

Chapter 1

Introduction

1.1 Introduction to Optimality Theory

1.1.1 Optimality Theory for my grandma

The history of *Optimality Theory* (OT) goes back to 1993 (Prince and Smolensky (2004), also referred to as Prince and Smolensky (1993) or Prince and Smolensky (2002)), and is the linguistic implementation of a very simple idea:¹

Imagine you drive into a major intersection. You have a number of possibilities of what to do, such as putting on the brakes, halting, turning left, right, etc. Let us call these possibilities *candidates*. You also have a number of factors determining your choice: traffic lights, signs, road marks, hand movements of a policeman, the position and the speed of your own car and of other cars, the presence of pedestrians. But also your own destination. These are *constraints* on the possibilities, since they *filter out* some of them. For instance, you are not going to turn left if it is prohibited by a traffic sign. Sometimes, constraints contradict each other: you have a green light, yet a policeman forces you to stop. The traffic code prescribes the *ranking* (the *hierarchy*) of the constraints: the sign given by a policeman overrules the traffic light, and the traffic light precedes traffic signs. Paradoxically enough, the ultimate goal of traffic—that is, reaching your own destination—is *ranked* the lowest: this constraint is applied only if more than one options (*candidates*) have survived the other filters. Otherwise, if you do not have any other option, you will turn left, even if you would like to reach a destination on your right.²

To rephrase, we have a given *set of constraints* (CON-1, CON-2,..., CON-*i*), which are ranked in a certain order. If CON-1 is the strongest, and CON-*i* is the weakest, we shall write:

$$\text{CON-1} \gg \text{CON-2} \gg \dots \gg \text{CON-}i \tag{1.1}$$

We also have a *set of candidates*: *A, B,...* Each of the constraints *evaluates*

¹For a short introduction to Optimality Theory, its background and its application for beginners, see among many others Gilbers and de Hoop (1998). The application of OT to traffic rules can be found for instance in Gilbers and Schreuder (2000) and Boersma (2004a), and I first heard it from Dicky Gilbers in 2001.

²This phenomenon is well-known to anybody who has ever driven in the city centre of Groningen.

each of the candidates. In the simplest example, a candidate either *satisfies* (the action is allowed) or *violates* the constraint (the given action is prohibited). The traditional way of representing such a situation is to use a *tableau*:

	CON-1	CON-2	...	CON- <i>i</i>
<i>A</i>	*!		...	
$\spadesuit B$...	*
<i>C</i>	*!	*	...	*
<i>D</i>		*!	...	

(1.2)

Here, a star (*) means that this candidate violates that constraint. As can be seen, candidates *A* and *C* violate the highest ranked constraint CON-1, so they are immediately out of the game. It does not help that candidate *A* satisfies all other constraints, unlike the competing candidates. The exclamation mark (!) shows where a candidate meets its Waterloo. Only two candidates survive the first constraint, *B* and *D*, the second of which is defeated at CON-2. In turn, candidate *B* wins—the hand symbol \spadesuit points to the winner—even though it violates lower ranked constraints. (If you have no other opportunity, you turn left, even if you do not want to.)

Observe that constraints are *violable*: the winner candidate does not have to satisfy *all* constraints, but has to satisfy them *better* than its competitors. If all candidates violate a certain constraint, all will survive. Hence the name *Optimality Theory*: we search for the *optimal* candidate, that is, the best candidate of the candidate set. Remember this last sentence, as it summarises my whole dissertation.

1.1.2 Optimality Theory as a scientific model

How can such a model serve scientific purposes? Most (empirical) scientific activities can be decomposed into the following three steps:

1. Collecting data
2. Systematising data (which includes some abstraction process)
3. Creating a model that describes the *systematised* data set

The data collection can be *ad hoc* (you catch all butterflies you can), or planned and controlled, motivated by some *a priori* theories or hypotheses. In the first case, systematisation already requires much intellectual work, as proven by the history of biological taxonomy or pre-Mendeleevian chemistry. Once a discipline has established a theory (a paradigm), data are collected in a systematic way in order to corroborate or falsify the given model.

The third step is the creation of a model that describes (“explains”) the data, that is, the *typology* obtained by the abstraction in the previous step. I claim that this third step is what makes science more than a knowledge base that can be found in whatever human activity (the knowledge required by a certain profession, stamp collection, knowing the currency of each country in the world,...). Namely, if a scholar or a community of scholars (working in a given Kuhnian paradigm) accepts a model describing the data set at hand as *convincing*, then they have the feeling that they have a *deeper understanding* of

the observed phenomenon. I wish I could explain what makes a model “convincing”, “explanatory” or “providing a deeper understanding” to a community of researchers.

In linguistics, field work and language description correspond to data collection, whether it be the descriptive linguistics of a well-known modern language, of a classical language, or of an “exotic” language. Describing a language involves describing its sound repertoire, its verbal system, its word order, its stress pattern, and so forth. We shall immediately use the example of word stress.

In the second step, *language typologies* can be set up. Suppose, for instance, that the languages of the world can be organised into the following three categories according to their (main) stress pattern:

- Stress on the *first* syllable:

According to Hayes (1995): e.g. Hungarian, Central Norwegian Lappish, Mansi (Finno-Ugric languages), Czech (Indo-European), Ono (New Guinea), Debu (Loyalty Islands), Diyari (South Australia). Gordon (2002) lists 57: e.g. Danish, Afrikaans, Latvian (Indo-European), Nenets (Uralic), Arawak, Arabela, Chitimacha.

- Stress on the *last* syllable:

According to Baković (1998): e.g. Uzbek, Yavapai. Gordon (2002) lists 59: e.g. Moghol (Altaic), Atayal (Austronesian), Guarani, Haitian Creole, Mazatec.

- Stress on the *penultimate* syllable:

According to Hayes (1995): e.g. Polish, Piro (in Peru), Cavineña (in Bolivia), Djingili (Australia), Warao (Venezuela). Gordon (2002) lists 53: e.g. Mohawk (Northern America), Albanian (Indo-European), Mussau (Austronesian), Shona (Bantu language in Zimbabwe), Jaqaru.

This is only a toy example for illustrative purposes, and a high number of languages—including English and Dutch—with more complex (e.g. syllable weight dependent) stress systems are ignored, similarly to secondary stress. Still, it seems to be true that there are no (in fact, only very few³) languages where the rule is to put the stress always on the second syllable. (The second syllable of a word in other language types may be stressed, though, if for instance the rule is to put the stress on the penultimate syllable and the word happens to have three syllables.) A high number of languages have been studied, so we hope that the lack of languages with a stress on always the second syllable is not only a random gap. Thus, if a model could describe this typology—that is, the existence of the existing types and the non-existence of the non-existing types—then we can claim that this model has “grasped” something from the essence of human language.

³Gordon (2002) cites only ten (including Basque, Tolai (New Guinea), Lakota or Koryak (Kamchatka)), as opposed to the more than fifty in each of the three listed types. In the present example, we shall ignore them, for we hope that a model predicting that a certain type does not exist may be the first step towards a more elaborate model that predicts that a certain type occurs significantly less frequently. The same applies to the seven languages mentioned by Gordon (2002) with a stress on the antepenultimate syllable (third from the end), such as Macedonian (Slavic), Cora (Uto-Aztecan, from the Americas) or some Austronesian languages.

A note for the non-linguist who is reading the introduction of my thesis. Linguistics has had several phases in its history. Up to the eighteenth century, it was most connected to literature, as it originally served as a tool or an aid for interpreting canonised literary and religious texts. Linguistics in the nineteenth century became a historical discipline: the history of and the “family relationship” between languages mirrored the history of and the “family relationship” between nations. After Saussure, in the first half of the twentieth century, language turned into a social construct: an arbitrary structure consented to by the society. Finally, the Chomskyan revolution resulted in seeing language as a biological (mental, cognitive) phenomenon.⁴

Consequently, if a model is able to describe some language typology—say, the observed stress patterns—then we hope nowadays that the model brings us closer to an understanding of how language works in the brain. (That is, for some, a better understanding of the human brain, in general.) Especially, if the same kind of models can be used for several independent phenomena: word stress can be described with the same repertoire of techniques as sound alternations, word order in a sentence or form-meaning matching. Additional arguments can also be made: a good model is able to reproduce not only observed language typologies, but also other language-related phenomena, such as those observed in language acquisition (child language), in language impairment and disorders (e.g., due to brain injury), and in language variation and change (dialects, sociolects, historical linguistics). For instance, I will argue for the cognitive relevance of my model (Chapter 5) by showing that it can also reproduce fast speech phenomena.

A practical aspect of language modelling is language technology. Can we use a certain model for building speaking computers? Recent products of language technology include spell checkers and grammaticality checkers, human-machine dialogue systems,⁵ reasonably working machine translation software, as well as automatic information extraction tools (question answering,⁶ text summarisation,...). I have to disappoint the reader: the model presented in the present dissertation does not aim at being readily usable in industrial applications. Many phenomena to be discussed can probably be implemented much more simply. There is no need for ten constraints to assign stress to the first syllable of each word.

Yet, one of the motivations is exactly applicability. Our starting point will be how a certain linguistic model can be implemented on computers, which is also interesting from a theoretical point of view. Although language technology nowadays can dismiss this linguistic model, the widely used linguistic model cannot dismiss its computational analysis (decidability, complexity, learnability,...). Additionally, we will be concerned with the psychological plausibility of that model, even if not with its contribution to language technology. It is like understanding the mechanics and dynamics involved in a human leg, while engineers still prefer realising a horizontal motion using wheels. Indeed, linguistics

⁴Observe that language typology (exemplified by word stress types) has nothing to do with language families. Genetically related languages frequently belong to different types, and unrelated languages may share many features.

⁵An example is when the user calls a phone service of the train company, and the computer answers questions concerning train schedules (Lendvai, 2004).

⁶See for instance the Imix project on *Question Answering for Dutch using Dependency Relations* of Gosse Bouma described at <http://www.let.rug.nl/~gosse/Imix/>.

has brought numerous arguments in favour of its models, and we shall argue for a specific implementation.

As the reader can guess, the model that will be used as a language model is Optimality Theory (OT). First, a non-linguistic example will demonstrate how OT may reproduce typologies (proving that the general idea is independent of linguistics), which is followed by a toy linguistic example.

A high number of chocolates can be found on the market, because different *types* of customers buy them.⁷ Chocolates not corresponding to any type of customers lack demand and are removed from the market. Different customers have different priorities: some go for quality, others for quantity, and others again for price. Suppose that the following four brands⁸ of chocolate are characterised by the following *tableau*:⁹

	QUALITY	QUANTITY	PRICE
<i>Mars</i>	excellent	55 g	0.50 EUR
<i>Túró Rudi</i>	excellent	30 g	0.30 EUR
<i>Côte d'Or</i>	good	200 g	1.40 EUR
<i>Milka</i>	medium	200 g	1.20 EUR

(1.3)

Here, the four brands of chocolate are the *candidates*, whereas the three characteristics act as the *constraints*. Unlike in the previous example on driving a car, constraints are not either *satisfied* or *violated*, but they assign different *evaluations* to each of the candidates. More levels are possible. Importantly, however, these evaluations can always be compared to each other: evaluations *a* and *b* are either the same, or *a* is better than *b*, or *b* is better than *a*. No fourth possibility exists, and we shall use this *Law of Trichotomy* in several occasions in the coming chapters.

Suppose that the *constraint hierarchy* of a customer is QUALITY \gg QUANTITY \gg PRICE. Similarly to (1.1) on page 1, the symbol \gg means again that the first constraint is more highly ranked (left in the tableau) than the second one. Consequently, our customer will first eliminate *Côte d'Or* and *Milka* from the set of candidates: they are not bad at all, but you can find better. In the next step, she will compare the quantity of the surviving two candidates, and, therefore, go for a Mars bar.

Other customers have different constraint rankings, driving them to different brands. Hierarchy QUALITY \gg PRICE \gg QUANTITY yields a *Túró Rudi*, similarly to the—quite different—hierarchy PRICE \gg QUANTITY \gg QUALITY. One can also simply check that QUANTITY \gg QUALITY \gg PRICE results in a *Côte d'Or*, whereas QUANTITY \gg PRICE \gg QUALITY in a *Milka* in our toy example. All four candidates are *winners* of some hierarchy, thereby they are preferred by some type of customers—as proven by the observable demand for them. The model also predicts the effect of changing the price of *Côte d'Or*: if its price is reduced to 1.10 EUR, those buying *Milka* would now purchase *Côte*

⁷As I was informed after having worked out this example for non linguists, Boersma (2000) uses a similar example (buying rucksacks and optimising for volume, weight and price), even if in a slightly different manner. The priority of using this example goes therefore to him. A further non-linguistic example will be brought in section 8.3 (Papert, 1980).

⁸*Túró Rudi* is one of the favourite brands of most Hungarians.

⁹In our toy example, we ignore the subjective factor, and suppose that QUALITY is as objective as the two other dimensions.

d'Or, but not those buying *Mars* or *Túró Rudi*. Altogether, *Optimality Theory* could account for customer typology and phenomena on the market.

Let us now turn back to our (oversimplified) linguistic example, stress typology (cf. Baković (1998), Gordon (2002)). The following constraints are simplifications of real constraints used by phonologists:¹⁰

- **EARLY**: number of syllables between the beginning of the word and the stressed syllable (i.e., the stress must occur as *early* as possible in the word).
- **LATE**: number of syllables between the stressed syllable and the end of the word (i.e., the stress must occur as *late* as possible in the word).
- **NON-FINAL**: 1, if the last syllable is stressed, otherwise 0 (the last syllable must not be stressed).

Which syllable of a, say, four-syllable word should be stressed? There are four options, which are the candidates, to be evaluated by the constraints just introduced. In the following tableau, the character *s* refers to a stressed syllable, and *u* to an unstressed syllable.

4-syllable word	EARLY	LATE	NON-FINAL
s.u.u.u	0 (excellent)	3 (worst)	0 (good)
u.s.u.u	1 (medium)	2 (bad)	0 (good)
u.u.s.u	2 (bad)	1 (medium)	0 (good)
u.u.u.s	3 (worst)	0 (excellent)	1 (bad)

(1.4)

One can simply verify that the three existing language typologies can be reproduced with different hierarchies. For instance:

- **EARLY** \gg **LATE** \gg **NON-FINAL** returns s.u.u.u (word initial stress)
- **LATE** \gg **EARLY** \gg **NON-FINAL** returns u.u.u.s (word final stress)
- **NON-FINAL** \gg **LATE** \gg **EARLY** returns u.u.s.u (penultimate stress)

Whereas no ranking yields u.s.u.u as the best candidate, which corresponds to its systematic absence in the observed typology.

Therefore, we have accounted for three positive observations (the existing types) and one negative observation (the lack of a type) using three constraints. If introducing a few more constraints increases the number of observations explained combinatorically, then the model has a strong reductionist power. On the other hand, the principle of *factorial typology* makes the strong prediction that the number of types cannot exceed the factorial of the number of constraints, while the fact that several hierarchies yield the same types further restricts the number of possible language types. For example, if five constraints account for why twenty or thirty types exist but not more, then many observations have been reduced to a few principles, on the one hand, and OT has also restricted the number of possibilities, on the other.

¹⁰For **EARLY** and **LATE**, cf. **EDGEMOST** of Prince and Smolensky (1993) and the alignment constraints of McCarthy and Prince (1993a). For **NON-FINAL**, cf. **NONFINALITY** in Prince and Smolensky (1993) and Hung (1994).

The explanatory power of OT is enhanced further if constraints are conceptually less complex than the resulting observations. The toy example presented might not be the most convincing example, even though I believe that not willing to stress the last syllable is simpler than what follows from it, namely, stressing the penultimate.

Summarising, we have seen in the present subsection how *Optimality Theory* can account for typologies (customer typology, language typology), and thereby become a scientific paradigm. It defines a set of candidates, all of which compete initially; as well as a set of constraints. The latter ones evaluate the candidates and act as filters: the best candidates survive, and the worse-than-best candidates are filtered out.

1.1.3 A slightly more formal definition of OT

Optimality Theory (OT), introduced in 1993 by Alan Prince and Paul Smolensky, has been an extremely popular model in linguistics in the last decade. In the present subsection, a more exact definition is presented.

It is useful to state at this point that the present thesis focuses first of all on phonology, although most of the proposals can be readily translated to other linguistic fields. This choice reflects the fact that Optimality Theory has been employed most often—yet not exclusively—by phonologists. Even though it claims to be applicable to any field, it has been most attractive to phonologists who wished to replace the SPE-style rules (Chomsky and Halle, 1968). Many linguists working on syntax or semantics are concerned with different types of problems (e.g. with representational issues), orthogonal to the answers offered by OT. In turn, similarly to most previous theoretical work on OT, I also have mainly phonology in mind. Concrete applications will also be taken from phonology. Although I claim that the ideas to be presented here are not exclusively related to phonology, the future will show whether they really can address a wider audience.

As in most models in generative linguistics, the goal is to map the *underlying representation* (UR) onto the *surface representation* (SR), the form observed in a particular language. Originally, the background idea is roughly that in language production the underlying representation is obtained by extra-linguistic processes: depending on the meaning of the “message” to be uttered, the UR is some list of elements taken from the mental lexicon.¹¹ At this point, no “real” linguistics has been involved, as the “message” is a function of social, contextual and cognitive factors, whereas the forms of the elements in the mental lexicon are arbitrary. Optimality Theory refers to this idea as the *Richness of the Base Principle* (Prince and Smolensky (2004) p. 220): “all inputs are possible in

¹¹The *mental lexicon* contains the list of morphemes in a given language, including their phonemic forms and all further information required. It has been supposed that the mental lexicon has to be minimised, redundancies have to be avoided, and the different forms of a morpheme have to be derived by the grammar, as much as possible. For instance, not all different forms of a word are stored, but only one form for each morpheme, which are then combined and submitted to phonological transformations. In fact, the human mind has always been pictured in function of the contemporary technology (as mechanical automata, steam engines, telegraph cables, etc.; cf. Daugman, 1990), and the idea of avoiding redundancies goes back to the early years of computers with a very restricted memory. Although nowadays, as a consequence of the development of computer memories, the mental lexicon is not required anymore to be minimal, the idea is still present and influences the way linguists build their models to explain regularities and analogies in language.

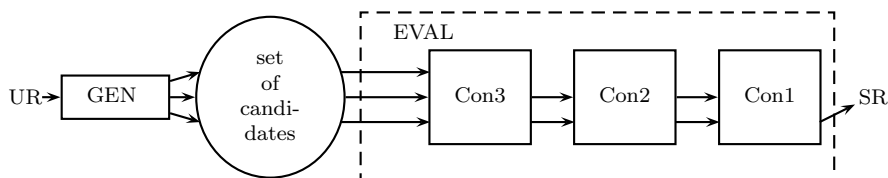


Figure 1.1: The basic architecture of an Optimality Theoretic grammar

all languages, distributional and inventory regularities follow from the way the universal input set is mapped onto an output set by the grammar”.

In brief, the differences among languages are accounted for by the mapping from the underlying representation onto the surface form. The task of a linguist is but to create a “convincing” model for this mapping.

How can this mapping be realised? Traditional generative grammars used rules. In phonology, the *Sound Pattern of English* (Chomsky and Halle, 1968) served long as the example with its (apparently) context-sensitive rewrite rules (but see also Johnson, 1972, on regular implementations of SPE rules). *Two-level morphology* (Koskeniemi, 1983) introduced a second type of architecture, and *Optimality Theory* proposed a third alternative.

The standard architecture of an OT grammar is shown in Figure 1.1. It is composed of two parts, two modules. Out of the input (the underlying representation UR), the GEN module *generates* a set of candidates ($GEN(UR)$). The elements of the latter are *evaluated* by the EVAL module, and the best element is returned as the output (the surface representation SR).

There are two ways of looking at EVAL. It is usually seen as a pipeline, in which the *constraints* filter out the sub-harmonic candidates. Each constraint assigns violation marks to the candidates in its input, and candidates that have more marks than some other ones are out of the game. This is the algorithm we have already used in previous examples to calculate which of the possibilities (candidates) is the best: you lose if another competitor is better than you.

Violation marks are the stars we used in tableau (1.2). There, a candidate either satisfied (no star) or violated (one star) the constraint. One can also imagine a constraint assigning more than one violation marks to a candidate. Indeed, in (1.3) and (1.4), we could replace *excellent* with zero star, *good* with one star, *medium* with two stars, *bad* with three stars, and *worst* with four stars. Many constraints used in linguistic models require that a substring of the candidate meet a certain criterion: each part of the input that fails to meet the criterion incurs an additional violation mark to the candidate.¹²

Alternatively, EVAL can also be seen as a function assigning to the candidates some (strange) *harmony value* derived from their behaviour on the

¹²Some call such constraints *gradient* ones, for they allow for more than two levels in the goodness of a candidate (e.g. Jäger, 2002). Other authors speak of *gradient constraints* only if any substring can violate a certain constraint on several levels, and the violation level of the candidate is the sum of these gradient local violations. (e.g. McCarthy, 2002; Bíró, 2003). In fact, the most interesting type of gradient constraints are those that can assign an unbounded number of violation marks to any locus in the string. For instance, the widespread ALIGN(Feat, Word, Left) assigns each metrical foot in the word as many stars as the number of syllables intervening between the left edge of the word and the left edge of the foot (e.g. Tesar and Smolensky (2000) p. 54-55 calls it ALL-FEET-LEFT; for its criticism and an alternative proposal, see McCarthy (2002)).

constraints. Additionally, EVAL also includes an optimisation algorithm that compares the harmony of the candidates, and finds the best one, for the most harmonic candidate is predicted to surface in the language. This *Harmony function* has, however, a remarkable property: being worse on a higher ranked constraint can never be compensated by good behaviour on a lower ranked constraint. That fact follows from the filtering approach: whoever is filtered out at an earlier stage, never comes back. This phenomenon is referred to as *strict dominance hypothesis*.¹³ We shall come back to this approach in Chapter 3, where we show that the Harmony function cannot be realised with a real valued function, and propose alternative approaches.

In fact, the success of Optimality Theory since 1993 is partly due to the idea of *strict dominance hypothesis*. Not only does it make the model more restricted, but also seems to be easier or more appealing to work with. Namely, in the pre-cursor of OT, *Harmony grammar* (Smolensky, 1986; Legendre et al., 1990a,b,c), severe violations of lower constraints could accumulate and become worse than the violation of a higher ranked constraint.

It should be noted that such cumulativity effects have recently come back to the foreground of research, and we shall return to them in subsection 1.3.5 (Jäger and Rosenbach, 2006). Further research has to decide how wide-spread cumulativity effects are, and whether ignoring them or incorporating them into the set of linguistic observations is the more fruitful for the development of science. Indeed, scientific progress requires neglecting some phenomena in order to be able to describe others. “To be able” in the phrase depends on the preferences of the scientists involved. Therefore, most adherents of OT feel fully legitimate in “postulating” that linguistic phenomena do not exhibit cumulativity effects, while others will reject that, and prefer Harmony grammar. Nonetheless, most general linguists use Optimality Theory, and form a well-organised community with a growing literature around ROA.¹⁴ My thesis aims at addressing this audience, and the model proposed here should be further developed based on their knowledge of particular linguistic phenomena.

To sum up, the following list of concepts play a central role in *Optimality Theory*:

- Underlying representation = input
- GEN
- Set of candidates
- EVAL
- Constraints, acting as filters
- Hierarchy (= ranking, ordering) of the constraints, which is categorical

¹³A related and widely used notion is *categorical ranking*. But, as Paul Boersma has pointed out, this latter notion refers to the non-variation of the constraint ranking. We shall soon see models (e.g. models by Anttila and by Boersma) in which constraints strictly dominate each other (lower constraints cannot help a candidate survive a higher ranked constraint), and yet, the ranking of the constraints may vary within a grammar.

¹⁴The *Rutgers Optimality Archive* at <http://roa.rutgers.edu> and the *Optimality List* are eminent examples for how a scientific paradigm of the 1990s should use the technology of the 1990s in order to become popular.

- Surface representation = output

According to the general philosophy of Optimality Theory, not only the set of possible inputs is universal (cf. the *Richness of the Base* principle mentioned earlier), but so is GEN and the set of the constraints present in a language. Constraints, in fact, should reflect universal tendencies in the world's languages, and vice versa, language universals correspond to some constraints. The basic claim of Optimality Theory is that the same determining factors are active in all languages, and only their relative influence differs. This is why, in our chocolate example, we used constraints that maximised quantity and quality and minimised price: we could have added a fourth constraint that minimises quality and rank it low, but this constraint would correspond to no observable phenomenon.

Many allow for some language specific parametrisation of the constraints. Furthermore, in practice, the set of candidates varies across articles. The goal set by current research is to determine the best set of constraints, and linguists propose different constraints, or reformulate previous ones, in order to account for more phenomena. The trick, as we shall see it soon, is the following rhetoric: a given model deals only with the highest ranked constraints, whereas all other constraints argued for by others may be ranked low so that they do not interfere with the choice of the best candidate. (See also section 1.3.)

The only language specific parameter, therefore, is the ranking of the constraints. The acquisition of a language, hence, means *learning the adequate hierarchy*, and a *grammar learning algorithm* is expected to return a hierarchy that produces the correct outputs for the given underlying forms.

Going back to the example of buying chocolate, we could illustrate the idea of *grammar learning* in the following way. Imagine you have a new girl friend, and you would like to know her better. You know what guidelines people consider universally (quality, quantity, price), yet you would like to know how she applies them. So, you take her to a shop (without telling her that you would pay). You propose her several sets of alternatives, and you observe which she chooses in different situations. Then, you can derive the hierarchy driving her choices. *Learnability* is a separate research line within the computational analysis of Optimality Theory (see e.g. Tesar and Smolensky, 2000; Boersma and Hayes, 2001; Pulleyblank and Turkel, 2000; Tesar and Prince, 2003; Ota, 2004; Goldwater and Johnson, 2003; Prince and Tesar, 2004; Pater, 2005b), which we shall touch upon here and there, especially in section 4.2.



1.2 Infinite candidate sets, implementing OT

In many Optimality Theoretical models advanced by theoretical linguists, the set of candidates is infinite. The reason for this is at least two-fold. First, most linguists working within the OT framework simply see GEN as a black box, producing literally *everything*. (Or almost everything, but most linguists are

not very explicit about it. I would like to urge linguists to be more exact about GEN.) And “everything” is infinite.

Second, in many linguistic phenomena, some structure—such as an epenthetical vowel, a default syllable onset or an expletive word—is inserted, and often more than one insertion is required. Therefore, the simplest way to proceed is to allow any (finite) number of insertions, *that is*, to allow recursive insertions, yielding an infinite set of possibilities. It is true that many of them have no chance to win under any constraint ranking: they are called *losers* in the OT jargon (e.g. Samek-Lodovici and Prince (1999), p. 3). Yet, it is simpler to include them into the model than to restrict GEN to the set of candidates that may win under some ranking. We allow, thus, an infinite set of candidates in order to save the simplicity, the homogeneity or the mathematical beauty of the model.

On the other hand, the infinity of the candidate set raises numerous questions. First of all, including losers into the model undermines the “philosophy” of Optimality Theory previously discussed. We have introduced OT as a model for language typology: the set of candidates includes the forms present in language typology, and each of the possible constraint rankings corresponds to a certain language type. Why should we include, then, forms that are *not* observable in language typology? Language typologies allow usually only for a very restricted set of possibilities, so what’s the business of all other (infinite number of) forms here? An interpretation of the model might claim that all forms generated by GEN are conceivable in some sense (for instance, as representations in the human brain), and yet, further restrictions (*i.e.*, the OT constraints) on human language exclude many of them from the set of possible surface forms. Indeed, for the proposal in section 6.5 the loser candidates are crucial: even though they do not surface in the language as grammatical forms, the model for the computing algorithm in the human mind makes use of them. They are like *Godot* in Samuel Beckett’s tragicomedy: an important character (like any other character), even if never appearing on the scene.

Additionally, the infinity of a character set poses a computational challenge to researchers who do not perceive theoretical linguistics as a discipline *per se*, rather in connection with language engineering, or with behavioural, cognitive and neurosciences. Could natural language technology make use of a model that first requires the generation of an infinite set? Does our brain really work with such huge data structures?

Different approaches have been proposed to spare the trouble of generating the whole candidate set. This work is important additionally because the computation in the case of a finite, though enormous set can also be not feasible, if the algorithm used is to compare each candidate with any other of them. Indeed, Optimality Theory as a framework allows for *intractable* problems (NP-hard—worst case exponential—in the size of the grammar, cf. Eisner (2000b)¹⁵). On the other hand, a clever algorithm can render the search in an infinite set ex-

¹⁵See Idsardi (2006a) for a simple proof adopting arguments from Eisner (2000b) that Optimality Theory as a framework is NP-hard. Kornai (2006a) criticised Idsardi (2006a) by arguing that the constraints employed by the latter are unattested in the phonologies of natural languages, to which Idsardi (2006b) answered in ROA. In response, Kornai (2006b) maintained his optim(al)ism by pointing to the fact that natural languages have very restricted phoneme inventories and large number of unbounded processes do not operate in parallel, and therefore real language OT does not blow computationally.

tremely simple for some problems: when interested in the smallest integer higher than n , you will use elementary school arithmetics, and not generate the whole infinite search space.¹⁶

Consequently, a major question is to work out computationally tractable implementations (algorithms) for Optimality Theory. The present dissertation discusses a novel approach, namely, simulated annealing. An alternative approach that I was also working with during my PhD scholarship is finite state technology (Bíró, 2003, 2005c), a research built especially on results by Frank and Satta (1998), Karttunen (1998), Gerdemann and van Noord (2000) and Jäger (2002).

Further approaches to handle a (possibly infinite) candidate set also exist. Chart parsing (dynamic programming) is probably the best known among them (chapter 8 in Tesar and Smolensky (2000) for syllabification, Kuhn (2000) for implementing OT LFG). It presupposes on the one hand that applying a recursive rule (usually insertion) incurs some constraint violation; and on the other, that “all constraints are structural descriptions denoting bounded structures”. The interplay of these two assumptions guarantees that the algorithm may stop applying the recursive rule after a finite number of steps, for no hope is left to find better candidates by more insertions.

The basics of another interesting implementation are presented by Turkel (1994). He uses *genetic algorithms* (e.g. Reeves (1995), Eiben and Smith (2003)), for both generation and learning, and claims that “*an OT system properly construed is a genetic algorithm.*”

Genetic algorithms are heuristic optimisation algorithms inspired by the idea of biological evolution. (For the concept of heuristic optimisation in general, see section 2.1.1.) In each step, we have a *population* of “chromosomes” (the algorithm starts with a random initial population), which are *evaluated* according to some fitness function, and which then participate in producing a new population (the next generation). Chromosomes with higher fitness are more likely to be chosen to participate in the generation of the new population (cf. *natural selection*). A few operations (such as crossover, mutation, etc.) are applied to the chosen chromosomes when they *generate* the next cohort. The idea is that the chromosomes with the highest fitness will be most likely to be selected, and thus the fitness in the pool of chromosomes will converge towards the optimum that is searched for.

When Turkel (1994) uses genetic algorithms for production in OT, it is GEN that realises the generation of the new population, and EVAL plays the role of the fitness function. A population of candidates enters GEN, which creates a new generation by applying basic operations (“mutation”, “crossover”) on the candidates entering it. Subsequently, EVAL selects the best ones from this new generation, which enters GEN again, and so forth. The idea that what GEN does is to map a candidate (here, in fact, a set of candidates) onto a set of “neighbouring” candidates by applying minimal modifications shall soon re-emerge in subsection 2.2.2, and has its parallels in the output-centric picture of Burzio (2002) discussed in section 4.1. In all these cases, the set of candidates can be walked across stochastically by applying these minimal modifications repeatedly.

The same genetic algorithm is then used to model language acquisition, that

¹⁶I am thankful for this example to an anonymous reviewer of a conference paper.

is, to learn the constraint hierarchy best fitting the observed language data (on learning cf. the end of section 1.1.3). A more matured version of this grammar learning algorithm, applied to vowel harmony, can be found in Pulleyblank and Turkel (2000).

1.3 Variation within OT

The primary aim of Optimality Theory is, thus, to account for language typology. The candidate set contains all the different types of the typology—and possibly further candidates. A given language, belonging to a given type, is described by the hierarchy of constraints that yields the grammatical candidate in that language as the only output (optimal form with respect to the hierarchy). Consequently, the standard philosophy behind Optimality Theory should allow each ranking to return only *one* candidate: the best one.

Nonetheless, Optimality Theory is, at the same time, a grammar that realises a mapping from the underlying representation to the surface form. Variation, a wide-spread phenomenon in languages that Optimality Theory certainly has to account for, may be seen as more surface forms corresponding to one underlying representation.¹⁷ But can an Optimality Theoretical model produce more than one output? In the present section, we shall present several approaches to this issue. Yet, before entering the discussion, we have to clarify what we expect from a model accounting for variation.

First of all, the term “variation” can be used in a number of senses. In sociolinguistic or dialectal variation, the distribution of the forms is defined by non-linguistic factors, and each speaker uses only one variant. In register dependent variation, a speaker can utter more than one form, yet the variation can be seen as if the same speaker switched between different languages. In free variation, no factor seems to play a role.

As sociolects, dialects and registers may be seen as different languages, an approach could be to assign them simply different grammars. And yet, these language varieties are clearly interconnected: they are genetically close, they are perceived as variations of the same language, and they influence each other. Thus, one would prefer a single grammar with some parameters that render switching between the varieties possible. Or, what is equivalent, a model that interconnects the elementary grammars into a larger meta-grammar. Note that here being able to control the variation is a very important requirement to a good model: we would like to put a hand also on the relation between the varieties. Free variation is a slightly different situation in that respect, supposing that really no factor is observable that would influence the variation.

Fast speech forms, a phenomenon we shall come back to later, is not exactly free variation, because speech rate is clearly a major influencing factor. It might be seen, then, as a special register. Nonetheless, I argue to perceive it rather as a dysfunction of the normal language production, due to the increased speed.

¹⁷Some variations can be analysed as the result of more underlying forms being present in the lexicon, but this approach would not work for productive phenomena, such as word order scrambling. Furthermore, conditioned variation may be accounted for by including further—for instance, pragmatic—constraints into the grammar. A Chomskyan linguist, however, may still wish to separate hard-core linguistics from pragmatics: she would prefer to allow more outputs from the core-grammar that can then enter pragmatics, OT-like filters in an additional module.

As opposed to, say, the hyper-correct or the official register of a language, the native speaker would not be able to decide whether a certain form belongs to some “fast speed register”. Additionally, the speaker and the hearer are not conscious of just having uttered or heard a fast speech form, unlike in the case of forms typical to some register. Lastly, fast speech is very often characterised not by a set of different forms, but by a gradual shift in the frequency of forms. Both the “correct” and the fast speech form is present in both normal and fast speech, but their frequencies differ. Consequently, predicting the *frequency* of the alternative forms becomes very important for fast speech models, even if a great many linguistic models content themselves with predicting which form is grammatical and which is ungrammatical.

In sum, we are in need of linguistic models—hence, of models within the Optimality Theoretic paradigm, as well—that can predict, or even control and fine-tune, the frequencies of alternative outputs corresponding to the same input. One may or may not like to apply them within sociolinguistics or for free variation. At least for fast speech, however, one cannot dispense with them.

Now, we have to turn back to our original question: can an Optimality Theoretical grammar return *more than one* output? Or, do we need to enrich the model, especially if we would like to account also for frequencies?

First, observe that if two candidates have different violation profiles, that is, if they behave differently with respect to at least one constraint, then one of them is more optimal than the other for a given hierarchy. We shall refer to this property of an OT-system as the *Law of Trichotomy*.¹⁸ Therefore, exactly one violation profile may be optimal with respect to a given ranking.¹⁹ The architecture of an OT-grammar suggests, thus, three different ways of returning more than one candidate within one language:

1. Two candidates are assigned exactly the same violation profile.
2. Not only the optimal candidate may emerge as a surface form.
3. A language includes more than one hierarchy (more than one grammar is present simultaneously).

¹⁸A formal proof of the *Law of Trichotomy* is provided in section 3.1. Informally, the proof is built on two standard assumptions in Optimality Theory. The first assumption is that the levels of violation (in practice, the number of violation marks assigned) are *fully ranked*: for a given constraint and a pair of candidates, candidate w_1 behaves either better or the same or worse than candidate w_2 (no fourth possibility is available). The second assumption, trivially true for a finite set of constraints, is that the constraints themselves are fully ranked, and each subset of constraints has exactly one *upper bound*, which is, furthermore, a member of that set. (Yet, see Tesar and Smolensky (2000) and Anttila and Cho (1998) for different proposals involving unranked constraints.)

The *Law of Trichotomy* states that for two violation profiles w_1 and w_2 , exactly one of the following three statements is true in a given hierarchy: 1.) w_1 is better than w_2 ; 2.) w_1 is worse than w_2 ; 3.) w_1 is the same as w_2 .

In order to prove it, take the set T of constraints for which the two profiles differ. This set is either empty (in case 3), or has exactly one upper bound. This upper bound is the highest ranked constraint for which the two profiles differ. Then, due to the first assumption (the violation levels are fully ranked), either w_1 or w_2 has to behave better with respect to this constraint, leading either to case 1 or case 2, respectively.

¹⁹At this point we see that one violation profile at most can be optimal, which is what we need in the present train of thought. In section 3.1, however, we demonstrate that one optimal violation profile always exists under the usual presuppositions.

In the following subsections, we shall discuss each of these cases separately. As we progress, the probability assigned to the candidates will become more important. We conclude this section by presenting a model in which the difference between candidates is no longer determined by whether a candidate surfaces or not, but exclusively by the probability of a candidate to appear in the language.

1.3.1 Forms assigned the same violation marks

First, can we describe alternations by assigning exactly the same violation profile to the alternating forms? In theory, it is possible, and yet, Anttila (1997a) calls it *the poor man's way of dealing with variation*. Notice that the two forms will be predicted to be totally free alternations, independently of further factors. We have absolutely no control over the variation. Furthermore, this approach does not allow one to predict frequencies, either, unless GEN is enriched so that it assigns frequencies to the candidates.

The second problem is the following: how to guarantee in an analysis that the two candidates are assigned exactly the same violation marks, while the number of constraints grows steeply with the number of papers published in OT? Such an analysis would rely heavily on restricting the number of constraints used, which is extremely dangerous. The usual way of sweeping the other constraints under the carpet is, namely, demoting them radically. The linguist presents an analysis of the phenomenon at hand based on a small number of proposed constraints, which filter out all but one candidate. Then, she adds—to save the idea of a universal set of constraints—that all other constraints argued for by her colleagues in other languages are indeed present in the given language; however, they are ranked low, hence not interfering with the presented analysis.

In the present case, this trick would not work. Even if we suppose that a constraint forgotten by the author of the analysis is very-very low ranked in the given language, it is active.

Take the following example. Standard Hungarian exhibits a variation [ɛ] ~ [ø] (e ~ ö) in many words, originating in the standardisation of two different dialectal forms, with a minimal preference for the [ɛ] forms in written language:

$$\begin{array}{llll}
 fel & \sim & föl & \text{'up'} \\
 felett & \sim & fölött & \text{'above'} \\
 seprű & \sim & söprű & \text{'broom'} \\
 tejfel & \sim & tejföl & \text{'sour cream'}
 \end{array} \tag{1.5}$$

An analysis could suppose underlying forms including an underspecified [round] feature, with GEN assigning some value to it. The two constraints proposed need no argumentation. FULLYSPECIFIED punishes underspecified features in the candidates, whereas HARMONY[ROUND] requires the [round] feature of a vowel match that of the previous vowel (Hungarian does exhibit roundedness-harmony for front vowels). Consequently, the tableau for *felett* ~ *fölött* looks as follows:

$/f[0round]l[0round]tt/$	FULLYSPECIFIED	HARMONY[ROUND]
$f[+round]l[+round]tt$		
$f[+round]l[-round]tt$		*
$f[+round]l[0round]tt$	*	
$f[-round]l[+round]tt$		*
$f[-round]l[-round]tt$		
$f[-round]l[0round]tt$	*	
$f[0round]l[+round]tt$	*	*
$f[0round]l[-round]tt$	*	*
$f[0round]l[0round]tt$	**	

(1.6)

The candidates $f[+round]l[+round]tt$ (i.e., *föLött*) and $f[-round]l[-round]tt$ (*felett*) seem to incur exactly the same violation marks, and are therefore equally optimal. Nonetheless, a constraint of the type $[\alpha back][\alpha round]$, preferring unrounded front (and rounded back) vowels to rounded front and (unrounded back) vowels, would differentiate between the two. This constraint, although not very prominent in Hungarian, which includes vowels $[\ø]$ and $[y]$ (*ö* and *ü*), is indeed part of the universal set of constraints, since it accounts for a linguistic universal. And, no matter how low a constraint is ranked, it will cause the less optimal candidate meet its Waterloo. This observation is called the *Emergence of the Unmarked* (McCarthy and Prince, 1994).

To sum up, assigning variation forms the same violation profile is not a safe, hence not a promising direction, for unseen constraints may spoil our model. Nevertheless, one may suppose a barrier beyond which constraints are not active anymore in filtering out candidates. Demoting constraints not required by one's analysis below this barrier may save such an approach. A problem arises only if that constraint still plays a role in a different phenomenon in the same language. Additionally, introducing such a barrier involves revising standard Optimality Theory more than what the fairly orthodox approach presented in the present subsection would permit. In fact, Coetzee's proposal, described below, can be seen as adding such a barrier to the standard OT architecture.

1.3.2 Non-optimal candidates emerging

Coetzee (2004) develops further the idea of the *Harmonic Ordering of Forms* introduced by Prince and Smolensky (1993). He proposes a rank-ordering model in which EVAL imposes a harmonic ranking on the *complete* candidate set. Standard OT is concerned exclusively with EVAL finding the optimal candidate with respect to this order, which will then surface as the output—the relative goodnesses of the other candidates are not of interest. In Coetzee's model, on the other hand, the losing candidates are also ordered with respect to each other, and most importantly, this order has linguistic significance. In his view, the second best candidate will be the second most frequently appearing variant of a certain form, the third best candidate may be predicted to be the third most frequent form, and so forth. Coetzee claims that the candidates that are still in competition after the so-called *critical cut-off point* can be variants of the optimal candidate:²⁰

²⁰Already the constraint M-PARSE of Prince and Smolensky (1993) acts as a cut-off point,

I propose that there is a critical cut-off on the constraint hierarchy that divides the constraint set into those constraints that a language is willing to violate and those that a language is not willing to violate. A candidate disfavoured by a constraint ranked higher than the cut-off will not be accessed as output if there is a candidate (or candidates) available that is not disfavoured by any constraint ranked higher than the cut-off. (p. 18.)

In fact, Coetzee’s proposal can be seen as a solution to the problem with the first approach, namely, assigning the same violation marks to alternative forms. As the $[\varepsilon] \sim [\emptyset]$ alternation in Hungarian has exemplified, its weakness was that a very low-ranked constraint still can dismiss one of the two forms. Now suppose that the constraints that are elements of the universal CON but “not really active” in the given language are demoted below Coetzee’s *critical cut-off point*: we may say that the alternating forms are assigned exactly the same violation profiles—as far as the constraints “really active” in that language are concerned. The lower constraints do not filter out candidates, but impose some preferences mirroring universal tendencies. In the Hungarian example, the slight preference for the $[\varepsilon]$ forms (at least in the written language) may be explained by the effect of the demoted constraint disavouring $[\emptyset]$. In fact, the $[\varepsilon]$ -forms are then predicted to be the *grammatical* ones, winning the competition. Yet, as the $[\emptyset]$ -forms are defeated only after the cut-off point, the latter ones emerge as free variants.

Nonetheless, Coetzee’s solution does not guarantee that one avoids the above mentioned problems. A constraint cannot be exiled beyond the critical cut-off point without consequences. As a given language is supposed to have only one cut-off point, some constraints in an analysis of an independent phenomenon of the same language may be required to overrank the critical cut-off point, and thereby spoil your proposal.

Even though he argues for the opposite, it is a further drawback of Coetzee’s model that it attempts only to give qualitative (“relative”, in Coetzee’s terminology), and no quantitative (“absolute”) predictions about the frequencies of the alternating forms (e.g. on pages 128-131, p. 226, and especially on p. 306). We have seen, however, that some phenomena (fast speech, in particular) are characterised only by a shift in the observed frequencies, so Coetzee could not account for them. A further criticism may be that ranking the whole candidate set—or at least compute its best subset—requires more computational power than finding the optimal element alone, often not a trivial task in itself.

Consequences of his model include that whenever the third best candidate is observed as an alternative form, then the second best one must also appear in the language. Furthermore, if the fourth best candidate is defeated by the same constraint as the third one, then the fourth one should also be an attested alternation form, else we cannot identify the critical cut-off point.

The model to be proposed in Chapter 2, *Simulated Annealing Optimality Theory* (SA-OT), although very different from it, resembles Coetzee (2004)’s approach—as opposed to all other proposals introduced and to be introduced in the present section—in that *SA-OT* also sees alternating forms as non-optimal

even if from the opposite perspective. This is the point where the Null Parse is eliminated by other candidates. Therefore, if all other candidates have fallen out previously, no surface form in the language corresponds to the input.

candidates still emerging. Variation forms will be modelled as *local optima* with respect to some *neighbourhood structure* on the set of candidates. Simulated annealing, the optimisation algorithm used, is prone to get stuck in such local optima, especially if optimisation is performed quickly, and this is why forms that are not globally optimal may be still returned in this approach. The art of SA-OT is to find a neighbourhood structure that is convincing (not *ad hoc*) on the one hand, and which turns the observed alternative forms—and, hopefully, only them—into local optima, on the other. Then, running the simulation and varying its parameters may or may not reproduce the observed data by returning the local optima with the expected frequencies.

In contrast to Coetzee (2004), Simulated Annealing OT aims at producing quantitative, (“absolute”) predictions about the frequencies of the forms. In this respect, SA-OT bears similarity to the models to be dealt with presently, which are based on the third possibility to have a grammar return more outputs.

1.3.3 Several hierarchies within one: reranking

The third way of dealing with alternative forms is to include more than one hierarchy into a language. This way might also be seen as an OT-style synthesis of the single route and the dual route approaches in the *Past Tense Debate* (see section 4.1): one grammar composed of more grammars.

More specifically, we may want to allow some rerankings, for instance by permuting neighbouring constraints. As it would be quite odd to stipulate two, very different hierarchies within one language, reranking neighbouring constraints helps minimising the “distance” of the hierarchies simultaneously present. Tesar and Smolensky (2000, p. 96) introduces the *h-distance* between some specific hierarchies, and along their line, we could define the distance of two hierarchies \mathcal{H}_1 and \mathcal{H}_2 as the minimal number of local permutations required to get from \mathcal{H}_1 to \mathcal{H}_2 . The simplest case, then, is if \mathcal{H}_1 differs from \mathcal{H}_2 in a single reranking of neighbouring constraints, whereas all other constraints are ranked in the same way, relative to each other and relative to these two constraints.

Ad hoc rerankings have been supposed in many phonological papers. For instance, in the example used in Chapter 5, Schreuder and Gilbers (2004) propose to account for fast speech phenomena in Dutch stress assignment by demoting a faithfulness constraint and promoting markedness constraints (Schreuder and Gilbers, 2004). Such an analysis has, however, some weaknesses: do you really claim that native speakers suddenly switch to a different grammar above a certain speech rate? If so, we predict form 1 being produced exclusively in slow speech, and only form 2 emerging above a critical speech rate—which contradicts observed data. In fact, the frequencies of the two forms change gradiently as a function of the speech rate. The fast speech form may also occur in relatively careful speech, whereas the first form is definitely present even at very high speech rates.

Three alternatives, three enlargements of the standard OT model have been proposed in order to allow reranking within one grammar in a systematic, more elegant way.

Anttila (1997b) and Anttila and Cho (1998) offer relaxing the strictness of a *fully* ranked hierarchy. See Anttila and Fong (2000) for an application in syntax-semantics, and Anttila (2002) for Finnish morphology.

So far, the set of constraints was *fully ranked*: for any two different constraints C_i and C_j , either $C_i \gg C_j$ or $C_j \gg C_i$. In a *partially ordered* set, on the contrary, two constraints may be not ranked relative to each other.

Formally speaking, a set S is a *partially ordered set* with some relation \prec if relation \prec is a subset of $S \times S$ such that the following properties are true:²¹

1. *Irreflexivity*: for all $a \in S$, $a \prec a$ does not hold.
2. *Asymmetry*: for all $a, b \in S$, if $a \prec b$ then $b \prec a$ does not hold.
3. *Transitivity*: for all $a, b, c \in S$, if $a \prec b$ and $b \prec c$ then $a \prec c$.

In a *totally ordered set* a fourth property also holds (rendering the first two axioms superfluous):

- 4 *Comparability* (aka the *Law of Trichotomy*): for all $a, b \in S$, exactly one of the following three statements holds: 1. either $a \prec b$; 2. or $b \prec a$; 3. or $a = b$.

A partial order \prec can be enlarged into another order \prec' on the same set S (its *refinement*, following Tesar and Smolensky (2000)'s terminology), such that for all $a, b \in S$ if $a \prec b$ then $a \prec' b$ (but not necessarily vice versa). In other words, relation \prec is a subset of \prec' within $S \times S$. Adding arbitrary (a, b) pairs in order to refine a partial order is not possible, nevertheless: the refinement also has to satisfy the above axioms.

Standard Optimality Theory requires the set of constraints to be totally ordered by the relation \gg . On the contrary, the grammar model proposed by Anttila and Cho (1998) involves only a partially ordered constraint set, and a surface form is predicted by such a grammar if and only if it wins for some fully ranked refinements of the partial order. Furthermore, at *evaluation time* (using the term of Boersma and Hayes, 2001), each of the refinements is chosen with equal probability, and then employed as in standard OT. This approach predicts the probability of a candidate to be the ratio of the number of refinements outputting this particular form to the total number of refinements.

For instance, suppose that the following three constraints are not ranked with respect to each other, and that they assign the following violation marks to candidates *cand1* and *cand2*:

	A	B	C
cand1		*	*
cand2	*		

(1.7)

These three constraints can be ordered in six different ways. Two of the rankings ($A \gg B, C$, which is an abbreviation for $A \gg B \gg C$ and $A \gg C \gg B$) yield *cand1* as the winner, whereas four of them return *cand2*. Consequently, Anttila and Cho's model will predict a frequency distribution of 33% *vs.* 67%. Additionally, Anttila and Cho (1998) propose to account for diachronic change

²¹The expression $a \prec b$ is an abbreviation of $(a, b) \in \prec$. The traditional way of defining an ordered set is to use the relation \leq that is reflexive, antisymmetric and transitive. All the same, the present formulation fits better with the use of the relation \gg in Optimality Theory, and follows the presentation of Anttila and Cho (1998).

and dialectal-sociolinguistic variations by enlarging and refining the partial ordered set of constraints.

As Boersma and Hayes (2001) correctly remark, however, certain frequencies can “be obtained only under very special circumstances.” For instance, a 99 to 1 ratio of two forms can be accounted for either by a single stratum in which 99 constraints prefer the first outcome and 1 favours the other; or, by a stratum of five constraints conspiring in such a way that only one of the 120 permutations yields the rare form. Furthermore, on a sociolinguistic level (that is, when the statistical model is used to reproduce the language production of a whole population, and not of an individual), such a model is unable to predict the gradual shift in frequencies observable either diachronically (e.g. cf. Hoeksema (1998)) or across language variation—dimensions that Anttila and Cho (1998) definitely aim at describing. Further factors can also cause a gradual frequency shift: we shall deal later on with the speech rate dependence of fast speech phenomena. In brief, a convincing model should be able to fine-tune the frequencies. The model advanced by Boersma (1997) (see also Boersma and Hayes, 2001), and of which Anttila’s model is a special case, will give a nice answer to these remarks.²²

Boersma (2001) calls our attention to the fact that what Anttila uses frequently (though, not exclusively) is a special type of partially ordered constraint sets, namely, *stratifiable partial orderings*. In such a grammar, constraints are grouped into *strata*, which are fully ranked relative to each other, and within which constraints are unranked. Hence, constraints within one stratum can be permuted freely:

Stratum 1 (undominated): $CON_{1,1}, CON_{1,2}, \dots$
 Stratum 2 (dominated only by Stratum 1): $CON_{2,1}, CON_{2,2}, \dots$
 Stratum 3 (dominated by Strata 1 and 2): $CON_{3,1}, CON_{3,2}, \dots$
 etc.

Anttila and Cho’s unranked hierarchies are not to be confused with the *stratified hierarchies* of Tesar and Smolensky (2000, and earlier versions) introduced for the sake of a learning algorithm. In the latter, the violation marks within one stratum are *summed up* (p. 38), and can also yield more outputs with different violation profiles simultaneously. In the following tableau:

	...	A	B	...
cand1			**	
cand2		**		
cand3		*	*	

(1.8)

²²William Reynolds proposed a further approach already in the early years of Optimality Theory (Nagy and Reynolds, 1997). A *floating constraint* is a constraint that is unranked relative to a span in a ranked constraint hierarchy, the *floating range* of the floating constraint. At evaluation time, that is, on every evaluation occasion, the floating constraint is anchored somewhere within its range, between two neighbouring constraints. If the range contains n constraints, the floating constraint has $n + 1$ docking sites (including the two ends of the range), resulting in $n + 1$ different possible hierarchies. These docking sites, that is, these hierarchies, are postulated to have equal probabilities. Thus, if a certain output form can be generated by m different hierarchies, then the predicted probability of this form is $\frac{m}{n+1}$. The critical remarks about Anttila’s model apply also to Reynolds’ proposal: it does not allow for fine-tuning the frequencies.

all three candidates will survive the stratum formed by constraints A and B , as all of them have two violation marks in sum, and no candidate has less. The constraints in one stratum form a “super-constraint” that we could call $A+B$,²³ and then traditional OT is used to evaluate the candidates with respect to the hierarchy formed by these super-constraints. Notice if cand3 is the best for lower constraints, it will win; whereas in Anttila’s model, cand3 could never win, for it was defeated by either cand1 or cand2 in the two possible permutations of the constraints A and B . The following tableau

	...	A	B	C	...
cand1			**	*	
cand2		**		*	
cand3		**			
cand4		*	*		

(1.9)

predicts an alternation $\text{cand1} \sim \text{cand3}$ in Anttila’s model, and an alternation $\text{cand3} \sim \text{cand4}$ for Tesar and Smolensky (2000).

Notice that a third construction is also possible, that is a mixture of the ideas of Tesar and Smolensky (2000) and of Anttila: seeing each stratum as a “super-constraint”, but which works according to Anttila’s model. That is, a candidate survives a certain stratum, iff it survives at least one of the mini-hierarchies formed by some permutation of the constraints in this stratum. In this approach, tableau (1.9) will return exclusively cand3, because the first three candidates survive the “super-constraint” formed by constraints A and B , out of which cand1 and cand2 are defeated at constraint C . This third approach may also yield more outputs with different violation profiles simultaneously: for tableau (1.8), both cand1 and cand2 will be returned, if they only differ for constraints A and B .

A stratified hierarchy Tesar and Smolensky (2000)-style can be seen as a traditional OT pipeline in which filters are the sum of the constraints within one stratum. Anttila, however, proposes a branching pipe-line, and the output of the different branches are collected only at the very end. The third proposal is a pipe-line which is forked and reconnected at each stratum. As tableau (1.9) has shown, these three—seemingly very similar—models may predict different outputs.

Nonetheless, Tesar and Smolensky (2000) introduced their mutation of Optimality Theory not in order to account for variation phenomena, but in order to introduce a learning algorithm. It is, among others, exactly the erroneous “alternation” forms generated which drive the *Error Driven Constraint Demotion* algorithm. I do not know about any analysis of linguistic variation which would use stratified hierarchies in the sense of Tesar and Smolensky (2000).

1.3.4 Several hierarchies within one: Stochastic OT

After having seen the changes proposed by Anttila, as well as by Tesar and Smolensky to standard Optimality Theory, let us turn to a third proposal. Boersma (1997)’s *Stochastic Optimality Theory* (see also Boersma and Hayes,

²³This notation especially makes sense if you see constraints as integer-valued (or real-valued) functions on the set of candidates.

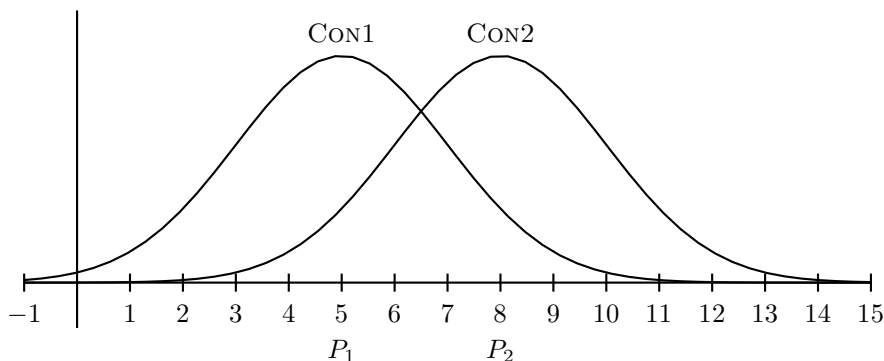


Figure 1.2: **Constraints in Stochastic OT:** Two Stochastic OT constraints (Boersma, 1997; Boersma and Hayes, 2001), CON1 and CON2, are associated with rank $P_1 = 5$ and $P_2 = 8$ respectively, corresponding to the unperturbed ranking $\text{CON2} \gg \text{CON1}$. Yet, the *selection points* p_1 and p_2 used at *evaluation time* are chosen by using a Gaussian noise with $\sigma = 2$. Therefore, the tails of the two distributions overlap, and the probability $\text{Prob}(p_1 > p_2)$ of reranking is not negligible.

2001) suggests a different solution to reranking constraints—that is, to have more than one hierarchy—within one grammar.²⁴

The key idea of *Stochastic Optimality Theory* is to add an *evaluation noise* to the constraint hierarchy. More concretely, the constraints are dispersed along a continuous scale: constraint CON- i is assigned a real number P_i , its *rank*. Ranking CON- $i \gg$ CON- j corresponds to $P_i > P_j$. Yet, whenever the candidates in a tableau have to be evaluated in order to determine a winner (in *evaluation time*, Boersma and Hayes (2001) p. 47), a Gaussian (normal) noise with a standard deviation σ around zero is added to the rank of the constraints (Figure 1.2). Each time, a random number π_i is generated, and the actual ranking of the constraint CON- i is determined by its current *selection point* $p_i = P_i + \pi_i$. That is, the current output is calculated with respect to the hierarchy gained from the p_i values. In the case of $P_i - P_j \gg \sigma$, the ranking $\text{CON-}i \gg \text{CON-}j$ can be seen as categorical. If, however, $|P_i - P_j|$ is on the order of magnitude of, or smaller than σ , then the probability of reranking is considerable, that is, $p_i < p_j$ may occur even if $P_i > P_j$.

Anttila’s model is obtained as a special case within Stochastic OT: constraints unranked by Anttila should be assigned the same (unperturbed) rank in Stochastic Optimality Theory. In contrast to Anttila’s model, however, Stochastic Optimality Theory is able to predict a larger spectrum of *any* frequency distribution by fine-tuning the real numbers assigned to the constraints using the *Gradual Learning Algorithm* (GLA).

Indeed, one of the key selling points of Stochastic OT has been the learning algorithm that comes with it. Without entering details at this point, what one should know is that the *Gradual Learning Algorithm* is fed by surface data following a certain statistical distribution, and it returns a hierarchy (that is the P_i rank of the *a priori* postulated constraints) that reproduces not only the same data, but also the same data *with the same distribution*. In addition, GLA

²⁴Check also Keller and Asudeh (2002) for an assessment and critical remarks.

is robust in the sense that it can handle noisy input data (*i.e.* data including erroneous forms), and is therefore more powerful than the *Constraint Demotion Algorithms* advanced by Tesar and Smolensky (2000).²⁵

Both approaches incorporating reranking within the model—Anttila’s grammars and Stochastic OT—make some very strong predictions. For instance, whenever a number of constraints must be unranked with respect to each other in order to predict a given variation, then all other forms produced by other permutations of these forms must also be attested variations. Take for instance the following example:

	<i>A</i>	<i>B</i>	<i>C</i>
cand1		*	**
cand2	**	*	
cand3		**	
cand4	**		**

(1.10)

In this case, cand1 is returned if and only if $A \gg B \gg C$, whereas cand2 is the winner for $C \gg B \gg A$. Two hierarchies ($B \gg A, C$) cause cand4 to win, and cand3 is the favourite of $A, C \gg B$. If cand1 and cand2 are observed alternating forms—and we have no better analysis—both Anttila and Cho, and Boersma and Hayes must draw the prediction that cand3 and cand4 are also alternatives appearing in the given language. If these strong predictions are confirmed by observation, then the model is corroborated.

Thus, Keller and Asudeh (2002) present an example from German syntax that suggests the following tableau:

	<i>A</i>	<i>B</i>	<i>C</i>
cand1	*		
cand2		*	
cand3		*	*

(1.11)

In this case, only cand1 and cand2 can emerge as the outputs of some hierarchy. The third candidate is *harmonically bounded* by cand2 (to be explained soon); therefore it is an eternal loser. As both approaches considered so far return only candidates that have won the competition for at least one ranking, cand3 is predicted not to emerge as an alternating form.

In the case dealt with by Keller and Asudeh (2002), ranking $A \gg B \gg C$ correctly predicts cand2 to be the best candidate, but cand1 and cand3 being equally wrong (never produced by StOT) does not match empirical findings: it is claimed that cand3 still has a significantly higher level of acceptability than cand1. Boersma (2004b) replies to the criticism of Keller and Asudeh (2002), and by differentiating between production and grammaticality judgements, he explains why a harmonically bounded candidate can be judged better than another candidate, even if it does not appear in production.

The notion of *harmonic bounding*, which will be frequently used in this thesis, is introduced by Prince and Smolensky (2004, p. 209-212)—attributed to Samek-Lodovici—as a strategy to prove that a certain structure of candidate

²⁵For the cognitive relevance of GLA, see for instance Broselow and Xu (2004), which demonstrates a relatively good match between the prediction of GLA and the observed second language acquisition of English final consonants by Mandarin Chinese speakers. For an example where GLA fails, see Pater (2005a).

can never win. it is sufficient to demonstrate that a better candidate exists always. A formal discussion of this concept can be found in Samek-Lodovici and Prince (1999), and the following definition is proposed:

Definition 1.3.1. Harmonic Bounding: *A candidate z is harmonically bounded relative to a constraint set Σ , if there exists a candidate β meeting two conditions:*

- **Strictness.** β is strictly better than z on at least one constraint in Σ .
- **Weak Bounding.** β is at least as good as z on every constraint in Σ .

Subsequently, Samek-Lodovici and Prince (1999) demonstrate that a candidate z that is harmonically bounded by another candidate β (or even by a *bounding set*) is a *loser candidate*. That is, z is suboptimal on every ranking, and hence, can never become an output. The advantage of such an argument is that one does not need to identify the winner in order to demonstrate that another candidate is suboptimal.

1.3.5 MaxEnt OT and cumulativity

This last observation of Keller and Asudeh (2002) brings us to the lack of *counting cumulativity* in these models. Cumulativity effects are the influence of lower ranked constraints on the probability of a candidate, a phenomenon that is required to account for some phenomena, as argued for by Jäger and Rosenbach (2006).

English has two ways to express possession, and—among other factors—the length of the possessor matters: short possessors prefer the *'s*-genitive (e.g. *Eastern's tickets*), while long possessors favour the *of*-genitive (e.g. *the rejection of the last minute French initiative*). In an OT account of this phenomenon, the competing candidates should be the *'s*-genitive and the *of*-genitive constructions of the same possessor-possesum pair. Neither is agrammatical; and yet, they display different frequencies, changing gradually in function of—among others—the length of the possessor (or, of the possessor's length).

Let a constraint assign a violation proportional to this length to the *'s*-genitive. *Counting cumulativity* in this case means that the worse an *of*-genitive is with respect to this constraint, the less probability it has to surface:

/Input1/	...	LENGTH('s)	...
's		**	
of			
/Input2/	...	LENGTH('s)	...
's		****	
of			

(1.12)

In the present case, Jäger and Rosenbach (2006) argue, an adequate model must return different frequencies: say, the *'s*-form should be predicted in 70% of the cases for /Input1/, and in 55% of the cases for /Input2/. Yet, neither Stochastic OT, nor its special subcase, Anttila's model, is able to account for this phenomenon using a single constraint, as both end up by using standard

OT at evaluation time. Some ranking is chosen with a certain probability, and this probability is independent of the input. If constraint $\text{LENGTH}(\text{'s})$ is, then, the highest ranked constraint where the two candidates differ, the 's -genitive will be defeated independently of its number of violation marks. Otherwise, if the *of*-genitive meets its Waterloo earlier, the 's -genitive wins, and the number of violation marks assigned by constraint $\text{LENGTH}(\text{'s})$ plays no role. The way Stochastic OT solves such problems is by introducing a series of binary constraints $\text{LENGTH}(\text{'s}) \leq n$, each of which is violated by 's -genitives longer than n .

Jäger and Rosenbach (2006), therefore, argue for the use of Maximum Entropy models (Goldwater and Johnson, 2003), a variation of Harmonic Grammar (Legendre et al., 1990a). If each constraint $\text{CON-}j$ is associated with some rank (weight) r_j , and output form o corresponding to input form i is assigned a violation level $C_j(i, o)$ by that constraint, then the *harmony value* (the ancestor notion of a violation profile) of that input-output pair is:

$$H(i, o) = - \sum_j r_j C_j(i, o) \quad (1.13)$$

As the values of $C_j(i, o)$ are considered usually positive punishments, this harmony function H is a measure of goodness, due to the negative sign in its definition. The higher (that is, the closer to zero on the negative side) $H(i, o)$ is, the more well-formed the given input-output pair (i, o) .

Maximum Entropy Optimality Theory (MaxEnt OT)—based on information theory (originating in statistical physics)—defines the probability of the grammar returning output o , upon condition of i being the input, as:

$$p(o|i) = \frac{e^{H(i,o)}}{Z(i)} \quad (1.14)$$

where $Z(i) = \sum_{o \in \text{GEN}(i)} e^{H(i,o)}$ is a simple normalisation constant to ensure that for all i ,

$$\sum_{o \in \text{GEN}(i)} p(o|i) = 1 \quad (1.15)$$

Even though the probabilities of the candidates are interconnected through $Z(i)$, the candidates do not compete with each other as directly and cruelly as it happens in traditional OT. If the harmony of a certain candidate is modified, then the probabilities of all other candidates usually change only mildly and uniformly.

Observe that the probability of an output is always higher than zero. Very ill-formed forms are going to have very low, but still positive probabilities. No form is predicted to have zero probability, supposing that GEN produces it. This fact raises a serious problem for MaxEnt OT: cannot it distinguish between low probability forms and totally absurd forms? I believe that a model should be able to draw this distinction, because we should not give up the idea of a linguistic competence totally rejecting some structures—even if in a very diluted form compared to a Chomskyan linguist. One can obviously restrict the production of GEN (in a language-specific manner), or argue for an *ad hoc* threshold under which probabilities are taken to be zero—neither seems to be a very promising way.

MaxEnt OT might be seen as a stochastic variant of Coetzee’s proposal. The core of both models is to introduce a direct connection between the Harmony function and the frequency: the higher the Harmony function $H(i, o)$, the higher the probability $p(o|i)$. Coetzee’s critical cut-off point can be realised here as a major jump in the ranks r_i : constraints ranked higher than this point have a very large rank, and lower ranked constraints have a very low rank. Then, candidates that are suboptimal for constraints ranked higher than the critical cut-off point have a significantly decreased $H(i, o)$ value, hence a very low probability. At the same time, candidates that survive the critical point will have a $p(o|i)$ probability that is larger with orders of magnitude.

MaxEnt OT, by definition, realises counting cumulativity. In (1.12),

$$C_{\text{LENGTH}('s)of}(\text{Input1}, 's) < C_{\text{LENGTH}('s)of}(\text{Input2}, 's)$$

Therefore, due to (1.14), the predicted probabilities will mirror the empirically observed frequencies: $p('s|\text{Input1}) > p('s|\text{Input2})$, supposing everything else is the same.

Similarly, it can be seen that *ganging-up cumulativity* also holds in the Maximum Entropy model. *Ganging-up cumulativity* is the joint effect played by several constraints ranked lower than the constraint where a certain decision is made. Take the following two tableaux with hierarchy $A \gg B \gg C$:

/Input1/	A	B	C
cand1		*	
cand2	*		*

/Input2/	A	B	C
cand1		*	*
cand2	*		

(1.16)

In standard OT, cand2 can never win, independently of its behaviour on lower ranked constraints. In Antilla’s approach, cand2 can emerge only if the constraints are unranked. In Stochastic Optimality Theory and in the Maximum Entropy approach, however, cand2 has a chance even if the three constraints are ranked relative to each other. In Stochastic OT, there is a chance that constraints A and B are reranked at evaluation time (the chance is significant if $P_A - P_B$ is not much larger than σ); whereas no candidate ever has absolute zero probability in the MaxEnt model. Furthermore, and this is *ganging-up cumulativity*, cand1 has more chance with /Input1/ than with /Input2/: in fact, due to its behaviour at the very low-ranked constraint C ! In MaxEnt, this follows directly from the definition (1.14). In Stochastic OT, there is some chance that C is promoted above both A and B , and this probability goes to cand1 in the case of /Input1/, and to cand2 in the case of /Input2/. Jäger and Rosenbach (2006) also bring examples where ganging-up cumulativity can be observed empirically.

As MaxEnt OT, but not Stochastic OT can account for *counting cumulativity*, Jäger and Rosenbach (2006) argue for the use of MaxEnt OT. However, Harmonic Grammar, a close relative of MaxEnt OT, has been unsuccessful among linguists; they prefer standard Optimality Theory, which requires less mathematics and whose more restricted framework produces categorical grammaticality

predictions. The situation may, nevertheless, change in the near future, under the influence of Smolensky and Legendre (2006), which was published when the finishing touches were added to this thesis. In any case, future research should bring further solid arguments in favour of some proposals and against other ones, so that scientific factors and meta-scientific ones (deriving from the sociology of science) will converge.

The model to be presented in the following chapter, Simulated Annealing Optimality Theory (SA-OT), bears much (superficial) resemblance to Harmonic Grammar and MaxEnt OT. Still, I hope, its formalism is closer to standard Optimality Theory, and therefore, may build a bridge between the two communities. Indeed, it is based on a standard Optimality Theoretical model, but adds to it a special application of the simulated annealing algorithm. We shall argue for this application to follow organically from the “philosophy” of standard OT, while Harmonic Grammar and MaxEnt OT employ a very different form of simulated annealing.

After we have considered a few examples from Chapter 5 onwards, Chapter 8 will confront SA-OT with the different approaches just presented and discuss the advantages and disadvantages of each of them.

1.4 Probabilistic linguistics?

The general goal of mainstream modern linguistics following the footsteps of *Noam Chomsky* is the description (modelling, understanding,...) of the linguistic knowledge encoded in the brain of the native speaker.²⁶ As Chomsky states in *Aspects*:

“Linguistic theory is concerned primarily with an ideal speaker-listener, in a completely homogeneous speech-community, who knows its language perfectly and is unaffected by such grammatically irrelevant conditions as memory limitations, distractions, shifts of attention and interest, and errors (random or characteristic) in applying his knowledge of the language in actual 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), pp. 3-4)

(Emphasis in the original.) Thus, Chomskyan linguistics takes interest in the linguistic *competence*, that is, “the speaker-hearer’s knowledge of his language.” *Linguistic performance*, on the other hand, “the actual use of language in concrete situations,” should be outside the scope of linguistics.²⁷

²⁶I am thankful to prof. Jay D. Atlas for a discussion which contributed importantly to rewriting this section. All flaws therein are nevertheless mine. A number of issues raised here are also discussed by Clark (2005), and further relevant points are added—thanks to Gerhard Jäger for suggesting me this interesting article.

²⁷The Chomskyan *performance* definitely parallels the Saussurian concept of *parole* (de Saussure (1974), p. 13), the actual manifestations of language in speech or writing. Yet, there is a difference between Chomsky’s *competence* and Saussure’s *langue* (*ibid.*, p. 9): Saussure sees *langue* as a system that is a social construct, whereas for Chomsky *competence* is a biological (mental, cognitive, psychological, neurological) phenomenon. Still, they share the view that the latter concepts should be the objective of linguistics.

Linguistic competence defines which form (word, sentence, etc.) is *grammatical* in a certain language. Already Chomsky (1957) sets the goal of linguistics to be the selection of the correct grammar for (or the correct theory of) each language (cf. *ibid*, p. 49), where a grammar (a theory) predicts whether a given form will be judged as grammatical by the competence of the native speaker. Consequently, the frequency or the probability of a form in the language should not concern the linguist: “[d]espite the undeniable interest and importance of ... statistical studies of language, they appear to have no direct relevance to the problem of determining or characterizing the set of grammatical utterances” (*ibid*, p. 17).

Recently, however, several linguists have turned back to the frequency of grammatical forms, partially due to the availability of large computational corpora. Additionally, several people have questioned the strict Chomskyan dichotomy of grammatical *vs.* ungrammatical forms: anyone (including Chomsky (1965), p. 11) who has ever tried to form or elicit grammaticality (acceptability) judgements knows that there is a large grey area in between. Both of these factors have motivated a recent turn (back) towards probabilistic (or stochastic) models (cf. e.g. the articles and references in Bod et al. (2003)), as opposed to the algebraic models in traditional Chomskyan linguistics.²⁸

The model to be presented in this dissertation (Simulated Annealing Optimality Theory) seems to contradict the Chomskyan research program in more aspects. Firstly, I shall argue that not only is it a model of linguistic competence, but it also covers parts of linguistic performance. Secondly, it is unapologetically probabilistic (stochastic). Therefore, it is important to reconsider the goals of linguistics at a deeper level, and not to content ourselves with a superficial understanding of Chomsky.

Chomsky (1957) is in fact not as negative towards probabilistic approaches as linguists usually think. It is true that he dislikes the models existing those days (Markov models), and allows statistical models only to describe performance, but not competence:

Given the grammar of a language, one can study the use of the language statistically in various ways; and the development of probabilistic models for the use of language (as distinct from the syntactic structure of language) can be quite rewarding ...

One might seek to develop a more elaborate relation between statistical and syntactic structure than the simple order of approximation model we have rejected. I would certainly not care to argue that any such relation is unthinkable, but I know of no suggestion to this effect that does not have some obvious flaws. (p. 17, n. 4)

The question is whether recent, more elaborate probabilistic models would be “unflawed enough” to Chomsky (1957) in describing the relationship between

²⁸I would not be surprised if a third motivation for many were that probabilistic models are easier to handle than algebraic models. With a few statistical knowledge and programming skills, one can easily create strong models that can be checked quantitatively. Whereas algebraic models require a very good training in mathematics in order to be able to produce new, non-trivial results. It is not a coincidence, furthermore, that the new generation of probabilistic models coincides with the spread of higher performance computers beyond the military and physical research institutes: nowadays, a linguist can also write and run probabilistic simulations easily.

competence and performance.

Recall *Stochastic Optimality Theory* proposed by Boersma (1997) and Boersma and Hayes (2001), introduced already in subsection 1.3.4. As we have already seen, in this approach, each constraint is assigned a real number defining their relative ranking, and the original hierarchy is perturbed by some random noise during evaluation, possibly leading to reranking. The closer the two constraints on the real-valued scale and the bigger the noise, the higher the probability of reranking the two constraints.

Notice the shift of the model's goal with respect to Chomsky's agenda. The objective is not simply to predict whether a form is grammatical or agrammatical, or to generate the set of grammatical forms. Some forms are indeed predicted not to be generated ever (the losers, which are harmonically bound; cf. the next section and Keller and Asudeh (2002)). Yet, the other forms come with a *probability*: the conditional probability of returning this form if the corresponding underlying representation enters the given model.²⁹

Even though some forms have vanishingly low probabilities (in the magnitude of the noise in the observed data), still there is no clear-cut border between improbable and probable forms. The prediction of such a model is not simply a set of grammatical forms, but a set of forms with a probability distribution on them. More precisely, a probability distribution on each set of realisable surface forms per underlying representation: the *conditional* probability $p(o|i)$ of producing output form o if the input form has been i .

How to interpret this probability? This is going to be the major issue. Stochastic OT is a *probabilistic* model, which does not necessarily mean that it is a *frequency-based* model, the target of Chomsky (1957)'s criticism. Statistically observable frequencies are not the only possible interpretations of probabilities.

Indeed, Boersma and Hayes (2001) propose to use Stochastic OT to model both

1. the frequency distribution of free variations;
2. as well as to model gradient grammaticality (well-formedness) judgements of alternative forms by native speakers.³⁰

At this point, we have turned back to the philosophical considerations. Both interpretations contradict the axioms of Chomskyan linguistics to focus on competence, that is, on grammaticality, which is categorical—not a sin in itself.

First, let us discuss the issue of frequency distributions. A typical argument against sentence probabilities goes as follows. It is undeniable that the sentence *I love you* is much more frequent (whatever “frequent” means) than the sentence *Let us now consider various ways of describing the morphemic structure of sentences* (Chomsky (1957) p. 18). And yet, both are equally grammatical.

²⁹To recapitulate: a model in Stochastic OT consists of a GEN, a set of constraints, the initial (noiseless) rank P_i of each constraint, as well as the standard deviation σ of the noise.

³⁰To model gradient grammaticality, Appendix B of Boersma and Hayes (2001) introduces a sigmoid transformation. Using this monotone function, the subjective gradient grammaticality judgements are transformed into data frequencies used to feed GLA. Then, the reverse of this transformation serves to map the frequencies produced by the learnt Stochastic OT model into the predicted well-formedness levels. See also e.g. Boersma (2005).

Concerning the interpretation of the stochastic component of Stochastic OT, see also the remarks in Keller and Asudeh (2002), replied to by Boersma (2004b).

The difference in frequencies is due to extra-linguistic factors, such as to the social embedding of the language.³¹

Nonetheless, one must be very careful when referring to frequencies: what is the *pool* in which we would like to determine the frequency of a certain event? Do we aim at predicting the frequency of a word form “in general” (in a given corpus), or, say, the frequency of a word form among its equivalent alternatives (synonyms, phonologically or morphologically alternating forms, etc.)? Stochastic OT claims the second: it only predicts the chance of outputting surface form o_1 —as opposed to the chance of returning o_2 —for a given input i . What is the chance that the speaker wishes to express somehow input i_1 , and not input i_2 ? This probability is indeed determined by extralinguistic factors, and does not belong to the scope of Stochastic OT.

Therefore, many contemporary probabilistic linguistic models—as exemplified by Stochastic OT—compare the probabilities of alternative forms *corresponding to the same input*, that is, when the extra-linguistic factors have been discarded.

As an example for this debate from within the early probabilistic OT literature, Anttila (1997a) cites Reynolds (1994) (who had proposed the first probabilistic account for variation within OT):

The claim I wish to emphasize here is that phonology itself should not be expected to provide us with [...] exact probabilities. These determinations must be made on the basis of empirical research, taking into account all of the various nonlinguistic factors – such as style, addressee, gender, age, and socioeconomic class – [...]

To which Anttila replies:

While this may be true in many cases, there seems little reason to decide a priori what the limits of phonological theory are. It is entirely possible that there exists variation which is not sensitive to style, addressee, gender, age or socioeconomic class, but is completely grammar-driven. To what extent extragrammatical factors are needed in deriving accurate statistics remains an empirical question. (p. 49)

We shall come back soon to this point, but first let us now turn to the second proposal of how to use Stochastic Optimality Theory, namely, how to model gradient grammaticality judgements of the native speaker.

Gradient grammaticality is explicitly opposed by Chomsky (1965) (p. 11). He distinguishes between *acceptability*, which can be gradient, and *grammaticality*, which is categorical. This distinction is rejected by many contemporary linguists who propagate gradient grammaticality. Maybe forgetting about the Chomskyan concept of *acceptability*, some of them claim that the native speaker cannot help but to judge certain forms on a gradient scale. They should speak of

³¹In the context of Optimality Theory, this argument has been brought by Keller and Asudeh (2002) against Stochastic Optimality Theory, and refuted by Boersma (2004b) using the example (sometimes attributed to Noam Chomsky): “*I’m from Dayton Ohio*” as opposed to “*I am from New York*”.

acceptability, and not *grammaticality*. Others may consciously refute the binary nature of Chomsky's *grammaticality*.³²

Here frequency distribution and gradient grammaticality meet again. For many, grammar (competence) may in itself influence the surface frequencies produced, as well as determine gradient grammaticality. Production and grammaticality or acceptability judgement are the two working modes of the same system, namely, language, the heart of which is competence—their categorical or probabilistic behaviour are thus interconnected.³³

Anttila and Cho (1998) interpret their own probabilistic theory in the following way:

[T]he partial ordering theory accommodates both categorical judgements and preferences without abolishing the distinction between grammaticality vs. ungrammaticality. One and the same grammar can predict both statistical preferences observable in usage data and categorical regularities of the familiar kind. Deriving quantitative predictions from grammars may at first appear to deviate from the standard assumption that a grammar is a model of competence, not performance. However, the distinction between competence and performance is clearly independent of the question whether models of competence are categorical or not. Insofar as usage statistics reflect grammatical constraints, such as sonority, stress and syllable structure, they reflect competence and should be explained by the theory of competence, which partial ordering permits us to do. Conversely, variable phenomena, including statistics, provide critical evidence for evaluating theories of competence.

Thus, is competence maybe assigning a scale, that is, (frequency, grammaticality) probabilities, to the linguistic forms?

It will turn out to be useful to distinguish between three levels, as opposed to the competence-performance dichotomy. The *surface level* is unquestionably performance, that is, what one can empirically observe: the set of produced forms and the acceptability judgements of the native speaker. All seem to agree that this level is probabilistic, the forms in a certain corpus have some frequency, and the judgements are gradient. Performance in a narrow sense includes only the outer phase of language production, and the influence of facts such as one having lost his teeth. But besides phonetics, factors influencing performance also include pragmatics and the structure of the world: certain words or sentences are more frequent simply because they contain messages to be uttered more often in a society.

The deepest level is the *static representation* of the language in one's brain. This level is Chomsky's *competence*, in a narrow sense. Between these two levels is situated the *functioning* of the brain: a dynamical process that produces some

³²Interestingly, Coetzee (to appear) argues that grammaticality is both categorical and gradient, depending on the task that the native speaker is confronted with. He then proposes a (non-quantitative) OT account for both.

³³Within OT, Boersma and Hayes (2001) demonstrates how to use the same system for production frequencies and gradient grammaticality judgements, as two working modes of the same system. In a later paper, however, Boersma (2004b) argues for very different approaches to predicting (conditional) corpus frequencies, on the one hand, and "paralinguistic tasks" (grammaticality judgements and prototypicality judgements), on the other.

output each time. This middle stage, where competence in some broad sense and performance in its broad sense overlap, is still strongly interrelated with competence in a narrow sense; hence the completely grammar-driven variations of Anttila (1997a), and hence the wish to account for it within linguistics. Nevertheless, it might also be seen as already part of performance by a Chomskyan reader. One may compare the *static representation* to anatomy, the *dynamical process* working on top of the static representation to physiology, whereas the *surface level* (performance in the narrow sense) corresponds to the outer appearance of an animal. Clearly, physiology depends on anatomy, and the outer appearance is a result of physiological processes. The animals' outer appearance is not the research topic of biology as a modern science, but physiology is unquestionably.

Stochastic OT, for instance, introduces two levels of description by differentiating between the unperturbed ranks of the constraints and their selection points at evaluation time. The selection points at evaluation time can be seen as a model of this middle level, for they describe grammar-driven variations and are thus closely related to the competence model; much closer than what would follow from Chomsky's traditional policy to exclude performance from linguistics.

Suppose that nobody questions the idea that linguistics is a science that aims at accounting for some observable data. These data, as explained, can be observed on the surface level. Three ways of proceeding can be imagined:

The first one remains on the level of the data, that is on the surface level. Although this approach can be useful—especially in practical applications, such as language technology—most linguists after Saussure and Chomsky are not satisfied with it. I concur, that is, I also would like to understand linguistic competence.

The second approach, on the opposite, concentrates solely on the competence, by insisting on certain axioms and turning competence into an esoteric concept. Such an attitude reminds me of medieval physics: only the celestial motions follow the ideal rules of physics, and therefore the sublunar motions are uninteresting. Additionally, as the celestial motions are ideal, they have to be described exclusively by using ideal concepts, such as circular motions. Even if not to such an extent, the linguist with an aversion towards performance may miss the scientific goal of linguistics, namely, to account for the observed data.

Therefore, I argue for a third approach, which aims at describing the observed performance data (including frequencies and gradient acceptability), and is simultaneously interested in better understanding all of the three layers. The agenda of medieval physics in focusing on celestial motions only had its reward at the end: the Newtonian laws could be most easily derived from these close-to-ideal phenomena. Being selective about phenomena, ignoring some observations, idealising and abstracting is not unscientific behaviour; on the contrary, it is the only method to ensure long-term advance. Nonetheless, one should also keep at least half an eye on the ignored data: after having successfully decomposed the sublunar motions into Newtonian motion and drag or friction, one must proceed and deal with the second factor, as well.

One may object that linguistics has not reached its Newtonian laws yet. I would answer that in order to appreciate Newton's mechanics in the sublunar world, convincing arguments are needed for the proposed decomposition into Newtonian motion versus friction. The physicist should be willing to deal with

friction, and not adhere to the idealisation in a medieval way. Similarly, besides preserving the competence-performance distinction, a successful model of competence has to point at least towards how to deal with the performance. Note that the competence-performance distinction is more than the decomposition of the problem into a first approximation and secondary terms: similarly to the decomposition in mechanics, it provides a better understanding of the factors yielding the data.

Having said that, we should also note that an *a priori* decomposition is certainly a good working hypothesis, but not necessarily the truth. The quote from Reynolds contradicted by Anttila has shown us above that several approaches are feasible about where frequencies and probabilities should enter the model. What many stochastic grammars, such as Stochastic OT, do—and what Simulated Annealing OT will also do—is to take a non-probabilistic grammar (for Chomsky: a non-probabilistic syntax) accounting for competence, and to *add* a stochastic component to it. The crucial question is the interpretation of the statistical distribution added by the stochastic component. Where is statistics located between competence and performance?

Some argue that even competence models, namely, the grammars, should produce probabilities—this is the case for Anttila’s model to be introduced in the next section, as well as for some interpretations of Stochastic OT. Simulated Annealing OT preserves a more Chomskyan concept of static competence in its narrow sense, and adds the stochastic component only to the second level: to the model of the dynamic working of the brain. I agree with Anttila (1997a) cited above, as opposed to Chomsky (1957) and Reynolds (1994): already the encoding of the language in one’s brain includes stochastic features. Still, I prefer to postpone it to the second level within the brain, as I claim that an adequate stochastic grammar must be able to make the distinction between competence in its narrow sense (first level) and the transition towards performance (second level).

For instance, free variations (or, “almost free” variations) are an integral part of language, as we shall see in the next section. A related phenomenon is the emergence of *fast speech errors* as the speech rate increases. Importantly, many have noticed that several phenomena banished to performance, such as the unequal distribution of equally grammatical forms, or the emerging of agrammatical forms as variation, are often related to linguistic factors and to concepts that also play a role in the grammar, that is, the competence. Stochastic OT and Simulated Annealing OT are just “more elaborate” models to account for the properties and frequencies of these alternations than the probabilistic models proposed by Reynolds or Anttila; which, in turn, are still much more elaborate ones than those criticised by Chomsky (1957).

As I shall argue, simulated annealing can on the one hand interpret the Chomskyan notions of “equally grammatical” forms (even though appearing with different probabilities) or forms that are “agrammatical, even if appearing”. These notions—grammatical and agrammatical—refer to the competence in its narrow sense. On the other hand, a probabilistic model of the second level (of the dynamic functioning of the brain) may account (partially) for frequencies: why are some grammatical forms rare, and why do some agrammatical forms (“performance errors”) appear? As this second level is a middle area between competence in its narrow sense and performance in its narrow sense, linguistic factors—supposedly related to competence and not to performance—may still

play a role in shaping probabilities. Nevertheless, I do not deny that further, unquestionably extra-linguistic factors also play a very important role on the third level (performance) in determining the observable frequencies.

1.5 Overview of the thesis

Chapter 2 first introduces the notion of heuristic optimisation techniques (based on Reeves, 1995) in general, and simulated annealing in particular. Afterwards, it argues for why and how simulated annealing could be used for *finding the best candidate of the candidate set in Optimality Theory*. The central result of this chapter, or even of this thesis, is the *Simulated Annealing for Optimality Theory (SA-OT) Algorithm* presented on Fig. 2.8 on page 64, as well as its embedding into a language production model shown in Table 2.1 on page 43. Finally, this chapter also presents a few toy examples demonstrating the use of this algorithm.

Chapter 3 introduces some possible formal approaches to Optimality Theory, proposes a formal definition and analyses its mathematical properties. As a consequence, it demonstrates—even twice—why the *SA-OT Algorithm* presented in Fig. 2.8 follows straightforwardly from the general idea behind standard Optimality Theory. Although the formal concepts employed are introduced, this chapter is heavily mathematical. The less mathematically oriented reader can skip it without losing anything from the rest of the present thesis.

The subsequent chapter touches upon a few issues that put SA-OT in a wider linguistic context. First, the connection between the lexicon and the grammar is dealt with, partially in order to introduce a novel definition for the constraint *Output-Output Correspondence* (OOC, or *Output-Output Faithfulness*), which plays an important role later, in Chapter 5. This section is followed by a few remarks on learnability, an issue unavoidable in formal discussions on Optimality Theory.

The rest of the dissertation presents different applications: stress assignment in Dutch fast speech (Chapter 5), voice assimilation of neighbouring Dutch stops (Chapter 6) and two issues in syllabification (Chapter 7).

The goal of these chapters, however, is less to account for specific linguistic phenomena. Sometimes, the exact nature of the data are unclear or the specific linguistic analysis (the constraints and the ranking used) might be subject to criticism. Certainly, more collaboration with general linguists should have been useful here or there, while I am thankful to those colleagues (primarily to Maartje Schreuder, Dicky Gilbers and Judit Gervain) who supplied me with empirical data or with linguistic models. Yet, all flaws in the linguistic analyses are exclusively mine.

My primary goal in these chapters has been more methodological: to demonstrate how SA-OT can be put into practice, what the roles of the algorithm's parameters are, and what further issues are raised when working with SA-OT. Hence, the models are presented in an order of growing complexity, and a summary is given in section 8.1.

Finally, Chapter 8 reviews the main results of this dissertation, before comparing SA-OT to alternative approaches to Optimality Theory. Finally, the arguments in favour of SA-OT are completed by demonstrating how well it fits into a more general cognitive framework.