, 2006), Anttila & Cho (1998), Stochastic OT (Boersma 1998), Noisy Harmonic Grammar (Coetzee & Pater (2008), Pater (2009), Potts et al. (2010)), and also Bíró (2006). In Stratal OT, constraints that are involved in the variation are not crucially ranked, so each time an evaluation occurs, the ranking may vary. Anttila (2006) assumes that frequency effects occur precisely in case the grammar cannot decide, i.e. when the constraints are not crucially ranked. Evidence comes from assimilation in Finnish. In bimoraic verbs, and not in other cases, the grammar is indecisive about the metrical parsing and can vary: either exhaustive parsing (CVVCV) or extrametricality (CVV)CV occurs, which depends on the metrical structure of the word. When the grammar is underdetermined about the metrical structure, frequency effects occur, according to Anttila (2006). In other cases, grammar overrides the frequency. A serious concern with respect to this model is that it is extremely unlikely that both variants are identically ranked by all constraints, in other words, there will (almost) always be another constraint somewhere in the grammar that favours one of the two variants. Notice also that the number of possible rankings and outcomes correspond to the ratio of the variants of a word, which is something different than word frequency as described above.
Stochastic OT and Harmonic Grammar (HG) are two models that are similar to a high extent and are based on weighting of the constraints, rather than ranking in order as in classical OT (in fact, HG is the predecessor of OT (Legendre et al. (1990), Prince & Smolensky (1997)). Constraint violations are the product of the constraint weight and the candidate violation. The value of the violations is summed and results in a harmonic weight for each candidate. The candidate with the highest weight is the winner. This is illustrated for, again, word-final -n deletion in Dutch, where I have assigned weight to the constraints so that the right output wins (16).

(16) Tableau in Harmonic Grammar of word-final -n deletion in Dutch.

<table>
<thead>
<tr>
<th>/lezən/</th>
<th>*ən</th>
<th>ₙ</th>
<th>NUC/vowel</th>
<th>MAX-IO</th>
<th>{RIGHT} ANCHOR-IO</th>
<th>Harmonic weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>[lezən]</td>
<td>-1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-4</td>
</tr>
<tr>
<td>[lezn]</td>
<td></td>
<td>-1</td>
<td>-1</td>
<td></td>
<td></td>
<td>-8</td>
</tr>
<tr>
<td>[lezɔ]</td>
<td></td>
<td></td>
<td>-1</td>
<td>-1</td>
<td></td>
<td>-3</td>
</tr>
</tbody>
</table>

An advantage of HG models is the generation of variable output by changing the weight of the constraints. Suppose variation occurs between [lezə]-[lezn],\(^4\) we can easily change the weight of the constraints so that [lezn] wins (see e.g. Boersma (1998)). If, hypothetically speaking, the forms [lezə]-[lezn] vary with a ratio of 40-60%, overlap of constraints makes it possible to account for that as well. Stochastic OT and HG (and also Biró (2006)) generate output with variation, which is to be understood as noise, and it is claimed that in this way these models can account for frequency effects—but the frequency effects mentioned in the introduction refer to a different concept. The variation that occurs by word frequency, Type I (analogy) or Type II (reduction) frequency effects, and moreover the systematicity of these frequency effects, is not accounted for by these models (this was first noticed by Coetzee & Kawahara (2013)). Since the constraint ranking and the constraint weight do not vary per word (which would of course be odd, since the grammar is supposed to be not word specific), these models treat all inputs identically, regardless of their frequency. Thus in Stochastic OT and HG, the ratio of [lezə]-[lezn] is predicted to be the same as of [wezə]-[wezn]. The frequency effect that is to be explained, however, is the ratio between the different variants of a word that is dependent on the frequency of that word. So, since the token frequency of the

\(^4\) As can be found in Groningen, in the northeast of the Netherlands.
Towards hybrid modelling

words *lezen* and *wezen* differs, we expect different ratios for the variants of these words. This is not accounted for in the aforementioned models.

Since stochastic OT/HG cannot account for the systematic behaviour of frequency effects, Coetzee & Kawahara (2013) developed a model, based on HG, in which frequency effects are accounted for by scaling of the constraints. The model does provide the correct output of frequency effects, but there are some theoretical complications. Coetzee & Kawahara (2013) propose that relevant constraints be scaled by scaling factors, which multiply the weight of the relevant constraints. HF words cause higher scaling factors and LF words cause lower scaling factors. The value of the scaling factor is part of the constraint, and the scaling factor itself represents the frequency value. Thus, frequency information is directly represented in CON, that is in the grammar—in my point of view, this is problematic since frequency as such is not a grammatical feature. Frequency is a property of a single word and therefore, variation on the basis of frequency is not a task of the grammar, because the grammar applies to all words equally. Moreover, it is unclear how the attribution of different scaling factors, reflecting frequency, can be justified in OT. The difficulty lies in the fact that in generative phonology, the level of activation is not represented in the lexicon. Since in generative grammar, the lexicon contains just one categorical underlying form per word, I fail to see how the lexicon can provide frequency information (as scaling factors or in any other way) to the OT grammar. In other words, variation can be captured by different constraint rankings for different individuals or different sociolinguistic strata, but “the place for statistics is outside and not inside the grammar” (Smith 1997).

In this section, I introduced Optimality Theory, which is the mainstream model for generative phonology. Its main characteristics are a fully fledged grammar and a lexicon without information that belongs to performance, such as frequency effects. OT naturally accounts for typology and TETU effects. I also discussed how frequency effects are treated in OT, namely by weighted constraints. Although variation can be generated by Stochastic OT and Harmonic Grammar, frequency effects are clearly more difficult to account for without a more developed model of the lexicon. We will now turn towards Exemplar Theory, a theory in which frequency effects receive a natural explanation.

### 1.2 From usage to grammar

In this section, I will introduce Exemplar Theory (§1.2.1) and explore the vision of Exemplar Theory regarding the grammar (§1.2.2).

#### 1.2.1 Exemplar Theory

Exemplar Theory (ET) is originally a psychological theory that models concept learning by categorization (Medin & Schaffer (1978), Nosofsky (1986)). ET is concerned with storage in the memory, activation of stored items, and retrieval of stored information. Storage of newly experienced items occurs by comparison to all other stored instances. The items are categorized as exemplars on the basis of similarity with the other exemplars: they are stored
in those categories which they resemble most. An item can be stored in more than one category, at different levels and in several subcategories. The activation of a category corresponds to the number of exemplars in that category. Retrieval of a stored item depends on activation levels: categories with many exemplars are retrieved faster. Retrieval also depends on recency: the most recently stored exemplars are activated faster than older exemplars, since exemplars are subject to memory decay. Exemplar models perform best for modelling group variability, since by remembering a single exemplar, the whole exemplar category is activated and its members can be directly compared to each other (Bodenhausen et al. 2003: 264).

ET has been adopted as a model for speech perception by Johnson (1997), extended to a model of speech recognition by Goldinger (1996), and elaborated on as a phonological model by Bybee (1999, 2007, 2010) and Pierrehumbert (2000, 2001). The core idea of ET as a phonological model is that memory is practically unlimited and that each linguistic item that is perceived is directly stored and categorized as an exemplar in the lexicon. The decisive factors in categorization are the meaning of a token (Bybee 2007: 717) and categorical perception (Pierrehumbert 2001). Since any token that is perceived is stored in the lexicon, fine-grained phonetic detail and variation are automatically part of the lexicon, i.e. any realization of the word *bird* word is stored in the category with the label *bird*, with all subphonemic detail. The ET lexicon is word-based. It mainly contains categories of exemplars of words although larger units, like fixed expressions and collocations, may also be stored as exemplars (e.g. Bybee 2010). Exemplars are stored with both linguistic and non-linguistic information. The linguistic information that is stored in an exemplar consists of e.g. pronunciation, word class, and meaning. The non-linguistic information may consist of e.g. the speakers’ voice/identity, sociolinguistic information (Pierrehumbert 2001), and orthography (see chapter 2). The words are stored in an analogical network to capture morphological relations, which is illustrated in Figure 1.1 (page 17).

The Exemplar model reflects frequency by the number of exemplars and the activation level of the category. Frequency information is not directly stored in the lexicon, but frequency effects are indirectly accounted for by the activation levels (Bybee (2001), Pierrehumbert (2001)). These activation levels are, of course, related to frequency (HF words cause higher activation levels than LF words (Morton 1969)), however, other usage-based factors also contribute to activation levels and interact with frequency effects, such as recency (e.g. Abramowicz (2007), Hay et al. (2006a)) and saliency effects (Rácz 2012). The relatively quick lexical access of HF words, the conservative behaviour of HF words in language change, and the susceptibility of HF words to reduction processes, naturally follow from either the exemplar modelling or the articulatory processes. First, HF words have relatively quick lexical access: they are represented by more exemplars than LF words, thus their mental representation is stronger, which means that their latent neural activation level is higher. Quick lexical access of HF words is also reflected in shorter reaction times in elicitation (Jurafsky 2003). Secondly, the relative strength of the lexical representation of the
word corresponds to its frequency and is negatively correlated with the likelihood of change: HF words have a stronger mental representations and as such resist language change more than LF words. Finally, HF words are more susceptible to reduction effects: because they are often used, the articulatory muscles are better trained and automation processes often cause reduction (cf. Gahl (2008)). Thus, in ET, frequency effects are naturally derived from the lexicon, without reference to actual frequency values.

Exemplar Theory is particularly strong in lexical storage of fine-grained phonetic detail, variation, frequency effects, recency, and saliency, i.e. a closer tie to psycholinguistics than OT. In addition, a production component has been added to ET by Pierrehumbert (2000). In the ET production model, the category is targeted through activation. Depending on the situation, different styles or social settings, different parts of the exemplar category may be targeted. Activation plays a key role in production, the higher the activation of a certain part, the higher the chance this is selected for production. However, production probably does not exactly mirror the target, since there is always deviation in the motor control. Thus each word will show a small amount of variability in production.
1.2.2 Usage-based grammar

The usage-based approach does not deny the existence of grammar, but regards grammar as the result of experience with language. This so-called usage-based grammar, as introduced by Bybee (2007), proposes that grammar is a cognitive abstraction which is formed by categorization over similarities and dissimilarities in the lexicon (see also Beckner et al. (2009) and references cited there). According to Bybee (2007), the grammar may change in different ways. For example, by the birth of prefabs, phonological reduction, and new constructions that are created out of existing constructions. Prefabs are complex expressions which are stored as one unit, like English go bananas. Phonological reduction occurs in HF words and constructions like be going to, which tends to be reduced to gonna. A new construction is, for instance, non-literal use of a sentence, like “What are you doing with that knife?”, which can have the literal meaning “What are you doing with that knife?”, but will more often be interpreted as “Why do you have that knife?” (Johnson 1997: 17). Beckner et al. (2009) also argue that grammaticalization occurs by usage, that is, phonetic changes and reorganization of categories change the grammatical form of the category. These examples support the point of view that certain categories may change, but it still seems to me that these changes occur in the lexicon, i.e. the formation of new words, rather than in the active grammar in the phonological sense. However, phonological grammars do undergo change, with a long period of lexical diffusion in which frequency effects are observed. Such changes often start as a phonetic rule, become phonologized over time, and subsequently may be morphologized (Bermúdez-Otero 2007). For instance, in German umlaut, the fronting of back vowels in the stem was initially caused by phonetic coarticulation of the stem vowel with a high front unrounded vowel [i] in the suffix. Umlaut developed to a phonological rule. Currently umlaut occurs, but only in particular morphological contexts and the trigger has disappeared, since the suffixes do no longer contain the high front unrounded vowel anymore (Bermúdez-Otero 2007). ET can make the generalization of the phonetic rule, the phonological rule, as well as the morphological rule, but grammar is more than a set of generalizations over the lexicon.

But why do we need a grammar, if it is the case that all linguistic information is stored in the lexicon? There are a number of reasons to maintain a fully fledged grammar in phonology. First of all, the grammar is not supposed to be a passive abstraction, but rather an active mechanism, that can account for the fact that new words, i.e. neologisms, loanwords, and new combinations that are the result of word formation processes, behave according to the rules. Grammar can also account for errors in L2 acquisition. Further, ET fails to account for typology and also TETU effects (van de Weijer 2012), which are the strong points of OT, as argued in §1.1. Besides, phonology is not restricted to words, but there are phonological processes above the word level, such as nasal place assimilation on word boundaries, raddoppiamento sintattico in Italian (see e.g. Napoli & Nespor (1979)) among many others) and intrusive r in English (e.g. Huddleston & Pullum (2002)). As pointed out by Boersma (2012), sentence level phonology is a problem for ET, since ET is mainly word-
based. So there are several motivations for maintaining the grammar in phonology, also if a richly specified lexicon with all phonetic detail is assumed.

In sum, ET proposes a model of a lexicon specified with much phonetic detail, which naturally explains many usage-based processes, including frequency effects. However, the concept of the grammar is rather passive and rudimentary. ET in its current form fails to account for underlying forms, the application of phonological rules when there is no lexical input, typology, and sentence phonology.

1.3 Towards hybrid modelling

In §1.1 and §1.2, I argued (along the lines of van de Weijer (2009, 2012)), that the mainstream models in phonology, OT and ET, have their own particular strengths and weaknesses: OT is a grammatical approach that is particularly strong in TETU effects and typology, but which seems not appropriate to account for frequency effects, while, on the other hand, ET can easily account for frequency effects, but lack a grammatical component. In this section, I will take the position that, in order to account for frequency-grammar interactions, a hybrid model of ET and OT appears to be an adequate option.

As mentioned in the introduction, in language change and in synchronic variation, frequency effects are often attested. Variation occurs in the language community, but it is also reflected in the individual lexicon. From an ET perspective, each time a variant is perceived, it is supposed to be stored as an exemplar, and the exemplar cloud is continuously updated. From a generative perspective, the underlying form and the grammar (i.e. the constraint ranking) change over time, which can even occur during a speakers’ lifetime. Therefore, it is not only important to investigate variation in the language community, although variationist studies necessarily begin there, but also to examine and model language change and reduction in ‘synchronic models’. In long-term change, we may distinguish clearly defined categorical changes of the grammar which must be accounted for. This is possible in OT, but it is at best a description and not an explanation (see also Gess (2003), McMahon (2000), Reiss (2003)). What occurs in between these stages is at least equally interesting and should be simultaneously accounted for (this will be shown in chapters 6 and 7).

As can be derived from the previous sections, it is clear that we process language in a categorical way, but also that we are able to store fine-grained detail in the lexicon. If we do not want to restrict ourselves to either the study of the lexicon, or the study of the grammar, but, instead, provide a fully fledged account of language, hybrid modelling of lexicon and grammar is necessary (Jackendoff (2007), Smolensky & Legendre (2006), van de Weijer (2009, 2012)). As mentioned above, there are different ways for hybrid modelling. One could extend generative grammar with lexical information, but we saw in §1.1 that there are a number of problems with this approach. On the other hand, it is possible to extend ET with a grammar, but although this has been proposed by Bybee (2007), usage-based grammar is not fully worked out. Another possibility is to construct other models, e.g. there are some
computational phonological models. Ernestus & Baayen (2011) provide an overview of these models, but argue that they are not yet fully articulated as well. However, given that the most successful models of generative and usage-based phonology are OT and ET, respectively, and that they are complementary in that OT provides a grammatical model and ET a lexical model, as described above, I prefer to combine the two into a hybrid model.

1.3.1 Prototypes

Nonetheless, the combination of OT and ET is, technically speaking, not straightforward. The lexical input for the grammar is a challenge for combined modelling, since OT allows only for underlying representations in the lexicon, whereas, in contrast, ET requires a richly specified lexicon. This can be solved in at least two different ways: either one assumes that an abstraction is made over exemplar categories, which serves as the input for the grammar, or one assumes that there is no input at all, but that exemplars are directly available for the grammar as candidates. The first approach, which will be discussed in this subsection, is common in psychology, the second approach is adopted by van de Weijer (2012) and will be discussed in §1.3.3.

As we saw in §1.1, the grammar is supposed to take an underlying form as an input and generates a phonetic output. What are the viewpoints of ET on the underlying form and grammar? According to Pierrehumbert (2001), ET is a theory in which prototypes are easily dealt with: “Another obvious success of the model is its treatment of prototype effects” (Pierrehumbert 2001: 143). Similarly, Bybee (2007) considers it as a positive fact that ET exhibits prototype effects.

“Exemplar clusters are categories that exhibit prototype effects. They are organized in terms of members that are more or less central to the category rather than in terms of categorical features”

(Bybee 2007: 717).

The positive aspect of the prototype to which both authors refer to is that a prototype representation might be linked to what is understood as an underlying form in generative phonology. Unfortunately, the viewpoint that Exemplar Theory makes use of prototypes is not supported in the psychological literature. On the contrary, prototype models are regarded to be essentially different from exemplar models. In prototype theory, the distance to the prototype is crucial in categorization. Prototype theories have the disadvantage that there are no relations between the category members (Ross & Makin 1999). In Exemplar Theory, on the other hand, the distance to all other representations is essential in storage. Concerning linguistics, prototype models seem better suited for underlying representations. On the other hand, exemplar models are stronger in explaining variation between category members in analogical networks. Note that this paradox is not only relevant for linguistics, but also in other learning processes. Therefore, some psychologists working within the ET framework argued that the assumption of some abstractions is necessary and cannot always
Towards hybrid modelling

be distinguished from exemplars (see Barsalou (1990) for an overview). What are these abstractions? Abstractions over exemplars may contain idiosyncratic information and information about correlations between different members of the exemplar category. They are not static and unchangeable like underlying forms, but are continuously updated, since the exemplar clouds are constantly changing. In exemplar-only models, abstractions can be made in production, but they are not stored, on the other hand, in so-called mixed models (ET with abstractions or prototypes), abstractions are stored as separate items in the memory (Juslin et al. 2003). Abstractions are not necessarily the same as prototypes, however. The crucial difference seems to be the arrangement of the exemplars: connected to each other or centralized towards a prototype.

There is evidence from psychology that, generally speaking, humans learn and categorize initially in an exemplar-based way and form prototypes in a later stage (rather than exemplar storage only). For instance, Sherman (1996), after testing stereotypes in social development in low-experienced and high-experienced test-settings, noticed that at the initial stage of experience, the character of the category is principally exemplar-based. In later stages, the representations become more abstract and stable. Sherman (1996), therefore, called for a mixed exemplar-prototype model for category learning in psychology. Such a mixed model would be beneficent for linguistics also, since it can capture prototypes as well as analogical networks, and thus provides the input for the grammar and still account for analogical networks. This would mean that in initial stage, a newly perceived word is stored in an exemplar-based way, and prototypes are constructed when the word is perceived more frequently.

Do we really need the notion of a prototype in phonology? And if so, what kind of information is stored in the prototype? Prototypes play a crucial role in cognitive phonology (Lakoff 1993), in which they contain more information than the underlying forms in generative phonology. First, in cognitive phonology, prototypes play an important role as phonemes. Moreover, within a particular category or abstraction, there may still be perceptually different and salient contrasts. The central members of these subcategories can become a prototype as well. In this way, even allophones can also have prototypes (Välimaa-Blum (2005: 70)). So, unlike what is called an underlying form in generative phonology, prototypes are not only constructed on a phonemic base. Rather, (sub)categories with their prototypes may also be formed on the basis of different varieties, stylistic variation, and social stratification (Kristiansen (2003: 93) and references cited there). Pierrehumbert (2001) hypothesizes that different subcategories of exemplars may be targeted for production on the basis of sociolinguistic or other criteria. Comparison between different (allophonic or allomorphic) alternants may lead to a prototype that is similar to what is known as the underlying form in generative phonology. Thus we arrive at the following definition for phonological prototypes:
(11) **Prototype (in phonology)**

A prototype is an abstraction across one or more related exemplar categories and/or subcategories.

So, the prototype is not necessarily simply an average of an exemplar category, which is in line with psychological research for decades that has shown that categorization does not only depend on just exemplar storage but also on different sorts of knowledge. Moreover, categorization can be adapted to the situation (see Juslin et al. (2003) and references cited there). The prototype may consist of phonemes, but it does not necessarily consist of only phonemes: the prototype may be an underlying form in the generative meaning, but also a sociolinguistic salient form. Therefore, although a perfect realization is almost never perceived, speakers are often able to tell what the most well-formed realization should be: they have a prototypical realization in their lexicon. Similarly, speakers are often much aware of a particular sociolinguistic or dialectal pronunciation, although they do not use it themselves: different prototypes may exist which relate to different sociolinguistic parameters. The question whether we do need exemplar storage as well as prototypes will be addressed in part II of this thesis, where I will show that there is evidence for two kinds of lexical storage, which nicely fits into the exemplar-prototype mixed model which has been proposed by Sherman (1996).

1.3.2 **EPOT (Exemplar-Prototype-Optimality Theory)**

Summarizing, in order to account for grammar as well as usage-based phonology and grammar-frequency interactions, I propose a combined model based on an exemplar-based lexicon, including prototypes, and a constraint-based grammar: Exemplar-Prototype-Optimality Theory (EPOT). This model combines the strengths of ET and OT, namely a richly specified lexicon with much phonetic detail and a fully fledged grammar. EPOT is constructed as follows. Any perceived linguistic word is stored in the lexicon as an exemplar. Exemplars are stored in word categories on the basis of their similarity and dissimilarity. These categories are constructed on the basis of categorical perception and differences in meaning. In the initial state of the construction of a category, i.e. in language acquisition or very infrequent words, the category consists only of a few exemplars. In later stages, prototypes are automatically derived from categories that consist of more exemplars. These prototypes may consist of phonemes, but may also contain specific information related to certain sociolinguistic parameters, particular varieties, word specific information, or other idiosyncratic information. This suggests that words which are stored in an exemplar-based way only, may behave differently than words which have a prototype. We will see examples of this in chapter 4 and 5. The prototype forms the crucial link between the lexicon and the grammar: it may be regarded as the output of the lexicon, which is simultaneously the input for the grammar. The grammar is constraint-based, categorical, and blind for fine-phonetic detail which enters the grammar with the prototype. The grammar may change the prototype according to the grammatical rules, which I regard to be generalizations over the lexicon.
However, these alternations are always categorical. Fine-grained information, as long as it is not affected by a constraint, are retained in the output. This is worked out in detail in chapter 7. EPOT, as will be shown in this dissertation, can account for categorical and non-categorical variation in synchronic language variation (e.g. chapter 4 on rendaku in Japanese), as well as diachronic language change (e.g. chapter 7 on the variation of the long vowel <ä> in German). In synchronic variation, EPOT accounts for usage-based as well as grammatical effects. In diachronic change, different grammatical stages are assumed, represented by tableaux, with in-between changes of the lexicon. In this dissertation, data will be presented that show that both the lexicon and the grammar should be modelled in detail, and that there seem to be evidence for different patterns of storage in the lexicon: exemplar-only and prototype-based. EPOT will be tested by the case studies presented in this thesis.

1.3.3 Alternative: Grammar as Selection (van de Weijer 2012)

EPOT builds on the proposal of van de Weijer (2012): “Grammar as Selection” (GS). This hybrid model also combines ET and OT, but there is a crucial difference. Whereas in EPOT the prototype forms the input of the grammar, in GS, there are no inputs at all. Thus no input-output correspondence occurs. Instead, candidates are formed by exemplars. Tokens of exemplars, which play the role of candidates, are evaluated on the basis of output-output correspondence, that is, exemplars are compared to each other as members of analogical networks (see Benua (1997, 2000), Burzio (1998, 2000), McCarthy & Prince (1995)).

\[(12) \quad \text{OO-MAX}\]

The output should correspond to the other output forms.

GS has a Harmonic Grammar component, in which not only are the constraints weighted, but the candidates are too. Candidate weight depends on their frequency. The summed products of candidate weight (in the rows of the tableau) and constraint weights (in the columns of the tableau) result in the harmonic weight. Violation of a constraint results in a negative value and satisfaction of a constraint results in a positive value. The winner is the candidate with the highest harmonic weight. A possible analysis for final -n deletion in Dutch in GS is provided in (13), where I added the weights and where I changed the constraint IO constraints into OO constraints, in order to make the comparison between the different candidates possible.
(13) Tableau in Grammar as Selection (van de Weijer 2012) of word-final -n deletion in Dutch. Frequency weights of the candidates are for illustration only.

<table>
<thead>
<tr>
<th></th>
<th>3</th>
<th>5</th>
<th>3</th>
<th>1</th>
<th>Harmonic weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>NUC/vowel</td>
<td></td>
<td></td>
<td>MAX-OO</td>
<td></td>
<td></td>
</tr>
<tr>
<td>{RIGHT}ANCHOR-OO</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 [lezn]</td>
<td>12</td>
<td>20</td>
<td>-12</td>
<td>-4</td>
<td>16</td>
</tr>
<tr>
<td>1 [lezn]</td>
<td>3</td>
<td>-5</td>
<td>-3</td>
<td>1</td>
<td>-4</td>
</tr>
<tr>
<td>2 [lezan]</td>
<td>-6</td>
<td>10</td>
<td>6</td>
<td>2</td>
<td>12</td>
</tr>
</tbody>
</table>

In stylistic variation, the relevant constraints have a different weight. For instance, in formal speech, it is likely that [lezn] would be the winner. The relevant constraint, which according to van de Weijer (2012) would be MAX-OO, which would be stronger so that [lezn] would have the highest harmonic weight.

Several comments are in order, regarding empirical evidence, stylistic variation, frequency values of the candidate, and (over)representation of frequency. Firstly, van de Weijer (2012) claims that GS is compatible with frequency effects, since ET is one of the two components of the model. This claim, however, comes without empirical tests. Secondly, GS is supposed to account for stylistic variation, by varying the weights of OO-correspondence constraints according to the register. In case there would be a prototypical or ‘ideal’ underlying form, it could be safely assumed that in formal speech styles, one tries to be closer to that ‘ideal’ form than in casual speech style. But in OO-correspondence, it is unclear why in formal speech the winning candidate should be “realized in an identical fashion across all instantiations” (van de Weijer 2012: 59). On the basis of exemplars only, one would expect that such a group mean as a winning candidate is formed only in the most frequent style—which is probably casual. Thirdly, in GS, frequency values are assigned to the candidates. That would not be possible if a candidate were the same as a single exemplar (since the frequency of a single exemplar is of course 1). A candidate in GS is therefore a token that consists of a mapping of exemplars that cannot be distinguished in perception. How this frequency type relates to word frequency remains unclear. Finally, in GS, faithfulness constraints emerge from generalizations across the lexicon and are represented by Output-Output constraints in order to make the grammar non-derivational (van de Weijer 2012 and references cited there). Similarly, markedness constraints are derived as abstractions over the lexicon (van de Weijer 2012: 59). These abstractions naturally contain
frequency information (represented by constraint weight). This suggests that, since constraint weight represents frequency information and the candidate weight represents frequency as well, frequency appears to be doubly represented in the harmonic weight. This overrepresentation of the grammar might not be problematic, but it goes against Occam’s razor, which favours the most economical explanation.

1.4 Research question and Hypotheses

This dissertation wants to provide a further contribution to the understanding how frequency effects interact with grammar. The leading research question of this thesis is:

(17) What is the exact nature of the interaction between frequency and grammar?

a. What is the content of the input
b. How are the input and the output derived?

Related to the main question:

(18) a. Which insights do frequency-grammar interactions provide into the relation between grammar and lexicon?

b. What do frequency-grammar interactions contribute to the modelling of language?

The goal of this thesis is thus two-fold. The practical goal is to collect data that show how exactly frequency may interact with grammar. The theoretical goal is to model these data in a combined Exemplar-Optimality model.

We may hypothesize that there are different possible interactions between frequency effects and grammar. A possible frequency-grammar interaction would be that different behaviour of LF and HF words only occurs within a particular grammatical context or a particular rule. For instance, -t/-d deletion in English largely depends on the variety and on phonological context, whether it is sentence-final or whether the following word starts with a vowel or a consonant (Tagliamonte & Temple 2005). A Type II frequency effect (reduction) typically occurs in -t/-d deletion in nouns, adjectives, past tense verbs, and monomorphemic words, but not elsewhere (Bybee (2002), Coetzee & Pater (2008)).

(19) Hypothesis I

Frequency effects within a particular variation pattern occur in particular grammatical contexts and are blocked in other grammatical contexts.
Introduction & Methodology

Alternatively, I hypothesize that it is also possible that frequency effects are not fully blocked, but still are dissimilar for different phonological and morphological contexts. For instance, since -t/-d deletion is a reduction process, I would hypothesize that reduction of final -t/-d would be more likely in pre-consonantal context than in pre-vocalic context, since in pre-vocalic position, resyllabification of the final -t/-d might occur. So, in both contexts Type II frequency effects may occur, but I hypothesize that the patterns differ.

(20) **Hypothesis II**

Frequency effects within a particular variation pattern occur in all grammatical contexts, but they are sensitive to the grammatical difference between these contexts.

Still another possibility is that frequency effects largely occur across the lexicon, independent of grammar.

(21) **Hypothesis III**

Frequency effects are independent of the grammar.

Finally, we expect all variation patterns to be sensitive to either Type I frequency effects or Type II frequency effects, since Type II frequency effects are typically involved in reduction and Type I frequency effects are typically found in other processes of phonological and morphological change.

(22) **Hypothesis IV**

Depending on the type of process of language variation and change, either frequency Type I or frequency Type II apply.

These hypotheses will be investigated for different languages and different processes of variation and change.

1.5 **Structure of this thesis**

Frequency effects are often attested in language variation and change. Another source where frequency effects are to be expected, but which is an understudied field of research in frequency studies, is loanword phonology. It might be expected that frequency effects play a role in gradual adaptation of loanwords in the native grammar. In this dissertation we will test the hypotheses (19)-(22) in frequency-grammar interactions that are attested in three empirical studies: synchronic variation in a categorical rule (rendaku) in Japanese, synchronic gradient variation in Standard German long vowel <ä>, a diachronic study in the categorical rule of coalescence in Dutch loanword integration in Indonesian, and diachronic variation in Standard German long vowel <ä>. In all four studies, frequency effects which interact with grammar are attested.
Table 1.1. Schema of the case studies in this dissertation, divided by type of variation (synchronic vs. diachronic and categorical vs. gradient).

<table>
<thead>
<tr>
<th>Type</th>
<th>Categorical</th>
<th>Diachronic</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Synchronic</strong></td>
<td>Chapter 4</td>
<td>Chapter 6</td>
</tr>
<tr>
<td></td>
<td>Japanese rendaku</td>
<td>Coalescence in Dutch loanwords in Indonesian.</td>
</tr>
<tr>
<td><strong>Gradient</strong></td>
<td>Chapter 5</td>
<td>Chapter 7</td>
</tr>
<tr>
<td></td>
<td>Standard German long &lt;ä&gt; (Alemannic variety)</td>
<td>Standard German long &lt;ä&gt; lowering</td>
</tr>
</tbody>
</table>

The different hypotheses are tested by investigating frequency effects in relation to grammatical processes, such that the relevant grammatical subcategories are taken into account. For instance, Dutch loanwords in Indonesian undergo coalescence, which affects the stem-initial consonant. Five different consonants /p t k f s/ form potentially input for coalescence with a preceding nasal. Instead of only investigating the relation between coalescence and frequency, I will also investigate whether, and how, frequency effects occur within the subclasses of words that begin with one of the consonants /p t k f s/. In this way we can observe whether frequency effects are blocked in a certain context (Hypothesis I), whether they are sensitive to the grammatical context (Hypothesis II), or whether they are insensitive to the grammar (Hypothesis III).

A relatively large part of this dissertation is concerned with the investigation of Standard German long vowel <ä>. The remainder of this, introductory, part I therefore contains a chapter on the background information of Standard German long vowel <ä>, in which we will look at the historical development of the vowel, the sources of the current variation, and the role of spelling and dialects in this variation. I will show that a former merger of the long vowel <ä> with long <e> is currently reversed, which underlies the variation and the frequency effects that we will investigate in later chapters. Chapter 3 provides the motivation for the methodology used in this thesis. It turns out that categorical ratings by native speakers probably do not constitute the most reliable methodology for analysis, hence, auditory data used in this thesis are all acoustically analysed. These chapters form the introduction for chapters 5 and 7 on the long vowel <ä> in Standard German.

Part II consists of two chapters: case studies that reveal an unexpected exceptional behaviour of LF words. Chapter 4 investigates a well-known morphophonological process in Japanese, rendaku, which involves voicing of the first segment of the right-hand member of a
compound. There are, however, a number of compounds that are unlikely to undergo rendaku. We investigate whether frequency effects may play a role in this variation. Indeed, it turns out that especially LF roots sometimes do not undergo voicing. Whether they undergo the rule of rendaku or not largely depends on the frequency of both the voiced as the unvoiced variant of the root in other compounds and the root in isolation. Rendaku thus reveals a frequency-grammar interaction, in which we observe that compounds with LF roots words are less likely to undergo the rule. Chapter 5 is an in-depth experimental study about the pronunciation of the long vowel <ä> in a regional variety of Standard German. Here we observe a frequency-grammar interaction in the sense that the pronunciation of the vowel, among other factors, depends on the interaction between frequency and the morphological category. This study confirms the finding in chapter 4: LF words may behave exceptionally with regard to the grammar. Besides, we find that frequency effects are sensitive to the grammar, and that frequency effects depend on relative values rather than absolute values.

Part III forms the core of this thesis and provides two clear and comparable interactions of frequency and grammar. Chapter 6 describes an interaction between frequency and coalescence in Dutch loanwords in Indonesian. This chapter confirms the results of chapter 4 and 5 in the sense that LF words behave exceptionally with regard to the grammar and that frequency effects are sensitive to grammatical context. In chapter 6, we also find that expected frequency effects are blocked in certain well-defined phonological contexts. Further, the frequency effect interacts with the first segment of the word. Finally, chapter 7 reports on an interaction between frequency effects and the pre-r vowel lowering rule in the change of the long vowel <ä> in Standard German. It will be shown that in the process of vowel lowering three different frequency effects are involved. Finally, part IV discusses the results and concludes.
Märchen

Ich weiß ein schönes Märchen.
Es war ein schönes Pärchen,
Hieß Hänselchen und Klärchen,
Die pflückten Blum' und Ähren,
Und aßen reife Beerchen.

Das Klärchen hatt’ ein Härchen,
Das Hänselchen ein Scherchen;
Das war ein goldnes Härchen,
Und das ein silbern Scherchen.

Das Hänselchen nahm Klärchen,
Schnitt mit dem Silberscherchen:
Ihr das goldne Härchen;
Da ging das goldne Härchen
Entzwei am Silberscherchen;
Da ging das Silberscherchen
Entzwei am goldnen Härchen.
Da weinte laut das Klärchen
Um ihr verlornes Härchen,
Und Hänschen mit dem Klärchen
Um sein zerbrochnes Scherchen;
Laut weinte das Pärchen
Um Härchen und Scherchen;
Gar viele, viele Zährchen.
Laut weinten Blum' und Ährchen
Und alle reifen Beerchen,
Zusammen mit dem Pärchen
Um Härchen und Scherchen.
Da saß im Busch ein Stärchen,
Das sah die vielen Zährchen,
Da sprach das kluge Stärchen:
Was weint ihr denn, ihr Närchen?
Das Härchen und das Scherchen,
Die Zährchen und die Ährchen,
Die Beerchen, und du Pärchen,
Und ich dazu, das Stärchen,
Sind alles nur ein Märchen.

Friedrich Rückert, 1788-1866