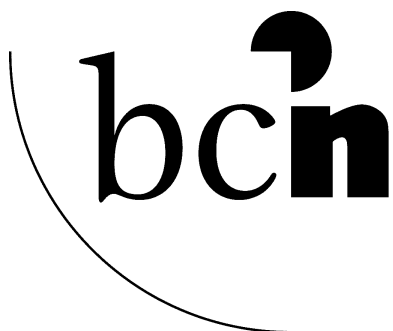


Finding the Right Words

Implementing Optimality Theory
with Simulated Annealing

Tamás Bíró



The work in this thesis has been carried out under the auspices of the Center for Language and Cognition Groningen (CLCG) and the School of Behavioral and Cognitive Neurosciences (BCN), Groningen, with the financial support of the University of Groningen's program for High Performance Computing and Visualization (2000).



Groningen Dissertations in Linguistics 62

ISSN 0928-0030

ISBN 90-367-2876-2

Document prepared with L^AT_EX 2_ε, using Stasinou Konstantopoulos's RuGthesis.cls.

Printed by Print Partners Ipskamp, Enschede.

RIJKS*UNIVERSITEIT* GRONINGEN

Finding the Right Words

Implementing Optimality Theory
with Simulated Annealing

Proefschrift

ter verkrijging van het doctoraat in de
Letteren
aan de Rijksuniversiteit Groningen
op gezag van de
Rector Magnificus, dr. F. Zwarts,
in het openbaar te verdedigen op
donderdag 7 december 2006
om 13.15 uur

door

Tamás Sándor Bíró

geboren op 24 november 1975
te Boedapest, Hongarije

Promotor:	Prof. dr. ir. J. Nerbonne
Copromotores:	Dr. G. Bouma Dr. D. G. Gilbers
Beoordelingscommissie:	Dr. A. Albright Prof. dr. P. P. G. Boersma Prof. dr. G. Jäger Prof. dr. J. Koster

Preface

Finding the right words... it is especially hard when one writes his or her first (and usually only) doctoral thesis. Finding the right topic might be even harder.

Originally, the goal was to write a dissertation on finite-state implementations of Optimality Theory (OT), possibly with some aspects on learning in OT. The noteworthy results of this research line can be found in B     (2003) and in B     (2005c), besides a few unpublished papers or research reports.

Meanwhile, I remembered that I had written a seminar paper in phonology (B    , 1997) many years earlier to *P     Rebrus*, the first person who introduced me to OT, in which I proposed to combine OT with simulated annealing, a technique I had learned about in a course on spin glasses with *Imre Kondor* as a student of physics. I was told then that Optimality Theory had been born exactly from simulated annealing and connectionism, even if you do not find any overt trace of it in most of contemporary OT literature. To be sure, my original seminar paper was very superficial, and it was during my Ph. D. research that I started to think more about the issue. I received a further impetus by reading the article of J       (2002) on bidirectional finite-state OT, which reminded me of the passion I used to have towards different types of “infinite numbers”.

Then, a discussion with Bal       and Kriszta Szendr     introduced the polynomials into the picture, while *Maartje Schreuder* and *Dicky Gilbers* supplied me with several phonological phenomena to work on. So the plan became to present in the thesis the combination of simulated annealing with Optimality Theory as an alternative to finite-state approaches to OT. But the topic grew larger and larger, and finally finite-state approaches were omitted from the dissertation altogether. And yet, the fact that many readers have found my arguments often less convincing, and certain decisions quite *ad hoc*, shows that the work is not finished yet: a more forceful framework and more compelling details have to be worked out in order to account convincingly for the observed phenomena. It is to be hoped that the present thesis is only the first, and not the last word pronounced on SA-OT.

Hearty thanks go to *Gosse Bouma*, my supervisor, *Gertjan van Noord*, who practically acted as a co-supervisor, and *John Nerbonne*, my promotor. The frequent discussions with them improved the content of my work, even if I not always listened to their advice. Indeed, you must be in a motivating environment in order to be fruitful, and they unquestionably provided me with such an environment on the highest level. Part of this environment were my colleagues, my roommates, who also helped me significantly (including solving technical problems): *Gerlof Bouma*, *Francisco Borges* and *Robbert Prins*. On a larger level, each member of the Alfa-Informatica Department, the Center for Language and Cognition Groningen (CLCG), the School of Behavioral and

Cognitive Neurosciences (BCN), ESSLLI and LOT have also contributed to this work, thanks to informal discussions, the reading groups, fascinating summer and winter courses or financial support. The research reported in this thesis has been primarily financed by the University of Groningen’s program for High Performance Computing and Visualization (2000), which I gratefully recognise.

As already mentioned, I am enormously thankful to *Maartje Schreuder* and *Dicky Gilbers*, whose empirical work on stress assignment in Dutch fast speech served as the first concrete example for SA-OT, and our further discussions in the subsequent almost two years have also been extremely fruitful for both Maartje (see the fruits in Schreuder, 2006) and me. Similarly has my dissertation profited from my cooperation with *Judit Gervain*, as I could use the results of her experiments. It is, however, not possible to emphasise sufficiently that I am alone to be blamed for all flaws (to be found especially in the phonological analyses) of the present thesis, which are often the result of not accepting their advice. Chronologically last, but far not least, I have to express my gratitude to the members of the reading committee—Adam Albright, Paul Boersma, Gerhard Jäger and Jan Koster—not only because I am proud of having them as my readers, but also because their critical remarks were very constructive in the last phase.

As language and style is an important aspect of any text, I am very much indebted to *Angela Ashworth* and *Ruben Comadina Granson* for their courses on academic writing in English and English for presentation—I dare to say, a must to any Ph. D. student who is non-native in English. The list of those who have shaped my way of thinking—such as my high school maths teachers, my professors in physics, linguistics and Judaic studies—and those who helped me in any other way in preparing this thesis is open ended, and any enumeration would unquestionably leave out somebody. Special thanks goes to the audience of each of my presentations, because without their questions my trains of thought would have been even less understandable.

Finally, here is a (certainly not complete) list in randomised order of those people who made my stay in the Netherlands pleasant: the secretaries of Cluster Nederlands, Pieter and Gezin Oegema (my landlord and landlady), the Czachesz family (István, Gyöngyi and Vica), the Hungarian and pseudo-Hungarian people in Groningen (Mónika Zempléni, Gábor Imre and Anikó Pausch, Andrea Szentgyörgyi¹ and Mihalis Kavaratzis, Szilárd Csiszár, István Back, Anikó Szpenatyi and Pieter, Janka Salát, András Káldi, etc.), the whole Folkingestraat community (Bep, Anette, Rami, Gershon, etc.), Bea Nink, Volker Nannen, Piroska Lendvai, Stefan van der Poel, Ela Polek, Géza Xeravits, Sophia Katrenko, Wout van Bakkum, Farah Berri, Paul and Liesbeth Gabriner.

My paranympths, *Lonneke van der Plas* and *Gerlof Bouma*, helped me out in the final stages before the defence whenever I needed, for which I am especially grateful. Yet, the largest credit goes to my parents and to my family for all of the last 30 years: *Péter, Márta, köszönök szépen mindent nektek!*

¹She is the cousin of the stepbrother of my sister’s husband.

Contents

1	Introduction	1
1.1	Introduction to Optimality Theory	1
1.1.1	Optimality Theory for my grandma	1
1.1.2	Optimality Theory as a scientific model	2
1.1.3	A slightly more formal definition of OT	7
1.2	Infinite candidate sets, implementing OT	10
1.3	Variation within OT	13
1.3.1	Forms assigned the same violation marks	15
1.3.2	Non-optimal candidates emerging	16
1.3.3	Several hierarchies within one: reranking	18
1.3.4	Several hierarchies within one: Stochastic OT	21
1.3.5	MaxEnt OT and cumulativity	24
1.4	Probabilistic linguistics?	27
1.5	Overview of the thesis	34
2	Optimality Theory and Simulated Annealing	35
2.1	Heuristic optimisation and simulated annealing	35
2.1.1	Heuristic optimisation for OT	35
2.1.2	A technique from statistical physics	37
2.1.3	Spin glasses in the brain	42
2.2	Simulated Annealing for Optimality Theory	44
2.2.1	How to combine simulated annealing with OT?	44
2.2.2	Topology on the search space	46
2.2.3	Temperature for OT	51
2.2.4	Introducing the SA-OT algorithm	62
2.3	Playing with SA-OT	66
2.3.1	When SA-OT works	66
2.3.2	When SA-OT <i>does not</i> work	69
3	Formal approaches to SA-OT	75
3.1	Towards a formal definition of OT	75
3.1.1	Constraints	76
3.1.2	Hierarchies	76
3.1.3	An order on violation profile-like vectors	77
3.1.4	Comparing candidates	80
3.1.5	The definition of Optimality Theory	81
3.1.6	Realisations of the Harmony function	82
3.2	Violation profiles as real numbers	84

3.3	Violation profiles as polynomials	88
3.3.1	Comparing polynomials	88
3.3.2	Simulated annealing with polynomials	92
3.4	Violation profiles as ordinal numbers	94
3.4.1	Ordinal numbers can realise violation profiles	95
3.4.2	SA-OT with ordinal numbers	97
3.4.3	Arguing more for the definition of $e^{-d/t}$	100
3.5	Summary of the formal approaches	103
4	The Linguistic Context of SA-OT	105
4.1	A few words about the lexicon	105
4.1.1	English Past Tense	106
4.1.2	Burzio's physical model of the mental lexicon	107
4.1.3	Burzio's Output-Output Correspondence	110
4.1.4	Burzio's model and SA-OT	112
4.1.5	Constituent-Output Correspondence	113
4.2	Learning with SA OT?	116
5	Stress in Dutch Fast Speech with SA-OT	121
5.1	The Schreuder-Gilbers model of Dutch stress	122
5.2	Fast Speech and different variations of OT	126
5.3	Fast speech and SA-OT: the building units	129
5.4	Experimenting with the Schreuder-Gilbers model	133
5.5	Further experiments	136
5.5.1	The role of T_{step}	136
5.5.2	The role of T_{max} and T_{min} (1)	139
5.5.3	The role of T_{max} and T_{min} (2)	142
5.6	Improving the Schreuder-Gilbers model	144
5.7	Getting rid of OOC: Biased initial state	156
5.8	Discussion	159
6	Dutch Voice Assimilation with SA-OT	161
6.1	The magic square	161
6.2	Voice assimilation in Dutch	165
6.3	The building blocks of Simulated Annealing	167
6.4	Model 1: Finite search space	171
6.5	Model 2: Infinite search space	175
6.5.1	Enlarging the search space	175
6.5.2	The landscape	176
6.5.3	Tuning the output of the model	179
6.5.4	The interaction of K_{max} with T_{step}	180
6.5.5	Experiments	186
6.6	What have we learnt from [voice] assimilation?	192
7	Word Structure and Syllable Structure with SA-OT	195
7.1	Th' article in Hungarian	195
7.1.1	The behaviour of the definite article	195
7.1.2	Constructing a model	196
7.1.3	Refining the model by changing the topology	200
7.1.4	Refining the model by demoting constraints	201

7.1.5	Conclusion	204
7.2	Syllabification (CVT) theory	205
7.2.1	Basic CV Theory	206
7.2.2	Syllabification with simulated annealing I.	207
7.2.3	Syllabification with simulated annealing II.	210
7.2.4	Conclusion	213
8	Conclusion: Is SA-OT, thus, Better?	217
8.1	Summary	217
8.2	Advantages (and disadvantages) of SA-OT	219
8.2.1	SA-OT and specific linguistic phenomena	219
8.2.2	SA-OT, competence and performance	222
8.3	SA-OT as a general cognitive model	224
	Bibliography	229
	Summary	239
	Samenvatting - Dutch summary	243
	Összefoglalás - Hungarian summary	247

List of Figures

1.1	Basic architecture of an Optimality Theoretic grammar	8
1.2	Overlapping constraints in Stochastic OT	22
2.1	Landscape produced by a real-valued energy or cost function	38
2.2	Traditional Simulated Annealing Algorithm.	40
2.3	An asymmetric landscape with three states	41
2.4	Optimality Theory and SA-OT	44
2.5	Schematic view of a search space	46
2.6	The $e^{-1/x}$ function	52
2.7	Visualising the domains traversed by temperature	61
2.8	The algorithm of Optimality Theory Simulated Annealing	64
2.9	An asymmetric landscape with three states	67
3.1	Levels of representing a violation profile	85
5.1	The interaction of three constraints in Stochastic OT	127
5.2	Topology for metrical stress with a three-syllable input	131
5.3	Tuning T_{step} in the Schreuder-Gilbers model	135
5.4	Zooming in on the interval $0.001 \leq T_{step} \leq 0.1$	138
5.5	Four different T_{max} - T_{min} pairs	143
5.6	*CLASH included	146
5.7	ALIGN(word, foot, left) included	148
6.1	The “magic square”	161
6.2	Search space used in the first model for voice assimilation	168
6.3	Search space used in the second model for voice assimilation	169
6.4	3-D landscape of the first model for voice assimilation	172
6.5	The second search space used for <i>op die</i>	177
6.6	“Channelling” effect in an infinite search space	179
6.7	Distribution of the random walker’s position	185
6.8	Frequency of [pt] when varying either K_{max} or T_{step}	187
6.9	Phase space for <i>op die</i>	189
7.1	Search space for the Hungarian article <i>az</i>	198
7.2	Frequencies of [az# ^m E] employing different K_{max} values (1)	199
7.3	Frequencies of [az# ^m E] employing different K_{max} values (2)	201
7.4	Search space for the Hungarian article <i>a</i>	202
7.5	The effect of demoting constraint KEEPSEGMENTSHORT	203
7.6	Varying the parameters of the <i>a priori</i> probabilities	211

List of Tables

2.1	The proposed three-level model of the human language	43
2.2	The way the transition probability $P(w \rightarrow w' T)$ is dependent on the temperature	53
2.3	An example for the difference of two violation profiles	57
5.1	Observed frequencies per type	123
5.2	Slow versus fast speech in the simplest Schreuder-Gilbers model .	137
5.3	Tuning T_{step} in the simplest Schreuder-Gilbers model	138
5.4	Varying T_{max} and T_{min}	141
5.5	Varying T_{max} and T_{min} , with $T_{step} = 1.5$	142
5.6	*CLASH included	145
5.7	ALIGN(word, foot, left) included	147
5.8	Type 1 words with constraint FOOTTYPE(trochaic)	148
5.9	Type 2 words	149
5.10	Outputs when using $OO_{\sigma=usus, z=3}$	151
5.11	Introducing constraint ALIGN(Word, Foot, right)	152
5.12	Results using hierarchy (5.9)	152
5.13	Type 3 words, $z = 0$	153
5.14	Type 3 words, $z = 1$	153
5.15	Summary for hierarchy (5.10) with $z = 0$	154
5.16	Summary for hierarchy (5.10) with $z = 1$	154
5.17	Observed vs. simulated frequencies	155
5.18	Biased initial candidate ([su][su])	157
6.1	Frequency of [pt] as a function of K_{max}	188
6.2	Frequency of [pt] as a function of T_{step}	189
6.3	Parameter settings producing [pt] with 25% chance	190
6.4	The turning point around $\tau = \kappa + 2$	191
7.1	Varying the parameters of the <i>a priori</i> probabilities	210
7.2	The role of ($P_{reparse}$ and $P_{postproc}$) in CV Theory	212
7.3	Different hierarchies in CV Theory	213
7.4	The dependence of precision on the hierarchy	214

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:

$$\text{CON-1} \gg \text{CON-2} \gg \dots \gg \text{CON-}i \quad (1.1)$$

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>	*!		...	
 <i>B</i>			...	*
<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  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.

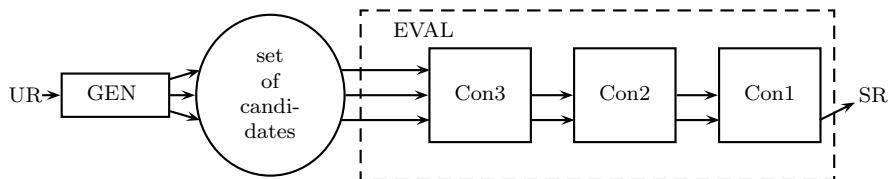


Figure 1.1: The basic architecture of an Optimality Theoretic grammar

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(Foot, 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 agrammatical.

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 $[\epsilon] \sim [\emptyset]$ ($e \sim \ddot{o}$) in many words, originating in the standardisation of two different dialectal forms, with a minimal preference for the $[\epsilon]$ forms in written language:

$$\begin{array}{llll}
 fel & \sim & f\ddot{o}l & \text{'up'} \\
 felett & \sim & f\ddot{o}l\ddot{o}tt & \text{'above'} \\
 sepr\ddot{u} & \sim & s\ddot{o}pr\ddot{u} & \text{'broom'} \\
 tejfel & \sim & tejf\ddot{o}l & \text{'sour cream'}
 \end{array} \tag{1.5}$$

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* \sim *f\ddot{o}l\ddot{o}tt* 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 $[\emptyset]$ 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): $\text{CON}_{1,1}, \text{CON}_{1,2}, \dots$
 Stratum 2 (dominated only by Stratum 1): $\text{CON}_{2,1}, \text{CON}_{2,2}, \dots$
 Stratum 3 (dominated by Strata 1 and 2): $\text{CON}_{3,1}, \text{CON}_{3,2}, \dots$
 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.

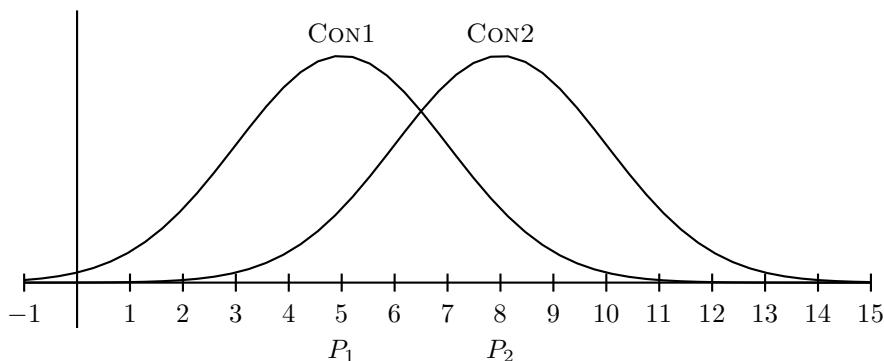


Figure 1.2: **Constraints in Stochastic OT:** Two Stochastic OT constraints (Boersma, 1997; Boersma and Hayes, 2001), CON1 and CON2, are associated with rank $P_1 = 5$ and $P_2 = 8$ respectively, corresponding to the unperturbed ranking $\text{CON2} \gg \text{CON1}$. Yet, the *selection points* p_1 and 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 CON- $i \gg$ 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-possessum 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 $\text{LENGTH}('s)$ 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 $\text{LENGTH}('s)$ plays no role. The way Stochastic OT solves such problems is by introducing a series of binary constraints $\text{LENGTH}('s) \leq n$, 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.

Chapter 2

Optimality Theory and Simulated Annealing

2.1 Heuristic optimisation and simulated annealing

2.1.1 Heuristic optimisation for OT

It has been mentioned that Optimality Theory can be seen in two different ways. According to the most frequent picture, the output of GEN is connected to a pipe-line of constraints acting as filters. Alternatively, we may see EVAL as a special function derived from the constraints: the goal is to find the candidate that optimises the Eval-function (also called the *Harmony function*).

Therefore, Optimality Theory belongs to the family of combinatorial optimisation problems. The world is indeed full of optimisation. Entrepreneurs maximise their benefit by minimising costs, football players maximise goals, whereas runners minimise time. Hedonists maximise pleasure, students minimise homework, and so forth. In physics, energy is to be minimised and entropy maximised. In all of these problems, a given function $f(x)$ has to be optimised (usually minimised, sometimes maximised) subject to a set of conditions, such as $g_i(x) \geq b_i$ (for $i = 1, \dots, m$) (Reeves, 1995): for what x satisfying these conditions is the quantity $f(x)$ optimal?

In Optimality Theory, the *Harmony function* $H(x)$ should be optimised, subject to the condition $x \in GEN(UR)$, and the solution is predicted to be the surface form corresponding to the input underlying representation UR . Famous members of the family of combinatorial optimisation problems include the *assignment problem* (minimise the costs, if a set of n people is available to carry out n tasks, and if person i performing task j costs c_{ij} units); different problems in logistics (planning the routing of vehicles); the *travelling salesman problem* (find a tour of minimum distance that passes through a given set of points, say, towns); or *node colouring* of graphs (find the colouring with the smallest possible number of colours, such that adjacent nodes are not given the same colour).

Certain problems are easy to solve, others are more complex. Sometimes it would require a huge amount of computational resources to find it with cer-

tainty. A great variety of problem-specific or more general solutions have been developed in the recent decades, and this is not the place to give a general overview of them. Here we focus on *heuristic techniques*, and in particular on one of them, namely, *simulated annealing*. Reeves (1995) defines *heuristic* (p. 6) as “a technique which seeks good (i.e. near-optimal) solutions at a reasonable computational cost without being able to guarantee either feasibility or optimality, or even in many cases to state how close to optimality a particular feasible solution is.”

Simulated annealing (Reeves, 1995) is one of the simplest and most popular among these heuristic techniques. Simulated annealing, similarly to other heuristic techniques, does not guarantee the correct answer. You do not even know if the answer returned is the optimal one! If you want the probability of obtaining the optimal solution to approach 1, you may require more iterations than exhaustive search (i.e., checking each possibility, supposing a finite search space).

Nevertheless, a heuristic technique such as simulated annealing may have its purpose within linguistics, in general, and within the Optimality Theoretic paradigm, in particular. What are the arguments for finding the optimal element of the candidate set—the element optimising the Harmony function—with simulated annealing, and not with other techniques mentioned in section 1.2?

First, the computation involved in heuristic techniques is simple, not involving a large working or storing capacity—an attractive feature both for cognitive models and for language technology.

Second, you are guaranteed to be returned *some* answer within constant time, even if not necessarily the correct one. In that respect, heuristic algorithms resemble human speech, which also produces outputs at a constant rate, and these outputs are not always fully correct. A counter-argument may be that the brain is an extremely powerful computer and the speech flow is only slowed down by the inertia of the speech organs, this is why we do not observe the time difference between processing simpler and more complex structures. Why then the higher rate of errors for complex structures? Why then the increased number of errors in fast speech, many of which are not due to the inertia of the speech organs? Consequently, I propose to view (some of the) *performance errors*—for instance, in fast speech—as errors introduced by the mental computation. In fast speech, the human mind is willing to give up precision in order to gain speed. And this is exactly a phenomenon that can be reproduced by simulated annealing: not only does simulated annealing return some output certainly within a constant time, but this constant time can be diminished (the simulation can increase its speed) at the cost of precision.

As a matter of fact, I have to acknowledge that sometimes it is hard to distinguish between performance errors that are consequences of computational difficulties and performance errors that are consequences of physical problems. Indeed, we run into the problem of how much phonology is grounded in phonetics: for instance, whether an assimilation process is related to the inertia of some speech organs. So, dropping some segments in fast speech might be argued to result from phonology (which is grounded in phonetics); but it can also be explained within phonetics, even if the phenomenon is influenced by phonological factors. However, progressive and regressive voice assimilations should be equally good solutions from a physical point of view, so if grammar requires regressive assimilation to take place, then a progressive assimilation can definitely

be seen as a result of computational difficulties. Finally, performance errors outside phonology (especially in syntax) cannot be derived from the inertia of the speech organs. In brief, my working hypothesis is that there exist performance errors that are due to computational difficulties.

Additionally, we have seen in section 1.2 that finding the optimal element of the candidate set can be a hard problem: Eisner (2000b) proves that it is NP-hard—worst case exponential—in the size of the grammar. Although many proposals have been advanced to use a series of simple OT-grammars,¹ the traditional view is still to view the language faculty as one huge OT-grammar. In turn, a huge grammar implies computational difficulties that heuristic techniques can overcome the most simply.

2.1.2 A technique from statistical physics

Our agenda is, thus, to employ simulated annealing in modelling real-time speech production, including variation or speech errors. In particular, to combine simulated annealing with Optimality Theory. This marriage is, however, far from obvious. In the present subsection, we introduce the key idea of *simulated annealing*, so that we can combine it with Optimality Theory in the next section. The implementations to be presented in chapter 5 will demonstrate the validity of the previous arguments.

Simulated annealing, also referred to as *Boltzmann Machines* or *stochastic gradient ascent* (descent), is a wide-spread stochastic technique for combinatorial optimisation, especially in the fields of neural networks (e.g. Reeves (1995), Spall (2003)). The idea originates in solid state physics (Metropolis et al., 1953),² and was first presented by Kirkpatrick et al. (1983), as well as, independently, by Černý (1985). It can be also related to the root finding algorithm of Robbins and Monro (1951) (cf. Spall, 2003, p. 97-98).

In linguistics, simulated annealing has been used for (context-free) parsing, by Selman and Hirst (1985), Selman and Hirst (1994) or Howells (1988). Kempen and Vosse (1989) present a cognitive architecture based on activation decay and simulated annealing, and compare the result of simulated annealing with different cooling schedules to aphasic data (see also Vosse and Kempen (2000) for a reconsideration of this model). The link of simulated annealing to connectionist foundations of Harmony Theory—the historical background of OT—goes back as early as Smolensky (1986). However, it has never been used with a concrete linguistic model within OT to my knowledge.

¹*Stratal OT*, a combination of Kiparsky (1982)'s *Lexical Phonology* with Optimality Theory, was already introduced by McCarthy and Prince (1993b). Further examples include, for instance, Bíró and Hamp (2002), who propose a similar model for Israeli Hebrew morphology.

²This paper, co-authored by the late Edward Teller, presented a modification of the Monte Carlo integration over the configuration space. The question was how to calculate the average of some quantity F over a large system. The original Monte Carlo algorithm randomly generated configurations x_i , in each of which the value of F was calculated ($F(x_i)$); then, the different $F(x_i)$ s were averaged with the probabilities $P(x_i)$ of the configurations, as weights. Nonetheless, the simulation is very likely to generate some of the many improbable configurations, and to avoid the few, high-probability equilibrium states. Therefore, the Metropolis algorithm proposes a random walk in the state space (or phase space), along which the values of F can be averaged. Indeed, the procedure to be described with respect to simulated annealing produces a series of states x_i that can be used as a representative sample, that is, in which the frequency of a state x_i is proportional to its theoretical probability $P(x_i)$.

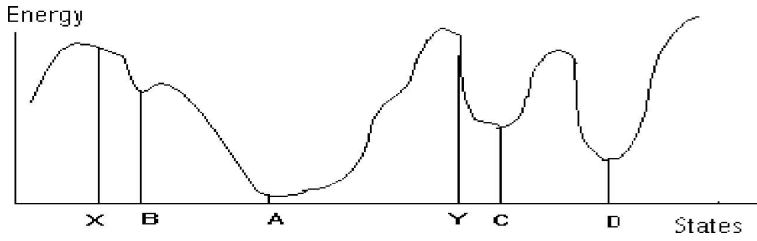


Figure 2.1: Landscape produced by a real-valued energy or cost function (Bíró, 1997).

Simulated annealing is a modification of the algorithm called *gradient descent* (*iterative improvement*; it is also called as *gradient ascent* or as *mountain climbing in the fog*, when used for maximisation). Gradient descent performs a random walk in the search space with the restriction that you may never move upwards, only horizontally, or, preferably, downhill. (For gradient ascent, one simply has to reverse the directions.) At each time step, the random walker picks a neighbour position, compares it with its actual place, and moves there only if the target position is better. Two strategies can be followed: either a neighbour is chosen randomly, or the walker picks its best neighbour, that is, it takes the steepest slope. The position of the random walker at the end of the walk is returned by the algorithm. In both cases, however, the random walker will easily be stuck into local minima—we need therefore a trick to avoid non-global local minima.

Simulated annealing, as we shall see presently, also allows upwards moves, in order to avoid getting stuck into these local minima. The trick originates in statistical physics (thermodynamics and solid state physics).

An interstitial defect in a crystal lattice, say, in the structure of some sort of metal, corresponds to a (non-global) local minimum in the energy of the lattice. Although the perfect lattice would minimise the energy level, the defect is stable, because any local change would increase the energy. In order to climb this energy barrier and to reach the global minimum, one needs either to globally restructure the lattice within one step, or to be permitted to temporarily increase the energy of the lattice. Heating the lattice means to go for the second option. The lattice is allowed “to borrow” some energy, that is, to transform thermic energy provisionally into the binding energy of the lattice, thereby climbing the energy barrier separating the local minimum from the global minimum.

At temperature T , the probability of a change that increases the lattice’s energy by ΔE is $e^{\frac{-\Delta E}{kT}}$. Here, $e \approx 2.7182$ is the base of the natural logarithm, and $k \approx 1.38 \cdot 10^{-23} JK^{-1}$ is Boltzmann’s constant, which connects energy (measured in *joules*, J) to the phenomenological measure of temperature (measured in *kelvins*, K). The higher the temperature, the smaller the absolute value of the exponent, and therefore the higher jumps in energy are reasonably probable. At relatively low temperature, the probability of a relatively high jump in energy is not likely to happen.

When annealing a metal, we increase its temperature, therefore the lattice can easily get out of a local minimum. Then, we slowly cool it down. The lower the temperature, the smaller the energy hills the system is able to climb, thus it gets stuck in some valley. Hopefully, this valley corresponds to the global

minimum, but at least to a minimum smaller than the initial local minimum. At the end of the annealing, the system probably arrives at the bottom of a deep valley.³

Now, the idea of simulated annealing is straightforward (Kirkpatrick et al., 1983). Suppose that we wish to find the state of a system for which the quantity E (say, energy or evaluation) is minimal. We define some sort of fictitious “temperature” T , a mere control parameter, and choose an initial state w_0 , the starting point of a random walk.

At each step of the simulation, when we are in state w , we randomly choose one of the neighbour states (w') of the actual state. This random choice is one of the main factors driving the stochastic behaviour of the algorithm. It presupposes some sort of *neighbourhood structure* (*topology*) on the search space that determines which states are the neighbouring states of w . Moreover, the topology also defines the *a priori* probability $P_{\text{choice}}(w'|w)$ of choosing w' if we are in state w . Unless one allows the system not to move at all, we stipulate that

$$\sum_{w' \in \text{Neighbours}(w)} P_{\text{choice}}(w'|w) = 1 \quad (2.1)$$

In the simplest case, a state has a finite number of neighbours, and each of them has equal *a priori* probability. We shall come back to further possibilities in section 2.2.2.

Once we have picked a neighbour, we have to decide whether to move there or not. If the neighbour state w' represents a lower level in E than w , we move there. Otherwise, we move only with probability $e^{\frac{-\Delta E}{T}}$, where ΔE is the increase in energy in the case we took that move:

$$P(w \rightarrow w') = \begin{cases} 1 & \text{if } E(w') \leq E(w) \\ e^{-\frac{E(w') - E(w)}{kT}} & \text{if } E(w') > E(w) \end{cases} \quad (2.2)$$

In other words, a random number with a uniform distribution is chosen from the interval $(0, 1)$, and if it is smaller than $P(w \rightarrow w')$, then we move from w to w' . Why an exponential function? One argument is the analogy from statistical physics, where the probability of state s at temperature T is proportional to $e^{-E(s)/T}$. Another argument is that the exponential function is the only non-trivial continuous function F that has the property $F(x+y) = F(x)F(y)$; hence, the probability of moving upwards $x+y$ is equal to the probability of an uphill step with a difference x followed by an uphill step with a difference y .

Please remember the difference between the *a priori* probability $P_{\text{choice}}(w'|w)$ of choosing state w' when the random walker is in state w , on the one hand; and the probability $P(w \rightarrow w')$, defined in Eq. (2.2), of really moving from state w to w' , once w' has been chosen, on the other. Clearly, the probability of

³In fact, the energy-structure of traditional lattices is too simple, too trivial, so that the global optimum is usually very well approached. This may be the reason why the original idea, already published by Metropolis et al. (1953), did not raise much interest beyond solid-state physics and related fields for thirty years. The behaviour of spin glasses in magnetic fields, however, yields quite a complicated energy-structure. The investigation of spin glasses led, therefore, both Kirkpatrick et al. (1983) and Černý (1985)—independently of each other—to applying Metropolis’ algorithm on optimisation problems in general, such as the *travelling salesman problem*. For more mathematical details, consult, among others, van Laarhoven (1987).

```

ALGORITHM: Simulated Annealing
Parameters:  w_init      # initial state (often randomly chosen)
             t_max       # initial temperature > 0
             alpha        # temperature reduction function

w := w_init ;
t := t_max  ;
Repeat
  Repeat
    Randomly select w' from the set Neighbours(w);
    Delta := E(w') - E(w) ;
    if Delta < 0
      then
        w := w' ;           # move to lower energy state
      else
        generate random r uniformly in range (0,1) ;
        if r < exp(-Delta / t)
          then w := w' ;    # move to higher energy state
        end-if
      end-if
    Until iteration_count = nrep      # usually simply: nrep = 1
    t := alpha(t)
  Until stopping condition = true     # usually: until t < t_min
Return w      # w is the approximation to the optimal solution

```

Figure 2.2: The algorithm of traditional Simulated Annealing.

actually moving from w' once in w is the product of the two probabilities. The chance of not moving from w at all in a certain step is:

$$1 - \sum_{w' \in \text{Neighbours}(w)} P_{\text{choice}}(w'|w) \cdot P(w \rightarrow w') \quad (2.3)$$

Equation (2.2) is the point where the control parameter called “temperature” T has come into play. Observe that if $\Delta E = E(w') - E(w) \ll T$, the probability of moving is practically 1. If, however, $\Delta E = E(w') - E(w) \gg T$, the move to w' becomes almost prohibited. In brief, the role of T is to define the order of magnitude of ΔE in which $P(w \rightarrow w')$ is neither very close to 1, nor very close to 0.

At the beginning of the simulation, T is assigned a high value, making any move very likely. The value of T is then decreased gradually according to some *cooling schedule*, yet T remains always positive. By the end, T is very close to zero: according to (2.2), even the smallest jump becomes highly improbable then. When the temperature has reached its lowest value, the algorithm returns the state into which the random walker is finally “frozen”.

The general algorithm of simulated annealing can now be introduced in Fig. 2.2. We follow Reeves (1995), with a few modifications and indications pointing to the way we shall use simulated annealing later.

Why does adding temperature improve the performance of this search algorithm compared to gradient descent? Imagine an asymmetric search space composed of three states, as represented in Fig. 2.3. Suppose that we would

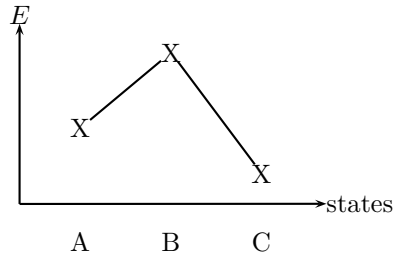


Figure 2.3: An asymmetric landscape with three states, two of which are local optima, but where only state C is a global optimum. State B is a neighbour of both A and C, whereas states A and C are not neighbours of each other.

like to minimise the function E .

Gradient descent (iterative improvement) would assign both local minima 50%. Namely, if each of the states is chosen with equal probability as the initial state, then there is 33% chance that the search begins in state A; then, as it is a local minimum, the random walker is stuck there. The same applies to state C. The remaining one third goes to the case when B is chosen as the initial state. Then, the two neighbours of B are chosen with equal probability: whichever is chosen, the random walker moves there, and get stuck there. Summing up, the probability of the search algorithm to terminate in either state A or state C is $1/3 + 1/3 \cdot 1/2 = 1/2$.

Suppose now that we perform *simulated annealing*. Then, the random walker can climb back from A or C to B. After having climbed to B, the random walker falls back to A or C with equal probability. It is, however, harder to climb to B from C than from A at a given temperature, because ΔE is higher.

Roughly speaking, the random walker will be confined to C, while it still can escape from A. In order to end up in A, it has to choose A always when in B, otherwise it gets locked into C. If the number of steps (number of iterations) is $2n$, then the random walker can be n times in state B: the chance of choosing always A is 0.5^n , which decreases quickly as n grows. With higher n , terminating in C has a probability $1 - 0.5^n$ very close to 1. The more iterations, the higher the chance to end up in C.

Although this train of thought has been only a rough illustration, the general morale still holds: *a slower cooling schedule, that is, increasing the number of the iterations, improves the precision of simulated annealing* (e.g. Reeves (1995), van Laarhoven (1987)).⁴ This observation will become very important in the argumentation in favour of Simulated Annealing Optimality Theory.

Obviously, nothing guarantees still that we have found the global minimum. Imagine, indeed, that the global minimum is situated at the bottom of a narrow valley: in such a case, the simulation will most often find a broader, but less deep basin, and the returned state will be the minimum of that one. One can, for example, run a few simulations in parallel, and choose the best of the results of these simulations.

Convergence results exist which prove the chance of finding the global optimum asymptotically converging towards 100% under certain circumstances.

⁴Interestingly, the exact manner of decreasing temperature influences the precision less than the rate at which temperature drops (Reeves, 1995).

Yet, these results imply solution times that are exponential in problem size, and often require more iterations than exhaustive search, so they are of limited use (Reeves, 1995).

2.1.3 Spin glasses in the brain

At the end of subsection 2.1.1, we have brought several arguments in favour of using heuristic optimisation techniques in linguistics in general, and for Optimality Theory, in particular. Now, we elaborate on some of them: how can we turn a seeming weakness—the lack of precision—into an advantage. Moreover, how to use it to contribute to a long-standing debate in linguistics, namely, to the issue of competence *vs.* performance, and categorical *vs.* gradient grammaticality.

Kirkpatrick et al. (1983), introduced simulated annealing as an analogy between physical systems and optimisation problems. In particular, they based their hope for the usefulness of simulated annealing upon the success of the Metropolis algorithm (Metropolis et al., 1953) in modelling *spin glasses* in statistical physics (Manrubia et al., 2004, cf. e.g.). Spin glasses are “highly frustrated systems”: magnetic interactions favouring different and incompatible kinds of ordering might be simultaneously present. Determining the optimal state of such a system is far a less trivial optimisation task than finding the optimal structure of a traditional magnetic lattice, and yet the Metropolis algorithm turned out to be useful. Furthermore, Kirkpatrick et al. (1983) write: “*The physical properties of spin glasses at low temperatures provide a possible guide for understanding the possibilities of optimizing complex systems subject to conflicting (frustrating) constraints*” (p. 673). But “conflicting constraints” is exactly the magic key word in Optimality Theory, as well!

A non-negligible difference between spin glasses and OT is, nonetheless, that “*systems like spin glasses have many nearly degenerate random ground states rather than a single ground state with a high degree of symmetry*” (*ibid.*). In other words, while most materials and conventional magnets have only one globally minimal configuration (ground state), which is realised if the particles form a highly symmetric crystal structure; spin glasses, on the other hand, can reach many different local minima, and these minima represent amorphous—glass-like—structures. Reaching a different minimum from some ground state requires considerable rearrangement, hence local minima are quite stable. Last, and most importantly, none of these local optima is significantly worse than the global optimum, thus “*it is not very fruitful to search for the absolute optimum*” (Kirkpatrick et al. (1983), p. 674). As an analogy, simulated annealing is claimed to be the most promising for problems where reaching a near-optimal solution is also satisfactory, such as the physical design of computers (Kirkpatrick et al., 1983) or the travelling salesman problem (Kirkpatrick et al., 1983; van Laarhoven, 1987; Spall, 2003).

In standard Optimality Theory, however, we definitely search for the *global* optimum: the candidate that *globally* optimises Harmony—that is, minimises the violation marks—with respect to the conflicting constraints. In turn, we either hope for a much simpler (less frustrated) search space with an easily reachable global optimum; or we accept the alternative “ground states” as linguistically meaningful solutions. Indeed, we shall see situations for both cases. In section 5, the search space will have a medium complexity: although the

Level	its product	its model	the product in the model
Competence in narrow sense: static knowledge of the language	grammatical form	standard OT grammar	globally optimal candidate
Dynamic language production process	acceptable or attested forms	SA-OT algorithm	local optima
Performance in its outmost sense	acoustic signal, etc.	(phonetics, pragmatics)	??

Table 2.1: The proposed three-level model of the human language

global optimum is easy to find with a slow cooling schedule, yet a fast cooling schedule may return other local optima, which are interpreted as fast speech forms. In section 6.1, however, some local optima will be interpreted as well attested (and acceptable) agrammatical forms. What is meant here?

Remember the Chomskyan distinction between *acceptability* and *grammaticality* (section 1.4): grammaticality refers to the linguistic competence or to its model, and not to the performance. As far as performance is concerned, a native speaker usually distinguishes between several levels of acceptability, even if some of the forms in the grey area are clearly grammatical or clearly agrammatical. Similarly, a corpus study will reveal that agrammatical forms are well attested, whereas grammatical ones may be very rare. This is why I urged in section 1.4 a model that makes the difference between the underlying static knowledge of the language (competence in its narrow sense), the dynamic production process (which may produce—or accept—forms that are agrammatical for the competence in its narrow sense) and the clearly extra-linguistic factors.

This is where the numerous near-optimal ground states of the spin glasses can be used as an analogy. I propose to see *Optimality Theory*, the model that formulates the optimisation problem, as the model of competence in its narrow sense. A form is *grammatical* if and only if it is the optimal solution to the problem posed by the OT grammar—that is, if it is *globally* optimal. On the other hand, further local optima also appear in the model, which will serve as pitfalls to *Simulated Annealing for Optimality Theory*, the model of the dynamic language production process. Consequently, the linguistic interpretation of the non-optimal ground states is that they model linguistic forms that are *agrammatical but acceptable* according to the judgement of the native speaker; alternatively, they are the forms that are *agrammatical but attested* in a corpus study (Table 2.1).

This approach—I believe—has two advantages. First, it may save a never-ending discussion on where the exact border between competence and performance is to be drawn. Second, it may help keep the competence model simple, as to be demonstrated by section 6: only the general rule has to be accounted for by the grammar (the model of the competence in the narrow sense), and (some) exceptions can be explained as being introduced by the language production process. Remember the history of mechanics: the decomposition of the problems into friction and motion following Newton’s laws helped grasping the sublunar motions, as well.

Now the question is how to implement Simulated Annealing to Optimality

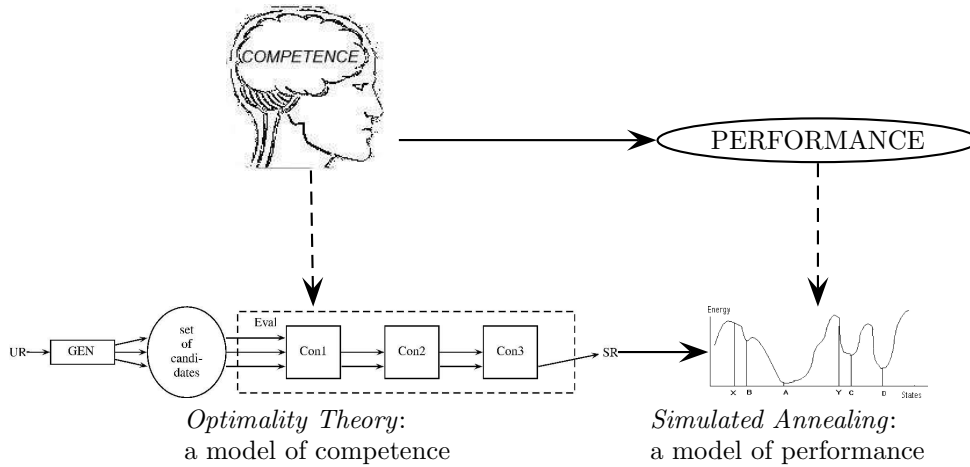


Figure 2.4: A simplified picture of the proposed relationship of OT and SA-OT.

Theory exactly. The search space is the set of candidates, but a real-valued energy function to be minimised cannot be defined in most cases (Prince and Smolensky (2004); see Prince (2002) for models where it can). Therefore, we have to find out how to implement simulated annealing to a non-real valued function. Section 2.2 introduces the general idea first. (Later, Chapter 3 will present a much more formal way to the same algorithm, by asking the question: if the function to be optimised is not real-valued, what is it then?) The algorithm of simulated annealing presented in Fig. 2.2 will have to be adjusted to the answer. After having constructed the SA-OT algorithm, section 2.3 contains a discussion of cases where SA-OT works and where it does not work as we might expect in anticipation.

2.2 Simulated Annealing for Optimality Theory

2.2.1 How to combine simulated annealing with OT?

Let us summarise where we are so far. Generating the winner candidate in Optimality Theory, as even suggested by its name, is a combinatorial optimisation problem. The goal is *to find the optimal candidate* in the candidate set, with respect to some *Harmony function* defined by the constraint hierarchy. It can be an NP-hard problem (Eisner, 1997, 2000b; Wareham, 1999), which motivates the use of heuristic algorithms in general, and the use of simulated annealing in particular.

The general idea of *Simulated Annealing for Optimality Theory* (SA-OT) is that the random walker roves around in the search space, which is the candidate set. The possible states of the system are thus the candidates. The function to be optimised is the Harmony function. Yet, the algorithm presented in Fig. 2.2 requires a few more non-trivial details so that we may implement it. What is $Neighbour(w)$? How to calculate the difference $E(w') - E(w)$? What should the cooling schedule (t_{max} , the stopping condition or t_{min} , and the function

$\alpha(t)$ be in SA-OT?

The procedure of applying simulated annealing to Optimality Theory will be decomposed into the following five steps:

- Step 1: Define the candidate set.
- Step 2: Define a neighbourhood structure (topology) on the candidate set.
- Step 3: Define the Harmony function to be optimised: what are the constraints and how are they ranked?
- Step 4: Define temperature and the transition probabilities.
- Step 5: Define the cooling schedule and perform the simulation.

Steps 1 and 3 are familiar to anybody who has worked with Optimality Theory. Yet, the formal definition of the candidate set and of the constraints may require some extra work. In most of the linguistic literature, Gen is vaguely seen as a “black box” that produces “everything”. For computational purposes, nonetheless, Gen (that is, the candidate set) has to be defined in an exact way: for instance, by explicitly listing the elements of the set, by specifying clear conditions on membership,⁵ or by using finite-state automata or context-free grammars. Moreover, constraints are usually defined by specifying the conditions that must be met by a candidate to *satisfy* the constraint. Yet, OT constraints are violable, and often the degree of violation plays a crucial role. Therefore, computational implementations require the formulation of constraints as functions that specify clearly how many violation marks are assigned to each candidate. For instance, a frequently employed, though only vaguely defined constraint—which we will reformulate for the sake of SA-OT—is Output-Output Correspondence: while the OT phonological literature is pleased by qualitatively demonstrating that a certain candidate is worse for OOC than the optimal one, SA-OT will require the exact difference of the marks assigned to any pair of candidates.

The hierarchy of the constraints should be settled by linguists working within the traditional OT paradigm, because the meaning of the hierarchy remains unchanged: to return the candidate that is supposed to be the grammatical one. The model underlying SA-OT is a traditional Optimality Theoretic grammar. It remains the task of the linguist to collect the data (that is, to find out which forms are grammatical), to make cross-linguistic comparisons, and to argue for specific constraints and specific hierarchies.

In brief, Gen, the set of constraints and the ranking—handed over by the theoretical linguist to the computational linguist—should yield as the optimal candidate the grammatical form, which is observed empirically by the descriptive linguist. What remains to clarify from the above agenda is step 2, that is, defining the neighbourhood structure (the topology); as well as, steps 4 and 5, the actual implementation of simulated annealing.

⁵As it is the case in all fields of mathematics, such a definition of a set must have the following form: “all members of set S [another well-defined set] that satisfy these well-defined conditions”. The goal of the exact definition is to be able to handle them in an algorithm.

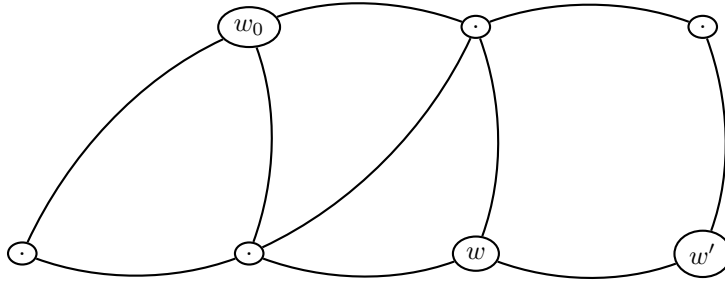


Figure 2.5: Schematic view of a search space—that is, the candidate set with a simple topology (neighbourhood structure)—in which simulated annealing realises a random walk. Candidate w_0 is the initial state of the random walk. Candidate w' is a neighbour of w , as they are connected by an edge.

2.2.2 Topology on the search space

We shall come back to step 4 in subsection 2.2.3. Our goal now is to work out the *neighbourhood structure* (the *topology*) of the candidate set (step 2), so that performing the simulation with different cooling schedules (step 5)—or with different parameter settings, in general—return interesting results.

The word *topology* refers to “the mathematical study of the properties that are preserved through deformations, twisting, and stretching of objects. Tearing, however, is not allowed.”⁶ The classical such property is *neighbourhood*: neighbouring points remain neighbours during twisting and stretching (but not during tearing), even though their distance may change. Thus, distance in absolute terms does not interest topology. A forerunner of topology was *graph theory*, which is exclusively interested in the connection between the vertices, but not in their spacial positions: moving the nodes of a graph does not alter it, as long as the edges are kept the same. If one prefers, one may employ the term *geometry of the candidate set*, as well.

When speaking about the *topology* or the *neighbourhood structure* of the candidate set (the search space), we have to imagine a graph-like structure in a first approximation (Fig. 2.5). Although the candidate set can include an infinite number of elements, its structure may be visualised as a set of points with the neighbouring points being connected by edges. In addition, the edges on this graph-like picture may be directed and labelled in order to represent the probabilities of picking a given neighbour.

In section 2.1.2, we have already seen what the topology on a search space consists of. For each state (now, candidate) w , we define the set $Neighbours(w)$ of its neighbours. A directed edge can be drawn from vertex w to vertex w' if and only if $w' \in Neighbours(w)$. Supposing that the neighbourhood relation is symmetric, one does not even need a directed graph. Importantly, this “graph” must be connected: a path of a finite length should connect any two vertices.

Furthermore, we have also seen that we also require a probability distribution on the set $Neighbours(w)$: the *a priori* probabilities $P_{choice}(w'|w)$ determine the choice of a particular neighbour in the first line of the core of the loop in the

⁶Eric W. Weisstein. “Topology.” From MathWorld—A Wolfram Web Resource. <http://mathworld.wolfram.com/Topology.html>.

simulated annealing algorithm (Fig. 2.2). These probabilities can be written on the directed⁷ edges of the graph—or graph-like structure, if the number of candidates is infinite. Due to equation (2.1), the sum of the weights leaving each vertex is 1.

Notice that the topology of the candidate space is the *horizontal structure* of the landscape in which the random walk takes place, and is independent of the landscape's *vertical structure*, defined by the Harmony function. That is to say that the same candidate set with a different hierarchy will yield the same $P_{choice}(w'|w)$ *a priori* probabilities, and different $P(w \rightarrow w'|T)$ transition probabilities. The former is defined by the candidate set, whereas the latter depends on the violation profile (that is, the “altitude”) of the two candidates.

It should be emphasised that adding a structure to the candidate set is new within Optimality Theory as seen by (almost) all linguists.⁸ I postulate that this structure is universal (innate), the same way as the set of possible underlying representations (*Richness of the Base*, Prince and Smolensky (2004) p. 225) and the Gen module—thus, the set of candidates—are claimed to be universal (innate). I suppose that the topology of the search space is a result of the way the candidates are represented, and neighbours differ only in a minimal component of their representations.⁹

Although the structure of the candidate set is assumed to be universally defined (if you wish, innate), it is still unknown to us. Research should discover it. That is to say, models have to be created to describe empirical data. Well, one may object that enriching the OT model by adding new components (such as the topology) to it will lead to *ad hoc* models that are less convincing. Indeed, there is such a danger: Ockham's razor advises simplifying scientific models, and not adding new concepts to them.¹⁰ Nonetheless, topology is not a superfluous addition, and two factors keep a tight rein on it. First, it should be convincing and cannot be *ad hoc*: a general principle has to define what simple basic operations (possible basic steps in the candidate set) transform one candidate into its neighbour. Second, the enriched model is required to account for an enriched set of data: not only for what is grammatical, but also for what the alternations or performance errors are, and under what circumstances they appear in speech.

⁷If the neighbourhood relation is symmetric, the graph is not directed in the sense that w_1 is connected to w_2 if and only if w_2 is connected to w_1 . However, $P_{choice}(w_2|w_1) = P_{choice}(w_1|w_2)$ does not necessarily hold even in this case.

⁸Paul Smolensky mentioned in a talk (October 2004, in Amsterdam) that variation forms are *local optima*—exactly what the present research line is about. Therefore, he must also suppose some topology on the candidate set, otherwise the expression *local optima* would be meaningless. Furthermore, the candidate set has been supposed to have the structure of a regular grammar in finite state approaches to OT since Ellison (1994).

⁹This proposal bears clear similarity with Paul Smolensky's connectionist approach to Optimality Theory. Yet, here we are free to embrace also representations different from those advanced by connectionism.

¹⁰*Ockham* was a nominalist in the medieval dispute between nominalists and realists. Realists (today, we would call them idealists) supposed that the Platonian ideals existed ontologically, whereas nominalists (e.g. Abelard) claimed that concepts are created only by the human intellect. According to Ockham, “plurality is not to be posited without necessity”, and referred to the razors used in those days to remove unnecessary ink from pergaments. Thus, nowadays, we should rather refer to *Ockham's eraser* or to *Ockham's delete button*, as proposed by Gábor Balázs—to whom I thank for this explanation. See also: Spade, V.S., *Ockham's nominalist Metaphysics*, In: Spade, V.S. (ed.), *The Cambridge Companion to Ockham*, 1999. pp. 101-102.

Simulated annealing makes a mistake when it gets stuck in a *local optimum*—supposing that the simulation has enough time to relax in its final phase. (Otherwise, if the simulation is terminated without waiting for it to arrive into some local optimum, basically any form may be returned.) Consequently, the topology should be defined so that the observed alternation forms be local optima: even if they are not globally optimal, they must be better than all their neighbours. The horizontal component (neighbourhood structure) and the vertical structure (Harmony function) of the landscape should jointly cause the simulation to sometimes fall into these traps.

The art of *Simulated Annealing Optimality Theory*, thus, consists of creating a landscape—with a simple and convincing definition of the neighbourhood structure and with arguably universal constraints—where the global optimum is the grammatical form and the other local optima are the observed alternations.

Later sections will introduce models demonstrating that even having such a landscape is not always sufficient. For instance, some local optimal traps are always avoided, or are never avoided. Sometimes, the logic of the representations forces us to also include other local optima: in such cases, we can only hope that they are avoided by the random walker, while the local optima corresponding to observed alternation forms are those where the walker is sometimes caught. By the end of the present thesis, the reader should be convinced that adding a topology to the candidate set does increase the explanatory power of Optimality Theory, because it may account for observations in a non-trivial way.

Now, let us elaborate on the concept of a topology on the candidate set. The goal of the neighbourhood structure, again, is to define how a next candidate w' is chosen as a *possible* next state of the random walker, when the walker is in candidate w . Obviously, after having chosen w' , it is still not sure yet that the walker really moves there: this depends on the transition probability $P(w \rightarrow w'|T)$, to be defined in the next section following equation (2.2).

As a matter of fact, we should distinguish between three steps when introducing the topology of a candidate set:

1. Define the set of candidates
2. Define the set of the neighbours $Neighbours(w)$ of each candidate w .
3. Define the *a priori* probability distribution $P_{choice}(w'|w)$ on $Neighbours(w)$.

What we really need is the probability distribution $P_{choice}(w'|w)$, yet the first two steps will lead to it. As it reflects some sort of probabilistic connection between states, one can compare this probability distribution to a Markov-model with two major differences: the number of states may be infinite, and choosing w' does not mean immediately moving there. Rather the product $P_{choice}(w'|w) \cdot P(w \rightarrow w'|T)$ is the probability of really moving from w to w' ; but then, this probability changes during the simulation as T changes, hence, we cannot speak of a Markov chain, either.

All introductions to simulated annealing emphasise the importance of an adequate neighbourhood structure in order to obtain an efficient optimisation algorithm. Too few neighbours, on the one hand, may result in too many local optima, and finding the global optimum becomes improbable. Too many neighbours, on the other, turns the algorithm practically into a random or exhaustive search, and does not deploy the structure of the search space.

A few strategies can be followed, and we turn now to comparing them. What has been suggested until now is that you should best define some *basic operations* which transform a candidate to a very similar other candidate. These operations typically alter only the candidate string at one single point, for instance by inserting, deleting or rewriting one atomic element of the string, or by flipping the value of a single feature. The number of *basic operations* should be minimal and the definition of such a *basic operation* (*basic step*, *basic transformation*) has to fit the way candidates are represented in the grammar (autosegmental phonology, context-free trees, AVMs, and so forth). Such *basic steps* would then lead to only a very restricted number of neighbours:¹¹

$$Neighbours(w) = \left\{ w' \in GEN(GEN^{-1}(w)) \mid w' = some_basic_transfo(w) \right\} \quad (2.4)$$

Nevertheless, even if the strategy allows only one basic operation per step, more options are available with respect to probabilities. One could assign *ad hoc* probabilities to the neighbours, but two further alternatives are more sound.

Either each of the neighbours has equal chance to be picked out—this proposal sounds reasonable if every candidate only has a few neighbours. We shall apply this approach in our treatment of Dutch stress in fast speech in Chapter 5, as well as in our example about Dutch voice assimilation in Chapter 6.1. If $\#$ denotes the cardinality of a set,

$$P_{choice}(w'|w) = \begin{cases} \frac{1}{\#Neighbours(w)} & \text{if } w' \in Neighbours(w) \\ 0 & \text{else} \end{cases} \quad (2.5)$$

Alternatively, performing each of the basic steps can be assigned some probability, and thus not all neighbours have equal chance. For instance, we first decide whether to insert, to delete or to rewrite an atomic segment, and then decide where in the string to perform this action. Even if each locus in the candidate string has equal chance, the probabilities p_{insert} , p_{delete} and $p_{rewrite}$ may vary. An example for this approach is found in our discussion of syllabification in section 7.

So far, $Neighbours(w)$ has been a relatively small subset of the candidate set. As already mentioned, the emerging “graph” must be a connected structure: each candidate should be reachable from any other candidate within a finite number of steps. In other words, for all w and w' , there must exist an integer n and a series of candidates w_0, w_1, \dots, w_n such that $w_0 = w$, $w_n = w'$, and for each $j < n$: $w_{j+1} \in Neighbours(w_j)$.

¹¹The implementation of Optimality Theory by Turkel (1994) uses a genetic algorithm, and proposes to see OT Gen as the generator of a new GA generation. He writes (p. 8):

[t]he standard assumption about the generator is that it takes a single representation and returns a set of representations consisting of modifications to the input. I will assume that the generator takes a set of representations and returns a set of representations. If the input set contains one element, then the generator returns a number of variations on that element (this is the standard operation). If the input set is empty, then the generator randomly creates a set of appropriate representations and returns that. ...

His standard operations (modifications, such as mutations, recombinations and crossovers) correspond to our *basic operations*: by applying them repeatedly, we can explore the search space. These operations, as proposed by several readers, could also be underpinned psycholinguistically. In any case, future work has to work out some general principles.

A radically different approach is to define the set of neighbours as the whole candidate set:

$$\text{Neighbours}(w) = \text{GEN}(\text{GEN}^{-1}(w)) \quad (2.6)$$

and to focus rather on the probabilities. In this case, some sort of similarity measurement may serve as $P_{\text{choice}}(w'|w)$. The more similar w and w' , the higher the probability $P_{\text{choice}}(w'|w)$. The similarity measurement must be normalised, so that $\sum_{w' \in \text{GEN}(\text{UR})} P_{\text{choice}}(w'|w) = 1$ hold (cf. equation (2.1)).

An approach based on (2.6) allows huge steps, which may be both useful and harmful. Clearly, if each candidate is equally reachable from a certain candidate, then simulated annealing turns into a very clumsy and ineffective way of trying out all possibilities: many candidates will be tried out repeatedly, whereas many other candidates will be ignored (let alone what happens in an infinite search space). Consequently, the $P_{\text{choice}}(w'|w)$ probabilities should make use of the properties of the search space in order to direct the search in a clever way. Indeed, a similarity-like $P_{\text{choice}}(w'|w)$ sounds promising, although further experimentations will be required.

In fact, a topology where every two candidates are neighbours, has no local optima, except for the global optimum. But notice that for all $\epsilon > 0$, the set $\{w' | P_{\text{choice}}(w'|w) > \epsilon\}$ defines a finite ϵ -neighbourhood around w .¹² This observation becomes important when applying a neighbourhood (2.6) to an infinite (or very large) set. Namely, if $1/\epsilon$ is in the magnitude of, or greater than the number of iterations performed, then the candidates beyond the ϵ -neighbourhood of w are practically unreachable from w within one step. In brief, such an infinite neighbourhood can be elegant, and still not significantly different from the finite neighbourhood structure defined by equation (2.4).

What do I mean by an elegant model? Sometimes, allowing one single basic operation per step, as in equation (2.4), may lead to problems. The example on syllabification in Chapter 7 will show that allowing exclusively the simplest basic steps may not prove very useful: the landscape will include too many local optima. There, one also has to be able to delete fully overparsed syllables—and not only single segments—in order to build a workable model. In similar cases, it is more fruitful to include *ad hoc* larger steps, as well. In turn, a well-designed model in which $P_{\text{choice}}(w'|w) > 0$ for any pair of candidates might prove to be both more elegant and useful.

Earlier in this section, I postulated the topology of the candidate set to be universal (alternatively, innate). Now, I should add that although the principles of the topology are claimed to be universal, yet some parameters may be language-dependent or speaker-dependent. Take the approach in which different basic transformations might have different probabilities: for instance, p_{insert} , p_{delete} and p_{rewrite} , which are not dependent on the argument of the operation. These three parameters are not independent of each other, $p_{\text{insert}} + p_{\text{delete}} + p_{\text{rewrite}} = 1$ should hold, because exactly one of them can be applied in one step. Now, the fact that three basic operations with some probabilities as parameters define the topology is claimed to be universal; nonetheless, the exact value of these parameters may vary across speakers or across languages.

¹²As $\sum_{w' \in \text{GEN}(\text{UR})} P_{\text{choice}}(w'|w) = 1$ must hold, fewer than $1/\epsilon$ candidates may be within such an ϵ -neighbourhood.

In this section, we have elaborated on the concept of a neighbourhood structure on the candidate set. The concept is new for main-stream Optimality Theory, even if not for connectionist approaches to OT. We have proposed several possibilities to define this topology, but only concrete applications will prove the usefulness of any of them. In the following section, we tackle the problem of how to define *temperature* for Simulated Annealing Optimality Theory.

2.2.3 Temperature for OT

As explained on page 45, creating an SA OT model consists of the following steps:

- Step 1: Define the candidate set.
- Step 2: Define a neighbourhood structure (topology) on the candidate set.
- Step 3: Define the Harmony function to be optimised: what are the constraints and how are they ranked?
- Step 4: Define temperature and the transition probabilities.
- Step 5: Define the cooling schedule and perform the simulation.

Out of these five steps, step 1 and step 3 depend on the traditional Optimality Theoretic system that serves as the underlying model. Subsection 2.2.2 has elaborated on step 2. Our present task is to work out step 4, so that we can experiment with different models (a candidate set with a neighbourhood structure, as well as a set of constraints with a hierarchy) and different cooling schedules.

The goal of defining a temperature is to define the transition probabilities $P(w \rightarrow w'|T)$ of moving in a counter-optimal direction. Remember that stepping to a more optimal candidate should always be possible: once a certain neighbour w' of the present position w of the random walker has been chosen, the random walker moves there if w' is better than w .

If $w' \in \text{Neighbours}(w)$ and $w' \succ w$, then $P(w \rightarrow w'|T) = 1$ for all T .

Here, and from now on, the relation $a \succ b$ denotes that candidate a is *better than* candidate b with respect to the current constraint hierarchy. A more formal definition follows in section 3.1.

In brief, the $w' \succ w$ case is simple. Yet, if $w' \prec w$, the probability of moving to w' should depend on the loss of harmony that this move would involve: the steeper the step, the less probable it is. Moreover, the probability of a particular step is gradually diminished from 1 to 0 during the simulation. The parameter called “temperature” is introduced to simulated annealing exactly in order to control the way how each of the probabilities $P(w \rightarrow w')$ are gradually reduced from 1 to 0.

Let us recapitulate the idea of simulated annealing. The notion of temperature is taken from thermodynamics and statistical physics. A physical system at a temperature above absolute zero has some inner random “vibration”, and this thermic energy “flows” between the particles of the systems as they interact with each other. Consequently, the energy of a given particle may randomly increase and decrease. Emitting an energy quantum is always possible (hence,

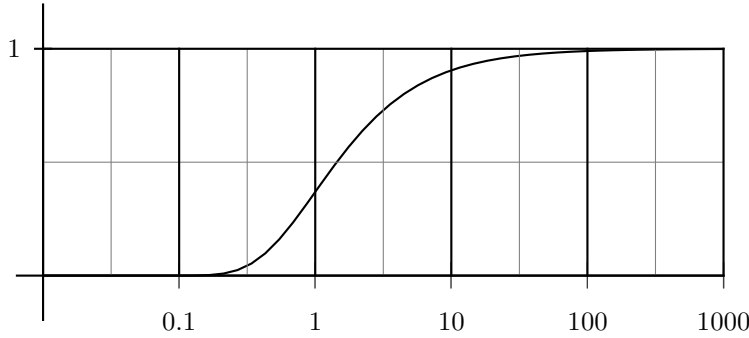


Figure 2.6: The $e^{-1/x}$ function with a logarithmic x -axis. Observe that the function is very close to 0 if $x < 0.1$, and very close to 1, if $x > 100$.

moving towards the better state has a probability of 1), whereas the chance of collecting, or borrowing, an energy package of ΔE depends on the temperature T of the system:

$$e^{-\frac{\Delta E}{k \cdot T}} \quad (2.7)$$

where k is *Boltzmann's constant* ($k = 1.3807 \times 10^{-23} J \cdot K^{-1}$; J stands for *joules* and K for *kelvin*), whose role is merely to connect energy to temperature, so that the exponent has no unit of measurement.

Let us have a closer look at the expression (2.7), as well as at Fig. 2.6. The meaning of temperature T is to define the *range* of energy transitions that have an intermediate probability. Large jumps in energy ($\Delta E \gg kT$) have a vanishingly small probability, very close to zero; whereas small changes in energy ($\Delta E \ll kT$) are almost certain to happen. For $\Delta E = kT$, however, the probability is $1/e \approx 0.37$. These observations are summarised in Table 2.2.

Therefore, defining temperature in simulated annealing means defining the transitions that we want to assign this medium probability to in a certain stage of the simulation. Thus, temperature has to belong to the same “type” as the changes in the function to be optimised. In physical terms, the exponent has to be dimensionless, that is, ΔE and $k \cdot T$ in (2.7) must have the same units of measurement (for example *Joule* or *kcal*).

This is also the reason for the introduction of Boltzmann's constant in physics. For historical and human reasons, the temperature scale has been defined independently of energy: we perceive temperature as a *qualia* in its own right, even though physics has reduced this concept to the concept of energy. In fact, why not call rather $k \cdot T$ the temperature? This is the way simulated annealing proceeds. In simulated annealing, $k = 1$ simply, so energy and temperature have the same dimension.

Our goal is to implement equation (2.2) for Optimality Theory in a form such as:

$$P(w \rightarrow w' | T) = \begin{cases} 1 & \text{if } w' \succ w \\ e^{-\frac{H(w') - H(w)}{T}} & \text{if } w' \prec w \end{cases} \quad (2.8)$$

$\Delta E < 0$	$P(w \rightarrow w' T) = 1$
$0 < \Delta E \ll T$	$P(w \rightarrow w' T) \approx 1$
$\Delta E \approx T$	medium $P(w \rightarrow w' T)$
$\Delta E = T$	$P(w \rightarrow w' T) = 1/e \approx 0.368$
$\Delta E \gg T$	$P(w \rightarrow w' T) \approx 0$

Table 2.2: The way the transition probability $P(w \rightarrow w'|T)$ is dependent on the temperature T and on the steepness $\Delta E = E(w') - E(w)$ of the transition $w \rightarrow w'$. By decreasing the positive control parameter T during the simulation, the same transition turns gradually from being highly probable into being highly improbable.

Consequently, we need to know the dimension of the Harmony function, in order to introduce temperature T to SA-OT. So, what is the *dimension* of the Harmony function?

The situation is even worse: not only does the Harmony function lack a proper dimension, but normally it is even not construed as a real-valued function! How can we divide the Harmony function with anything, and then compute its exponential?

Luckily enough, however, we do not need the dimension of the Harmony function proper. What we really need in equation (2.8) is the *difference* $H(w') - H(w)$ of two violation profiles. This seemingly additional step—subtraction—will help us. For this reason, our action plan to define $P(w \rightarrow w'|T)$ for the case w' is worse than w will be the following:

- Firstly, we represent the violation profiles in an appropriate way.
- Secondly, we define the *difference* of two violation profiles.
- Thirdly, we define *temperature* in a similar format.
- Fourthly, we define the exponential of their quotient.
- Lastly, we introduce the SA-OT algorithm.

We shall perform this action plan several times. In this section, we try to build up an intuitive idea. Therefore, we shall represent violation profiles as vectors, whereas the difference of two violation profiles will be a pair of numbers. A violation profile as a vector is but a shorthand for a row in a traditional tableau, most probably familiar to the reader. Subsequently, the following section presents two alternative mathematical models. Yet, all three approaches will lead to the same SA-OT algorithm.

Difference of violation profiles

We follow the action plan just presented, and begin with the representation of a violation profile. In traditional terms, a violation profile is a set of tokens of violation marks. For instance, candidate w_1 incurs two marks from constraint C4, one mark from C2, five marks from C1, and one mark from constraint C0.

Such a set can be simply visualised by a tableau, many of which we already saw in Chapter 1.

If candidate w_1 has incurred five violation marks from constraint C1, then we simply write $C_1(w_1) = 5$: candidate w_1 has a *violation level* of 5 on constraint C1. In short, constraints are functions mapping from the set of candidates onto the set of non-negative integers.

Suppose that the constraints form the following hierarchy: $C_N \gg C_{N-1} \gg \dots \gg C_0$. Then, a row in a tableau is:

$$\begin{array}{|c|c|c|c|c|} \hline & C_N & C_{N-1} & \dots & C_0 \\ \hline w & C_N(w) & C_{N-1}(w) & \dots & C_0(w) \\ \hline \end{array} \quad (2.9)$$

Frequently, cells with value $C_i(w) = 0$ are left empty. Now, a violation profile can be seen as a vector formed by the levels of violation:

$$H(w) = (C_N(w), C_{N-1}(w), \dots, C_0(w)) \quad (2.10)$$

It is simply a shorter way to write a row of a tableau. We shall call the $H(w)$ thus introduced *the violation vector* corresponding to the *violation profile* of w . It will be also called the *vector representation* of the Harmony function.

Note that the indices *decrease* within the vector representation. The first component of the vector has the highest index, and it represents the violation level of the highest ranked constraint. Although this notation might look awkward at this point, it will become handy later on to assign higher indices to higher ranked constraints. Additionally, this notation clearly parallels an OT tableaux, in which the highest ranked constraints appear on the left.

Now, we proceed to the *difference* of two violation profiles. Let us take two violation profiles represented in the form of a traditional tableau:

$$\begin{array}{|c|c|c|c|c|c|} \hline & C4 & C3 & C2 & C1 & C0 \\ \hline w_1 & ** & & * & ***** & * \\ \hline w_2 & ** & & *** & & * \\ \hline \end{array} \quad (2.11)$$

The reader familiar with Optimality Theory will immediately observe that the crucial constraint that *differentiates* between the behaviour of the two candidates w_1 and w_2 is constraint C2. We shall call a constraint playing this important role the *fatal constraint*, the *critical constraint*, or, following Prince and Smolensky (2004), *the highest ranked constraint with uncanceled marks*.

Even though candidates w_1 and w_2 do not necessarily satisfy the two highest ranked constraints, C4 and C3, these constraints do not play any role in the comparison. If the question is whether to move from w_1 to its neighbour w_2 , the two violation marks incurred by both candidates with respect to constraint C4 just “elevate the baseline”: hiking from 500 m above sea-level to 800 m above sea-level is the same as hiking from 1500 m to 1800 m above sea-level. In brief, these highly ranked, but shared violations will not influence the difference in the Harmony value of the two candidates.

Furthermore, standard Optimality Theory teaches us not to look further in the hierarchy once we have found a difference between the violation profiles (for counter-examples and the notion of *cumulativity*, see subsection 1.3.5). Candidate w_2 is defeated by w_1 at constraint C2, and the many violation marks assigned to w_1 by C1 do not make any *difference*. Hence, the *difference* of these

two violation profiles will be “two violations of constraint C2”. That is, the difference of two violation profiles will have the form of a pair: a constraint followed by the difference of the violation levels of this constraint.

By generalising this concrete example, we now define the *difference of two violation profiles*, by following the theoretical foundations of OT. Prince and Smolensky (2004) introduce the concept of *mark cancellation* and prove the *Cancellation Lemma* (p. 258):

Cancellation Lemma. Suppose two structures [violation profiles—*T.B.*] S_1 and S_2 both incur the same [violation] mark $*m$.¹³ Then to determine whether $S_1 \succ S_2$, we can omit $*m$ from the list of marks of both S_1 and S_2 (‘cancel the common mark’) and compare S_1 and S_2 solely on the basis of the remaining marks. Applied iteratively, this means we can cancel *all* common marks and assess S_1 and S_2 by comparing only their unshared marks.

A consequence of this lemma is the *Cancellation/Domination Lemma* (Prince and Smolensky (2004) p. 261):

Cancellation/Domination Lemma. Suppose two parses [candidates] A and B do not incur identical sets of marks. Then $A \succ B$ iff every mark incurred by A which is not cancelled by a mark of B is dominated by an uncanceled mark of B .

It is this second lemma that teaches us that the crucial point in comparing two candidates is the highest constraint C_{fatal} where the two profiles differ. All violation marks assigned by constraints higher than C_{fatal} are cancelled, whereas lower violation marks are dominated by some violation of C_{fatal} . Consequently, constraints ranked lower than C_{fatal} can be ignored.

For the sake of simulated annealing, however, we require not only the *comparison* of two violation profiles, but also their *difference*. The general “philosophy” of Optimality Theory just presented motivates us to neglect what happens at constraints lower than the fatal constraint:

Definition 2.2.1. DIFFERENCE OF TWO VIOLATION PROFILES: *Suppose that for candidates A and B , constraint C_{fatal} is the fatal constraint, that is, the highest ranked constraint assigning either A or B an uncanceled mark following mark cancellation.*

Then, the difference of the violation profiles of candidates A and B is the pair $\langle C_{fatal}, C_{fatal}(A) - C_{fatal}(B) \rangle$, that is: “the difference¹⁴ $C_{fatal}(A) - C_{fatal}(B)$ at constraint C_{fatal} ”.

Furthermore, the difference of the violation profiles is defined to be zero $\langle 0, 0 \rangle$ if the two candidates incur exactly the same violation marks (i.e. there is no uncanceled mark).

For instance, in tableau (2.11), the difference of the violation profiles of w_1 and w_2 is the pair $\langle C2, -2 \rangle$ (“−2 violations of C2”).

¹³Here the mark $*m$ is meant to be a token of violating constraint m .

¹⁴In linguistic applications, the levels of violation are non-negative integers, thus their difference is always an integer. Generalising to real-valued violation levels is straightforward. Problems may arise, however, if the constraints map to a fully ordered set in which subtraction is not defined. Yet, we do not deal with this case.

Let us introduce this definition now in a slightly more formal way. The point-wise difference of two profiles seen as vectors (equation (2.10)) is still a *violation profile-like vector*:

$$H(w') - H(w) = (C_N(w') - C_N(w), \dots, C_0(w') - C_0(w)) \quad (2.12)$$

This difference, however, does not have yet a form that can be used in (2.8).

When one compares two candidates, what matters according to the *Cancellation/Domination Lemma* is the leftmost non-zero component in this difference vector. This is the component corresponding to the fatal constraint: the highest ranked constraint that assigns a different number of marks to the two candidates. In standard OT, only its *sign* (positive or negative) matters. The crucial step in the present proposal is to take its *value* (as opposed to only its sign) and its *place* in the vector, but no further information.

Consequently, two difference vectors are *equivalent* for the present purpose if their leftmost non-zero component is the same and in the same column:¹⁵

Definition 2.2.2. Two vectors, $\underline{a} = (a_N, \dots, a_0)$ and $\underline{b} = (b_N, \dots, b_0)$ are *equivalent*

$$\underline{a} \cong \underline{b}$$

if and only if there is a $k \in \{N, \dots, 0\}$ such that a_k is the leftmost non-zero component of \underline{a} , b_k is the leftmost non-zero component of \underline{b} , and $a_k = b_k$.

Additionally, $(0, 0, \dots, 0) \cong (0, 0, \dots, 0)$.

In other words, we require both vectors to begin with the same number of zeros, and their first non-zero element to be also equal. They may differ in their further components.

It can be easily shown that \cong is indeed an *equivalence relation*:¹⁶ it is reflexive ($\underline{a} \cong \underline{a}$), symmetric ($\underline{a} \cong \underline{b}$ implies $\underline{b} \cong \underline{a}$) and transitive (if $\underline{a} \cong \underline{b}$ and $\underline{b} \cong \underline{c}$ then $\underline{a} \cong \underline{c}$). Consequently, \cong defines *equivalence classes* on the set of the vectors: \underline{a} and \underline{b} belong to the same equivalence class iff $\underline{a} \cong \underline{b}$. The equivalence classes are disjoint (their intersection is empty) and cover the whole set of vectors. An equivalence class can be specified by the index of the left-most (that is, highest ranked) component, as well as by the value of this component.

What really interests us is not the difference as defined in Eq. (2.12), but the equivalence class to which this difference belongs. The philosophy of standard Optimality Theory, namely the *Cancellation/Domination Lemma*, does not differentiate between two difference vectors that belong to the same equivalence class with respect to equivalence relation \cong .

This is why we define the *magnitude* of a violation profile-like vector as the equivalence class to which the vector belongs:

Definition 2.2.3. The *magnitude* of a violation profile-like vector (a_N, \dots, a_0) is

$$\|(a_N, \dots, a_0)\| = \langle k, a_k \rangle$$

¹⁵This definition introduces the equivalence \cong of two vectors in general. This relation will be used on vectors representing the *difference* of two violation profiles. It is not to be confused with the equivalence relation in Definition 3.1.6, according to which two violation profile-like vectors are equal ($H(w_1) = H(w_2)$)—the candidates are equivalent ($w_1 \simeq w_2$)—iff they incur the same number of violation marks by each of the constraints.

¹⁶Cf. e.g. Eric W. Weisstein. "Equivalence Relation." From MathWorld—A Wolfram Web Resource. <http://mathworld.wolfram.com/EquivalenceRelation.html>

	C_N	C_{N-1}	...	C_{k+1}	C_k	C_{k-1}	C_{k-2}	...
w'	2	0		1	2	3	0	
w	2	0		1	3	1	2	
$H(w') - H(w)$	0	0		0	-1	2	-2	
$\ H(w') - H(w)\ $	0	0		0	-1	-	-	

Table 2.3: **An example for the difference of two violation profiles:** given the profiles of w and w' , $|w' - w| = \|H(w') - H(w)\| = \langle C_k, -1 \rangle$, and C_k is the fatal constraint. The differences in the violation levels of the constraints ranked lower than C_k are ignored in the magnitude of a vector.

where k is the lowest (rightmost) element of $\{N, \dots, 0\}$ such that $\forall j > k$: if $j \leq N$ then $a_j = 0$

Thus, $\langle k, a_k \rangle$ is the name of the equivalence class to which all vectors belong whose first components (with indices $N, N-1, \dots, k+1$) are zero, and whose k th component is a_k . If $k = N$, there are no zero components on the left.

Then, we define the *difference of two violation profiles* as simply the magnitude of the difference of the violation profile vectors:

Definition 2.2.4.

$$|w' - w| = \|H(w') - H(w)\| = \langle k, C_k(w') - C_k(w) \rangle$$

Here, C_k is the fatal constraint, that is, the highest ranked constraint with uncanceled violation marks (the highest k such that $C_k(w') - C_k(w) \neq 0$):

$$\begin{aligned} \|H(w') - H(w)\| &= \|(0, 0, \dots, C_k(w') - C_k(w), \dots, C_0(w') - C_0(w))\| = \\ &= \langle k, C_k(w') - C_k(w) \rangle \end{aligned} \quad (2.13)$$

An example is given in Table 2.3.

Note that the difference of two violation profiles is a pair of *numbers*: the first number is the index of a constraint, whereas the second number is a difference in violation levels. So far, both have been integers, but later we shall be more flexible.

We have already introduced the *Law of trichotomy* in Chapter 1, and we shall return to it soon. It states that for any two violation profiles A and B , exactly one of the following statements holds: 1. A is better than B ($A \succ B$); 2. A is equivalent to B (same violation profile, $A \simeq B$); 3. and A is worse than B (that is, $A \prec B$). By using the *Cancellation/Domination Lemma*, it is easy to demonstrate that these three cases correspond to the second component in the difference of the violation profiles being negative, zero or positive, respectively. Consequently, it is well-founded to represent *violation profiles* by *violation vectors*. Section 3 will introduce two further ways of representing violation profiles, prove their well-foundedness, and derive ways of applying simulated annealing to Optimality Theory.

Even if we could not reduce the difference of two violation profiles, the expression $H(w') - H(w)$ appearing in pseudo-definition (2.8), to a real value, we have now a pair of numbers instead of an N -dimensional vector.

We shall soon need—for the introduction of temperature—the comparison of two such pairs. Which jump is greater: moving from w_1 to w_2 , or from w_3 to w_4 ? Increasing the violation level of a higher ranked constraint is a steeper step than increasing the number of marks assigned by lower ranked constraints only. If both of two steps increase the violation level of constraint C_k (without increasing the violation level of higher ranked ones, and we do not care about lower ranked ones), then the step that adds more violation marks is the steeper one, although the steepness of the two steps do not differ dramatically.

Consequently, we propose the following relations (not all of which exclude the other ones):

Definition 2.2.5. *Let K_1 and K_2 be real numbers, while t_1 and t_2 be positive real numbers.*

1. *Two pairs are equal:*
 $\langle K_1, t_1 \rangle = \langle K_2, t_2 \rangle$, iff $K_1 = K_2$ and $t_1 = t_2$.
2. *One of the pairs is greater:*
 $\langle K_1, t_1 \rangle > \langle K_2, t_2 \rangle$, iff either $K_1 > K_2$ or $K_1 = K_2$ and $t_1 > t_2$.
3. *One of the pairs is much greater:*
 $\langle K_1, t_1 \rangle \gg \langle K_2, t_2 \rangle$, iff $K_1 > K_2$.
4. *Two pairs are approximately equal:*
 $\langle K_1, t_1 \rangle \approx \langle K_2, t_2 \rangle$, iff $K_1 = K_2$.

After this slight detour, let us turn back to our agenda set up earlier. We have argued for a definition of the difference of two violation profiles. How does it help us in defining temperature?

Defining temperature and the transition probabilities

Recall the slight move of replacing the fatal constraint by its index in Definition 2.2.3: we shall use $\langle k, C_k(w') - C_k(w) \rangle$, and not $\langle C_k, C_k(w') - C_k(w) \rangle$. The goal of this slight change has been to help defining temperature.

As discussed earlier, the “temperature” in simulated annealing determines the range of change in energy (harmony, or some other function to be optimised), above which counter-optimal moves are prohibited, and under which counter-optimal moves are allowed. Therefore, temperature has to have the same “form” (dimension, structure) as changes in the function to be optimised. As a difference of two violation profiles is now a pair $\langle k, t \rangle$, we define temperature also as a pair of numbers:

$$T = \langle K_T, t \rangle \in \mathcal{R} \times \mathcal{R}^+ \quad (2.14)$$

If $K_T = i$ such that there is a constraint C_i , then the temperature T is said to be *in the domain of* constraint C_i . The temperature $T = \langle K_T, t \rangle$ is *above* (*below*) constraint C_i if $K_T > i$ ($K_T < i$). Using Definition 2.2.5, temperature is *in the domain of* constraint C_i iff $T \approx \langle i, 1 \rangle$; and temperature *above* the domain of constraint C_i iff $T \gg \langle i, 1 \rangle$. Temperature being *far above* the domain of C_i (if one wishes, $T \gg \gg \langle i, 1 \rangle$) will denote $K_T \gg i$, in an informal sense.

$t \leq 0$ is not allowed by definition, because, as we shall see, that would lead to mathematical problems in the simulation. The situation is similar to the

one in physics, where zero and negative absolute temperatures (Kelvin's scale) are also prohibited. On the other hand, K_T can be both positive and negative, and behaves like relative temperature scales in physics (Celsius, Fahrenheit, Réaumur). In practice, K_T will be an integer, although nothing prohibits having non-integer values for K_T . In addition, K_T will range between K_{max} , usually (far) above the highest ranked constraint, and K_{min} , always far below the lowest ranked constraint. Assigning index 0 to the lowest ranked constraint causes the system to freeze exactly when the first component of the temperature $K_T < 0$, as the continental reader might expect.¹⁷

Notice that if K_T is the index of some constraint—and this case will be the most interesting one—, then temperature $\langle K_T, t \rangle$ corresponds to some equivalence class on the set of possible violation profile vector differences. Namely, to the vector differences whose first (leftmost) components (with indices $N, N-1, \dots, K_T+1$) are zero, and whose K_T th component is t (by Definition 2.2.3). In other words, such a temperature is equivalent to a move which increases the number of violation marks assigned by constraint C_{K_T} by t , leaves the violation marks of higher ranked constraints unchanged, and might change the violation marks assigned by lower ranked constraints in any way.

However, K_T does not necessarily correspond to some constraint. This is why the structure of temperature is said to be a *generalisation* of the violation profile differences. And yet, we can also apply the comparison relations introduced by Definition 2.2.5 to such generalisations. Using this definition, a *decreasing series of temperature values* (a *cooling schedule*) can be specified. Moreover, it is possible to compare a change in the harmony function to the actual temperature, in order to define the transition probabilities.

Remember Table 2.2, based on equation (2.7): temperature in simulated annealing draws a smooth border line between allowed and prohibited counter-optimal transitions. The chance of increasing the energy function with T is $1/e$ at temperature T .

Combining Definition 2.2.5 with Table 2.2, we can easily formulate now the definition of the transition probability for the $w \succ w'$ case:

$$P(w \rightarrow w'|T) = \begin{cases} 1 & \text{if } \|H(w') - H(w)\| \ll T \\ 1/e & \text{if } \|H(w') - H(w)\| = T \\ 0 & \text{if } \|H(w') - H(w)\| \gg T \end{cases} \quad (2.15)$$

In words, if temperature is in a domain above that of the fatal constraint, then the loss of Harmony (the increase in violation marks) incurred by the move is negligible compared to the temperature, and the move is always taken. If, on the other hand, the temperature is very cold compared to the increase in violation marks, then the move is never taken.

According to Table 2.2, if the increase in the cost function is comparable to the temperature, then the probability of moving has a medium value. Specifically, if the two are equal, the move has a chance of $1/e$. By Definition 2.2.5, this translates to the case when the temperature is exactly in the domain of the fatal constraint. In this special case, equation (2.7) can be simply copied: augmenting the violation marks of the fatal constraint by d has a probability of $e^{-d/t}$, where t is the second component of the temperature in equation (2.14).

¹⁷Even if weird, nothing prohibits the Anglo-Saxon reader from assigning index 32 to the lowest ranked constraint.

In particular, if $\|H(w') - H(w)\| = T$, that is, $d = t$, the transition probability is indeed $e^{-1} = 1/e$.

In summary, the transition probabilities for the $w \succ w'$ case are:

$$P(w \rightarrow w'|T) = \begin{cases} 1 & \text{if } \|H(w') - H(w)\| \ll T \\ e^{-d/t} & \text{if } \|H(w') - H(w)\| \approx T \\ 0 & \text{if } \|H(w') - H(w)\| \gg T \end{cases} \quad (2.16)$$

where d is the second component of $\|H(w') - H(w)\|$, and t is the second component of T .

Equation (2.16) differs from the behaviour of the traditional transition probabilities summarised in Table 2.2 only in one respect. Namely, in traditional simulated annealing (and in physics), the $\Delta E \gg T$ and $\Delta E \ll T$ cases were informal notions, and consequently, the statements $P(w \rightarrow w'|T) \approx 0$ and $P(w \rightarrow w'|T) \approx 1$ were also to be taken informally. SA-OT, however, implements a non-real valued optimisation (due to the Strict Domination Hypothesis of OT), therefore the \gg relation could be introduced exactly in Definition 2.2.5 (again, due to the Strict Domination Hypothesis of OT). This is the reason why the probabilities in the $\Delta H \gg T$ and the $\Delta H \ll T$ cases are postulated to be exactly 0 and 1, respectively.

At this point, we are already getting very close to the introduction of the SA-OT algorithm. We have defined temperature and the transition probabilities. The very last step is to introduce the cooling schedule, that is, a series of decreasing temperature values. The way to do this is already implicit in Definition 2.2.5.

The cooling schedule is realised in standard simulated annealing as a single loop gradually decreasing the temperature (Fig. 2.2). Thereby, the probability of jumps increasing the cost function are gradually decreased from very close to 1 to very close to 0. At each moment, temperature defines the jump that has a probability of $1/e$.

What should the cooling schedule do in SA-OT? Initially, it should allow any transitions. Then, it should prohibit transitions increasing the violation level of highly ranked constraints. Then, prohibit also the transitions that would only increase the violation marks assigned by lower ranked constraints, etc. In the final phase of the simulation, no move to a worse state should be allowed.

According to equation (2.16), a temperature that would allow any move has a first component K_T that is higher than the index of the highest ranked constraint. A temperature that prohibits augmenting the violation level of the highest ranked constraint, but allows augmenting the violations of lower ranked constraints, has a K_T value that is lower than the index of the highest ranked constraint, but higher than the indices of the lower ranked constraints. Finally, if K_T is lower than the index of the lowest ranked constraint, no transition to a worse candidate is allowed. Consequently, we have to diminish K_T .

Furthermore, we also want to diminish t in a more fine-grained way. At a given point, C_{K_T} is the highest ranked constraint whose violation marks can increase. Still, equation (2.16) prefers increasing the number of these violation marks by $d = 1$ over by $d = 2$. If initially, $t = 5$, taking both steps are relatively easy. As t is decreased, the chances of both moves are smoothly turned off, although the $d = 2$ jump will be always less likely than the $d = 1$ one.

In sum, we add an embedded loop diminishing t into the loop diminishing K_T . Such a double loop will mimic traditional simulated annealing. At each time of the simulation, temperature shows what is the jump that has a probability of $1/e$. Steeper jumps are less probable, and much steeper jumps have a zero probability. Smaller jumps have a higher probability, and much smaller jumps have a probability of 1. If the value of temperature corresponds to some equivalence class on the set of profile differences, its meaning can be simply translated as “increase the violation marks assigned by constraint C_{K_T} by t , do not change the violation marks assigned by higher ranked constraints, and do what you want with the violation levels of lower ranked constraints”. Otherwise, the temperature cannot be translated into a change in the violation profile, but can be compared to such changes, by using Definition 2.2.5.

The following picture may help to visualise the idea better. Take a scale which is composed of intervals called *domains*, denoted by $K = 5$, $K = 4, \dots$, $K = 0$, $K = -1, \dots$ ¹⁸

		C_2		C_0	
\dots	$K = 3$	$K = 2$	$K = 1$	$K = 0$	\dots
\dots	... 2.5 2.0 1.5 1.0 0.5	... 2.5 2.0 1.5 1.0 0.5	... 2.5 2.0 1.5 1.0 0.5	... 2.5 2.0 1.5 1.0 0.5	\dots

Figure 2.7: Visualising the domains traversed by temperature

Each domain is an interval open on the left and closed on the right, so that the interval $(+\infty, 0]$ can be projected onto it. Temperature $T = \langle K, t \rangle$ is represented on this scale by the point that is the projection of t onto the domain K . Moving to the right on this scale means decreasing temperature: remember, $T_1 = \langle K_1, t_1 \rangle > T_2 = \langle K_2, t_2 \rangle$ iff either $K_1 > K_2$, or $K_1 = K_2$ and $t_1 > t_2$.

Furthermore, some of the domains correspond to constraints—in the present example, constraint C_2 corresponds to $K = 2$, and constraint C_0 corresponds to $K = 0$. The higher ranked a constraint, the higher the domain K it is associated with (in the present case, thus, $C_2 \gg C_0$). Notice that here we already enjoy the advantages of the initially surprising notation that associates higher indices with higher ranked constraints. Temperature may be assigned values that are in domains corresponding to some constraints, or values that are in domains above / between / below some constraints.¹⁹

Now, decreasing temperature can be realised by decreasing t within one domain, and then jumping to the next domain (or to another domain further below). In practice, embedded loops will be used: the outer cycle decreases the domain of the temperature (K descends from K_{max} to K_{min} , by steps K_{step}), and the inner loop decreases t within a given domain (from t_{max} to t_{min} , by steps t_{step}). Importantly, the borders of the domains are taboo: $t > 0$ always, similarly to absolute temperature in physics.

¹⁸This approach, similarly to the applications to follow, supposes that the first component of the temperature $T = \langle K, t \rangle$ may take only integer values.

¹⁹In what follows, we shall place constraints into the domains $K = 0$, $K = 1$, $K = 2$, etc. Nothing prohibits us, however, from leaving out some domains. Then, in some phase of the simulation, temperature may take values from the domain lying between the domains associated to two successive constraints.

2.2.4 Introducing the SA-OT algorithm

At this point we are able to formulate the way simulated annealing will be implemented for Optimality Theory. We will soon introduce the precise algorithm for *Simulated Annealing Optimality Theory* (Fig. 2.8).

We begin the random walk in the space of the candidates by choosing (randomly or semi-randomly) an initial candidate w_0 (recall Fig. 2.5). In the case of a finite candidate set, we shall use each candidate as a starting point with equal probability. In practice, we shall run the simulation many times from each of the candidates. In the case of an infinite candidate set, however, a more useful solution is to start the simulation with equal probability from the elements of a small finite subset: for instance, from the candidates without the recursive insertion that results in an infinite candidate set. A third option is to launch the simulation always from the candidate that is arguably the default one with respect to the underlying representation.²⁰

In the beginning, temperature $T_0 = \langle K_{max}, t_{max} \rangle$ is high, that is, K_{max} is higher than the domain of the highest ranked constraint. A lower initial temperature can also be chosen, but if K_{max} is higher then the index of the highest ranked constraint, the simulation may have an initial phase in which the transition probability $P(w \rightarrow w')$ is 1 for all w and w' . The advantage (or disadvantage, depending on what your goal is) of setting K_{max} high enough is to reduce the influence of the initial candidate's choice.

At each time step of the simulation, temperature is decreased, and one step may be performed in the search space. Traditional simulated annealing sometimes allows for more steps before reducing the temperature (cf. the parameter $nrep$ in Fig. 2.2). Instead of introducing an additional parameter, we rather set $nrep = 1$, and prefer to reduce the temperature in smaller steps. The difference between the two approaches should not be significant, but a proliferation of parameters might render our analysis more complex. Nonetheless, further research could analyse the role of this factor, as well.

At a given moment in the simulation, the random walker is located in the position represented by candidate w . Temperature then is $T = \langle K_T, t \rangle$. A neighbour w' of candidate w (cf. Fig. 2.5) is randomly picked. The choice is determined by the *topology* of the search space: the neighbourhood structure provides the set $Neighbours(w)$, from which w' is chosen using the *a priori* probability distribution $P_{choice}(w'|w)$ on $Neighbours(w)$. As discussed in section 2.2.2, the elements of $Neighbours(w)$ may have equal or different chance to be picked, and several strategies exist to define $Neighbours(w)$.

There follows a comparison of the violation profiles of w and w' . If w' is more harmonic than w ($w' \succ w$), or they are equally harmonic ($w' \simeq w$: they incur exactly the same violation marks), the random walker automatically moves to w' . Otherwise, moving to w' depends on the temperature. How? The likelihood is defined in equation (2.16).

To sum up, if $T = \langle K_T, t \rangle$ and $\|H(w') - H(w)\| = \langle k, d \rangle$, then the transition probability is:

²⁰Suppose for instance that the stress pattern of a compound word has to be predicted. Then the default candidate could be the one whose stress pattern is the concatenation of the stress patterns of the two parts of the compound. We shall return to this idea in section 5.7.

$$P(w \rightarrow w' | T) = \begin{cases} 1 & \text{if } d \leq 0 \\ 1 & \text{else if } k < K_T \\ e^{-d/t} & \text{else if } k = K_T \\ 0 & \text{else} \end{cases} \quad (2.17)$$

We can reformulate this equation in words:

RULES OF MOVING: Let the crucial constraint (the highest ranked constraint with uncanceled marks) when comparing w to w' be C_k , and temperature be $T = \langle K_T, t \rangle$. Then the following options are available:

- If w' is better than w ($w' \succ w$, that is, $C_k(w') < C_k(w)$), then move from w to w' .
- If w' loses due to the critical constraint $C_k > K_T$: don't move!
- If w' loses due to the critical constraint $C_k < K_T$: move!
- If w' loses due to the critical constraint $C_k = K_T$: move with probability $P(w \rightarrow w') = e^{-d/t}$, where $d = C_k(w') - C_k(w)$.

The case $k < K_T$ corresponds to $\Delta E \ll T$ in standard simulated annealing: the transition is highly probable. Similarly, $k > K_T$ means in the OT philosophy that $\Delta E \gg T$, and the transition is prohibited. The middle case, $\Delta E \approx T$, corresponds here to $k = K_T$, and the exponential function ensures a smooth transition between the two extremes. In this last case, as temperature and the violation profile difference are in the same domain, the second (real-valued) components of both play the main role. Otherwise, these second components— t and d —do not get to the stage.

Temperature $T = \langle K_T, t \rangle$ is decreased in a double loop. The outer one diminishes K_T from K_{max} to K_{min} , in linear steps of K_{step} . Similarly, the inner cycle reduces t from t_{max} to t_{min} , in linear steps of t_{step} . The linearity of the outer cycle follows directly from the picture presented in Fig. 2.7.

However, the component t of the temperature could be decreased in several other ways, as well, for instance on a logarithmic scale.²¹ Still, according to the literature on standard simulated annealing Reeves (1995), the exact way of decreasing temperature does not have a major influence on the precision. Consequently, we opt for the simplest way, and leave alternatives for future research. Section 5 will examine the role of t_{max} and t_{min} , and will conclude that different ways of running the inner loop may result in minor—though statistically significant—results. Thus, I suppose that a logarithmic scale would have similar consequences.

The algorithm in Fig. 2.2 also includes a *stopping condition*. Frequently, a simulated *specific heat* is measured, that is, the improvement in energy (in the cost function) per unity temperature change ($\partial E / \partial T$). The stopping condition then requires this measure to drop below a certain level, that is, the algorithm terminates when decreasing the temperature does not lead to much decrease in the cost function anymore.

²¹A logarithmic scale involves multiplying t each time with a constant factor smaller than 1: $t_{n+1} = t_{step} \cdot t_n$. Whereas a linear scale adds a constant negative value to it: $t_{n+1} = t_n - t_{step}$.

```

ALGORITHM: Simulated Annealing for Optimality Theory
Parameters: w_init, K_max, K_min, K_step, t_max, t_min, t_step
            # t_step: number of iterations / speed of simulation
w <-- w_init ;
for K = K_max to K_min step K_step
  for t = t_max to t_min step t_step
    choose random w' in neighbourhood(w) ;
    calculate < C , d > = ||H(w')-H(w)|| ;
    if d <= 0 then w <-- w'
    else
      w <-- w' with probability
      P(C,d;K,t) = 1      , if C < K
                  = exp(-d/t) , if C = K
                  = 0      , if C > K
  end-for
end-for
return w

```

Figure 2.8: **The algorithm of Optimality Theory Simulated Annealing:** Moving with probability $P(C,d;K,t)$ is a short-hand for generating a random number $0 \leq r \leq 1$, and moving iff $r \leq P(C,d;K,t)$.

We could have included a similar condition: the algorithm stops whenever the random walker has not moved for a certain number of time steps. For instance, if the random walker has not left candidate w for $c \cdot |\text{Neighbours}(w)|$ steps (with $c = 5$, say), we may safely conclude that w is a local optimum, and temperature has become too cold to be able to escape it. Running further the algorithm makes no sense, for the likelihood of such an escape will further drop with decreasing temperature. If c is set too low, the algorithm may frequently stop in candidates that are not local optima, but have a few neighbours that are even worse. If c is set too high, the simulation may run unnecessarily long. Nonetheless, it is not difficult to come up with a reasonable compromise.²² The danger only arises in the case of a search space that has “locally optimal valleys” formed by neighbours that incur exactly the same violation marks and are more harmonic than their environment. Such a system is prone to run into an infinite loop, so one either has to employ also a K_{\min} limit, or to count horizontal moves as if the system were stuck in a state (that is, count time steps without improvements in the Harmony function).

The algorithm of *Optimality Theory Simulated Annealing* (OT-SA) is finally presented in Fig. 2.8 in the form we shall use it.

²²Suppose that a candidate has n neighbours, only one of which is more harmonic than this candidate. Temperature is low enough, so that the random walker cannot move to a less harmonic neighbour. Suppose, furthermore, that each neighbour has equal *a priori* probability. Then, the likelihood of choosing a neighbour to which the system will not move is $\frac{n-1}{n}$. Trying $c \cdot n$ times, and still not leaving this non-local optimum has a probability of

$$\left(\frac{n-1}{n}\right)^{cn} < e^{-c}$$

because the series $(1 + \frac{1}{n-1})^n$ is a monotonically decreasing series with e as its limit. Consequently, $c = 3$ guarantees you to terminate the algorithm in a local optimum with probability 95%, and $c = 5$ with a probability higher than 99%. Moreover, this likelihood is even much higher in most search spaces.

Compare it to the standard simulated annealing algorithm in Fig. 2.2 in section 2.1.2. The two main changes are the way temperature is decreased and the transition probability is calculated. Both are connected to the more complex, non-real valued character of the function to be optimised (the cost function). Temperature is decreased in a double cycle, as explained when temperature was introduced: the outer loop decreases its first component and the inner loop its second component. Furthermore, the transition probability can also be determined in a more complex way, which follows equation (2.17).

As can be seen, the parameters of the algorithm are the initial candidate (w_0) from which the simulation is launched, as well as the parameters of the cooling schedule: K_{max} , K_{min} , K_{step} , t_{max} , t_{min} , t_{step} .

Typically, K_{max} will be higher than the highest ranked constraint, so that there will be a phase of the simulation when the random walker can freely increase the violation marks of even the highest ranked constraints. One may however add constraints higher ranked than K_{max} : the resulting picture will be as if we restricted GEN, since the candidates sub-harmonic for these “hyper-strong” constraints would not play a role at all in the model. Section 6.5 introduces a model whose success depends on tuning K_{max} , a parameter that does not influence the model in Chapter 5 (other than trivially).

If the role of K_{max} is to supply additional layers above the highest ranked constraint in order to allow the random walker to move unhindered in the initial phase of the simulation, then the role of K_{min} is to define the length of the final phase of the simulation. Namely, by having K_{min} (much) below the lowest ranked constraint, the system is given enough time to “relax”, to reach the closest local optimum, that is the bottom of the valley (the top of the hill, if you maximise the function) in which the system is stuck. Without such a final phase, the system will return any candidate, not only local optima, leading to an uninteresting model. Consequently, K_{min} has to be chosen such that the number of iterations in this frozen state ($K_T < 0$, if the lowest ranked constraint is C_0) be enough to reach the closest local optimum by a hindered random walk.²³

Not much attention will be given to the parameter K_{step} . Although other options are also possible, the standard way we shall proceed is the following: if we have $N + 1$ constraints, we always place them into the domains $K = 0$, $K = 1, \dots, K = N$, and we set $K_{step} = 1$. Further, if not specified otherwise, $K_{max} = N + 1$. Moreover, the system is said to be *frozen* if the temperature T drops below domain $K = 0$, that is, $T = \langle K_T, t \rangle$ with $K_T < 0$. In a frozen system, no move can lead to a less harmonic candidate.

The parameters t_{max} , t_{min} and t_{step} drive the inner loop of the algorithm, that is the decreasing of the second component t of temperature $T = \langle K, t \rangle$. This second component plays a role only in the expression $e^{-d/t}$, used when the crucial constraint at which w' is defeated by w coincides with the domain of the current temperature. Because the neighbouring candidates w and w' typically differ only minimally (a *basic operation* transforms w into w'), their violation profile is also quite similar, thus the difference d in violating the crucial constraint is expected to be a low number (usually $1 \leq |d| \leq 2$). Consequently, the $e^{-d/t}$ vanishes if $t \gg 3$, and so the default values used will be $t_{max} = 3$,

²³In our models—with $T_{step} = 1$ —a good choice might be $K_{min} = -100$. If $T_{step} = 0.01$, however, $K_{min} = -1$ is enough, and $K_{min} = -100$ will make the simulation unnecessarily long. Remember that the constraints are located in the domains $k = 0, \dots, N$. My scripts automatically adjust K_{min} to T_{step} .

$t_{min} = 0$.²⁴

Chapter 5 will also present experiments tuning t_{min} and t_{max} . Increasing or decreasing the half-closed interval $[t_{min}, t_{max})$ crossed by t has a measurable, though minor effect on the outcome of the simulation. Based on this result, we argue that t and the exponential expression $e^{-d/t}$ in the definition of SA-OT does have an effect on the outcome of the simulation.

The most interesting parameter is t_{step} , for it is inversely proportional to the number of iterations performed—if the other parameters are kept unchanged. In this way, it directly controls the speed of the simulation, hence, indirectly, its precision. Therefore, most of our experiments will vary this parameter. Actually, the other parameters also may change the number of iterations performed, but their effect is more complex, so tuning t_{step} is the most straightforward way—at least, for me—to change the speed. On the other hand, when measuring the role of t_{max} and t_{min} in Chapter 5, the parameter t_{step} will help in controlling for the number of iterations.

In the next chapter, additional arguments are brought in favour of the SA-OT Algorithm in Fig. 2.8: different formal approaches will lead to the same proposal. Before that, however, let us perform some first experiments using the *Simulated Annealing Optimality Theory Algorithm* in order to get an impression of it.

2.3 Playing with SA-OT

2.3.1 When SA-OT works

We have defined the *Simulated Annealing for Optimality Theory Algorithm* (SA-OT), so it is now high time to try out whether it really works, and to see under what circumstances it “fails”.

In the present section, we are trying out toy models in order to obtain a better understanding of what SA-OT really does in practice. The reader is welcome to implement these examples personally on the demo SA-OT available on my web site at <http://www.let.rug.nl/~birot/sa-ot/>. One can simply vary the different parameters of the algorithm, and examine its behaviour. Furthermore, one can specify a “verbose” output, which explains all the details during the simulation.

In section 2.1.2, we introduced a search space with three states and with an asymmetric lambda-shape neighbourhood structure (Fig. 2.3) in order to show why traditional simulated annealing works. Now, we shall try out analogous cases in Optimality Theory and analyse the performance of SA-OT.

Among the three candidates, B is a neighbour of A and C, whereas A and C have the singleton set {B} as their neighbour set. In other words, A and C are not neighbours of each other. Candidate C is always the global optimum, to which the usual \models symbol points, whereas candidate A is a local optimum, annotated by the variation symbol \sim (Fig. 2.9). If A and C were neighbours, A could not become a local optimum.

²⁴Notice that for the sake of convenience, t_{min} is always understood as the lower border of t , *exclusively*. The inner loop runs while $t > t_{min}$. Nevertheless, K_{min} will be understood as the lower border, *inclusively*: the outer loop is meant to run while $K \geq K_{min}$ —whenever rarely K_{min} will be mentioned. This slight incongruity should render the notations otherwise simpler.

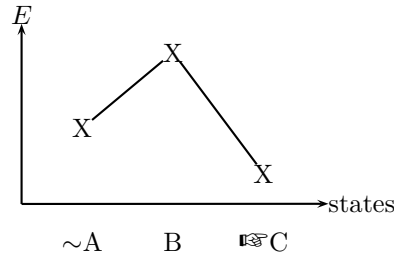


Figure 2.9: An asymmetric landscape with three states, two of which are local optima, but only state C is a global optimum. State B is a neighbour of both A and C, however states A and C are not neighbours of each other.

The first tableau to be examined is the following:

		C1
~	A	*
	B	**
⊞	C	

(2.18)

It is easy to check that $C \succ A \succ B$. The only constraint $C1$ is assigned index (a *K-value*) of 0, that is, temperature will be in its domain whenever the first component of T is $K_T = 0$. Let us set $K_{max} = 1$, $K_{step} = 1$, $t_{max} = 3$, $t_{min} = 0$, and K_{min} low enough, say, $K_{min} = -100$. The only parameter we vary is t_{step} .

As K_{max} is higher than the rank of the only constraint, the initial candidate of the simulation will not really matter. In practice, however, we launch the simulation in 1/3 of the cases with A as the initial candidate, in 1/3 of the cases with B as the initial candidate, and in 1/3 of the cases with C as the initial candidate.

The algorithm is stochastic, and is prone to return different outputs. If K_{min} were higher than the rank of the constraint, even B would be returned in 1/3 or 2/3 of the cases, depending on the parity of the number of the iterations. With different parameter settings, the algorithm terminates only in A and C, or exclusively in C. Therefore, what will interest us is not the output of a certain simulation, but the *proportions* of the different outputs of several simulations. In other words, we seek to know the *likelihood* of the simulation to return a certain candidate—similarly, to other stochastic linguistic models referred to in chapter 1.²⁵

The following table contains the absolute frequencies returned by a number of experiments performed on the demo web site, with the simulation launched 100 times from each of the three candidates:

²⁵By running the simulation many times, the *proportion* of a certain output (the relative frequency) should approximate the theoretical *probability* of returning that candidate.

t_{step}	Frequency of A	Frequency of C
3	141	159
3	146	154
3	145	155
3	129	171
3	131	169
1	128	172
1	122	178
1	134	166
1	130	170
1	127	173
0.1	79	222
0.1	86	214
0.1	82	218
0.1	72	228
0.1	90	210
0.01	32	268
0.01	33	267
0.01	40	260
0.01	40	260
0.01	27	273
0.001	9	291
0.001	10	290
0.001	12	289
0.001	6	294
0.001	14	286

(2.19)

Based on these data, the probability and the standard deviation (n) of returning the globally optimal candidate, C, in function of t_{step} is:

t_{step}	Frequency of C
3	0.537 ± 0.0237
1	0.573 ± 0.0130
0.1	0.727 ± 0.0207
0.01	0.883 ± 0.0167
0.001	0.967 ± 0.0087

(2.20)

The results speak for themselves. A very fast cooling schedule, such as $t_{step} = 3$, returns the local—but non-global—optimum A in almost half of the cases. Slowing down the cooling schedule, that is, increasing the number of iterations performed, will increase the likelihood of returning the global optimum in a highly significant level. For $t_{step} = 0.001$, the chance of returning A is below 5%, and an even smaller t_{step} would further reduce the frequency of A. This case, in which the frequency of the global optimum increases gradually to 100% as t_{step} decreases, will be considered a success of SA-OT.

One can also check simply that the choice of the initial candidate does not influence significantly the simulation, if K_{max} is large enough. For instance, here are the absolute frequencies for the outputs, when we launched the simulation 500 times from each of the three candidates, with $t_{step} = 1$ (the data in

parentheses are the results of repeating the experiment twice):²⁶

Initial candidate	Returning A	Returning C
A	211 (218, 231)	289 (282, 269)
B	191 (198, 195)	309 (302, 305)
C	198 (227, 218)	302 (273, 282)

(2.21)

We can now proceed to cases where SA-OT does not work in the sense that decreasing t_{step} will not increase the likelihood of returning the globally optimal (that is, the grammatical) candidate. In such cases, the topology of the candidate set forces the simulation to always return local optima with a constant likelihood. However, empirical research might find phenomena that can be accounted for by these systematic failures, as we shall argue for in section 6.4. The main reason for these systematic failures will be that in our definition of violation profile difference we neglect constraints that are below the fatal constraint.

2.3.2 When SA-OT *does not* work

Remember that a crucial step in introducing SA-OT was neglecting the constraints below the *fatal constraint* (the highest ranked constraint with uncanceled marks). The difference of two violation profiles is the difference in their levels of violating the fatal constraint, independently of how they behave with respect to lower constraints. Consequently, the difference of w and w_1 is the same as the difference of w and w_2 —violation C_1 once—in the following tableau:

	C_2	C_1	C_0
w	*	**	
w_1	*	*	
w_2	*	*	**

(2.22)

We argued for this definition based on the foundations of Optimality Theory, following the claim that cumulativity effects should be avoided. Nevertheless, neglecting lower ranked constraints—not making any difference between the difference of w and w_1 on the one hand, and the difference of w and w_2 on the other in tableau (2.22)—leads us to cases where the algorithm for OT-SA just presented does not work.

Imagine again the asymmetric lambda-shaped landscape with three candidates, two of which are local optima, such as the one presented in figure 2.9. This is the repetition of figure 2.3, which we used in section 2.1.2 to understand why traditional simulated annealing works.

Remember the argumentation there. Moving to both directions from the middle state B has equal chance: both neighbours are chosen with equal probability, and the random walker moves there certainly, once chosen. Yet, it is more difficult to escape from the global optimum C than from the local optimum A, because moving to B involves a greater step uphill. So a slower cooling schedule with more time steps n makes it more probable that you will escape from A,

²⁶Notice that $t_{step} = 1$ and $K_{max} = 1$ allows only for three unhindered steps, namely when $T = \langle 1, 3 \rangle$, $T = \langle 1, 2 \rangle$ and $T = \langle 1, 1 \rangle$. Changing the parameter setting would increase the number of unhindered steps, and would therefore decrease further the role of choosing the initial candidate.

and at least once choose to go to C, where you will be caught.²⁷ In order not to get confined in C, you have to choose always moving to A from B, which has a probability of 0.5^n vanishing with high ns . The chance of choosing C—and get stuck there—at least once in n time steps is $1 - 0.5^n \approx 1$ for high ns , that is, for slow cooling schedules.

Consequently, although gradient descent would assign the same probability of ending in A and C, the stochastic process introduced by simulated annealing increases the chance of finding the global optimum. The slower the cooling schedule, the higher the chance to find it.

Can our proposed simulated annealing for Optimality Theory exhibit the same behaviour? It depends on the profile of the three candidates, A, B and C. Without changing the topology, let us observe the behaviour of OT systems with different violation profiles. The simplest example was the one analysed in the previous subsection, namely, tableau (2.18), and there, simulated annealing worked perfectly. The effect just observed in the real-valued case also works if the three candidates are characterised by the following tableau:

	$C1$	$C2$
\sim A	*	
B	*	*
\mathbb{E} C		

(2.23)

Using the demo web site at <http://www.let.rug.nl/~birot/sa-ot/>, we can observe that finding the global optimum—to which the hand points—is even easier in this system than in the one defined by tableau (2.18). Namely, at $t_{step} = 0.5$, the likelihood of terminating in candidate A—the alternative form marked with the \sim symbol—is already below 10%; whereas $t_{step} = 0.1$ will return exclusively the globally optimal candidate C.

What happens in this case? When temperature T is in the range of the constraint $C2$, moving from C to B is already impossible, for such a move would increase the number of violations of constraint $C1 \gg T$. However, escaping from A to B is still possible, and thus a slower schedule will result in a higher probability of escaping from A and falling into the trap of C. The more time steps are allowed while T is in the domain of $C2$, the higher the probability of ending up in C.

Thus, tableau (2.23) represents a second case where SA-OT behaves the same way as traditional simulated annealing. Take now the following tableau:

	$C1$	$C2$
\sim A		*
B	*	
\mathbb{E} C		

(2.24)

Here, stepping to B involves incurring an extra violation of constraint $C1$, both from A and C. Candidate C being better than A does not matter for they are not neighbours—otherwise they would not be local optima. The situation is fully symmetric from the viewpoint of the probabilities, because escaping from

²⁷Here, I am presenting a caricature of the situation, because escaping from state C always has some minor positive probability in a real-valued simulated annealing. The train of thought should nevertheless be clear.

C has always the same chance as escaping from A: the violation profile difference incurs one violation of C1 in both cases. If you succeed to climb from A to B, you will also succeed to climb from C to B. If you are caught in C, you are also caught in A. The difference between A and C is invisible, and the trick that worked before does not work now. The distribution of the outputs will only depend on how candidate B distributes the probabilities between A and C—independently of the cooling schedule.

Can we make the chance of moving from A to B higher than that of moving from C to B in this last situation, too (at least in some phase of the simulation)? Remember that the transition probability $P(A \rightarrow B|T)$ cannot depend on the violation profile of candidate C.

Changing the *a priori* probabilities of moving from B to A or C results in a distribution different from 50%-50%, because the chance of going to A or to C from B is not symmetric anymore; and yet, this distribution is still independent of the cooling schedule. Indeed, the *a priori* probabilities define the horizontal structure of the landscape, whereas our problem here concerns its vertical structure, to which the cooling schedule is related, as well.

This question is still an open research question, and may lead to altering the whole concept of simulated annealing for OT. Meanwhile, I can only imagine solutions that would not only contradict the general philosophy of OT, but also introduce further problems.

For instance, take the following new definition of the difference of two violation profiles. Start with the mark cancellation procedure from the highest ranked constraint. Initial violation marks shared by both candidates do not interest us. Suppose C1 is the fatal constraint where the better candidate (candidate A) has fewer violation marks than candidate B. Until here, the difference of the two violation profiles was defined as the constraint C1 and the number of uncanceled marks by C1. Now, we also want to take into consideration the highest violation mark of candidate A that is below C1. Eyeballing tableau (2.24), we can see that this is how one may capture the fact that moving from C to B involves a bigger upward step than from A to B.

Nonetheless, how to implement this idea in the case of the following tableau?

		C1	C2	C3
~	A		*	*
	B	*		
⊗	C		*	

(2.25)

Both A and C have their first violation mark at the same constraint C2, and their difference shows up only deeper in the hierarchy. Hence, this new definition would make no difference between moving from A or from C to B. Remember also that OT suggests not to look at further constraints if you have found the needed differences when comparing two candidates.

As a side remark, as SA-OT may predict the same probability to candidates A and C here, we have just shown that SA-OT does not have ganging-up cumulativity (cf. tableau (1.16) in section 1.3.5). Namely, the probability of candidate A (and similarly to C) does not change from tableau (2.25) to the following tableau, notwithstanding its different behaviour for constraint C3:

		$C1$	$C2$	$C3$
\sim	A		*	
	B	*		
\sim	C		*	*

(2.26)

Further, the new definition proposed would work incorrectly in this case:

		$C1$	$C2$	$C3$	$C4$
\sim	A	*			*
	B	*	*		*
\sim	C		*		*

(2.27)

Now, when comparing A to B, we have to descend two constraints in order to find the first violation mark incurred by A below the critical constraint $C2$. Two levels is a higher difference than the one level needed when we compare C to B. Obviously, we may consider again the highest constraint where the two profiles differ, and then the given differences can be said to be “two levels from $C2$ ” as opposed to “one level from $C1$ ”.

In addition, the number of violation marks found at a lower level by the better candidate needs also to be taken into consideration, as shown by the following tableau:

		$C1$	$C2$
\sim	A		**
	B	*	
\sim	C		*

(2.28)

We may speculate further. Nevertheless, I have no idea how to capture all these observations into a single elegant model. Do not forget that the only information at hand when determining the transition probabilities are the two violation profiles, and we cannot refer to other neighbours of the target state. Probably, the phenomenon discussed in the present subsection is an inevitable consequence of applying simulated annealing to a non-real valued function, such as is the case for the harmony function in Optimality Theory.²⁸

One may also add a small bit of memory to the random-walker: while wandering around in the search space, the best candidate found so far is always remembered and compared to the present position. Then, the algorithm returns not necessarily the final position, but the best candidate found during the walk. This trick would work, especially in small search spaces as those seen in this section, but not necessarily in large ones. Nonetheless, we shall leave

²⁸As proposed by Balázs Szendrői, a solution may be the approximation of the harmony function of OT with polynomials $P(w)[q] = C_N(w)q^n + \dots + C_1(w)q + C_0(w)$ the coefficients of which are the constraint violation levels (using exponential weights; cf. subsection 3.3, and e.g. Smolensky’s pre-OT *Harmony Theory* or Prince (2002)). By increasing q , we may better approach the OT harmony function, but an unreasonably high value for q will simply reproduce the described situations. An additional argument for keeping q relatively low is that if the neighbouring candidates do not differ much, their violation profile is also similar, consequently smaller qs will also correctly account for their difference in harmony. Yet, one can also reproduce cumulativity effects (cf. subsection 1.3.5) this way—supposing one wishes to. As higher qs require higher ranges for the temperature, further research may also consider the possibility of gradually increasing q instead of decreasing the temperature.

this possibility for future work, and rather focus on what the implementation of traditional simulated annealing to OT can propose us.

Additionally, we shall later turn these failures of SA-OT into an advantage by claiming that such failures correspond to empirically attested agrammaticalities. Language—even slow and careful speech—can display irregular forms that contradict the general rules of the grammar, and yet, they are inevitably produced in speech and attested in corpora. Instead of turning the grammar much more complex so that it account for these irregularities, we shall argue to keep the grammar model (the underlying OT-system) simple, and explain these forms on the language production level.

The OT-grammars discussed in this section have been very simple, as they included only a few candidates and a few constraints. One may ask then how the precision is influenced by the number of candidates and the number of constraints, but no simple answer can be given. A major disadvantage of SA-OT is that the interactions between the neighbourhood structure, the constraint hierarchy and the cooling schedule is so complex that it is often impossible to find out the behaviour of the system without running the simulations. The second part of my dissertation introduces different models, involving larger but finite, as well as infinite candidate sets, and displaying very different behaviours. If there are only few local optima, then the system's precision may be similar to the precision of those analysed so far; if, however, the huge candidate set is full of local optima, as will be the case with the different models of Chapter 7, then precision can drop drastically.

Before that, let us turn to a few theoretical, formal and mathematical issues related to Optimality Theory in general, and SA-OT in particular.

Chapter 3

Formal Approaches to SA-OT

The goal of this chapter is to underpin the *Simulated Annealing Optimality Theory Algorithm* (Fig. 2.8) in general, and the definition (2.17) of the *transition probability* $P(w \rightarrow w' \mid T)$, that is, the *Rules of moving* (page 63) in particular. We demonstrate that this definition follows directly from the *Strict Domination Hypothesis*, which constitutes the basis of OT. Therefore, we introduce several formal representations of Optimality Theory. Each of them is first proved to realise the *Strict Domination Hypothesis*, and then to lead to the same definition of temperature and to the same transition probabilities $P(w \rightarrow w' \mid T)$ —independently of each other. Before introducing different representations (real numbers in section 3.2, followed by polynomials in section 3.3, and finally ordinal numbers in section 3.4), however, we have to formally define what a representation is in Optimality Theory (section 3.1). Fig. 3.1 on page 85 might serve to the reader as a road map to the present chapter.¹

As the formal models developed here result in the same algorithm, latter chapters, as well as further implementations of SA-OT, can certainly be understood without the mathematically demanding details of the present chapter.

3.1 Towards a formal definition of OT

The *violation profile* as introduced by Prince and Smolensky (2004) (Prince and Smolensky, 1993) is a list of violation marks—or, rather, a set of tokens of violation marks. The *Harmony function* is the mapping that assigns a violation profile to a candidate. Nonetheless, Prince and Smolensky’s “list of violation marks” is a less convenient construction. Therefore, here we (re-)introduce the *vector representation* of a *violation profile*, a straightforward translation (and generalisation) of Prince and Smolensky’s idea. We repeat the idea already presented in section 2.2.3, and elaborate more on this approach. Then, we consider it as the standard for introducing two further representations: polynomials and ordinal numbers.

¹A summary of the present chapter has been published as B     (2005b).

As our starting point, we are given GEN, a mapping from the set of possible underlying representations to the set of possible candidates. Let $GEN(UR)$ denote the *set of candidates* corresponding to a specific underlying representation UR . The set \mathcal{PC} of *all possible candidates* is the union of $GEN(UR)$ for all possible UR .

3.1.1 Constraints

Let $C_i(w)$ be the number of times candidate w violates constraint C_i . In general, we shall call $C_i(w)$ the *level of violation* incurred by candidate w with respect to constraint C_i , and in specific models we may speak of the *number of violation marks* assigned by the constraint.

Indeed, C_i most often takes non-negative integer values in linguistic practice, conform to the original idea of a “list of violation marks” in Prince and Smolensky (1993). Yet, here we generalise the concept:

Definition 3.1.1. *Constraint C_i is a function on the set \mathcal{PC} of all possible candidates, such that for each possible UR : the set $\{C_i(w) \mid w \in GEN(UR)\}$ (the image of the candidate set corresponding to UR) is a totally ordered set with some ordering relation $\mathcal{R}_{i,UR}$, and any of its subsets has a lower bound contained by the subset.²*

Notice that different constraints may have different ranges. Moreover, the same constraint with different underlying representations could also have very different ranges in theory. Moreover, the requirement of a lower bound is only important to ensure that an optimal form always exists. The set of non-negative integer values with the simple *greater than* relation used in practice for C_i clearly satisfies all our requirements.

Although it has been already mentioned, it may be useful to repeat here:

Definition 3.1.2. *Let S be a set, and $>$ a binary relation on S .³ Then, the pair $(S, >)$ is a totally (fully) ordered set, iff:*

1. *The law of trichotomy: for all $x, y \in S$ exactly one of the following three statements holds: 1. $x > y$, 2. or $y > x$, 3. or $x = y$.*
2. *Transitivity: for all $x, y, z \in S$, if $x > y$ and $y > z$ then $x > z$.*

3.1.2 Hierarchies

First, let us introduce the notion of *isomorphism*.⁴

Definition 3.1.3. *The totally ordered sets $(A, <)$ and $(B, <)$ are ORDER ISOMORPHIC, iff there is a bijection⁵ f from A to B such that for all $a_1, a_2 \in A$, $a_1 < a_2$ iff $f(a_1) < f(a_2)$.*

²In brief, the set $\{C_i(w) \mid w \in UR\}$ is well-ordered with the relation $\mathcal{R}_{i,UR}$.

³That is, $<$ is a subset of $S \times S$.

⁴Cf. e.g. Eric W. Weisstein: “Bijection”, from MathWorld—A Wolfram Web Resource, <http://mathworld.wolfram.com/Bijection.html>; Holz et al. (1999, p. 11).

⁵A bijection f is “a transformation which is one-to-one and onto”. That is, its domain covers A and its inverse f^{-1} is also a function ($f(a_1) = f(a_2)$ iff $a_1 = a_2$) whose domain covers the entire set B .

In other words, the *isomorphism* f translates the order $<$ on set A into the order $<$ on set B .

The subsequent key concept in OT is a *constraint hierarchy*:

Definition 3.1.4. A CONSTRAINT HIERARCHY \mathcal{H} , is a finite set of totally ordered constraints $\{C_N, C_{N-1}, \dots, C_1, C_0\}$ with an ordering relation \gg .

Any two totally ordered sets with finite k elements (for any nonnegative integer k) are order isomorphic.⁶ Consequently, the above hierarchy \mathcal{H} is order isomorphic to the ordered set $(N, N-1, \dots, 0)$, so it can be easily represented as a vector:

$$\mathcal{H} = (C_N, C_{N-1}, \dots, C_1, C_0) \quad (3.1)$$

In turn, the VECTOR REPRESENTATION of the *Harmony function* of a candidate w with respect to a constraint hierarchy $\mathcal{H} = (C_N, C_{N-1}, \dots, C_1, C_0)$ is defined as the vector

$$H_{\mathcal{H}}(w) = (C_N(w), C_{N-1}(w), \dots, C_1(w), C_0(w)) \quad (3.2)$$

In subsection 2.2.3 (equation (2.10)), we already saw that this vector representation is but a shorthand for a row in a traditional tableau.

Most frequently the hierarchy will be constant, so the subscript \mathcal{H} may be left out. Notice that the subscripts of the constraints are written in a decreasing order: this minor inconvenience at this point will help us later in keeping our notations simple.

The way Prince and Smolensky's original "list of violation marks" can be translated into this vector representation is straightforward: first, if the list of violation marks incurred by candidate w contains n tokens of violation marks $*C_i$, then let $C_i(w) = n$; second, the ranking of the constraints can simply be mapped onto a vector using the order isomorphism. Therefore, this representation of a violation profile will serve as the formalisation of standard Optimality Theory.

3.1.3 An order on violation profile-like vectors

Our goal is to formulate the central idea of Optimality Theory in (3.3) that says that the surface representation is the candidate that maximises the Harmony function. Therefore, our next step is to define an order between two *violation profile-like vectors*. First we introduce

Definition 3.1.5. A VIOLATION PROFILE-LIKE VECTOR with respect to underlying form UR is an element of the following Cartesian product:

$$Range_{UR}(C_N) \times Range_{UR}(C_{N-1}) \times \dots \times Range_{UR}(C_0)$$

Here, $Range_{UR}(C_i) = \{C_i(w) \mid w \in GEN(UR)\}$. We shall, however, omit the reference to UR for the sake of simplicity.

Now, we define an order \succ on two, violation profile-like vectors:

⁶Stated for instance in the MathWorld of Eric W. Weisstein. "Ordinal Number" at <http://mathworld.wolfram.com/OrdinalNumber.html>.

Definition 3.1.6. Let $A = (a_N, a_{N-1}, \dots, a_0)$ and $B = (b_N, b_{N-1}, \dots, b_0)$ be two violation profile-like vectors (with respect to the same UR). Then, A is MORE HARMONIC THAN B , $A \succ B$ if and only if there is an integer $k \in \{N, N-1, \dots, 1, 0\}$ such that

1. $a_k < b_k$
2. and for all $i \in \{N, N-1, \dots, 1, 0\}$: if $i > k$ then $a_i = b_i$

Moreover, $A = B$ iff for all $k \in \{N, N-1, \dots, 1, 0\}$, $a_k = b_k$.

The set $\{N, N-1, \dots, 1, 0\}$, which is going to recur very frequently, could have been replaced by an arbitrary finite set of indices \mathcal{I} with some order $>$. This more general and shorter notation is however equivalent to the more transparent notation we use, since, as noted, any two totally ordered finite sets of equal cardinality are order isomorphic.

We may call C_k the *critical*⁷ or *fatal constraint*: this is the constraint that determines the relative harmony of two violation profile-like vectors. This definition of the binary relation \succ may be also called *lexicographic ordering* (Eisner, 2000b).

It may be confusing that the *more harmonic than* relation has an opposite direction to the *bigger than* relation on the number of violation marks: *more harmonic* ($A \succ B$) corresponds to *fewer marks* ($a_k < b_k$). In the following sections, the rule is that the Harmony function is to be optimised or maximised, whereas the energy function (cost function, the violation marks) minimised.

In what follows, we demonstrate that relation \succ is a total order on any set of violation profile-like vectors: that is, both trichotomy and transitivity hold.

Theorem 3.1.7. TRANSITIVITY: Suppose that $A = (a_N, a_{N-1}, \dots, a_0)$, $B = (b_N, b_{N-1}, \dots, b_0)$ and $C = (c_N, c_{N-1}, \dots, c_0)$ are violation-like vectors (with respect to the same UR). If $A \succ B$ and $B \succ C$, then also $A \succ C$ holds.

Proof. Suppose that $A \succ B$ with C_k being the crucial (fatal) constraint ($a_k < b_k$; for all $i > k$: $a_i = b_i$). Furthermore, suppose $B \succ C$ with C_l as crucial constraint ($b_l < c_l$; for all $i > l$: $b_i = c_i$). Now, we have to demonstrate that $A \succ C$. Let us distinguish between three cases: $l > k$, $l = k$ and $l < k$. If $l > k$, then, by the definition of \succ , $a_l = b_l < c_l$, and for all $i > l > k$: $a_i = b_i = c_i$. Thus, $A \succ C$, and the crucial constraint is C_l . Secondly, if $l = k$, then $a_l = a_k < b_k = b_l < c_l$, and for all $i > l = k$: $a_i = b_i = c_i$. Again, $A \succ C$, with the crucial constraint being $C_l = C_k$. Finally, whenever $l < k$, $a_k < b_k = c_k$, and $a_i = b_i = c_i$ for all $i > k > l$. In this case, $A \succ C$ because the crucial constraint is C_k . \square

Theorem 3.1.8. TRICHOTOMY: Suppose that $A = (a_N, a_{N-1}, \dots, a_0)$ and $B = (b_N, b_{N-1}, \dots, b_0)$ are violation-like vectors (with respect to the same UR). Then, exactly one of the following three relations hold:

- $A \succ B$
- $B \succ A$

⁷This notion of critical constraint should not be confused with the *critical cut-off point* in Coetzee (2004)'s proposal (cf. section 1.3.2).

- $A = B$

Proof. Suppose that $A \neq B$. By the last part of definition 3.1.6, two vectors are not equal if at least one of their components is different. Take the set $S = \{i \in \{N, N-1, \dots, 1, 0\} \mid a_i \neq b_i\}$. Observe that S is a finite set, thus it has a maximum k , and therefore contains it: $k = \max(S) \in S$. We are demonstrating now that C_k is the crucial constraint. Because k is the maximum of S , it is true that for all $i \in [N, \dots, 1, 0]$: if $i > k$ then i is not in S , so $a_i = b_i$. As for the first requirement in the definition of \succ : because $k \in S$, either $a_k > b_k$ or $a_k < b_k$. (Note that here becomes important that the range of the constraints are also fully ranked sets.) In the first case $B \succ A$, and in the second case $A \succ B$. \square

In sum, we have shown that any set of violation profile-like vectors are totally ordered with respect to the relation \succ . Now, we demonstrate that any set of violation-like vectors has a minimum, and also contains it:

Theorem 3.1.9. THE MAXIMUM-THEOREM ON VIOLATION PROFILE-LIKE VECTORS: *Let S be a non-empty set of violation-like vectors (with respect to the same UR). Then, there is exactly one violation-like vector $A_0 = \max(S)$ such that: 1. $A_0 \in S$; and 2. for all $A \in S$, if $A_0 \neq A$ then $A_0 \succ A$.*

Proof. We shall find A_0 the same way as a linguist finds the best candidate in a tableau.

Let $S_{N+1} = S$. Further, for all $i \in \{N, N-1, \dots, 0\}$: suppose that $m_i = \min\{w_i \mid W \in S_{i+1}\}$ and $S_i = \{W \in S_{i+1} \mid w_i = m_i\}$, where we use the abbreviation $W = (w_N, w_{N-1}, \dots, w_0)$. In other words, m_i is the lowest violation level for constraint C_i attested among the elements of S_{i+1} ; whereas S_i is the subset of S_{i+1} containing the elements which have exactly violation level m_i for constraint C_i . Observe that the definition of m_i makes crucially reference to a property in Definition 3.1.1 of a constraint: a subset of $\text{Range}(C_i)$ (here $\{w_i \mid W \in S_{i+1}\}$) always has a lower bound. Not only does it have a lower bound, but the subset also contains its bound. Consequently, there is at least one $W \in S_{i+1}$ such that $w_i = m_i$, which is why S_i is not empty. In brief, S_i is the set of violation profile-like vectors that have “survived” the filtering effect of constraint C_i .

Now, we show that S_0 has exactly one element. First, we have just seen that S_0 is not empty, similarly to all S_i s. Second, suppose both $A \in S_0$ and $B \in S_0$. Then $A \in S_0 \subseteq S_1 \subseteq \dots \subseteq S_i$, that is, $a_i = m_i$, for all $i \in \{N, N-1, \dots, 0\}$. Similarly, $b_i = m_i$, by definition of S_i . Consequently, all components of A and B are equal, that is, by definition 3.1.6, $A = B$.

Last, we show that the only element A of S_0 is the minimum predicted by the theorem. Clearly, $A \in S_0 \subseteq S_{N+1} = S$. Moreover, take any $B \in S$ that is different from A (hence, not a member of S_0). Let k be such that B is an element of S_{k+1} , but not an element of S_k . Such a k exists, because $S_0 \subseteq S_1 \subseteq \dots \subseteq S_{N+1} = S$, and B is element of S , but not of S_0 . Now, as $A \in S_0 \subseteq S_k$, $a_k = m_k < b_k$, by definition of m_k . Yet, for all $i > k$, both A and B are in S_i : in other words, $a_i = m_i = b_i$. Therefore, we have demonstrated, by definition 3.1.6, that $A \succ B$. In sum, the only element A of S_0 is a maximal element of S .

Finally, S cannot have two different maximal elements. Suppose that both A_1 and A_2 were maximal elements. Because A_1 is a maximal element, and A_2 is

a different element of S , then $A_1 \succ A_2$. Similarly, $A_2 \succ A_1$ should hold, which contradicts the law of trichotomy (theorem 3.1.8). \square

3.1.4 Comparing candidates

The following definition follows closely the proposal of Prince and Smolensky, also called *strict domination*:

Definition 3.1.10. *For a given hierarchy $\mathcal{H} = (C_N, C_{N-1}, \dots, C_1, C_0)$ and candidates w_1 and w_2 , w_1 is MORE HARMONIC THAN w_2 , or $w_1 \succ_{\mathcal{H}} w_2$, if and only if there is an integer $k \in \{N, N-1, \dots, 0\}$ such that*

1. $C_k(w_1) < C_k(w_2)$
2. and for all $i \in \{N, N-1, \dots, 1, 0\}$: if $i > k$ then $C_i(w_1) = C_i(w_2)$

Moreover, two candidates w_1 and w_2 are EQUIVALENT, $w_1 \simeq w_2$ if and only if for all $i \in \{N, N-1, \dots, 1, 0\}$: $C_i(w_1) = C_i(w_2)$.

The reference to the hierarchy \mathcal{H} may be omitted whenever obvious. The expression *strict domination* refers to a very important property of this definition: if a candidate meets its Waterloo at a given constraint, it can never come back to the battle field. Even by satisfying all lower ranked constraints, behaving with respect to them much better than all surviving candidates, it is definitely defeated.

Observe the following properties:

Corollary 3.1.11. *The relation \simeq is an equivalence relation on the set of candidates. That is, if w_1 , w_2 and w_3 are candidates, then*

1. $w_1 \simeq w_1$ (reflexivity)
2. $w_1 \simeq w_2$ iff $w_2 \simeq w_1$ (symmetry)
3. $w_1 \simeq w_2$ and $w_2 \simeq w_3$ the $w_1 \simeq w_3$ (transitivity)

Furthermore, for a given hierarchy \mathcal{H} , if $w_1 \simeq w_2$ and $w_2 \succ_{\mathcal{H}} w_3$ the $w_1 \succ_{\mathcal{H}} w_3$.

By comparing the definition 3.1.10 of the \succ and \simeq relations on candidates to the definition 3.1.6 of \succ and $=$ on violation profile-like vectors, we immediately see by equation (3.2) that

Corollary 3.1.12. THE EQUIVALENCE OF STRICT DOMINATION AND VIOLATION PROFILES: *The vector-representation of the violation profiles realises Prince and Smolensky's definition of strict domination:*

1. $H_{\mathcal{H}}(w_1) \succ H_{\mathcal{H}}(w_2)$ if and only if $w_1 \succ_{\mathcal{H}} w_2$;
2. $H_{\mathcal{H}}(w_1) = H_{\mathcal{H}}(w_2)$ if $w_1 = w_2$;

Moreover, $w_1 \simeq w_2$ if $H_{\mathcal{H}}(w_1) = H_{\mathcal{H}}(w_2)$.

In other words, the function $H_{\mathcal{H}}$ is a *homomorphism* (Holz et al., 1999, p. 10-11) with respect to relations $\succ_{\mathcal{H}}$ on the set of candidates and \succ on the set of violation profile-like vectors.

In the following two subsections, we shall demonstrate that the alternative representations to be proposed are also equivalent to the violation profiles, hence, to strict domination.

Now, relation \succ is almost a total ordering on the candidate set corresponding to a given underlying representation UR . Transitivity holds, as a consequence of the transitivity on the set of violation profile-like vectors. Yet, the Law of Trichotomy only holds in a weaker modified version: whenever neither $w_1 \succ w_2$ nor $w_2 \succ w_1$, then w_1 and w_2 are *equivalent* ($w_1 \simeq w_2$), that is, they incur the same violation level by all constraints ($H(w_1) = H(w_2)$). To prove it, one has to apply the law of trichotomy on violation profile-like vectors to the vectors $H(w_1)$ and $H(w_2)$, and use corollary 3.1.12.

Similarly, the consequence of the *Maximum-theorem on violation profile-like vectors* is the following:

Theorem 3.1.13. THE MAXIMUM-THEOREM ON CANDIDATES: *Let S be a set of candidates (corresponding to the same UR). Then, for a given hierarchy of constraints \mathcal{H} , S has a unique subset $S_0 = \min_{\mathcal{H}}(S) \subseteq S$ such that*

1. *if $w_1 \in S_0$ and $w_2 \in S_0$, then $H_{\mathcal{H}}(w_1) = H_{\mathcal{H}}(w_2)$;*
2. *if $w_1 \in S_0$ and $w_3 \in S \setminus S_0$, then $w_1 \succ_{\mathcal{H}} w_3$.*

Proof. To prove this statement, one has to apply the *Maximum-theorem on violation profile-like vectors* to the set $\{H_{\mathcal{H}}(w) \mid w \in S\}$. This is a set of violation profile-like vectors, and has exactly one maximal element A_0 . Now, the maximum subset S_0 of S is formed by the elements $w \in S$ such that $H_{\mathcal{H}}(w) = A_0 = \min\{H_{\mathcal{H}}(w) \mid w \in S\}$. In other words: $S_0 = \operatorname{argmin}_{w \in S}(H_{\mathcal{H}}(w))$.

Set S_0 is not empty, because the maximal element $A_0 \in \{H_{\mathcal{H}}(w) \mid w \in S\}$, that is, for at least one $w \in S$, $H_{\mathcal{H}}(w) = A_0$. If both w_1 and $w_2 \in S_0$, then $H_{\mathcal{H}}(w_1) = A_0 = H_{\mathcal{H}}(w_2)$. Finally, if $w_1 \in S_0$ and $w_3 \in S \setminus S_0$: $H_{\mathcal{H}}(w_3) \in \{H_{\mathcal{H}}(w) \mid w \in S\}$, but $H_{\mathcal{H}}(w_3) \neq A_0$ (otherwise, $w_3 \in S_0$), thus $H_{\mathcal{H}}(w_1) = A_0 \succ_{\mathcal{H}} H_{\mathcal{H}}(w_3)$. Then, by corollary 3.1.12, $w_1 \succ_{\mathcal{H}} w_3$.

Finally, we show that the maximum subset S_0 is unique. Namely, suppose that two such maximum subsets, S_0 and S'_0 exist at the same time. If the two subsets are different, then there exist an element $w \in S$ such that either $w \in S'_0$ and $w \notin S_0$, or $w \notin S'_0$ and $w \in S_0$. As the two cases are symmetrical, let us take the former case. Then, $w \in S \setminus S_0$, hence $w_1 \succ_{\mathcal{H}} w$ for any $w_1 \in S_0$. As S'_0 is also a maximum subset, all its elements $w_2 \in S'_0$ are equivalent to w ($H_{\mathcal{H}}(w) = H_{\mathcal{H}}(w_2)$). Consequently, for any $w_1 \in S_0$ and $w_2 \in S'_0$, it is true that $w_1 \succ_{\mathcal{H}} w_2$, and therefore S_0 and S'_0 are disjoint sets (by the Law of Trichotomy). Furthermore, S'_0 cannot be a maximum subset, for its elements are less harmonic than some elements of $w \in S \setminus S'_0$, namely, the elements of S_0 , which fact would contradict the Law of Trichotomy. \square

3.1.5 The definition of Optimality Theory

Finally, we are able to formulate what Optimality Theory is about, and see the soundness of this formulation.

Remember that GEN is a function that maps each underlying representation (UR) to a set of candidates. The central idea of Optimality Theory is that the surface representation is the optimal (maximal) candidate of the candidate set with respect to an ordering defined by the given hierarchy:

$$\begin{aligned} SR &= \max_{\mathcal{H}}(GEN(UR)) = \\ &= \operatorname{argmax}_{w \in GEN(UR)} H_{\mathcal{H}}(w) \end{aligned} \quad (3.3)$$

The first line of this definition makes sense because of theorem 3.1.13, and is equal to the second line by corollary 3.1.12.

In words: the surface representation(s) maximise(s) the Harmony function. Whenever more candidates $w \in GEN(UR)$ maximise $H(\cdot)$, all of these candidates are predicted to appear as a grammatical form on the surface (Prince and Smolensky (2004) p. 82): this is the approach mentioned in section 1.3.1.

Sometimes, the candidates include information not present in the overt linguistic form, such as parsing brackets. Nevertheless, the surface form can be readily arrived at from the winning candidate by a simple function (e.g. by erasing these brackets). Note that because the inverse of this transformation is not always a function, and more candidates can correspond to the same form, learning algorithms face extra difficulties. Tesar and Smolensky (2000) propose using *Robust Interpretive Parsing* in order to decide which candidate corresponding to the overt learning data form to employ in the learning algorithm.

3.1.6 Realisations of the Harmony function

Subsection 2.2.3 showed how Prince and Smolensky (2004)'s concept of a set of violation mark tokens can be translated into the *vector representation* of the Harmony function. There, we referred crucially to Prince and Smolensky's *Cancellation/Domination Lemma*. The present subsection has demonstrated formally that this representation makes sense, and that the formulation of an OT grammar as equation (3.3) is well-founded.

In the following subsections, we introduce two new representations of the Harmony function. We do that in order to carry out again the agenda of introducing SA-OT:

- Firstly, we represent the violation profiles in an appropriate way.
- Secondly, we define the *difference* of two violation profiles.
- Thirdly, we define *temperature* in a similar format.
- Fourthly, we define the exponential of their quotient.
- Lastly, we introduce the SA-OT algorithm.

Before launching this program, however, let us define what an appropriate representation of a violation profile is. Corollary 3.1.12 has already stated the (almost) equivalence of the ranking \succ on the candidate set and the ranking \succ on the set of vector representation of the candidates' violation profiles. Thus, the vector representation of a violation profile is a typical example of *order isomorphism* as introduced in definition 3.1.3.

To be more exact, based on corollary 3.1.11, we introduce the *set of violation profiles*, which is the set of equivalence classes on the set of candidates with respect to equivalence relation \simeq . Again by corollary 3.1.11, but by its second part this time, a total order $\succ_{\mathcal{H}}$ on the violation profiles may be introduced: an equivalence class is *more harmonic* than another equivalence class, if some element of the first class is more harmonic than some element of the second class (by definition 3.1.10). Now, the set of violation profiles (equivalence classes on the set of candidates) with order $\succ_{\mathcal{H}}$ is *order isomorphic* to the set $\{H_{\mathcal{H}}(w) | w \text{ is a candidate}\}$ with the order \succ on violation profile-like vectors (def. 3.1.6).

The new representations of the Harmony function must be isomorphic to the set of the violation profiles, too. Only this can ensure that the new representations will yield the grammar defined by equation 3.3. In other words:

Definition 3.1.14. A REALISATION of the Harmony function $H(w)$ is a mapping $E(w) : \mathcal{PC} \rightarrow X$ (from the set of all possible candidates to some set X), such that:

- a total ordering relation \prec and an equivalence relation $=$ is defined on the set X ;
- for all candidates w_1 and w_2 : $H(w_1) \succ H(w_2)$ iff $E(w_1) \prec E(w_2)$;
- for all candidates w_1 and w_2 : $H(w_1) = H(w_2)$ iff $E(w_1) = E(w_2)$.

Observe that the new representation E is compared to the vector representation H , which can be done because isomorphy is a transitive relation between ordered sets. We know that the set of violation profiles is order isomorphic to the set of vector representations; hence, a new representation is isomorphic to the set of violation profiles if and only if it is isomorphic to their vector representation.

Besides being clearer, an additional advantage of using the vector representation as the starting point—as opposed to a set of violation mark tokens—is that definition 3.1.1 allows for more flexibility concerning the range of the constraints.⁸

Note that the ordering will be reversed: while the Harmony function H is to be maximised, the new representations—seen as energy or cost function—will be minimised.⁹ The advantages of reversing the \succ relation are manifold. It is simpler to derive formally the new representations from the constraints seen as non-negative valued functions in a way that results in this reversed relation. Intuitively, the minimisation approach parallels better the idea of minimising the violation marks—that is, the punishment symbols. Moreover, simulated annealing is traditionally formulated for minimising the cost function. Observe that subsection 2.2.3 already defined the difference of two violation profiles so that it is positive if less violation marks is subtracted from more violation marks:

⁸Most often in practice, violation levels are non-negative integers. Nonetheless, equation (4.8) introduces a constraint whose possible range is $\mathbb{N}_0 + z \cdot \mathbb{N}_0$ with $z \in \mathbb{R}$ —a different set, which still meets the requirements of definition 3.1.1. Note that the polynomial approach will allow for this more general type of constraints, whereas the ordinal number approach requires this set be mapped by an isomorphism to the set of non-negative integers.

⁹Adding a minus sign would be possible in the case of the polynomial representation, but not possible for the ordinal numbers. Furthermore, maximising negative numbers is probably less intuitive than minimising positive numbers.

that is, the goal was there also to minimise the violation marks. Similarly, even definition 3.1.6 would be simpler (the usual definition of *lexicographic ordering*) if we reversed the \succ sign. Yet, there the motivation was to follow the OT concept of Harmony maximisation.

Fig. 3.1 summarises the different levels of representations.

Definition 3.1.14 requires us to add new items on our agenda. In turn, the introduction of the two new representations will follow this agenda:

1. Introduce the representation $E(w)$.
2. Define the relations \prec and $=$ on the range of E .
3. Prove that for all candidates w_1 and w_2 : $H(w_1) \succeq H(w_2)$ if and only if $E(w_1) \preceq E(w_2)$.¹⁰
4. Define the *difference* of two violation profiles.
5. Define *temperature* in a similar format.
6. Define the exponential of their quotient.
7. Introduce the SA-OT algorithm.

3.2 Violation profiles as real numbers

In what follows, we introduce two further representations of a violation profile. The goal of both approaches is to interpret Eq. (2.2) in the context of Optimality Theory, and thereby to implement simulated annealing.

Surprisingly or not, both approaches will result in the SA-OT algorithm already presented in Figure 2.8.

As an introduction, we try to realise violation profiles as real numbers. If it worked, SA-OT could be implemented as a real-valued optimisation problem. As it does not, we will have to proceed to the realisations using polynomials and ordinal numbers.¹¹ Yet, it is educative to understand why $H_{\mathcal{H}}(w)$ cannot

¹⁰We shall also demonstrate the law of trichotomy on the range of E , and therefore the two latter points of definition 3.1.14 can be summarised as above.

¹¹Gerhard Jäger pointed out that it is possible to define order preserving mappings from violation profiles into the real numbers. Their applicability to SA-OT should be tested in the future, even if these functions do not preserve the magnitude of differences between violation vectors, as defined in the previous chapter.

Jäger's proposal is based on the fact that the function $f(x) = 1 - (x+2)^{-1}$ is order preserving on non-negative reals and maps all positive numbers into the interval $(0, 1)$. Therefore, he proposes the following recursive definition, supposing that the violation levels $C_i(w)$ are non-negative integers:

$$\begin{aligned} g_0(w) &= C_0(w) \\ g_{i+1}(w) &= C_{i+1}(w) + f(g_i(w)) \\ E(w) &= g_n(w) \end{aligned}$$

A similar solution is $E(w) = \sum_{i=0}^n f_i(C_i(w))$, where $f_i(x) = 2^{2^i} - \frac{2^{2^i}}{x+1}$. Observe that $f_i(x)$ grows monotonously, $f_i(0) = 0$, $f_i(1) = 2^{2^i-1}$ and $\lim_{x \rightarrow \infty} f_i(x) = 2^{2^i}$, whence it is easy to show that constraints ranked lower than C_k can never accumulate a sum larger than the weight of a single violation of constraint C_k , for $\sum_{i=0}^{k-1} 2^{2^i} = \frac{4^k - 1}{4 - 1} > 2^{2^k-1}$ if $k > 0$.

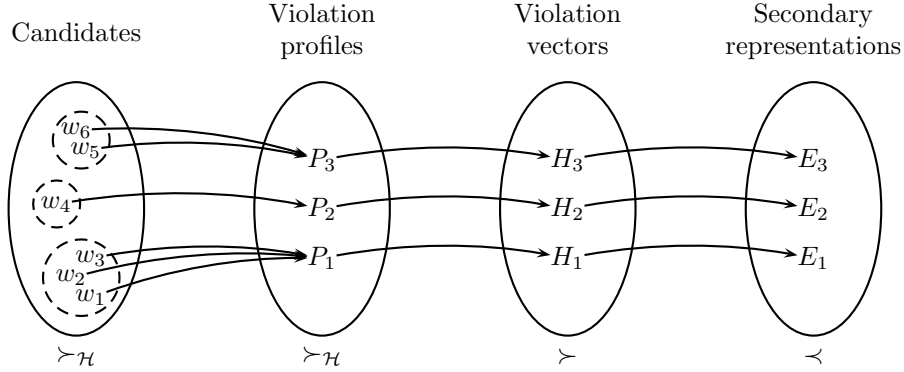


Figure 3.1: **Different levels of representation:** A candidate w incurs a certain number of violation marks from each of the constraints. Its violation profile $P(w)$ can be seen as a set of tokens of these marks. Given a certain hierarchy \mathcal{H} , the violation profile $P(w)$ corresponds to a vector $H_{\mathcal{H}}(w)$, introduced by equation (3.2). Finally, $H_{\mathcal{H}}(w)$ will be translated to different secondary representations $E_{\mathcal{H}}(w)$.

Definition 3.1.10 has introduced the order $\succ_{\mathcal{H}}$ and the equivalence relation \simeq on the candidate set. The relation \simeq defines equivalence classes on the candidate set (the dashed circles on the figure), which are then identified with the violation profiles: all elements of such an equivalence class have the same profile. Relation \simeq depends only on the definition of the constraints, which is universal, similarly to the candidate set. Thus, the set of violation profiles is also universal. The language dependent hierarchy \mathcal{H} determines the ranking $\succ_{\mathcal{H}}$ both on the candidate set and on the set of violation profiles.

This distribution of tasks is reversed in the right half of the figure. Different hierarchies map the same profile to different vectors. Therefore, different languages involve different subsets of the set of all possible vectors (that is, \mathbb{N}_0^{N+1} if the violation levels of the $N + 1$ constraints are the non-negative integers). The total order relation \succ used is however universal, as introduced in definition 3.1.6.

The secondary representations to be introduced will behave similarly. Each vector H_i will be mapped by an isomorphism onto some E_i . The order \prec on the E_i s is universal, but the hierarchy \mathcal{H} determines the specific value $E_{\mathcal{H}}(w)$ associated with the candidate w .

be realised as a real number, and this train of thought will lead us in a natural way to the subsequent proposals.

In the present section, as well as in section 3.3 (but not in 3.4), we could suppose that the violation levels are non-negative real numbers ($C_i(w) \in \mathbb{R}_0^+$ for any w and i) from the point of view of the definitions. Some of the theorems will, however, require that they are non-negative integers—which requirement is met by most applications in practice.

As mentioned in section 3.1, a crucial feature of Optimality Theory is *strict domination*: if a candidate is suboptimal for a higher ranked constraint, it can never win, even if it satisfies the lower ranked constraints best. Losing a battle means definitely being out of the game. Prince and Smolensky (2004) present on page 236 why a harmonic function $H(w)$ satisfying strict domination cannot be realised with a real-valued function.

Suppose first that there exists an upper bound $q - 1 > 0$ on the violation level a candidate can incur: for all $i \in \{N, \dots, 1, 0\}$ and for all $w \in \text{GEN}(UR)$, $0 \leq C_i(w) \leq q - 1$. (Note that this is exactly the condition required by the finite state approach of Frank and Satta (1998).) In such a case, the following real-valued Energy function $E(w)$ realises the Harmony function $H(w)$ perfectly:¹²

$$E(w) = C_N(w) \cdot q^N + C_{N-1}(w) \cdot q^{N-1} + \dots + C_1(w) \cdot q + C_0(w) \quad (3.4)$$

Following definition 3.1.14, we mean by $E(w)$ *realising* $H(w)$ that for all w_1 and w_2 , $E(w_1) \leq E(w_2)$ if and only if $H(w_1) \succeq H(w_2)$. In other words, optimising the Harmony function is equivalent to minimising the Energy function.

Indeed, equation (3.4) assigns candidate w a number $E(w)$ in a number system of base q whose digits are the violation levels. Informally speaking, this observation already proves that $E(w)$ defined accordingly realises strict domination.

Formally, we demonstrate this fact in two steps.

Lemma 3.2.1. *Given a hierarchy \mathcal{H} ($C_N \gg \dots \gg C_1 \gg C_0$), suppose that some $q \in \mathbb{R}$ exists such that for all constraints C_i and for all candidates w , the inequality $0 \leq C_i(w) \leq q - 1$ holds. Moreover, let $C_i(w) \in \mathbb{N}_0$, while $E_{\mathcal{H}}(w) = \sum_{i=0}^N C_i(w) \cdot q^i$. Then, with the Harmony function $H_{\mathcal{H}}(w)$, as defined in (3.2), for any two candidates w_1 and w_2 : if $H_{\mathcal{H}}(w_1) \succ H_{\mathcal{H}}(w_2)$, then*

$$E_{\mathcal{H}}(w_1) < E_{\mathcal{H}}(w_2) .$$

Proof. Following the definition 3.1.6, if $H_{\mathcal{H}}(w_1) \succ H_{\mathcal{H}}(w_2)$, then there exists a $k \in \{N, \dots, 0\}$ such that

$$E_{\mathcal{H}}(w_1) - E_{\mathcal{H}}(w_2) = \sum_{i=0}^k (C_i(w_1) - C_i(w_2)) \cdot q^i \quad (3.5)$$

¹²On page 61 we mentioned that the domains (the first component of the temperature, the indices of the constraints) are not necessarily consecutive integers, but can be arbitrary real numbers. Hence, a more general formulation of the real valued representation of a violation profile could be thus:

$$E(w) = \sum_{i \in \mathcal{I}} C_i(w) \cdot q^i$$

where \mathcal{I} is a finite set of real valued indices. We could but we shall not use this notation.

Moreover, $C_k(w_1) - C_k(w_2) < 0$. As the violation levels are integers,

$$C_k(w_1) - C_k(w_2) \leq -1 \quad (3.6)$$

Recall the sum of a geometric series:

$$\sum_{i=0}^{k-1} q^i = \frac{q^k - 1}{q - 1} \quad (3.7)$$

Therefore, and because 0 and $q - 1$ are lower and upper bounds on the number of violation marks ($C_i(w_1) - C_i(w_2) \leq q - 1$):

$$\sum_{i=0}^{k-1} (C_i(w_1) - C_i(w_2)) \cdot q^i \leq (q - 1) \frac{q^k - 1}{q - 1} \quad (3.8)$$

Summarising, from (3.5), (3.6) and (3.8), we obtain:

$$E_{\mathcal{H}}(w_1) - E_{\mathcal{H}}(w_2) \leq -q^k + q^k - 1 < 0 \quad (3.9)$$

That is, $E_{\mathcal{H}}(w_1) < E_{\mathcal{H}}(w_2)$. \square

Next, we can prove that the Harmony function can be realised with real numbers under some specific conditions:

Theorem 3.2.2. *Given a hierarchy \mathcal{H} ($C_N \gg \dots \gg C_1 \gg C_0$), suppose that some $q \in \mathbb{R}$ exists such that for all constraints C_i and for all candidates w , the inequality $0 \leq C_i(w) \leq q - 1$ holds. Moreover, $C_i(w) \in \mathbb{N}_0$. Then, the energy function*

$$E_{\mathcal{H}}(w) = \sum_{i=0}^N C_i(w) \cdot q^i$$

realises the Harmony function $H_{\mathcal{H}}(w)$, as defined in (3.2). That is, for any two candidates w_1 and w_2 ,

- $H_{\mathcal{H}}(w_1) = H_{\mathcal{H}}(w_2)$ iff $E_{\mathcal{H}}(w_1) = E_{\mathcal{H}}(w_2)$
- $H_{\mathcal{H}}(w_1) \succ H_{\mathcal{H}}(w_2)$ iff $E_{\mathcal{H}}(w_1) < E_{\mathcal{H}}(w_2)$

Proof. This theorem comprises four statements. We have already demonstrated in lemma 3.2.1 that if $H_{\mathcal{H}}(w_1) \succ H_{\mathcal{H}}(w_2)$ then $E_{\mathcal{H}}(w_1) < E_{\mathcal{H}}(w_2)$.

Furthermore, if $H_{\mathcal{H}}(w_1) = H_{\mathcal{H}}(w_2)$ then $E_{\mathcal{H}}(w_1) = E_{\mathcal{H}}(w_2)$. Namely, due to the second part of definition 3.1.6, each coefficient in the definition of $E_{\mathcal{H}}(w_1)$ and of $E_{\mathcal{H}}(w_2)$ are equal.

Suppose now that $E_{\mathcal{H}}(w_1) = E_{\mathcal{H}}(w_2)$. By the law of trichotomy (theorem 3.1.8), either $H_{\mathcal{H}}(w_1) \succ H_{\mathcal{H}}(w_2)$ or $H_{\mathcal{H}}(w_2) \succ H_{\mathcal{H}}(w_1)$ or $H_{\mathcal{H}}(w_1) = H_{\mathcal{H}}(w_2)$. We have already demonstrated that the first two possibilities would involve $E_{\mathcal{H}}(w_1) \neq E_{\mathcal{H}}(w_2)$, which leaves us the only possibility of $H_{\mathcal{H}}(w_1) = H_{\mathcal{H}}(w_2)$.

Similarly, suppose now that $E_{\mathcal{H}}(w_1) < E_{\mathcal{H}}(w_2)$. Because the law of trichotomy also applies on the set of real numbers with the usual $>$ relation, this supposition would be contradicted if $H_{\mathcal{H}}(w_1) \succ H_{\mathcal{H}}(w_2)$ did not hold, but one of the two other possibilities in theorem 3.1.8. \square

3.3 Violation profiles as polynomials

However, nothing in the general theory of Optimality Theory guarantees that such an upper bound $q - 1$ exists.¹³ The behaviour of an energy function (3.4) with some q only approximates the behaviour of the Harmony function.¹⁴

Then, why not consider the behaviour of this polynomial as q goes to infinity?¹⁵ We propose to see the violation profiles as a polynomials of $q \in \mathcal{R}^+$ ($q > 0$):¹⁶

$$E(w)[q] = C_N(w) \cdot q^N + C_{N-1}(w) \cdot q^{N-1} + \dots + C_1(w) \cdot q + C_0(w) \quad (3.10)$$

This equation defines the *polynomial representation* of a violation profile: each candidate, or each violation profile is realised as a real-valued polynomial of q . The energy or the Eval-function $E(w)$ is not any more a real number, but a function mapping \mathbb{R} to \mathbb{R} . It is $E(w)[q]$, but not $E(w)$ which is in \mathbb{R} .

This proposal is opposed to seeing a violation profile as a real number, as a vector, or as some other construct. Namely, equations (2.10) and (3.2) introduced the *vector representation* of a violation profile. Equation (3.4), for a constant q , attempted to introduce a *real valued representation* (different qs would correspond to different representations); even though we have just seen that this approach would not work in the general case. The next section presents how to realise a profile as an ordinal ("infinite") number (cf. equation (3.19)). All these representations correspond to different ways of defining the rightmost set ($\{E(w) \mid w \in \text{GEN}(UR)\}$) on Fig. 3.1.

3.3.1 Comparing polynomials

This new representation now requires us to introduce the relations \prec and $=$ on the range of the representation E . As the $=$ relation is simply the identity relation, introducing the order \prec is always the less trivial task. So far, we used the lexicographic order on the vector representations, and the everyday "less than" relation on the real valued representation. How shall we deal now with the polynomial representations?

Obviously, $E(w)[q]$ goes to infinity as q grows without bound:

$$\lim_{q \rightarrow +\infty} E(w)[q] = +\infty$$

Therefore, observing directly the limit of $E(w)[q]$ will not work, whatever we would like to do with the violation profiles. (First, we will aim at defining the \prec relation in order to prove the soundness of our approach. And then, we

¹³Notice that this problem arises only if the candidate set corresponding to a certain input is infinite. Otherwise the real valued representation would work, even if different inputs required different qs . In unidirectional Optimality Theory, the candidate sets of different inputs may overlap, but do not interact with each other.

¹⁴For cases when any monotonically decreasing series of weights can be used, see Prince (2002).

¹⁵The idea of using polynomial arithmetics originates from Balázs Szendrői.

¹⁶The more general formulation mentioned in footnote 12 looks as:

$$E(w)[q] = \sum_{i \in \mathcal{I}} C_i(w) \cdot q^i$$

shall interpret Eq. (2.2) for SA-OT.) The trick will always be *first* to perform an operation, or *first* to check the behaviour of the energy function, and only *subsequently* bring q to the infinity. In using *continuous* operations, it makes sense to change the order of the operation and of the limit to infinity.

First, how shall we compare two violation profiles seen as polynomials? The following definition—comparing the limits—is useless: $E(w_1) \prec E(w_2)$ if and only if

$$\lim_{q \rightarrow +\infty} E(w_1)[q] < \lim_{q \rightarrow +\infty} E(w_2)[q]$$

We may, however, consider the limit of the comparisons, instead of the comparison of the limits. The following definition works consequently perfectly, that is, it realises the harmony function:

Definition 3.3.1. $E(w_1) \prec E(w_2)$ if and only if either

$$\lim_{q \rightarrow +\infty} (E(w_2)[q] - E(w_1)[q]) > 0$$

or

$$\lim_{q \rightarrow +\infty} (E(w_2)[q] - E(w_1)[q]) = +\infty$$

Furthermore, $E(w_1) = E(w_2)$ if and only if $E(w_1)[q] = E(w_2)[q]$ for all $q \in \mathbb{R}^+$.

By using the definition of the limits and the properties of polynomials,¹⁷ we may reformulate the first part of this definition thus:

Corollary 3.3.2. $E(w_1) \prec E(w_2)$ if and only if there exists a $q_0 \in \mathbb{R}$ such that for all $q \in \mathbb{R}^+$: if $q > q_0$ then $E(w_2)[q] - E(w_1)[q] > 0$.

In other words, for any two candidates one can choose a q that is high enough so that we can simply compare the “energies” as real values. The problem with the real valued representation was that no single q exists that would always work perfectly. But the polynomial representation allows for choosing different q s for any two candidates w_1 and w_2 , and hence we have circumvented the problem.

Indeed, energy-polynomials with this definition realise the Harmony function: $E(w_1) \prec E(w_2)$ if and only if $H(w_1) \prec H(w_2)$. Similarly, $H(w_1) = H(w_2)$ if and only if $E(w_1) = E(w_2)$ (that is, $E(w_1)[q] = E(w_2)[q]$ for all $q \in \mathbb{R}^+$). We are going to prove this *equivalence* of the Harmony function to the energy polynomials in three steps.

First we demonstrate the *law of trichotomy* on the set $\{E(w) | w \in \text{GEN}(UR)\}$ with respect to the relation \prec . Namely:

Theorem 3.3.3. LAW OF TRICHOTOMY FOR THE ENERGY POLYNOMIALS: *for all candidates w_1 and $w_2 \in \text{GEN}(UR)$, exactly one of the following three statements hold: either $E(w_1) \prec E(w_2)$, or $E(w_2) \prec E(w_1)$, or $E(w_1) = E(w_2)$.*

Proof. First, recall first that polynomials are continuous functions, and that a basic property of continuous functions is that they map an interval onto an interval. In other words, if the continuous function $f(x)$ is defined on the interval $[a, b]$, and X is in the interval $[f(a), f(b)]$ (or $[f(b), f(a)]$, depending on whether

¹⁷Namely, the fact that if a real valued polynomial $P(x)$ is not constant, then it converges to infinity: $\lim_{x \rightarrow +\infty} P(x) = \pm\infty$.

$f(a) \leq f(b)$ or $f(b) \leq f(a)$ holds), then there exists an $x \in [a, b]$ such that $f(x) = X$.

If $E(w_1) = E(w_2)$, then by definition, $\lim_{q \rightarrow +\infty} (E(w_2)[q] - E(w_1)[q]) = 0$, so neither $E(w_1) \prec E(w_2)$ nor $E(w_2) \prec E(w_1)$.

Suppose now that $E(w_1) \neq E(w_2)$, thus we have to demonstrate that exactly one of the first two statements applies. In this case, $P[q] := E(w_1)[q] - E(w_2)[q]$ is a polynomial, which is not constant zero, and whose order is maximally N (the order of $E(w_1)[q]$ and $E(w_2)[q]$). Such a function may have maximally N different roots, that is $q_{(i)}$ values rendering it zero: $P[q_{(i)}] = 0$.

If no such real valued root exists, $P[q]$ is either positive or negative for all q s. Otherwise, $P[q_1] > 0$ and $P[q_2] < 0$ for some q_1 and q_2 would force $P[q]$ to have a root, due to the property of the continuous functions mentioned at the beginning of this proof. In turn, any $q_0 \in \mathbb{R}$ can be chosen to show by corollary 3.3.2 that $E(w_1) \prec E(w_2)$ if $P[q]$ is always negative, and that $E(w_2) \prec E(w_1)$ if $P[q]$ is always positive.

If, on the other hand, $P[q]$ does have at least one root, let q_0 be the greatest root. Now, for all $q > q_0$ the value of $P[q]$ has the same sign (always positive, or always negative): if there existed a $q_1 > q_0$ such that $P[q_1] > 0$ and another $q_2 > q_0$ such that $P[q_2] < 0$, then $P[q]$ would have a root greater than q_0 , between q_1 and q_2 , again because $P[q]$ is a continuous function. Consequently, either $P[q] = E(w_1)[q] - E(w_2)[q] > 0$ for all $q > q_0$, proving that $E(w_2) \prec E(w_1)$; or $P[q] = E(w_1)[q] - E(w_2)[q] < 0$ for all $q > q_0$, and then $E(w_1) \prec E(w_2)$, by corollary 3.3.2. \square

In the next step, we demonstrate that

Lemma 3.3.4. *If $H(w_1) \succ H(w_2)$, then $E(w_1) \prec E(w_2)$.*

Proof. If $H(w_1) \succ H(w_2)$, then, by definition, there exists an integer $k \in [N, N-1, \dots, 1, 0]$ such that

1. $C_k(w_2) - C_k(w_1) > 0$, and
2. for all $i \in [N, N-1, \dots, 1, 0]$: if $i > k$, then $C_i(w_2) - C_i(w_1) = 0$.

If $k = 0$ then for all q

$$E(w_2)[q] - E(w_1)[q] = \sum_{i=0}^N [C_i(w_2) - C_i(w_1)]q^i = C_k(w_2) - C_k(w_1) > 0$$

Therefore $E(w_1) \prec E(w_2)$, and any $q_0 \in \mathbb{R}$ may be chosen.

In the case, however, when $k > 0$, let us define c such that for all $i < k$: $c > C_i(w_1), C_i(w_2) \geq 0$. Such a c exists because a finite number of violation levels always have a finite upper bound. First note that for all $i < k$:

$$c > C_i(w_2) - C_i(w_1) > -c \quad (3.11)$$

Second, remember the sum of a geometric series ($q \neq 1$):

$$\sum_{i=0}^{k-1} q^i = \frac{q^k - 1}{q - 1} \quad (3.12)$$

Now let $q_0 = \max(\frac{2c}{C_k(w_2) - C_k(w_1)}, 2)$. For all $q > q_0$, then

$$\begin{aligned} E(w_2)[q] - E(w_1)[q] &= \sum_{i=0}^N [C_i(w_2) - C_i(w_1)]q^i = \\ &= [C_k(w_2) - C_k(w_1)]q^k + \sum_{i=0}^{k-1} [C_i(w_2) - C_i(w_1)]q^i \end{aligned} \quad (3.13)$$

due to the definition of the fatal constraint C_k . Because $q > q_0 \geq \frac{2c}{C_k(w_2) - C_k(w_1)}$, in the first component of the sum we can use $C_k(w_2) - C_k(w_1) > 2c/q$. For the second component, we may use equations (3.11) and (3.12). In turn, we obtain:

$$\begin{aligned} E(w_2)[q] - E(w_1)[q] &> \frac{2c}{q}q^k - c\frac{q^k - 1}{q - 1} = \\ &= \frac{c}{q - 1}[q^k - 2q^{k-1} + 1] > 0 \end{aligned} \quad (3.14)$$

because $q > q_0 \geq 2$.

In sum, either $k = 0$ or $k > 0$, we have shown that there exists a q_0 such that for all $q > q_0$: $E(w_2)[q] - E(w_1)[q] > 0$. Therefore, $E(w_1) \prec E(w_2)$. \square

Observe that the present proof did not require $C_i(w)$ be an integer, unlike the proof of the corresponding lemma for the real-number representation (Lemma 3.2.1). The reason of this difference is that now, if $C_k(w_2) - C_k(w_1) < 1$, we could simply increase q_0 . Similarly, Lemma 3.2.1 (and hence, Theorem 3.2.2) could be generalised, if a positive lower bound existed for the difference of different violation levels of the constraints. Nevertheless, the real-number representation requires a universal upper bound on the violation levels and a global lower bound on the difference of the violation levels in order to specify some q , the base of the exponential weight system. The advantage of the polynomial approach is that q is handled in a flexible way, and thus a different q_0 (or any $q > q_0$) can be used for any pair of candidates. A pair of candidates has a finite number of violation levels, which guarantees the existence of the required upper and lower bounds.

Third, we can formulate and prove that energy polynomials realise Harmony function, if using definitions 3.1.6 and 3.3.1:

Theorem 3.3.5. ENERGY POLYNOMIALS REALISE THE HARMONY FUNCTION:
 $E(w_1) = E(w_2)$ if and only if $H(w_1) = H(w_2)$;
 $E(w_1) \prec E(w_2)$ if and only if $H(w_1) \succ H(w_2)$.

Proof. This statement includes four substatements. First, if $H(w_1) = H(w_2)$, then, by definition, $C_i(w_1) = C_i(w_2)$ for all $i \in [N, N-1, \dots, 1, 0]$. Consequently, for all $q \in \mathcal{R}^+$, $E(w_1)[q] = E(w_2)[q]$.

Second, if $H(w_1) \succ H(w_2)$, then $E(w_1) \prec E(w_2)$, as demonstrated by the previous lemma.

Third, if $E(w_1) = E(w_2)$, then $H(w_1) = H(w_2)$. This is true, because either $H(w_1) = H(w_2)$, or $H(w_1) \succ H(w_2)$, or $H(w_2) \succ H(w_1)$, due to the law of trichotomy on vectors (theorem 3.1.8). Using an indirect proof, suppose $H(w_1) \succ H(w_2)$. As just shown, $E(w_1) \prec E(w_2)$ would follow, which would

contradict the law of trichotomy for the energy polynomials (theorem 3.3.3). Similarly, $H(w_2) \succ H(w_1)$ is also impossible, leaving us the only possibility $H(w_1) = H(w_2)$.¹⁸

Fourth, if $E(w_1) \prec E(w_2)$, then $H(w_1) \succ H(w_2)$. This can be demonstrated similarly to the third case, by referring to the laws of trichotomy for the Harmony function and for the energy polynomials. Namely, $H(w_2) \succeq H(w_1)$ would require $E(w_1) \preceq E(w_2)$ by the statements of the present theorem already demonstrated, which would contradict the law of trichotomy on polynomials (Theorem 3.3.3). \square

3.3.2 Simulated annealing with polynomials

So far, we have seen that energy polynomials can be used to model the behaviour of the Harmony function of Optimality Theory: their definition is sound and they realise the Harmony function. The trick was first to compare two candidates, and only then take the $q \rightarrow \infty$ limit.

Can we use energy polynomials to formulate simulated annealing for Optimality Theory? The recurring problem has been how to define the transition probability $P(w \rightarrow w')$ from candidate w to a worse candidate w' in a form that is reminiscent of the traditional expression $e^{(E(w')-E(w))/T}$. How would we define the transition probabilities in the polynomial representation of violation profiles?

Using the polynomial representation of two violation profiles, translating the expression $E(w') - E(w)$ is straightforward. It is simply another polynomial of q , namely $P[q] = E(w')[q] - E(w)[q]$. The difference of two real valued function is given for free by elementary school arithmetic. Observe that if C_k is the fatal constraint, the highest ranked constraint—the constraint with the highest index—that assigns different violation levels to w and w' , then the dominant component in $P[q]$ is q^k .

Temperature in OT simulated annealing, as explained in section 2.2.3, should have a structure similar to that of the difference of two violation profiles. If presently the difference $E(w') - E(w)$ is a polynomial of $q \in \mathbb{R}^+$, so must the temperature $T = \langle K, t \rangle$ be, as well:

$$T[q] = \langle K, t \rangle [q] = t \cdot q^K \quad (3.15)$$

The attentive reader will notice that this formulation allows K to be a real number, not only an integer, if $q > 0$.¹⁹ Nonetheless, we shall not really exploit this opportunity, besides the fact that we theoretically allow any real K values in the outer cycle of the SA-OT algorithm (Fig. 2.8).

¹⁸An alternative proof of this third substatement would employ the fact that a constant zero polynomial—such as $P[q] := E(w_1)[q] - E(w_2)[q]$ in the case $E(w_1) = E(w_2)$ —must have but zero coefficients. Thus, $C_i(w_1) - C_i(w_2) = 0$ for all $i \in [N, N-1, \dots, 1, 0]$, yielding $H(w_1) = H(w_2)$ by Definition 3.1.6.

¹⁹The polynomials proposed to represent violation profiles have been defined on the domain of positive real numbers ($q \in \mathbb{R}^+$). Although this restriction might have appeared to be unnecessary, now we can see its advantage. Besides, this restriction also allows generalising the polynomial representation to the case if the indices of the constraints are real numbers. Furthermore, q was originally the base of an exponential weight system in (3.4), which makes sense only if $q > 1$. In any case, as only the $q \rightarrow +\infty$ limit will be of interest, we can always remove a lower subset of q 's domain.

The last step is to formulate the probability of moving from candidate w to a neighbour candidate w' . If $w' \succeq w$, the probability is 1. Otherwise, we shall repeat the trick: *first* perform the operation, and only *afterwards* take the $q \rightarrow \infty$ limit.

Thus, the probability of moving from a candidate w to a worse candidate w' shall be defined as:

$$P(w \rightarrow w') = \lim_{q \rightarrow +\infty} e^{-\frac{E(w')[q] - E(w)[q]}{T[q]}} \quad (3.16)$$

Analysing the defining equation (3.16), one can quickly check that this definition yields the RULES OF MOVING on page 63, which we have been hoping for:

- If w' better than w : move $w \rightarrow w'$!
- If w' loses due to constraint $C_k > K$: don't move ($P = 0$)!
- If w' loses due to constraint $C_k < K$: move ($P = 1$)!
- If w' loses due to the constraint $C_k = K$: move with transition probability $P(w \rightarrow w') = e^{-(C_k(w') - C_k(w))/t}$.

This is so because the mathematical operations involved are continuous. Further, as q grows very large, the dominant component in $E(w')[q] - E(w)[q]$ will be the highest non-zero component, which is $(C_k(w') - C_k(w))q^k$ where C_k is the fatal constraint when comparing these two candidates:

$$\begin{aligned} P(w \rightarrow w') &= \lim_{q \rightarrow +\infty} e^{-\frac{E(w')[q] - E(w)[q]}{T[q]}} = \\ &= \lim_{q \rightarrow +\infty} e^{-\frac{(C_k(w') - C_k(w))q^k}{tq^K}} \\ &= \left[\lim_{q \rightarrow +\infty} e^{(-q^{k-K})} \right]^{\frac{C_k(w') - C_k(w)}{t}} \\ &= \begin{cases} 0 & \text{if } k > K \\ e^{-\frac{C_k(w') - C_k(w)}{t}} & \text{if } k = K \\ 1 & \text{if } k < K \end{cases} \quad (3.17) \end{aligned}$$

For a visualisation, recall Fig. 2.6. The expression $e^{(-q^\alpha)}$ is equal to e^{-1} if $\alpha = 0$. If however $\alpha < 0$, then it converges to 1 with $q \rightarrow +\infty$, similarly to the function $e^{-1/x} = e^{(-x^{-1})}$ on Fig. 2.6. In the third case, that is when $\alpha > 0$, the expression e^{-q^α} converges to 0, because this case corresponds to the $x \rightarrow +0$ limit of the function $e^{-1/x}$ (replace x with $q^{-\alpha}$).

In (3.15), we could have used a more complex expression as the definition of $T[q]$, but the form $t \cdot q^K$ will be good enough. As we take the $q \rightarrow +\infty$ limit, where only the highest component of a polynomial plays a role, adding lower components would not influence the behaviour of the system. Temperature could also have been defined not as a polynomial, but as a different function of q . Nevertheless, if $T[q]$ did not converge like some polynomial ($T[q] = \mathcal{O}(q^K)$), it would not turn useful in the equation 3.16 defining the transition probability, for the latter would always be 0 (if $T[q] = o(q^K)$) or 1 (if $T[q]/(q^K) \rightarrow \infty$). This

is the reason why we required $T[q]$ to be a polynomial of the form appearing in (3.15). In sum, the polynomial approach also advocates temperature to be a pair $T = \langle K, t \rangle$, similarly to the approach proposed in section 2.2.3, on page 58.

3.4 Violation profiles as ordinal numbers

In the present section, an alternative way is presented to introduce *Optimality Theory Simulated Annealing*. Instead of considering real-valued polynomials $E(w)[q]$ in the limit $q \rightarrow +\infty$, we immediately take *infinite weights* for q .

As demonstrated, no finite weights can reproduce, in the general case, the *strict constraint ranking* postulated by Optimality Theory. A series of exponential weights, such as in

$$E(w) := \sum_{i=0}^N C_i(w) q^i \quad (3.18)$$

realises the constraint hierarchy $C_N \gg \dots \gg C_1 \gg C_0$ only if each constraint can assign at most $q - 1$ violation marks to any candidate.

However, the number of violation marks assigned by most constraints used in linguistics does not have any upper bound theoretically. Even if one argues that performance usually limits the length of words and sentences that can be uttered, still, linguistic models can require generating never winning candidates of an unbounded length. It would be nice consequently to allow unbounded weights in (3.18). In other words, to let equation (3.18) introduce a value $E(w)$ in a number system of infinite base.

Axiomatic Set Theory proposes a solution to carry out this idea in a mathematically sound framework. When the possible levels of violation formed the well ordered set $\{0, 1, 2, \dots, q - 1\}$ —which is the definition of the integer q (Holz et al., 1999, p. 19).—, we used q as the base of an exponential weight system. In the case of unbounded violations, the possible levels of violation most often form the ordered set $\{0, 1, 2, \dots\}$. This well ordered set is called ω , the first limit ordinal (Suppes, 1972; Holz et al., 1999). In other words, ω is the upper limit of the set of the natural numbers \mathcal{N} .

Arithmetic can be defined on ordinal numbers, including comparison, as well as addition and multiplication (Holz et al., 1999).²⁰ These latter operations are associative, but not commutative. Therefore, we can redefine the E function as:²¹

²⁰See also references under: Eric W. Weisstein: *Ordinal Number*, From MathWorld—A Wolfram Web Resource, <http://mathworld.wolfram.com/OrdinalNumber.html>.

²¹Footnotes 12 and 16 proposed a more general formulation, which would translate now as:

$$E_{\mathcal{H}}(w) = \sum_{i \in \mathcal{I}} \omega^i \cdot C_i(w)$$

with the important caveat that the elements of the finite set of indices \mathcal{I} are ordinal numbers (practically: non-negative integers). Further, as ordinal addition is not commutative, we have to specify that the elements of \mathcal{I} are read in a decreasing order. This formulation naturally allows us not using certain numbers as indices; whereas in (3.19) one has to stipulate $C_j = 0$ in the case we would like to associate no “real” constraint with the index j .

$$\begin{aligned}
E_{\mathcal{H}}(w) &= \omega^N C_N(w) + \dots + \omega C_1(w) + C_0(w) \\
&= \sum_{i=N}^0 \omega^i C_i(w)
\end{aligned} \tag{3.19}$$

This expression introduces the *ordinal number representation* of a violation profile for hierarchy \mathcal{H} . We will nonetheless dismiss the index \mathcal{H} , as long as we work with a constant constraint ranking. Observe that unlike in the polynomial representation, the violation levels must be ordinals, such as non-negative integers, in order to 3.19 be meaningful.

Because ω is the upper limit of the natural numbers, $\omega^i n < \omega^{i+1}$ for any finite n . This property will guarantee that if candidate w_1 is less harmonic than candidate w_2 then $E(w_1) > E(w_2)$. In other words, ordinal arithmetic furnishes us with the relation $<$ and $=$ for free in the ordinal number representation of a violation profile.

3.4.1 Ordinal numbers can realise violation profiles

We heavily rely on results demonstrated by Holz et al. (1999), while we are proving trichotomy and representation:

Lemma 3.4.1. TRICHOTOMY ON ORDINAL NUMBERS: *Let σ and τ be ordinal numbers. Then exactly one of the following three statements hold: 1. $\sigma < \tau$; 2. $\sigma = \tau$; 3. $\tau < \sigma$.*

Proof. Lemma 1.2.3 in Holz et al. (1999, p. 16) demonstrates that for any two ordinal numbers at least one of the three statements holds. Lemma 1.2.1.c states that $\sigma \not< \sigma$; hence statements 1 and 2, as well as statements 2 and 3 cannot simultaneously hold. Similarly, by the latter lemma and by the transitivity of the $<$ relation, statements 1 and 3 cannot hold in the same time. \square

Lemma 3.4.2. *Let w_1 and w_2 be two candidates, and let $E(w_1)$ and $E(w_2)$ be the ordinal number representation of their violation profile with respect to some hierarchy \mathcal{H} . If $H(w_1) \succ_{\mathcal{H}} H(w_2)$, then $E(w_1) < E(w_2)$.*

Proof. As $H(w_1) \succ_{\mathcal{H}} H(w_2)$, and violation levels are integers, there is a constraint C_k such that

- [1] $C_k(w_1) + 1 \leq C_k(w_2)$, and
- [2] $C_i(w_1) = C_i(w_2)$ for all $i > k$.

Lemma 1.4.3 in Holz et al. (1999, p. 33) contains among others the following properties, if α , β and γ are cardinal numbers:

- [3] if $0 < \alpha$ and $\beta < \gamma$, then $\alpha \cdot \beta < \alpha \cdot \gamma$
- [4] if $\beta < \gamma$, then $\alpha + \beta < \alpha + \gamma$
- [5] $\alpha^{\beta+\gamma} = \alpha^{\beta} \cdot \alpha^{\gamma}$
- [6] if $1 < \alpha$ and $\beta < \gamma$, then $\alpha^{\beta} < \alpha^{\gamma}$

$$[7] \quad \alpha \cdot (\beta + \gamma) = \alpha \cdot \beta + \alpha \cdot \gamma$$

$$[8] \quad \alpha \cdot 1 = \alpha = 1 \cdot \alpha$$

From properties [1] and [3] it follows that $\omega^k(C_k(w_1) + 1) \leq \omega^k C_k(w_2)$. Therefore, and due to [2], [4], [7] and [8],

$$\sum_{i=N}^k \omega^i C_i(w_1) + \omega^k \leq \sum_{i=N}^k \omega^i C_i(w_2) \quad (3.20)$$

Furthermore, from $0 \leq C_i(w_1) < \omega$, by [3] and [5], follows that $\omega^j \cdot C_i(w_1) < \omega^j \cdot \omega = \omega^{j+1}$ for any i and j . Hence, due to [6], $\omega^j \cdot C_i(w_1) < \omega^k$ for any $j < k$ (that is, $j + 1 \leq k$).

From Lemma 1.4.7.b of Holz et al. (1999, p. 37) follows that for all j and $\alpha < \omega^j$, $\alpha + \omega^j = \omega^j$. Now, if both $\alpha < \omega^j$ and $\beta < \omega^j$, property [4] ensures that $\alpha + \beta < \alpha + \omega^j = \omega^j$. That is, the sum of any two ordinals smaller than ω^j is smaller than ω^j . Using this observation recursively in the case of $j = k$, we obtain:

$$\sum_{i=k-1}^0 \omega^i C_i(w_1) < \omega^k \quad (3.21)$$

From (3.21) and (3.20), by using repeatedly [4]:

$$\begin{aligned} E(w_1) &= \sum_{i=N}^k \omega^i C_i(w_1) + \sum_{i=k-1}^0 \omega^i C_i(w_1) < \\ &< \sum_{i=N}^k \omega^i C_i(w_1) + \omega^k \leq \\ &\leq \sum_{i=N}^k \omega^i C_i(w_2) \leq \\ &\leq \sum_{i=N}^0 \omega^i C_i(w_2) = E(w_2) \end{aligned} \quad (3.22)$$

□

Now, we can formally prove that the representation of a violation profile using ordinal numbers is isomorphic to the vector representation:

Theorem 3.4.3. ORDINAL NUMBERS REALISE VIOLATION PROFILES: *Let w_1 and w_2 be two candidates, and let $E(w_1)$ and $E(w_2)$ be the ordinal number representation of their violation profile with respect to some hierarchy \mathcal{H} . Then,*

- $E(w_1) = E(w_2)$ if and only if $H(w_1) = H(w_2)$;
- $E(w_1) < E(w_2)$ if and only if $H(w_1) \succ_{\mathcal{H}} H(w_2)$.

Proof. This theorem contains four substatements. If $H(w_1) = H(w_2)$, then by definition, $C_i(w_1) = C_i(w_2)$ for all i , and therefore $E(w_1) = E(w_2)$ follows directly.

If $H(w_1) \succ_{\mathcal{H}} H(w_2)$, then we have just demonstrated in the previous lemma that $E(w_1) < E(w_2)$.

If $E(w_1) = E(w_2)$, then Theorem 1.4.6 of Holz et al. (1999, p. 36) (Cantor Normal Form for the base ω) ensures that $C_i(w_1) = C_i(w_2)$ for all i s. Namely, from the theorem follows that if

$$\omega^N \cdot a_N + \omega^{N-1} \cdot a_{N-1} + \dots \omega^0 \cdot a_0 = \omega^N \cdot b_N + \omega^{N-1} \cdot b_{N-1} + \dots \omega^0 \cdot b_0$$

then $a_i = b_i$ for all i s. Consequently, $H(w_1) = H(w_2)$.

Last, if $E(w_1) < E(w_2)$, then $H(w_1) \succ H(w_2)$. Suppose this does not hold. Then $H(w_2) \succeq H(w_1)$ should be true, because of the trichotomy on the set of violation profile-like vectors (Theorem 3.1.8). This, however would involve $E(w_2) \leq E(w_1)$ due to the previously proven parts of the present theorem, which in turn would contradict the trichotomy on the class of ordinal numbers (Lemma 3.4.1). (A similar proof is also possible for the third substatement of the present theorem, if you would like to avoid the Cantor Normal Forms; cf. the relevant part of the proof of theorem 3.3.5.) \square

3.4.2 SA-OT with ordinal numbers

The next step towards SA-OT is the definition of the difference of two $E(w)$ values, which will pave the way for the introduction of temperature, necessary to interpret the expression $e^{-(E(w')-E(w))/T}$ in the context of Optimality Theory.

On the class ON of all ordinal numbers, subtraction is not defined as it is defined on the set \mathbb{Z} of the integers, or on the set \mathbb{R} of the real numbers. The class ON of all ordinal numbers can be seen as a generalisation of the natural numbers (non-negative integers), and observe that the difference $a - b$ of two natural numbers a and b is defined on the set \mathbb{N} only if $a \geq b$.

(Holz et al., 1999, p. 34) proves the following

Lemma 3.4.4. SUBTRACTION LEMMA *If $\alpha \leq \beta$ are ordinal numbers, then there is a unique ordinal γ such that $\alpha + \gamma = \beta$.*

Based on this lemma, we introduce the notation $\Delta(a, b)$ for ordinals $a \geq b$, to denote the unique ordinal x that satisfies $a = b + x$. As addition is not commutative on the class of ordinals ON, $a = \Delta(a, b) + b$ does not follow (and usually does not hold) from $a = b + \Delta(a, b)$. The notation $a - b$ and the term “subtraction” will be avoided in order to remind us this caveat, as well as the fact that $\Delta(a, b)$ is defined only if $a \geq b$.

Violation profiles are represented with a subset of ON, namely, with ordinals of the form $a = \sum_{i=N}^0 \omega^i a_i$, where $a_i \in \mathbb{N}_0$ (a_i is a non-negative integer). Thus, the elements of the set $\sum_{i=N}^0 \omega^i \mathbb{N}_0$ will be referred to as *violation profile-like ordinal numbers*.

The following proposition sheds light on how ordinal numbers represent violation profiles:

Proposition 3.4.5. *Given violation profile-like ordinals $a = \sum_{i=N}^0 \omega^i a_i$ and $b = \sum_{i=N}^0 \omega^i b_i$, such that $a > b$,*

$$\Delta(a, b) = \sum_{i=N}^0 \omega^i \delta_i$$

where for all $0 \leq i \leq N$

$$\delta_i = \begin{cases} a_i - b_i & \text{if } a_j = b_j \forall j. (j > i \wedge j \leq N) \\ a_i & \text{otherwise} \end{cases}$$

Proof. As the Subtraction lemma 3.4.4 proves uniqueness, it is satisfactory to show that $a = b + \sum_{i=N}^0 \omega^i \delta_i$.

Recall that ordinal addition is associative (Lemma 1.4.3.a.(v) in Holz et al. (1999, p. 33)), as well as that $\omega^i a + \omega^j b = \omega^j b$ if $i < j$.²²

Let k be the lowest index to which $\forall j > k : a_j = b_j$ holds (in the case of violation profiles, this is the index of the fatal constraint). Such a k exists, because the set $\{0, \dots, N\}$ is finite, hence well-ordered: each set, for instance the set $\{i \in \{0, \dots, N\} \mid \forall j \in \{0, \dots, N\} : (j > i) \Rightarrow (a_j = b_j)\}$, has a least element. Then,

$$\begin{aligned} b + \Delta(a, b) &= \sum_{i=N}^0 \omega^i b_i + \sum_{i=N}^0 \omega^i \delta_i = \\ &= \sum_{i=N}^k \omega^i b_i + \left(\sum_{i=k-1}^0 \omega^i b_i + \omega^k \delta_k \right) + \sum_{i=k-1}^0 \omega^i \delta_i = \\ &= \sum_{i=N}^{k+1} \omega^i a_i + \omega^k b_k + \omega^k (a_k - b_k) + \sum_{i=k-1}^0 \omega^i a_i = a \quad (3.23) \end{aligned}$$

□

In the case of violation profile-like ordinal numbers, the co-efficient δ_k of the highest non-zero term in $\Delta(a, b)$ is the difference of the respective terms in a and b . In OT, this co-efficient will reflect the difference of violation marks (i.e. the uncanceled marks) of the constraint C_k where the fatal violation takes place when comparing these two candidates. All the lower terms $\omega^i \delta_i$ are equal to the respective terms in a .

By neglecting the lower terms, which are negligible compared to the highest one, we can define another difference-like function, which better reflects what is relevant for OT. In addition, its use saves us from some unnecessary calculation.

Definition 3.4.6. Given $a = \sum_{i=N}^0 \omega^i a_i$ and $b = \sum_{i=N}^0 \omega^i b_i$, where $a > b$, let be $\Delta'(a, b) = \sum_{i=N}^0 \omega^i \delta'_i$ such that

$$\delta'_i = \begin{cases} a_i - b_i & \text{if } a_j = b_j \forall j. (j > i \wedge j \leq N) \\ 0 & \text{otherwise} \end{cases}$$

This function returns the difference of violations of the constraint where the fatal violation takes place when we compare the two candidates. It is still somehow a sort of difference, because $b + \Delta'(a, b)$ differs from a only in lower terms than what is relevant when comparing the two violation sets.

First, SA-OT will be introduced by using some intuitive conventions, as a short cut, and then we argue for using this conventions.

²²By Lemma 1.4.3.c.(iii) in Holz et al. (1999, p. 33), $\omega^i < \omega^j$. Furthermore, due to Lemma 1.4.7.b in Holz et al. (1999, p. 37), $\alpha + \omega^j = \omega^j$ for all $\alpha < \omega^j$.

Thus, I propose the following notations, reflecting the idea that ω is a form of “infinity”:

$$e^{-\frac{\omega^i a}{\omega^j b}} := e^{-\omega^{i-j} \frac{a}{b}} := \begin{cases} 1 & \text{if } i < j \\ e^{-\frac{a}{b}} & \text{if } i = j \\ 0 & \text{if } i > j \end{cases} \quad (3.24)$$

$$e^{-\frac{x+y}{z}} := e^{-\frac{x}{z}} e^{-\frac{y}{z}} \quad (3.25)$$

where a, b, i and j are positive natural numbers, while x, y and z are ordinal numbers.

Employing these notational conventions, we can directly introduce the transition probabilities required by simulated annealing:

$$\begin{aligned} &\text{If } E(w) \geq E(w') \text{ then } P(w \rightarrow w' \mid T) = 1, \text{ otherwise} \\ &P(w \rightarrow w' \mid T) := e^{-\frac{\Delta(E(w'), E(w))}{T}} = e^{-\frac{\Delta'(E(w'), E(w))}{T}} \end{aligned} \quad (3.26)$$

Therefore, temperature T is also an ordinal number of the form:

$$T = \langle K_T, t \rangle = t\omega^{K_T} \quad (3.27)$$

One can simply check that both notions of difference, $\Delta(E(w'), E(w))$ and $\Delta'(E(w'), E(w))$, define the same probability. Using the second notion is somewhat farther from the traditional idea in SA (it is not exactly the difference of the energy levels), but it is closer to the philosophy of OT (ignore the constraints below the fatal constraint), and it is simpler to calculate.

By representing the Harmony function as an ordinal-valued energy function, we could formulate equation 3.26, which has a form that is fully analogous to the traditional transition probability equation used in real-valued simulated annealing:

$$P(w \rightarrow w' \mid T) = e^{-\frac{\Delta E}{T}} = e^{-\frac{E(w') - E(w)}{T}} \quad (3.28)$$

The interpretation of equation 3.26, in turn, leads to the same rules determining transition probabilities (the *Rules of moving* on page 63) that we have formulated earlier, in section 2.2.3. Namely, if temperature is $T = \langle K_T, t \rangle$, then:

- If w' is better than w ($w' \succ w$, that is, $C_k(w') < C_k(w)$), then move from w to w' .
- If w' loses due to the critical constraint $C_k > K_T$: don't move!
- If w' loses due to the critical constraint $C_k < K_T$: move!
- If w' loses due to the critical constraint $C_k = K_T$: move with probability $P(w \rightarrow w') = e^{-d/t}$, where $d = C_k(w') - C_k(w)$.

3.4.3 Arguing more for the definition of $e^{-d/t}$

Conventions (3.24) and (3.25), on the one hand, “make sense” because ω is but a mathematically sound way of saying “infinite”, and these proposals lead directly to a formulation of SA-OT in the ordinal representation of the violation profiles. On the other hand, they might nonetheless seem to the reader as *ad hoc*, and therefore spoil the mathematically precise underpinning of SA-OT. In the remaining pages of the present section we argue for this short-cut.

First, we quote another lemma from Holz et al. (1999, p. 34). Not only does the *Subtraction Lemma* holds on the class of ordinals, but also the *Division Lemma* and the *Logarithm Lemma*. The former states the following:

Lemma 3.4.7. DIVISION LEMMA: *Let a and b be ordinals. If $b \neq 0$, then there are unique ordinals q and m satisfying $a = b \cdot q + m$ and $m < b$.*

For ordinals a and $b \neq 0$, let $q(a, b)$ and $r(a, b)$ therefore denote the unique ordinals such that $a = b \cdot q(a, b) + r(a, b)$ and $r(a, b) < b$. $q(a, b)$ will be referred to as the *quotient*, and $r(a, b)$ as the *remainder* of a and b .

Our goal is to translate the expression $e^{-\frac{E(w')-E(w)}{t}}$ into ordinal arithmetic. The quotient $\frac{E(w')-E(w)}{t}$ can be easily rewritten as $q(\Delta(E(w'), E(w)), T)$ if $E(w') \geq E(w)$ —that is, exactly in the case we actually need this expression for the transition probabilities (if $w \succeq w'$). But we are still not able to interpret the expression $e^{-(E(w')-E(w))/t}$, because of the negative sign and because e is not an integer. Not surprisingly, for the value of this expression, a real number between 0 and 1, is unquestionably beyond the scope of ordinal arithmetic.

However, the following two observations can help us overcome this difficulty.

First, observe that the expression $e^{-d/t}$ can be replaced by the expression $a^{-d/t}$ for any real number $a \neq 1$, by simply rescaling temperature (or the violation levels), because²³

$$e^{-\frac{d}{T}} = a^{-\frac{d}{T \ln a}} \quad (3.29)$$

The concept of *rescaling* originates from physics. One can measure a quantity using different scales—e.g., metres, kilometres, feet, yards, lightyears, etc. for distance—and the difference is but a constant multiplicative factor. Now, $T' = T \ln a$ will replace the earlier T , and then the form of the equations can be kept unchanged.

In turn, ordinal exponentiation can be used by replacing e with an arbitrarily chosen *integer* base $a > 1$ —for instance $a = 2$ or $a = 3$ in order to remain close to the original exponentiation of base $e \approx 2.71$.

Second, we can also get rid of the $-$ sign in the exponent. A transition probability $p = e^{-d/t}$ means that first we generate a random number r in the interval $]0, 1[$ with an equal distribution, and then we move the random walker iff $r < p = e^{-d/t}$. The transition probability is the *measure* of the set of the r values that result in moving—that is, of the r values that satisfy this inequality. Now, the negative sign can be removed by rewriting this inequality. We can say therefore that we move iff $r^{-1} > e^{d/t}$; that is, if and only if for all $\alpha > 0$

$$r^{-\alpha} > e^{\frac{d \cdot \alpha}{T}} \quad (3.30)$$

²³Recall that $\log_a b = \frac{\log_e b}{\log_e a}$, that is, $\log_a e = \frac{1}{\ln a}$, if $a \neq 1$.

If P is a probability measure on $\{r | 0 < r < 1\}$ (specifically, we use an equal distribution: $P(\{r | a < r < b\}) = b - a$), then

$$P(w \rightarrow w' | T) = P(\{r | \forall \alpha > 0 : r^{-\alpha} > e^{\frac{d \cdot \alpha}{T}}\}) \quad (3.31)$$

Introducing the arbitrary multiplier *alpha* will help us in the case d is not dividable by T in integer arithmetic.

In order to be able to compare formally a real number, such as $r^{-\alpha}$, with an ordinal derived from the representation of the violation profiles, we introduce the following

Definition 3.4.8. Let $\mathbb{R}^\infty := \mathbb{R} \cup \{\infty\}$, the enlargement of the set of the real numbers with $+\infty$.²⁴ Let the relation $>'$ be the enlargement of the usual order $> \subset \mathbb{R} \times \mathbb{R}$ on \mathbb{R}^∞ :

$$>' := \{(a, b) \in \mathbb{R} \times \mathbb{R} \mid a > b\} \cup \{(\infty, a) \mid a \in \mathbb{R}\}$$

If O is a set of ordinal numbers, then the function $R : O \rightarrow \mathbb{R}^\infty$ is defined as

$$R[a] := \begin{cases} a & \text{if } a < \omega \\ \infty & \text{if } a \geq \omega \end{cases}$$

One can simply demonstrate that the relation $>'$ is indeed a total order on \mathbb{R}^∞ . The symbol $>$ is usually used both on the set \mathbb{R} and on the class of ordinals, whereas we rather use $>'$ on \mathbb{R}^∞ in order to avoid confusion.

The definition of the function R makes use of the fact that the integers are defined as ordinals less than ω , and then they are injected into \mathbb{R} . The class of all ordinals is not a set, yet we can for instance take the set $O = \omega^\omega$ for our purposes, which contains all violation profile-like ordinals.

After all these remarks and definitions, we can reformulate within ordinal arithmetic how to decide in SA-OT whether to move from candidate w to candidate w' , if $w \succ w'$, that is, if $E(w) < E(w')$.

The straightforward solution would be to generate a real number r between 0 and 1 with equal distribution, and then move if and only if

$$\frac{1}{r} >' R \left[2^q \left(\Delta(E(w'), E(w)), T \right) \right] \quad (3.32)$$

The problem with this proposal is that the division is performed in a coarse way, similarly to division in integer arithmetic. Suppose for instance that $T = \omega^k \cdot 3$. Then, no distinction is made between $\Delta(E(w'), E(w))$ being $\omega^k \cdot 5$ or $\omega^k \cdot 3$, for in both cases $q \left(\Delta(E(w'), E(w)), T \right) = 1$ and only the remainders are different. Even though the empirical predictions of such a model might be worth investigating, this is probably not what we want. Namely, this model could not make the difference between a step that increases the violation level of constraint C_k by 3 or by 5.

The problem is that we cannot make use of the remainder of the division. How do you get a higher precision if you are forced to use exclusively integer division? You multiply the numerator by 10 or by 100, and then you consider

²⁴Compare to the addition of the *point at infinity* to each line in projective geometry.

the last digits of the quotient as being beyond the decimal point. Applying this trick has been the purpose of introducing the arbitrary positive multiplier α in equation (3.30).

Therefore, we rather propose moving from w to w' if and only if

$$\forall \alpha \in \mathbb{N}^+ : r^{-\alpha} >' R \left[2^{q \left(\Delta(E(w'), E(w)) \cdot \alpha, T \right)} \right] \quad (3.33)$$

What follows from this definition? Let $T = \omega^K \cdot t$, and $\Delta'(E(w'), E(w)) = \omega^k \cdot d$. In words, the fatal constraint is C_k , and $d = C_k(w') - C_k(w) > 0$. Then, we consider the following cases:

1. Suppose $K > k$ (informally, $TT \gg \Delta(E(w'), E(w))$, cf. definition 2.2.5). Then, for all $\alpha \in \mathbb{N}^+$, $q \left(\Delta(E(w'), E(w)) \cdot \alpha, T \right) = 0$, because the divider is always larger than the numerator ($\omega^j \cdot \alpha < \omega^{j+1}$). It follows that for all $r < 1$, rule (3.33) prescribes to move to w' . That is, the measure of the set of the r values resulting in move—the transition probability—is $P(w \rightarrow w'|T) = 1$.
2. Suppose now $K = k$ (informally, $TT \approx \Delta(E(w'), E(w))$, cf. definition 2.2.5). Then for all $\alpha \in \mathbb{N}^+$,

$$R \left[q \left(\Delta(E(w'), E(w)) \cdot \alpha, T \right) \right] = \left[\frac{d\alpha}{t} \right] \leq \frac{d\alpha}{t} \quad (3.34)$$

where $\left[\frac{d\alpha}{t} \right]$ denotes the integer part of $\frac{d\alpha}{t}$.

Hence, (3.33) holds if and only if $r < 2^{-d/t}$. Namely, if $r < 2^{-d/t}$, then for all $\alpha \in \mathbb{N}^+$,

$$r^{-\alpha} > 2^{\frac{d\alpha}{t}} \geq 2^{\left[\frac{d\alpha}{t} \right]} = R \left[2^{q \left(\Delta(E(w'), E(w)) \cdot \alpha, T \right)} \right] \quad (3.35)$$

Further, if $r \geq 2^{-d/t}$, then choose $\alpha = ct$ (for any positive integer c) to show that the condition for moving is not satisfied anymore:

$$r^{-\alpha} \leq 2^{d\alpha/t} = 2^{q \left(\Delta(E(w'), E(w)) \cdot \alpha, T \right)} \quad (3.36)$$

As r is chosen with an equal distribution, the measure of the set of the r values causing the system to move is thus $2^{-d/t}$. By rescaling temperature, we obtain the usual rule for this case: “move with probability $P(w \rightarrow w'|T) = e^{-d/t}$ ”

3. Suppose finally $K < k$ (informally, $TT \gg \Delta(E(w'), E(w))$). Then

$$q \left(\Delta(E(w'), E(w)) \cdot \alpha, T \right) \geq \omega^{k-K} \cdot (d\alpha) \geq \omega \quad (3.37)$$

Hence, the exponentiation in (3.33) returns an infinite ordinal ($2^\omega = \omega$) in the present case. As $r^{-\alpha} \not>' \infty$ by the definition of $>'$, the consequence is that no r ever results in moving. Thus, the measure of the set of r values causing the system to move is zero: $P(w \rightarrow w'|T) = 0$.

Summarising, we have again derived the *Rules of moving* from page 63.

3.5 Summary of the formal approaches

When we introduced the idea of applying simulated annealing to Optimality Theory, many different options could have been followed. Indeed, we saw in section 2.3.2 that the proposed solution does not always work. Due to the Strict Domination Hypothesis, some models are always stuck in local optima, and the algorithm's precision—the likelihood of finding the global optimum—does not converge to 1 as the number of iterations grows infinite. There, we speculated about possible ways to solve this problem, without much success.

I will therefore argue that these failures are inevitable in SA-OT, and actually we can make use of them in building linguistic models. Yet, before making these statements, I have to convince the reader that the proposed SA-OT is indeed the most appropriate implementation of simulated annealing for Optimality Theory. This has exactly been the goal of the present chapter. Already section 2.2.3 contained a train of thought that introduced SA-OT, whereas the present chapter formally showed how the Strict Domination Hypothesis leads directly to the same *Rules of moving*—twice, at that.

In sum, we may conclude that the transition probabilities driving OT-SA are well-founded: we have seen several ways in which they may be derived from the basic ideas of Optimality Theory. The bottom line was in both cases the same *Rules of moving*.

One may ask what the polynomial approach and the ordinal number approach can contribute to each other. The answer is manifold. Firstly, the two approaches are based on very different mathematical concepts, and yet, they led to the same algorithm. Secondly, the mathematical beauty of a model is a very subjective feature, hence, different readers may prefer one or the other approach. For instance, one may not like the way calculating the limit is proposed in (3.16), or not be convinced of the necessity of introducing an arbitrary α in (3.33). Additionally, the beauty of transfinite arithmetic (analysed as conceptual blending by Núñez, 2005) may arguably be an additional subjective value of the cardinal approach in the eyes of some readers. Indeed, the contradicting reviews I received to my article (Bíró, 2005b) demonstrated that different people are more convinced by one or the other approach.

Formal arguments can also be made why to introduce both approaches in this dissertation. In most linguistic models, violation profiles are non-negative integers (represented as a certain number of stars in a tableau). These are exactly the values allowed the $C_i(w)$ s to take by the representation of violation profiles as ordinal numbers in (3.19). If the range of $C_i(w)$ is some other well-ordered set (such as the set of the consonants ordered according to sonority in the Berber example of Prince and Smolensky, 2004), an isomorphy could be applied to map this set onto some ordinal numbers. Indeed, definition 3.1.1 (page 76) requires the set $\{C_i(w) \mid w \in UR\}$ be a well-ordered set in order to make definition 3.3 introducing the main idea of OT (page 82) well-founded. This observation invites the generalisation to allow constraints that can take any ordinal numbers as values. Such an approach would naturally prohibit ganging up effects for most phenomena (if $C_i(w) < \omega$), and allow them in some special cases by stipulating $C_i(w) \geq \omega$.²⁵

²⁵Similarly to the way we propose here to generalise the range of the constraints from \mathbb{N}_0 to larger sets of ordinal numbers, further research might also enlarge the set of constraints. Indeed, the main restriction of the constraint set is that it must be a totally ordered set such

Nonetheless, the polynomial representation in (3.10) furnishes us with more flexibility, as the only requirement is that $C_i(w)$ take real values—even if the formal definition of OT (3.3) was based on the definition 3.1.1 of a constraint requiring its range to be well-ordered. And indeed, the SA-OT algorithm in Fig. 2.8 follows the idea that the violation levels are real numbers.

Even though both approaches have been introduced in order to faithfully realise the Strict Domination Hypothesis, both of them point towards a possibility of representing situations that do not satisfy this hypothesis. In the polynomial approach, one may decide not to perform the $q \rightarrow \infty$ limit, but to replace it by stipulating a high value for q , as an approximation of the $q \rightarrow \infty$ limit. Then, even if most constraints follow the Strict Domination Hypothesis, some special cases can display cumulativeness effects. In the ordinal approach, one can argue for using non-finite violation levels ($C_i(w) \geq \omega$) if forced to account for cumulativeness phenomena.

The advantage of both approaches—especially of the ordinal approach—over the traditional real valued realisation is that the Strict Domination Hypothesis can be saved as categorically true for the cases where it really applies; and towards the cases where it does not, the border is sharp.

that each subset have a unique *maximal* element. Take for instance Prince and Smolensky's "bag of violation marks" approach and their *Cancellation Lemma*: after cancelling the shared violations, the task is to identify the *unique* highest ranked constraint that still has a violation in one of the bags.

By reversing the direction of the ranking relation among the constraints, we could therefore propose simply a constraint set $\{C_i \mid i \in \mathcal{I}\}$ where \mathcal{I} can be any well-ordered set (hence, even larger than ω). The indices i of the constraints would then be ordinal numbers. The problem is that the *inverse* of the indices would be needed in equation (3.19) (page 95) due to the reversion of the constraint ranking relation, which operation is not defined among ordinal numbers. Nevertheless, this observed "duality" of the range of the constraints and the constraint set is probably worth analysing further, and might have consequences for the relationship of generation and learning in OT in general (cf. Turkel, 1994).

Chapter 4

The Linguistic Context of SA-OT

4.1 A few words about the lexicon

The goal of the present section is three-fold. First, it aims at saying something about the way the lexicon can be seen from the point of view of Simulated Annealing Optimality Theory. Linguistics has failed to get round the questions related to the lexicon such as lexical exceptions, and interest has recently increased in lexicalist approaches. Within OT, the language specificity of the lexicon seems at first view to conflict with the *Richness of the Base* principle (all inputs are possible in all languages, cf. Prince and Smolensky, 2004, p. 225). According to another principle, *Lexical Optimisation* (*ibid*), the language learner should choose the input that corresponds to the most harmonic output among the possibilities given the surface form observed (cf. also the *Robust Interpretive Parsing* of Tesar and Smolensky, 2000).¹

In particular, and this is the second goal of the present section, we point to the way that the complex lexicon model of Burzio (2002) could be realised in practice using SA-OT. Concrete realisation of this model drawn on physics is left to future work, nevertheless.

The main component of this model, *Output-Output Correspondence* (OOC, *Output-Output Faithfulness*) proposed by Benua and Burzio has been a widely used constraint within Optimality Theoretic phonology, and yet, it lacks a precise workable definition to my best knowledge. To be more precise, Burzio (2002) seems to have not fully worked out the details of his proposal, so that even though most linguists use OOC (successfully) in an even less formal way, this practice does not work for SA-OT which requires the exact number of violation marks assigned to any candidate. Consequently, and this is the section's third goal, a more formal definition of an Output-Output Correspondence-like constraint will be introduced, in order to employ it in Chapter 5.

Subsequently, the second section of the present chapter includes a few notes on learnability in SA-OT. Learnability issues have been successfully tackled both in standard OT (Tesar and Smolensky, 2000), as well as in Stochastic Optimality

¹For the role of the lexicon in OT syntax, see for example van der Beek and Bouma (2004) and references therein.

Theory (Boersma and Hayes, 2001), a fact that provides a strong argument in favour of Optimality Theory, as opposed to many previous linguistic models. Therefore, the reader would naturally ask whether SA-OT has anything to say about learnability.

4.1.1 English Past Tense

One of the most investigated issues related to rules, *minor rules* and lexical exceptions is the case of English past tense. As is well known, the productive *major rule* is to add the suffix <ed>,² while *minor rules* may prescribe changing <ing> to <ang> (like in *sing* – *sang* and *ring* – *rang*) or the last coda to <ought> (e.g. in *bring*, *think*, *seek*). Some cases, such that of the verb *to be* or *to have* are fully irregular.

Several approaches have been presented to tackle the problem, starting with connectionist approaches (Rumelhart and McClelland, 1986), to output-output correspondence (Burzio, 2002) or ACT-R models (Taatgen and Dijkstra, 2003). Within OT, the first approaches have been proposed by Boersma (1998b) and by Burzio (1999). The latter was finally published as Burzio (2002), and we shall turn back to it soon.

In fact, a major debate emerged from this phenomenon, the so-called *Past Tense Debate*, when Pinker and Prince (1988) reacted to Rumelhart and McClelland (1986): the connectionist camp (McClelland, Plunkett, Seidenberg,...) argued for a single mechanism for both regular and irregular verbs, whereas the proponents of symbolic computation (Pinker, Ullman,...) fought for a dual route mechanism. For a recent, two-sided overview of the debate, see both Pinker and Ullman (2002) and McClelland and Patterson (2002). For recent neurolinguistic arguments for the dual route model based on double dissociation, see the work of William Marslen-Wilson and his colleagues (e.g. Tyler et al. (2002) and Stamatakis et al. (2004)). This debate goes much beyond the issue of English past tense, the latter being only a test case: the question of debate is the role of symbolic rules as opposed to connectionist approaches in language processing, or even in cognition in general.

Although its details may have been debated, a pattern called *U-shaped development* can be (more or less) observed in children's acquisition of English past tense forms (Brown (1973), pp. 333; Kuczaj (1977); Harley (2001), p. 96 and 125-126, and references therein, including an introduction to the Past Tense Debate). Even if using only a very restricted vocabulary, the youngest children perform quite well in producing the past tense of verbs. In a second stage, however, performance drops, before improving again in the third stage. Roughly speaking, we may say that the child memorises all forms in the first stage; later, the growing vocabulary allows making generalisations, and the drop in performance is due to over-generalisation, so forms such as **bringed* or *singed* appear beside the correct ones; in the third stage, nevertheless, these cases of over-generation are learned to be errors, that is, exceptions are (re)-learned.

A further interesting phenomenon is the acquisition of the so-called *minor rules*. For instance, such a minor rule, inferred from *sing* – *sang* and *ring* – *rang* may require changing the coda of a monosyllabic verb from <ing> into <ang>.

²For the sake of ease, I use the written form of the segment strings, and not the underlying representation or some surface allomorphs.

Child speech indeed produces, although with a very low frequency, forms such as **brang*, which can be seen as the result of overgeneration from this minor rule (Xu and Pinker, 1995; Taatgen and Dijkstra, 2003).

One may speculate about minor rule forms corresponding to local optima in the SA-OT search space; but only future work can tell whether such an approach to account for these phenomena—including those in acquisition—will turn to be fruitful.

4.1.2 Burzio's physical model of the mental lexicon

Burzio (2002) attempts at giving an Optimality Theoretical compromise to the English Past-Tense Debate, by using a model based on an idea taken from (classical) physics; namely, on the concept of *forces* and *fields*. In what follows, I am making these physical analogies more explicit than as found in Burzio (2002) itself.

His aim is to explain why “[l]exical sectors that are morphologically irregular tend to be phonologically regular, and vice-versa”. He proposes that whenever morphology is irregular and phonology is regular (“level 1” affixes in Kiparsky’s Lexical Phonology), then the phonological markedness constraints dominate. For instance, the vowel of the verb *keep* is shortened in the past tense form *kept* in order to meet the limitations on syllable size. But in other cases, morphology is regular and phonology turns to be irregular (“level 2” affixes): for instance, the regular past tense form *beeped* includes a syllable that is so long that it would be otherwise prohibited. Then, the analogy in the paradigm acts as an attraction between the forms, overranking phonological well-formedness requirements. This attraction is described by *Output-Output Correspondence* (or *Faithfulness*), and is seen as some sort of gravitational force between the lexical items.

In physics, bodies with a mass create gravitational fields around themselves, bodies (or particles) with an electric charge create additionally an electric field, and so on. The fields thus created by the individual bodies are summed up to form the field in which (the same or different) bodies follow their trajectories. The movements of the bodies are driven by the forces derived (literally) from the overall field, whereas this field in each moment is a function of the location (and speed, for magnetism) of the bodies. Two additional forces can be present: friction hinders any changes of position, whereas external forces (e.g. a gravitational field) favour some positions over others.

A field can be seen either as a scalar-valued function (*energy*, or rather *potential*) or a vector-valued function (*force*) of space (and time). If you put a given body at a given point in space and time, the properties of that body (e.g. its mass in the case of gravitation, its electric charge for electric interaction, its charge and speed for magnetic interaction, etc.) and the field (as a function of all the bodies or particles around) will determine what the energy of that body is, and what force the field exerts on that body. Moreover, the force is the *negative gradient* of the energy: a vector pointing into the direction in which energy declines the most, and the length of the vector is proportional to the steepness of the energy function in that direction. Indeed, the idea is that the physical force influences the body to move towards the minimal energy state. In other words: it is sufficient to define the energy (potential) as a scalar function in space, for its negative gradient (a spatial derivative) in each point gives the

force acting upon some particle there.

We can now summarise the picture thus far: the position and speed of a particle in the next moment is determined—besides its mass, position and speed in the previous moment—1. by friction, 2. by the external forces, as well as 3. by the aggregate force exercised by the other bodies. The latter can be calculated from the position of the bodies and their mass or charge. If x_i and m_i are the position and mass of particle i , while $F(j \rightarrow i)$ and $V(j \rightarrow i)$ are the force acting on particle i and the energy of particle i in the field created by particle j (the influence of j on i), then the differential equation describing the trajectory of particle i is by Newton's second law:

$$\begin{aligned} m_i \cdot \frac{d^2}{dt^2} x_i &= F_{friction} + F_{external} + \sum_{j \neq i} F(j \rightarrow i) = \\ &= F_{friction} + F_{external} - \frac{d}{dx} \sum_{j \neq i} V(j \rightarrow i) \end{aligned} \quad (4.1)$$

In Burzio's model, lexical items are the bodies in a multidimensional space, whose dimensions correspond to phonological, syntactic and semantic features. The distance of two words can be measured in the number of features they differ from each other. As Burzio writes (p. 11): “[t]his model performs a simple calculation in which the input is the position at which the object is originally placed, and the output is the ultimate resting position”. Thus, friction will correspond to *Input-Output Correspondence*, the force that acts against changing position. External forces correspond to the markedness constraints: independently of the position of the different bodies, they pull each of the bodies towards some preferred positions. Finally, the force exercised by the other bodies translates to *into Output-Output Correspondence* (OOC)—we shall return to this point in the next subsection.

So for instance, most constraints used in linguistics can be seen as external factors, such as the Earth's gravitational field in which everyday objects with mass follow a certain trajectory. Similarly to gravitation, which favours some positions over other ones, markedness constraints favour certain feature combinations, that is, specific positions in the space. Remember that the linguistic features (the phonological content, the syntactic class, semantic properties) of lexical items are encoded as the dimensions of the space. If, for example, some constraint disfavours front rounded vowels, harmony improves by moving towards [-round] in the [round] dimension—just like gravitation, which prefers the butter side of slices of bread and butter to be lower in the vertical dimension.

In OT terms, the points of the multidimensional space are the candidates, while the output, the “ultimate resting position” is the winning candidate where the forces neutralise each other. In physics, such a stable resting point is a local minimum of the energy: there the spatial derivative (the gradient) of the energy is zero, so no force acts upon the body, and moving away from that point would increase the energy. Consequently, the OT Harmony function will correspond to the (negative) energy in the physical analogy, and the goal is to find the position (the candidate) that minimises energy (maximises harmony).

Here, energy includes not only the energies from the interactions with each of the other particles, but the external forces and friction are also integrated

(literally) into energy. Actually, friction should rather be replaced by springs, also mentioned by Burzio. The more the spring is pulled, the larger its energy, which corresponds to a larger force pushing the particle back to the origin. In turn, candidates or lexical items are strings stretching between the input form and the output form. A candidate's energy (or harmony) is the sum of the spring's energy (Input-Output Faithfulness or Input-Output Correspondence), of the energy from the external field (markedness constraints) and of the energy from the interaction with the other lexical items (Output-Output Correspondence).

It is unclear how precisely this sum has to be calculated in Burzio's model. As he later employs an OT-model referring to strict constraint ranking, I suggest the polynomials or ordinal numbers as representations, following Chapter 3. Thereby, it will be possible both to interpret the physical analogy (involving sums and derivatives), and to save the connection to Optimality Theory.

Burzio does not elaborate either on what the "ultimate resting position" is, he simply supposes that it is the global minimum of the energy (harmony), following the principles of standard Optimality Theory. Indeed, in quantum physics, a non-global local minimum is only a metastable position, as sooner or later (this time range is called the *half-life*) the particle jumps to some lower minimum. But if the half-life is very long, as well as in classical mechanics, local minima can also be quite stable. Therefore, if the "ultimate resting position" is only required to be some local minimum (following the physical analogy), we obtain a similar picture to that used in SA-OT: possible surface forms are local optima, among which the global optimum is (usually) the most frequent one. Indeed, "local optimum" is the central concept, and the global optimum is but a special local optimum.

Additionally, the parallel between Burzio's model and the topology in SA-OT becomes even stronger if we make explicit that in Burzio's model neighbours—a concept required in the definition of local optima—are points whose distance is 1, that is, candidates that differ exactly in one feature, in one basic transformation. Alternatively, a quantum physics-like model, in which non-global local optima may be metastable if the half-life is very long, corresponds to another type of SA-OT topology: to the definition in which any two candidates are neighbours, but the *a priori* probability diminishes with distance. In this case, a candidate can be attested because it is a "metastable local optimum" in the sense that jumping to a better one is extremely improbable, because better candidates are very far away (so SA-OT will be stuck there); similarly to radioactive isotopes found in nature whose half-life is comparable or longer than the age of the universe, so that they have not decayed yet.

In brief, Burzio's search space is a special case for the search space employed in SA-OT. A special case, but a very self evident and general one. He does not specify the way he would perform the search for the "ultimate resting position". (Would he calculate step by step the trajectory of each item from the input form to the output form? Does anything guarantee that such a trajectory ends in a resting position?). And yet, the physical systems that motivated simulated annealing (including the $e^{-\Delta E/T}$ factor) are the same as those inspiring Burzio. Hence, the close connection between the two proposals, I believe, is worth further research.

4.1.3 Burzio's Output-Output Correspondence

Let us turn our attention now to the way Burzio (2002) introduces the most interesting type of force present in his model, Output-Output Correspondence, that is, the interaction between particles. This “gravitational force” between pairs of words is argued to be responsible for phenomena such as analogical effects.

To sum up what we have discussed so far, the lexicon of a language is composed of lexical items that optimise locally their “energy” (*i.e.*, their harmony function). The energy of an output form depends on its well-formedness (phonological markedness constraints), on its distance from the input form (Input-Output Correspondence), as well as on its interaction with the other output forms (OOC). Hence the Saussurian concept of the language as a complex system: altering one surface form influences all other outputs through their interaction.

Burzio introduces the notion of *representational entailments* (on p. 176), which, he argues, is cognitively plausible. The position vector of some word $A = (a_1, \dots, a_n)$ can be seen as a set of entailments of the form “if position i has a_i , then position j has a_j ” for all possible i 's and j 's. Take now a second lexical item, B , whose coordinates equal those of A in k out of the n dimensions (features), and differ in $n - k$ dimensions. Given this, B violates $k(n - k)$ entailments of A : there are k different positions i and $n - k$ different positions j , such that B has a_i in position i , and yet, not a_j in position j . Hence, Burzio's proposal—the way I interpret the August 1999 version of his paper, which is slightly more explicit (Burzio, 1999)—defines the “gravitational” potential $V(A \rightarrow B)$ exercised by word A upon word B as the number of entailments of A violated by B . This potential as a function of the non-Euclidian distance k is:

$$V(k) = k(n - k) = nk - k^2 \quad (4.2)$$

The “gravitational” force with which A attracts B is the spatial derivative of this potential:³

$$F(k) = \frac{\partial V(k)}{\partial k} = n - 2k \quad (4.3)$$

The direction of this force points towards word A .

It becomes clear that the closer the two words (that is, the smaller the k), the stronger they attract each other. In that property, Burzio's inter-word force resembles vaguely gravitation and electrostatic force. If the two words are very far, attraction vanishes; even further ($k > n/2$), the force turns into repulsion (“anti-gravitation”).

Subsequently, a trick often used in physics is employed by Burzio. A body composed of many particles can be replaced by its mass centre (“centre of gravity”) for the purpose of calculating its gravitational attraction. This is so, because the gravitational force exerted by each particle can be decomposed into two components: when summing up the forces exerted by all the particles, the first components cancel each other, whereas the second components sum up as if all the mass were concentrated in the mass centre. This trick helps Burzio to understand the effect of a group of words on a particular word.

³To be more precise, the negative derivative is the repulsive force. To increase clarity, we concentrate on attraction, however. Cf. the right hand side of equation (4.1).

Suppose that a group of words share some representational entailments: in Burzio's example, *parental*, *natural*, etc. all share the entailment according to which "the ending *al* must be preceded by a noun". Other entailments that are not shared include "the ending *al* must be preceded by the string *parent*" or "the ending *al* must be preceded by the string *atur*". When summing up the entailments of these words, (hence, the force, since the derivative in Eq. (4.3) is additive), the effect of the latter entailments will neutralise each other. Yet, the group of words will yield the *macro-entailment* "the ending *al* must be preceded by a noun". This is the way Burzio hopes to explain paradigmatic effects.

Consider an arbitrary word. The gravitational force exerted on it by most of the lexicon is negligible, partially because most of the words are far enough away (having many different features), and partially because their effects neutralise each other—unless the word is located "outside" of the majority of the lexicon, such as in the case of a foreign word whose phonological features have not been assimilated into the general phonological system of the language. In the latter case, the lexicon as a whole exerts some attracting force. In the most frequent case, however, only particular words in the neighbourhoods will exert attraction: assimilation to a set of similar words, paradigmatic levelling, etc. Furthermore, the closest existing lexical items to a derived word are its root and the outputs of the previous cycles of the derivation. Through this idea, Burzio's Output-Output Correspondence is able to account for phenomena previously accounted for by cyclical derivation.

Burzio is even able to explain why the root has more influence on the derived form than vice versa. He argues that more representational entailments of the shorter root are satisfied by the derived form than vice-versa. For example, *parent* violates *parental*'s entailment "if the word's first segment is a *p* then its eighth segment is an *l*", while all entailments referring to the segments of *parent* are satisfied by *parental*. In turn, *parental* is more influenced by *parent* than vice versa. The only problem with this argument is that we become uncertain about the exact representation of a lexical item as an n dimensional vector, with always n_{segm} dimensions corresponding to phonological segments.

Finally, turning back to the Past-Tense Debate, what is Burzio's explanation of the different behaviour of Level 1 (highly irregular morphology, highly regular phonology) and Level 2 (highly regular morphology, yet often irregular phonology) word derivations? The difference is the place where Output-Output Correspondence is ranked, relative to phonological markedness constraints and to Input-Output Correspondence. Moreover and most importantly, the different ranking results from the significantly different numbers of stems belonging to a certain derivational paradigm (Burzio (2002), p. 195). Level 1 affixes take relatively few stems, and therefore gravitation's morphological levelling effect is weak: Output-Output Correspondence is ranked below phonological markedness, yielding irregular morphology and regular phonology. On the other hand, the possibly infinite number of stems to which Level 2 affixes can be applied boosts the effect of Output-Output interaction over the phonological markedness constraint—resulting in a regular morphology, and an irregular phonology.

By (literally) deriving grammatical effects (output-output constraints) from the words in the lexicon, Burzio reverses—as he himself remarks—the one-way relation from the (adult) grammar to the output dominant in the generative tradition.

4.1.4 Burzio’s model and SA-OT

A very important question is still open, however. How precisely are the different forces summed up? Equation (4.3) gives the “force” with which lexical item A attracts word B . Without asking crucial details about the exact number and nature of the dimensions, and supposing that the different forces are simply summed up, it is still unclear how this effect is translated into a position of Output-Output Correspondence within the hierarchy. And this last issue seems to be the major point in assessing Burzio’s proposal.

Although we are not going to come up with concrete proposals, the different ways of representing the Harmony function introduced earlier allow us to speculate about possible directions of future work.

Notice that Burzio’s proposal gets tangled up where it has to accommodate a traditional Optimality Theoretical framework. Supposing that representational issues—the exact form of the feature vectors—are solved, and accepting the neural plausibility of representational entailments, the potential introduced in Eq. (4.2), as well as the derived force in (4.3) are well-founded and elegant. And yet, are we sure and certain how many stars to assign to a particular candidate in any case?

This question might be avoidable in SA-OT, however. Do we really need to translate Burzio’s formalism into terms of standard Optimality Theory? Observe that what we did in earlier chapters was the opposite translation: transforming constraint violations into some energy (potential, harmony) function to be optimised. As a simple real-valued function would not work for Optimality Theory in the general case, we have introduced the polynomial representation and the ordinal number representation of the Harmony function—both having the form of a sum.

Consequently, Burzio’s “gravitational” potential, once well formulated, can be added directly to some formulation of the Harmony function. This new addend does not necessarily have to have the exact form of the addends obtained from the traditional constraints: we may give up on seeing the gravitational effect as a constraint. Yet, the gravitational potential should be formulated within a similar formalism, so that it can be added to the representation of the constraints. Not bad news in itself, as probably Burzio’s “gravitational” potential is not really suited for a real-valued representation, for the simple reason that it requires the sum over an indeterminably large lexicon. Furthermore, although the gravitational effect in the harmony function will not have the form of a constraint, yet its magnitude within the summands probably can be estimated—and be interpreted as Output-Output Correspondence being ranked higher or lower than markedness constraints.

As a speculation, remember how Burzio explained the different ranking of Output-Output Correspondence for Level 1 and Level 2 derivations: in the first case the effect of at most a few hundred words are summed up, whereas the summation in the second case takes place on an open class of words. If the mean potential obtained from a single word in the class is \bar{v} , then one hundred words provide a potential of $\bar{v} \cdot 100$. Yet, in the case of fully productive morphological processes in Level 2 derivations, the open set has the cardinality of a countably infinite set (\aleph_0): in turn, is the summed up potential $\bar{v} \cdot \omega$? If so, the corresponding addend is of a higher magnitude in ω and we have understood why Burzio argued for promoting Output-Output Correspondence higher in the

case of Level 2 morphology.

Let us now step back from speculations. Burzio's model is undoubtedly attractive—at least to a person with a background in physics. However, the model is very hard to implement. In practice, the phonologists using Output-Output Correspondence define *a priori* which other output the given form must be compared to, and do not demonstrate that the interactions with *all* the other words in the lexicon are negligible, and indeed extinguish each other.

Consequently, we recommend replacing Output-Output Correspondence with correspondence constraints that refer explicitly to the process of morphological derivation. One such constraint could be Kenstowicz's BASE-IDENTITY (Kenstowicz, 1995). However, in the following section, we demonstrate that very often it is not the *base*, but the *output of the previous cycle of the derivation* that is relevant, a fact well known in the *Lexical Phonology* of Kiparsky (1982). Therefore, we recommend introducing a constraint named COMPONENT-OUTPUT CORRESPONDENCE / CONSTITUENT-OUTPUT CORRESPONDENCE (COC). If I keep the original name OOC in the next chapter while I mean in fact COC, it is because I would like to retain readability for phonologists.

4.1.5 Constituent-Output Correspondence

In this subsection, we define COC, so that it can serve us in Chapter 5 on metrical stress in Dutch fast speech.

By way of introduction, I must express my reservations with regard to the general way of defining a constraint in the OT literature. It is true that originally constraints were requirements that a linguistic form either met or did not, and therefore, introducing a constraint meant defining the condition that a form had to meet (for instance: “each syllable has an onset”, “no syllable has a coda”). Nevertheless, with the advent of violable constraints, and, especially since more levels of violation could be distinguished, a constraint is rather seen as a function mapping each linguistic form to a numerical value (usually, the number of violation marks). Consequently, the definition of a constraint must tell how many violation marks are assigned to a given candidate (for instance: “the number of codas in the word”, or “one star per syllable with a coda”).

We particularly have to emphasise this here, because this task is especially difficult in the case of OOC and COC. The authors of most articles are lucky enough to be able to point intuitively to the fact that the optimal candidate is “clearly” better with respect to OOC than its competitors, so they can eschew giving an exact definition of OOC. Yet, Simulated Annealing Optimality Theory has to be able to compare the violation levels of any two neighbouring candidates. In turn, the number of violation marks incurred by a candidate should be defined exactly; or, at least, the difference in the violation levels ought to be given for any pairs of neighbouring candidates. The second way, undoubtedly challenging, assigns a violation difference to each of the possible basic steps.⁴

⁴For instance, in the case of metrical stress assignment to be presented in Chapter 5, moving the unobservable foot borders should not introduce any changes with respect to OOC (COC). Nonetheless, deleting and inserting a stress (a foot), as well as moving the position of a stress (changing the head syllable of a foot) may involve some changes in violating OOC (COC). One parameter will define the possible change due to deletion or insertion, and another parameter will describe the role of changing the place of a stress. These two parameters, nevertheless, correspond to the parameters used in the approach described presently.

In the following, nonetheless, we follow the first way, for its being simpler and consistent with the general claim of defining each constraint as a function.

Correspondence Theory was introduced in the early years of Optimality Theory by McCarthy and Prince (1993b) (p. 67) for the sake of reduplicative phenomena. (We shall use it also in section 6.3.) In later developments, the *correspondence relation* \mathcal{C}_w maps the segments of the underlying form to the “corresponding” segments of the candidate string w . Then, constraints may require that each underlying unit have a corresponding image in the candidate (constraint MAX—originally called PARSE with a different philosophy—prohibiting the underparsing, *i.e.* the deletion of parts of the input); each surface element be the correspondent of an underlying segment (constraint DEP—FILL in Prince and Smolensky (1993)—punishing epenthesis); and that input and output segments be the same (further types of faithfulness constraints).

Unlike in the general case, pairing the basic units of the input and the output string is easy in stress assignment, for GEN adds some structure (namely, the metrical structure) on the top of the input string without altering the latter. Hence, the input string and the output string are composed of the same number of syllables, and the n th syllable of the input string *corresponds* to the n th syllable of the output string.

Thus, we focus on the correspondence of the metric structure (stress pattern). Yet, we employ *Output-Output Correspondence*, or *Component-Output Correspondence*, and not *Input-Output Correspondence*. When assessing a candidate w with respect to *Output-Output Correspondence* (*Component-Output Correspondence*), we will compare it to a string σ of the same length. In the case of *Output-Output Correspondence*, σ has to be derived from the stress pattern of any word in the lexicon, which does not necessarily has the same length as w . Yet, as previously argued, I propose to replace *Output-Output Correspondence* with *Component-Output Correspondence*, and in this case σ is the stress pattern derived from the stress patterns of the morphological constituents of w .

In the simplest case, if w (actually, $GEN^{-1}(w)$, the corresponding underlying representation) is the concatenation of a number of morphemes, then σ is the concatenation of their stress patterns. To be more precise, the candidate (the output form-to-be) is compared to the way its components are realised as independent words (output forms) in the language—hence the name of the constraint. Affixes are not independent words of the language with some stress pattern, yet they may act as if they were: in Burzio’s approach, all the words with a given affix and a given stress pattern on that affix would jointly have such an analogy effect.

Burzio’s paradigmatic example is *condensation* as opposed to *compensation*. The word *còmpensàtion* is derived from *cómpensàte*, and the vowel of the unstressed second syllable may be reduced to a schwa. Yet, *còndensàtion* is derived from *condénse*, and the stressed second syllable in the root adds a tertiary stress to the second syllable of *condensation*, prohibiting its reduction to schwa.

A similar example has been proposed by Dicky Gilbers and Maartje Schreuder (personal communication). The six-syllable-long Dutch words *sèntimentàlitéit* (‘sentimentality’) and *ìndivìduàlíst* (‘individualistic person’) have seemingly very similar syllable structure: only their third syllables differ in weight, but if the weight-to-stress principle were active, it would predict the opposite pattern. Nevertheless, their morphological derivation is different:

Cycle 1	sèn.ti.mént	ìn.di.vi.dú	
Cycle 2	sèn.ti.men.téel	ìn.di.vì.du.éel	(4.4)
Cycle 3	sèn.ti.men.tà.li.téit	ìn.di.vì.du.a.líst	

Observe that it is cycle 2 which determines the stress pattern of cycle 3. If the root were the decisive factor, *sentimentaliteit* should have a stress on its third syllable, and *individualist* on its fourth one, but the change of the stress pattern in cycle 2 causes the opposite constellation. Interestingly, the native speaker of Dutch observes that the misplacing of the stress changes the semantic field of the (non-existing) word form: *sèntimèntalítéit* is conceived of as some kind of *mèntalítéit* ('mentality'), whereas *indivìdualíst* sounds as some sort of *dualist*.

Consequently, the stress pattern to which the different parsings of the input form *individualist* are to be compared to is *sususs* (s meaning stressed syllable, u referring to unstressed syllable): the stress pattern *susus* of *individueel* followed by the pattern *s* of the suffix (for the *ist* ending attracts stress). Similarly, *sentimentaliteit* is compared to the concatenation of the stress patterns *suus* from *sentimenteel* and of *us* from *-iteit*.

After these preparations, we are ready to define the constraint COMPONENT-OUTPUT CORRESPONDENCE. The number of violation marks assigned to a candidate w is the number of mismatches with the corresponding string σ , after a pairwise comparison of the corresponding elements of the (equally long) strings:

$$\text{COC}_\sigma(w) = \sum_i \Delta(w_i, \sigma_i) \quad (4.5)$$

where w_i and σ_i represent the i th element (in the present case, whether the i th syllable is stressed or not) of the candidate w and of the string σ used for the comparison; and where:

$$\Delta(w_i, \sigma_i) = \begin{cases} 1 & \text{if } w_i \neq \sigma_i \\ 0 & \text{if } w_i = \sigma_i \end{cases} \quad (4.6)$$

The definition of COC (or, OOC) is thus complete, but not satisfactory. The result is maybe not exactly what we wish. Intuitively speaking, misplacing one stress seems to be a smaller difference than missing a stress entirely, or having extra stresses. If the target string is $\sigma = \text{suus}$, then $w_1 = \text{susu}$ seems to be closer than $w_2 = \text{suuu}$ or $w_3 = \text{suss}$.⁵ Yet, the above definition will assign two violation marks to w_1 , because there is a mismatch in both the third and in the fourth syllable, whereas only one violation mark will be assigned to w_2 and to w_3 . Candidate w_1 violates constraint COC_σ on the same level as the "totally misconceived" candidate $w_4 = \text{ssss}$. Is this situation that we wanted?

In turn, a modification of the constraint may assign additional violation marks to the difference in the number of stressed syllables. Let $\|\alpha\|$ denote the number of stresses in the string α :

$$\|\alpha\| = \sum_i \Delta(\alpha_i, s) \quad (4.7)$$

⁵Again, from this point onwards, *s* refers to a stressed syllable with either a primary or a secondary stress, whereas *u* represents an unstressed syllable.

Subsequently, the new definition of COC is:

$$\text{COC}_{z,\sigma}(w) = \sum_i \Delta(w_i, \sigma_i) + z \cdot \left| \|w\| - \|\sigma\| \right| \quad (4.8)$$

Notice that the present definition introduces a new parameter, namely z , which determines the relative weight of the two parts, pointwise mismatch *vs.* difference in the global number of stresses. As pointed out by several readers, here I have combined two standard OT constraints. The first addend corresponds to IDENT(stress), and the second one to MAX(stress). Instead of having these two constraints in a strictly dominating rank order, we have just created a weighted sum in a Harmony Grammar-style. By varying z and keeping it small, the two addends, that is, constraints IDENT(stress) and MAX(stress), can create different interesting landscapes, as the experiments to be described in the next chapter shall demonstrate.

Last, one would define *Component-Output Correspondence* as COC_σ (or, $\text{COC}_{z,\sigma}$) with σ being always the concatenation of the immediate morphological components (in the present case, their stress pattern), and not the concatenation of deeper components. This would be how OT could account for the *bracket erasing convention* in Kiparsky (1982)'s Lexical Phonology.

On the other hand, one can make use of the above definition of COC_σ (or, $\text{COC}_{z,\sigma}$) when defining Burzio's OUTPUT-OUTPUT CORRESPONDENCE. Then, σ can be any element of the lexicon, and the definition should also define how to sum up the different COC_σ s:

$$\text{OOC}(w) = \sum_{\sigma \in \text{Lexicon}} d(w, \sigma) \cdot \text{COC}_\sigma(w) \quad (4.9)$$

with $d(w, \sigma)$ being some distance measure between the elements of the lexicon, which acts here as a weighting factor.

In Chapter 5, we shall make use of the *Component-Output Correspondence* constraint in the way we just have defined it, including also the z weight. Nonetheless, we shall call it Output-Output Correspondence, in order to make the discussion comprehensible to the reader familiar with past and current phonological literature, in which the term *Output-Output Correspondence* is used rather in the sense of *Component-Output Correspondence*, and not really following Burzio's original proposal.

4.2 Learning with SA OT?

The idea of learning a grammar has already been introduced roughly at the very end of section 1.1.3. The interest in learning is twofold: from the viewpoint of psycholinguistics, the question is whether a certain grammar model can reproduce language acquisition observations, such as those in child language, second language learning, post-traumatic language recovery, etc. The adequacy of a grammar model is clearly questionable if it cannot be acquired. On the other hand, natural language processing (NLP) may require machine learning algorithms (Mitchell, 1997) that can—at least partially—automate the construction of complex, high-coverage grammars.

In both cases, the goal is to find a grammar that reproduces the observed data (as well as possible). The problem is reversed compared to what we have

been dealing with so far: *grammar implementation* is concerned with producing the linguistic forms for a given grammar, whereas *grammar learning* aims at creating a grammar for given linguistic forms.

The basic philosophy is defined by Chomsky's *Principles and Parameters* (*P&P*) approach. Both acquisition in psycholinguistics and machine learning require a framework, otherwise the search for a grammar would be ill-defined. At this point, we cannot enter discussions about how much this framework has to be restricted, in what sense it is innate, and how poorly or amply a child is supplied with input data about her native tongue. What is usually supposed by linguists is that *some* of the grammar is *universal* (these are the *principles*), and it is already given to the learner at the beginning of the learning process. These principles reflect, as a matter of fact, features that are, arguably, characteristic of *all* languages of the world, as all human children inherit the same framework. The cross-linguistic differences are accounted for by the different values assigned to the *parameters*, and the task of the learner is to find the parameter setting reproducing the observed data.⁶

As discussed in section 1.1.3, traditional Optimality Theory postulates GEN, as well as the set of constraints to be universal. Applying *Principles and Parameters* to standard Optimality Theory means, therefore, that one searches for the constraint ranking that accounts for the input data, because it is the hierarchy that is supposed to be the only source of cross-linguistic variation, corresponding to the notion of “parameters” in *P&P*.

Grammar learning algorithms within standard Optimality Theory, *Recursive Constraint Demotion* (RCD) and *Error Driven Constraint Demotion* (EDCD), have been developed by Bruce Tesar (Tesar and Smolensky, 2000). A linguistically more informed version of RCD is *Biased Constraint Demotion* (BSD) (Prince and Tesar (2004), Tesar and Prince (2003)), used by Ota (2004) and by Pater (2005b) to learn lexically indexed faithfulness constraints.⁷ Constraint Demotion, however, lacks robustness: it presupposes that the data are produced by an OT grammar, the target of the learning algorithm, and that no noise infiltrates the data set. Cases requiring *Robust Interpretive Parsing* (Tesar (1999), Tesar and Smolensky (2000)), which inevitably introduce some sort of noise, may be unlearnable. Eisner (2000b) proposes a generalisation for RCD.⁸

The most popular of the learning algorithms for variations of OT is the *Gradual Learning Algorithm* (GLA) closely connected to *Stochastic Optimality Theory* (Boersma and Hayes, 2001), and widely used in recent years.⁹ Addi-

⁶Actually, many algorithms rather aim at reproducing the observed data only as well as possible. The data set may include *noise*, inconsistencies, errors, etc., and therefore finding a model that fits all the observed data perfectly is not always feasible. Furthermore, one may want to avoid *overfitting* (Mitchell, 1997): the goal, then, to be more precise, is to correctly predict the behaviour of the system on unseen data.

⁷Bruce Hayes lists the following learning algorithms with their earliest references in the manual of *OTSoft: A Constraint Ranking Software* (available at <http://www.linguistics.ucla.edu/people/hayes/otsoft/>, version of January 12, 2004): *Classical Constraint Demotion* (Tesar and Smolensky, 1993), *Gradual Learning Algorithm* (Boersma, 1997), *Low Faithfulness Constraint Demotion* (Hayes, 1999) and *Biased Constraint Demotion* (Prince and Tesar, 1999). Both of the later two are similar to Classical Constraint Demotion, but they attempt to place all faithfulness constraints as low as possible.

⁸For the application of EDCD to a heterogeneous data set, see an early manuscript at http://www.let.rug.nl/~birot/publications/t_biro_clin2002.pdf.

⁹For example Jäger (2003a) combines GLA with bidirectional OT (Blutner, 2000) in order to create a language evolutionary model.

tionally, stratified grammars are learned in Ota (2004) and Pater (2005b).

All these results showing that Optimality Theory is a learnable framework have significantly contributed to the success of Optimality Theory. The obvious question arising now is what *Simulated Annealing Optimality Theory* has to say about grammar learning. The question is open to further research yet, and here we can only speculate about the possibilities.

The most important contribution of SA-OT to the Optimality Theoretic paradigm is probably the *topology* (neighbourhood structure) of the candidate set. In section 2.2.2, it has been suggested that the topology should be universal and reflect the “logic” of GEN and of the inner structure of the candidates. If this is so, the structure on the candidate set does not have to be learned; rather, it is given to the learner initially. In a second approach, however, one could include a few parameters determining the details of the topology. In Chapter 7, for example, the basic operations transforming a candidate into its neighbours are supposed to be universal, and yet, the probability of applying a particular operation may vary. In such a model, a fine-tuning of the parameters is required to reproduce frequencies similar to those appearing in the learning data set. Details are postponed to further research.

Concerning the hierarchy, Simulated Annealing Optimality Theory uses a traditional approach, so the learner may want to use one of Tesar’s constraint demotion algorithms (EDCD or RCD). Do not forget that SA-OT deals exclusively with the way of calculating the optimal form in a standard OT model. Hence, you can also propose to build a Stochastic OT model and to learn with GLA; then, SA-OT is used to produce quickly an output at evaluation time for each hierarchy that is derived from the current ranks of the constraints by including noise. In both cases—constraint demotion and GLA—simulated annealing solves a seemingly elementary step cheaply, unimportant from the viewpoint of the learning algorithms. And yet, if generating the winner for a certain hierarchy is otherwise a costly operation, then learning algorithms calculating the optimal forms for different hierarchies many times would incur computational troubles. Consequently, a learning algorithm may be speeded up by using a heuristic technique.

SA-OT is not guaranteed to return the optimal candidate, however, and this fact introduces some noise into the learning algorithm. Does this observation disfavour less robust algorithms, such as EDCD? In fact, it most probably does not. Both EDCD and GLA generate the optimal candidate with respect to the current hierarchy in order to compare it to the piece of learning data. If the piece of learning data turns to be suboptimal, then the present hierarchy is altered in order to get closer to the target hierarchy, which would produce the observable data. Otherwise, no change is made. What happens, now, if SA-OT fails to find the optimal candidate for the current hierarchy? If the returned candidate is still better than the learning data, the detected error helps drive the learning algorithm (EDCD or GLA)—hopefully, towards the target hierarchy. Else, if the candidate returned happens to be worse than the piece of learning data, the learning algorithm mistakenly derives that the present hierarchy can account for the learning datum: in fact, the algorithm has just missed an opportunity to learn, and goes further to the next piece of data (unless this misconception causes the algorithm to stop). In sum, the (relatively low) noise introduced by SA-OT most probably has no other effect than to increase the number of learning steps required by the learning algorithm. Further experimentation may

compare the gain in speed due to the use of a heuristic technique to the increase in the number of steps caused by this noise.

A real SA-OT learning task would be the following: the learning data are produced using an SA-OT model (with known or unknown parameter setting), and a hierarchy is sought that reproduces the same distribution of outputs. Suppose that the topology and the set of constraints are given, and the goal is to find the association of the constraints with certain indices (domains of temperature) such that the landscape created by the model has the same local optima. Either a traditional learning algorithm would work, and once the global optimum (the grammatical form) is reproduced, the other local optima (the performance errors) are given for free by the topology; or the performance errors are also informative, and they provide further information for distinguishing between hierarchies that return the same global optimum. An additional task will be then to fine-tune the frequencies.

A third direction for combining simulated annealing, learning and Optimality Theory is to use simulated annealing not for production, but for learning. SA-OT performs a random walk in the structured candidate set searching for the best candidate with respect to a certain hierarchy. The dual (inverse) problem would be to search the (structured) set of possible hierarchies in order to find the best hierarchy for a certain set of learning data. Each hierarchy is scored by the number of learning data it generates correctly, yielding an integer-valued function to be maximised. Minimal permutations of the hierarchy could be the operation defining the neighbourhood structure. In fact, already Turkel (1994) observed the duality of production and learning in Optimality Theory, and he proposed to use genetic algorithms for both problems, an optimisation technique not very far from simulated annealing (cf. also section 1.2). Nonetheless, applying simulated annealing to the two, dual problems, have only few things in common: very different type of functions have to be optimised on a very different type of search space. I think that the similarities are too few, actually, to have a guilty conscience if I also leave that to future research.

The dual problems are much more closely related in the Maximum Entropy model advanced by Goldwater and Johnson (2003).¹⁰ Although simulated annealing and MaxEnt Optimality Theory are closely related at first sight, the two originate in very different approaches. Yet, some connection could be possible to be worked out through the polynomials used in section 3.3.

As superficially introduced in section 1.3.5, MaxEnt OT defines the probability of form o (derived from input i) as

$$p_{\{r_j\}}(o|i) = \frac{e^{-\sum_j r_j C_j(i,o)}}{Z_{\{r_j\}}(i)} \quad (4.10)$$

Here, r_j is the real-valued rank of constraint C_j , which, in turn, assigns $C_j(i, o)$ violations (not necessarily a non-negative integer) to the input-output pair (i, o) . $Z(i)$ is a normalisation factor, not important for us presently.

Observe that the exponent in this expression is a sum with addends composed of two factors. The dual problems, generation and learning, interconnect at this point. In production, the ranks $\{r_j\}$ of the constraints are fixed, and we search for the output o that maximises $p_{\{r_j\}}(o|i)$ for a certain input i . The

¹⁰See for instance Mullen (2002) for using MaxEnt for parse selection in Dutch, and for further references in the field.

grammar learner, however, varies the ranks $\{r_j\}$, so that the observed input-output pairs have the highest probability. See Jäger and Rosenbach (2006) for an implementation of a simulated annealing-like algorithm to learning in Max-Ent OT, called there *stochastic gradient ascent*, and argued to be a modification of GLA (Jäger, 2003b).

Chapter 5

Stress in Dutch Fast Speech with SA-OT

The goal of the second half of my dissertation—the goal of this chapter, as well as of the following two ones—is to present a few applications of Simulated Annealing Optimality Theory. The previous chapters have introduced the algorithm, and argued for it on theoretical grounds: by showing that different approaches derived from the philosophy behind Optimality Theory lead to the same algorithm. We have also seen how SA-OT connects to different aspects of language, such as the plausible computational complexity of tasks performed by the brain and possible errors introduced by the production process, as well as the proposal’s connections with the lexicon and learnability issues. Nonetheless, “a theory without an example is like a car without an engine: it may look gorgeous, but will bring you nowhere.”¹

Consequently, three models will be introduced, each of them having some additional illustrative goal. Indeed, the motivation is more methodological, to demonstrate how to work with SA-OT, than to account for specific linguistic phenomena.

This chapter demonstrates that SA-OT can indeed be used as a model of speech production, and outperforms the existing models. By varying parameter T_{step} , one can achieve different levels of precision, and if the algorithm is run faster, then fast speech phenomena emerge in the model. Besides, we shall look into the role of further parameters of the algorithm, in order to get a better grasp of it.

Subsequently, Chapter 6 applies SA-OT to voice assimilation. The problem seems to be simple, and yet, the first model to be proposed will not satisfy us, even if arguments can be brought in favour of it. Therefore, a second model will be introduced. The latter will show us how Simulated Annealing Optimality Theory can make use of doomed, never realised, seemingly totally redundant candidates in order to produce various distributions of the winning forms. With this second model, an infinite search space will also be introduced for the first time.

Thirdly, Chapter 7 shows how to syllabify inputs. First, the behaviour of the definite article in Hungarian is reproduced using SA-OT. Then, we turn to

¹Source: advice to the young researcher at ESSLLI 2002, Trento, Italy.

Basic Syllable Theory, which has been *the* classical example for OT combinatorial typology since Prince and Smolensky (1993). Different implementations of Optimality Theory have been demonstrated through this model: Tesar and Smolensky (2000) employ dynamic programming (chart parsing), whereas Gerdemann and van Noord (2000) uses finite state OT for the same problem. Thus, the performance of SA-OT on the same task is definitely an interesting question. Additionally, the model requires inherently an infinite candidate set—a challenging situation we will have encountered in connection with the second model of Chapter 6, but not earlier. Syllabification will also give us the opportunity to examine the role of the *topology* of the search space, the new concept introduced to Optimality Theory in subsection 2.2.2.

Finally, Chapter 8 closes the discussion of SA-OT by comparing it to the alternatives mentioned in section 1.3.

5.1 The Schreuder-Gilbers model of Dutch stress

Schreuder and Gilbers (2004) analyse the influence of speech rate on stress assignment in Dutch (see also Schreuder, 2006). The compound word *fototoestel* ('photo camera') is assigned in normal (slow, andante) speech a primary stress on its first syllable and a secondary stress on its third syllable (*fóto tòestel*). This output is faithful to the stress patterns of the component words: *fóto* and *tóestel*. However, in a laboratory experiment forcing the participants to produce fast (allegro) speech, Schreuder and Gilbers observed a stress shift in similar words: the secondary stress moved in a number of cases from the third syllable to the fourth one, which would yield *fótotoestèl* in our example.

The words used in the experiments of Schreuder and Gilbers (2004) can be partitioned into the following groups, exhibiting the following slow speech (andante) and fast speech (allegro) stress patterns:²

Type 0: andante: susu, allegro: suus, OO-correspondence to: susu
fo.to.toe.stel 'camera'

Type 1: andante: susuu, allegro: suusu, OO-correspondence to: susuu
stu.die.toe.la.ge 'study grant'
weg.werp.aan.ste.ker 'disposable lighter'
ka.mer.voor.zit.ter 'chairman of Parliament'

Type 2: andante: usus allegro: suus, OO-correspondence to: usus
per.fec.tio.nist 'perfectionist'
a.me.ri.kaan 'American'
pi.ra.te.rij 'piracy'

Type 3: andante: ssus allegro: suus, OO-correspondence to: ssus
zuid.a.fri.kaans 'South African'
schier.mon.nik.oog name of an island
uit.ge.ve.rij 'publisher'

²I follow closely the data of Schreuder and Gilbers (2004). Still, I have added as *Type 0* the word that they only use to introduce the phenomenon, but did not employ in their experiments. It will turn out to be useful also for our purposes.

Type 1.: right shift <i>stu.die.toe.la.ge</i> ‘study grant’	Type 2.: left shift <i>per.fec.tio.nist</i> ‘perfectionist’	Type 3.: beat reduction <i>uit.ge.ve.rij</i> ‘publisher’
susuu	usus	ssus
<i>stú.die.tòe.là.ge</i> slow: 0.81 fast: 0.38	<i>per.fèc.tio.níst</i> slow: 0.20 fast: 0.13	<i>ùit.gè.ve.ríj</i> slow: 0.96 fast: 0.67
<i>stú.die.toe.là.ge</i> slow: 0.19 fast: 0.62	<i>pèr.fec.tio.níst</i> slow: 0.80 fast: 0.87	<i>ùit.ge.ve.ríj</i> slow: 0.04 fast: 0.33

Table 5.1: **Observed frequencies per type:** The relative frequencies of the *andante* and of the *allegro* forms at fast and slow (normal) tempo, as observed by Schreuder (2006, pp. 80-82).

Here and henceforth, *s* refers to a stressed syllable and *u* to an unstressed one. We shall return to the role of Output-Output Correspondence presently, and the above table contains these notes only in order to save us space later.

Importantly, both forms of each word mentioned in the table occur in both *andante* and *allegro* speech (Schreuder, 2006). It does not even necessarily hold that one of the two forms dominates speech at slower rates whereas the other form outnumbers the first one at a faster tempo. The *andante* form of Type 1 (called *right shift* by Schreuder) is observed in 2/3 of the cases even in *allegro* style; whereas the fast speech form of Type 2 words (*left shift*) is pronounced four times more often than the slow speech form even in slow speech (Table 5.1 and Schreuder, 2006, pp. 80-82).

The phenomenon is rather a statistically significant shift in the observable frequencies. What is called the fast (*allegro*) speech form becomes more frequent at higher tempo, while the slow (normal, *andante*) form is characterised by a *relatively* higher frequency at a lower pace than at a higher pace. Only in this sense can we distinguish between *andante* and *allegro* forms; and only with this caveat do we refer to “errors in fast speech”. Indeed, some of these “errors” are also made in normal speech—very frequently at that, even though less frequently than in fast speech.

A second argument justifies why we refer to these forms as “erroneous forms”: why, for instance, the *suus* stress pattern of Type 2 words can be seen as “fast speech error”, even if it also dominates normal speech. Namely, the *andante* forms can be described with the same simple grammar, which will consequently be claimed to be *the* grammar describing the given language. The *usus* form of type 2 words corroborates the simple and convincing proposal that will account for the *andante* forms of the words in the other categories—the latter being really observed in more than 95% of the cases in *andante* speech. However, the attempt to create a model that reproduces the most frequent form of each type has not succeeded. Even if it did—referring to the previous argument—how could we explain the fact that the frequency of the grammatical forms according to this grammar changes inversely in function of speech rate for Type 2 words?³

In brief, the basic grammar to be proposed has to account for the stress

³As pointed out by Paul Boersma, there are also non-alternating *suus* forms such as *economie* ‘economy’.

patterns called the *andante* forms. Then, a second component of the model explain why the *allegro* forms also appear and become more frequent as the speech rate increases.

Having anticipated thus the existence of a grammar and a second component (which will obviously be the language production model, implemented as SA-OT), we turn now to a closer analysis of the different observable forms.

What does the experiment show? In slow (*andante*) speech, the words are pronounced in a way that reflects their inner structure. Types 0, 1 and 3 are compound words, and they keep the stress pattern of their components unchanged (e.g.: *fóto*+*tóestel* or *stúdie*+*tóelage*). Further, the examples in Types 2 and 3 end in a morpheme (a suffix, with the exception of *Schiermonnikoog*) that must bear stress. In sum, the stress patterns of the *andante* forms are fully derivable from morphology. On the other hand, all of the fast speech (*allegro*) forms display the *suus* pattern, followed by an extra unstressed syllable in the case of the five-syllable words of Type 1.

The faithfulness of the slow speech forms to their morphological structure can be accounted for by *Output-Output Correspondence* (also called as *Output-Output Faithfulness*)—or, Component-Output Correspondence, the constraint introduced in section 4.1.5—to a string derived from the morphology of that word. On the other hand, the pattern *suus* emerging in all of the types represents the form that is “easy to pronounce”; hence, it is accounted for by the corresponding markedness constraints. Faithfulness and markedness compete, slower speech emphasising faithfulness, while faster speech promoting the role of markedness.

Consequently, the solution proposed by Schreuder and Gilbers involves the *reranking* of the two crucial constraints, namely Output-Output Correspondence and the markedness constraint FOOT REPULSION ($*\Sigma\Sigma$). The latter corresponds to constraint $*\text{FTFT}$ introduced by Kager (1994) in order to account for similar cases of *ternary stress*. The latter punishes two adjacent feet⁴ that are not separated by an unparsed syllable, thereby making sure that stressed syllables are not too close to each other:

Definition 5.1.1. *The number of violation marks assigned by constraint FOOT REPULSION ($*\Sigma\Sigma$) to a candidate w is the number of syllable pairs (σ_1, σ_2) in w such that:*

1. σ_1 precedes σ_2 ;
2. the two syllables are adjacent (without any intervening syllable);
3. there exists metrical feet ϕ_1 and ϕ_2 in w , such that ϕ_1 contains σ_1 , ϕ_2 contains σ_2 , and ϕ_1 is a different foot from ϕ_2 .

If this constraint dominates the still importantly ranked FOOTBINARITY and PARSE- σ , then the ternary pattern resulting will be either “...[su]u[su]u[su]u...” (in case of a trochaic language) or “...[us]u[us]u[us]...” (if that language prefers iambic feet). No intervening syllable between two feet violates FOOT REPULSION, while not parsing more syllables in a foot does not improve the candidate for FOOT REPULSION and makes it worse for PARSE- σ .

⁴A *metrical foot* is an intermediate level between a syllable and a prosodic word, a parsing unit widely used in metrical stress theory (Hayes, 1995). Note that not all syllables have to be parsed into some foot: syllables not contained by any foot are referred to as *unparsed syllables*. A more formal definition will be given in section 5.3.

Indeed, the well-known constraint $\text{PARSE-}\sigma$ punishes unparsed syllables (cf. e.g. Tesar and Smolensky, 2000, p. 54) and will be soon used, as well:

Definition 5.1.2. *The number of violation marks that constraint $\text{PARSE-}\sigma$ assigns to candidate w is the number of syllables σ such that σ is not contained by any metrical foot ϕ in w .*

Schreuder and Gilbers (2004) present the following tableaux in order to explain why one of the candidates emerge in andante speech, whereas the second one in allegro speech:⁵

Slow (andante) speech:

fototoestel	OOC	* $\Sigma\Sigma$	$\text{PARSE-}\sigma$
$\mathbb{E}\text{S}$ (fóto)(tòestel)		*	
(fóto)toe(stèl)	*!		*

(5.1)

Fast (allegro) speech:

fototoestel	* $\Sigma\Sigma$	OOC	$\text{PARSE-}\sigma$
(fóto)(tòestel)	*!		
$\mathbb{E}\text{S}$ (fóto)toe(stèl)		*	*

(5.2)

The argument behind the explanation based on reranking is that the speaker is more constrained by the rhythmic beat enforced by constraint * $\Sigma\Sigma$ in fast speech than in slow speech. (See Kirchner (1998), who also accounts for casual speech phenomena by lowering the ranking of the faithfulness constraints and augmenting the base effort costs of gestures.) Therefore, in fast speech * $\Sigma\Sigma$ may overrank previously undominated constraints, such as Output-Output Correspondence.

Yet, this explanation raises at least two questions. First, fast speech errors are usually seen as a performance phenomenon. If an OT hierarchy is a competence model and the competence of the speaker is not altered, why would one explain these phenomena by using a totally new grammar? One may give up the idea of OT being a model of competence—but then, how is one to distinguish between competence and performance?

Second, suppose that one accepts that a sudden drastic change takes place in the grammar above a certain speech pace. Then, how can one explain the fact that the fast speech form appears only in some percentage of the cases—even if with an growing percentage as speech rate increases? If the grammar is altered, then the new form should *always* appear, which is definitely not the case, as discussed earlier (Schreuder, 2006).

⁵The candidate *(fó)to(toestèl)* is discarded by the high-ranked constraint FOOTBINARITY_μ (personal communication), but we shall not make use of this constraint in what follows. Constraint FOOTBINARITY_μ assigns a violation mark to each foot containing a light syllable only, but permits a foot of two moras: composed of two syllables or of one, but heavy syllable, such as *(stèl)* (Gilbers and Jansen, 1996). Otherwise, the fast speech form would be harmonically bounded by this form, and hence, the latter could never emerge. We shall return to this fact soon with respect to tableau (5.4). Output-Output Correspondence compares the candidate to the string *susu*, and a star simply shows a mismatch.

5.2 Fast Speech and different variations of OT

Besides reranking, most of the other models for variation presented in Section 1.3 are not very helpful either; primarily because they do not allow for fine-tuning the frequencies of the different outcomes.

Alternating outputs that are assigned exactly the same violation marks? Not even worth mentioning! As explained in section 1.3, this approach does not predict frequencies, and the analysis is extremely vulnerable to future research introducing new constraints. Coetzee (2004) refuses to give quantitative frequencies of the different forms, and hence, he cannot account for a phenomenon that involves the shift of frequencies. Suppose that your model has both *andante* and *allegro* forms optimal for all constraints ranked higher than the critical cut-off point. Coetzee suggests that the relation of their harmony predicts the relation of their frequency. And yet, as Table 5.1 shows, the phenomenon discussed often does not involve a reversal in relative frequency, but a simple shift in the quantitative values. Such observations are outside the scope of Coetzee’s model

In the different versions of stratified hierarchies, the frequencies are strictly defined, and the only imaginable way to proceed would be to *add* a new constraint to some stratum, which does not sound very promising, either. Even if we could argue for a new constraint emerging suddenly in fast speech, the possible predicted frequencies would be too rigid.

In Maximum Entropy models, the ranks assigned to the constraints could be varied—but this again would be a change in the *competence* model, whereas fast speech phenomena are rather changes on the level of the performance. Additionally, too many independent parameters would be required to describe how the rank of each of the constraints changes in the function of the speech rate: the explanation obtained might be little convincing.

A solution can be the use of *Stochastic Optimality Theory* (cf. also Boersma, 1998a). (Boersma, 1997) proposes, for instance, to raise the rank of all faithfulness constraints by a certain value in careful speech. But then, again, the competence model is altered for the sake of a performance phenomenon. Moreover, again, the parameters of the model are doubled, because Boersma and Hayes (2001) propose to introduce a new parameter for each constraint that measures the correlation of the rank of that constraint with style.

Nevertheless, StOT allows a second solution. By increasing the standard deviation σ of the evaluation noise in the function of the speech tempo (which may be done with a single monotonous function), the underlying competence model is kept intact, and yet, the likelihood of reranking two constraints increases.⁶

Let the three constraints introduced be distributed on the *continuous ranking scale* (i.e. before perturbation takes place) in such a way that the hierarchy is $\text{OOC} \gg * \Sigma \Sigma \gg \text{PARSE-}\sigma$, but the two highest ranked constraints are still relatively close to each other (Fig. 5.1). Now, suppose that fast speech results in increasing the *evaluation noise*: greater σ means enlarging the bell-shape of the

⁶Adam Albright has proposed a third solution that solves many of the problems of the first two solutions. Suppose that before pronouncing a word we run not one but several Stochastic OT evaluations. The different outputs for several randomly perturbed rankings are produced, then compared, and the best or the majority winner is uttered. Suppose also that more such “samples” are collected in normal speech than in fast speech. Therefore, the optimal candidate for the unperturbed hierarchy increases its chances over the winner of the reversed hierarchies more in normal speech than in fast speech.

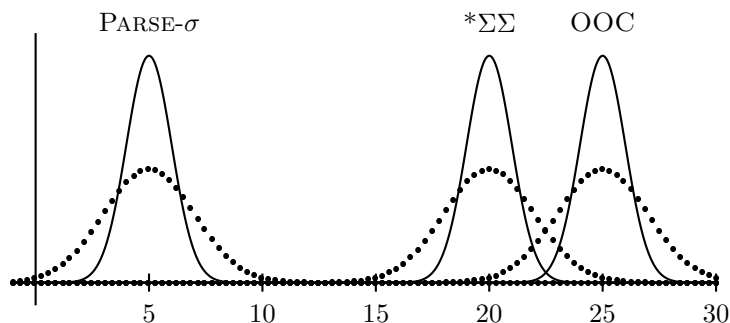


Figure 5.1: **The interaction of three constraints in Stochastic OT:** The three constraints proposed by Schreuder and Gilbers (2004) are assigned ranks along a real-valued scale, as proposed by Boersma (1997). As the two highest ranked constraints are relatively close to each other, they overlap. A lower random noise level results in a smaller overlap (solid lines), whereas more noise—a wider bell-shape distribution—increases the probability of reranking these two constraints (dotted lines).

distribution. High evaluation noise, subsequently, augments the probability of reranking temporarily the close constraints, OOC and $*\Sigma\Sigma$. As such a reranking causes the model to return the fast speech form with the shifted stress, we have created a model for fast speech.

The only hypothesis we need is to postulate that increasing the speech rate corresponds to increasing the σ defining the normal distribution of the evaluation noise. As speech rate grows, so does σ , causing the two constraints to be reranked more frequently, due to which the model correctly returns the fast speech form with a higher frequency. Future empirical research could formulate a more exact connection between speech rate and σ .

The advantage of this approach is that the distinction between competence and performance remains clear. The underlying competence model—the ranks associated with the constraints—is kept intact. This underlying model corresponds to the $\sigma \rightarrow 0$ limit, at which no performance error is predicted. There is no need for a theory about how the ranks of each constraint changes in function of speech rate. Only one parameter depends on speech rate, σ , the parameter that is decoupled from the static knowledge of the language, and which plays a role in the actual production process. And yet, the question still remains open to speculation: *why* should the evaluation noise (σ) change in function of the speech rate?

Moreover, further problems also arise. One such problem is that the present model predicts that the proportion of the fast speech form may never exceed 50%. Because the unperturbed rank of constraint OOC is higher than that of $*\Sigma\Sigma$, the chance of selecting points such that $\text{OOC} \gg *\Sigma\Sigma$ is always higher than the opposite ranking at evaluation time. In the $\sigma \rightarrow \infty$ limit, the probabilities of the two subhierarchies—hence, the predicted frequencies of the two forms—converge to 0.5 each.

This observation presents a problem, because the frequencies of fast speech forms may grow higher than 50% (Schreuder, 2006). With the help of a third constraint, Stochastic Optimality Theory could solve this problem.

fototoestel	OOC	* $\Sigma\Sigma$	CONSTRAINT-X
$\text{f}^{\text{f}}\text{to}(\text{t}^{\text{t}}\text{estel})$		*	*
$(\text{f}^{\text{f}}\text{to})\text{toe}(\text{st}^{\text{t}}\text{el})$	*!		

(5.3)

Imagine first that the three constraints are assigned exactly the same ranks. This case corresponds to Anttila’s model with unranked strata: out of the six permutations of the three constraints, two yield $(\text{f}^{\text{f}}\text{to})(\text{t}^{\text{t}}\text{estel})$ as the winner, and four return $(\text{f}^{\text{f}}\text{to})\text{toe}(\text{st}^{\text{t}}\text{el})$. Thus, the predicted frequency of $\text{f}^{\text{f}}\text{otot}^{\text{t}}\text{estel}$ is only 33%. Now, let us demote constraints * $\Sigma\Sigma$ and CONSTRAINT-X only minimally, that is, let us decrease their ranks with some values that is much smaller than σ . The resulting model should be interpreted as $\text{OOC} \gg \text{*}\Sigma\Sigma, \text{CONSTRAINT-X}$. Nevertheless, the probability of returning $\text{f}^{\text{f}}\text{otot}^{\text{t}}\text{estel}$ has not increased much, because this form—the only grammatical one according to the underlying model—is returned exclusively if the selection points are such that OOC overranks both of the two other constraints (Jäger and Rosenbach, 2006).

The problem now is that even though Stochastic OT is able to predict less than 50% frequency for the grammatical form, a third constraint is required—what should be that third constraint? The set of constraints appearing in the phonological literature is ample, and yet, adding a new, highly ranked constraint to a model is always risky.

The following argument against Stochastic Optimality Theory could also be refuted by an alternative proposal including different constraints. Still, SA-OT will spare you the time of hunting for a new phonological model. Namely, notice that in some of the types, all possible parses of the allegro form are harmonically bounded:⁷ the first two rows by the third row, and the fourth row by the fifth one in tableau (5.4). The fast speech forms can never win for any constraint hierarchy, since there is always a candidate that violates each constraint fewer or equal times.⁸

	OOC	* $\Sigma\Sigma$	PARSE- σ
$[\text{su}]\text{u}[\text{s}]$			
$[\text{s}]\text{u}[\text{us}]$	*		*
$[\text{s}]\text{uu}[\text{s}]$	*		**
Type 0: $[\text{s}]\text{u}[\text{su}]$ Type 2: $[\text{us}]\text{u}[\text{s}]$			*
$[\text{su}][\text{us}]$	*	*	
Type 0: $[\text{su}][\text{su}]$ Type 2: $[\text{us}][\text{us}]$		*	

(5.4)

The most serious criticism follows, however. Observe that for a given σ the probability of reranking the two constraints is constant. Therefore, Stochastic

⁷For the definition of *Harmonic Bounding* see Definition 1.3.1, on page 24.

⁸Concerning OOC, no exact analysis is necessary. Here, we content ourselves with the observation that the andante forms satisfy OOC, unlike the allegro forms. This fact is the key to the allegro forms being harmonically bounded.

Optimality Theory predicts that the probability of stress shift is the same for all of the four types, because all of them are the consequences of the same phenomenon, the reranking of faithfulness and markedness. The experiment of Schreuder and Gilbers (2004) shows, however, very significant differences in fast speech form frequencies. Two solutions remain for Stochastic OT: either the introduction of new constraints that differentiate between different types (different words), or some explanation why σ depends not only on speech rate, but also on the input word.⁹

But then again, Simulated Annealing will not require any further manipulation: it predicts correctly which word is more likely to undergo stress shift at a given speech rate. This difference is not introduced by an additional factor (new constraints, σ dependent upon the input), but by something that is already in the model, namely, the structure of the search space.

5.3 Fast speech and SA-OT: the building units

Let us apply simulated annealing to stress assignment. In building the model, we follow closely the five steps described on page 45:

- Step 1: Define the candidate set.
- Step 2: Define a neighbourhood structure (topology) on the candidate set.
- Step 3: Define the Harmony function to be optimised: what are the constraints and how are they ranked.
- Step 4: Define temperature and the transition probabilities.
- Step 5: Define the cooling schedule and perform the simulation.

The input is a word composed of a number of syllables, say, $\sigma\sigma\sigma\sigma$. The set of candidates corresponding to this input is formed by all its possible correct parses. A parse of a word groups syllables into units called *feet*. Hayes (1995) writes:

*One of the seminal ideas in metrical stress theory is this: the best way to express stress rules might not actually be the most direct one, that is, to place stress on a particular syllable. The alternative is to state the possible structures for metrical constituents and construe stress placement as the parsing of a word into such constituents. These constituents, the minimal bracketed units of metrical theory, are called **feet**. (p. 40, emphasis in the original)*

While ignoring the difference between primary and secondary stress, a correct parse meets the following criteria in metrical stress theory (cf. Hayes (1995), Tesar and Smolensky (2000)):

- It contains the same number of syllables as the input.

⁹An analogous train of thought will also appear in connection to Simulated Annealing Optimality Theory in later chapters. Still, there the connection between the parameters (the simulation speed) and the speech rate is more straightforward, and therefore the argumentation might be more convincing.

- It contains at least one foot.¹⁰
- Feet do not overlap.
- A syllable not belonging to any foot (an unparsed syllable) is unstressed.
- Each foot contains one or two syllables.
- Each foot contains exactly one stressed syllable (called the *head* of the foot).

For instance, a four-syllable input word $\sigma\sigma\sigma\sigma$ may be parsed as $u[s]uu$, $[su]uu$, $[us]u[s]$, $[s][s][s][s]$, etc. Here, brackets represent foot borders, whereas u and s refers to unstressed and stressed syllables, respectively. The set of possible parses is finite, and can be easily formulated as a regular expression (cf. B    , 2003, 2005c).¹¹

The set of candidates having been defined, we can now proceed to define the *topology* (neighbourhood structure) of the search space. Among the different strategies proposed in section 2.2.2, we chose the simplest one, described by equations (2.4) and (2.5): each candidate has a small number of neighbours with equal probability.

In particular, the *neighbours* of a candidate are the candidates which can be reached within one *basic step*, where a *basic step* is defined as performing exactly one of the following actions:

- Insert a monosyllabic foot: turn an unparsed u into $[s]$.
- Remove a monosyllabic foot: turn $[s]$ into an unparsed u (unless the resulting form would contain no foot).
- Move one foot border: enlarge a foot by taking an unparsed syllable into a monosyllabic foot.
- Move one foot border: narrow a foot by taking an unstressed syllable out of a foot.
- Change which is the head (stressed) syllable within a bisyllabic foot: $[su]$ to $[us]$, or vice versa.¹²

These basic steps are minimal changes in the metrical structure of a parse, and follow very simply from the logic of what a correct parse is. Observe that the neighbourhood relation is a regular mapping on the set of candidate strings, because it is the union of very simple regular operations. Moreover, it is important that any parse can be transformed into any other parse with some series of basic steps, the neighbourhood structure obtained is a connected graph.

¹⁰When not ignoring the difference between primary and secondary stress, we require the word to contain exactly one *main* / *head* foot, instead of this rule. The stress in that foot is the primary stress of the word. All the other feet, which contain secondary (tertiary,... in some theories) stresses, are optional. Note that all these rules apply only to *prosodic* words, which are defined as having exactly one main stress. Clitics and further, unstressed linguistic units belong to some adjacent prosodic word.

¹¹See especially an early manuscript at <http://www.let.rug.nl/~birot/publications/t-biro.clin2002.pdf>.

¹²The notations have not been chosen to refer to the Soviet Union and to the United States...

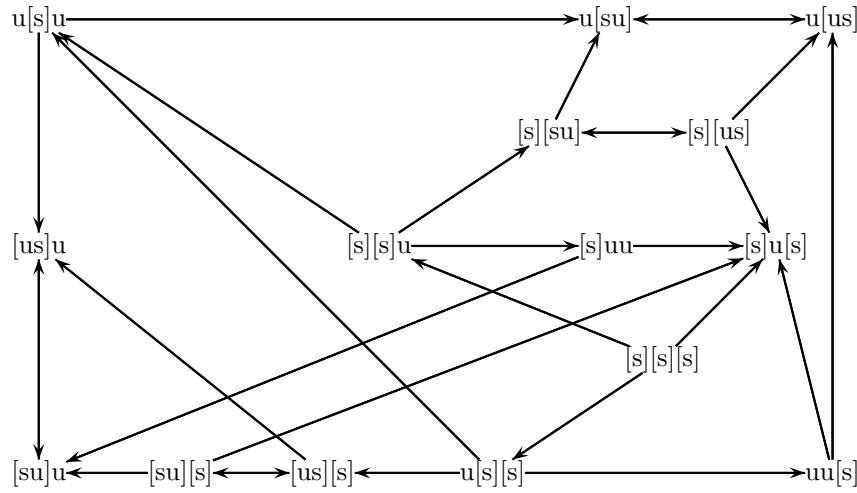


Figure 5.2: **Topology for metrical stress formed by a three-syllable input:** The arcs of the graph visualise the neighbourhood relations. The *a priori* probabilities are omitted, for they follow from equation (2.5). An arrow points to a neighbour that is not less harmonic with respect to the toy hierarchy $^*\Sigma\Sigma \gg \text{PARSE-}\sigma$.

According to this definition, uu is not adjacent to $[us]$, although the difference is only the existence of a foot, as has been formulated by the readers. I viewed the insertion of a bisyllabic foot as a more complex operation than those listed above, and therefore did not include it among the basic steps; but future work should indeed investigate the consequences of an altered definition of neighbourhood. Alternatively, one might propose to have the set of basic steps *minimal* in the sense that removing one of the permissible actions may result in a disconnected graph. Then, the insertion of a bisyllabic foot can be missed, for it can be replaced by two permissible steps ($uu \rightarrow u[s] \rightarrow [us]$), but then the same applies also to changing the head syllable of a bisyllabic foot ($[su] \rightarrow [s]u \rightarrow [s][s] \rightarrow u[s] \rightarrow [us]$).

Defining the topology of the search space, finally, includes specifying the probability measure according to which one of the neighbours is chosen at each step of the simulation. Thus, we will have accomplished the three steps of defining a topology mentioned on page 48. For the sake of ease, we shall assign equal probability to each neighbour. So, if candidate w has three neighbours, each will be chosen with probability 33%; if, however, w has four neighbours, they receive 25% each.

The graph in Fig. 5.2 shows the topology of the search space in the case of a three-syllable input. Although our simulations will be based on inputs with four or five syllables, the graph formed by the 43 candidates generated from a four-syllable input would be too complex to display here, let alone the 119 candidates of a five-syllable input. The arcs of the graph connect candidates that are neighbours, that is, a single basic step will transform one into the other. The arrows on the arcs point towards the candidate which is at least as harmonic, with respect to the ranking $^*\Sigma\Sigma \gg \text{PARSE-}\sigma$.

This brings us to the next step in Simulated Annealing, step 3 from the five steps repeated at the beginning of the present section. Once we have the topology of the search space, that is, the horizontal map of the landscape in which the random walk will take place, we proceed to build the vertical structure of the landscape. Each point of the search space is situated “higher” or “lower”, where more harmonic candidates are lower than the less harmonic ones. An arc on the graph with only one arrow points downhill, whereas an arc with two arrows represents a horizontal move. If there is an arrow from candidate w to candidate w' then the move from w to w' will be possible with a probability of 100% during the entire simulation.

Eyeballing the graph, we can point to some phenomena. Candidate [s][s][s] represents a summit, so that you go downhill, no matter which direction you take. In fact, it is the global maximum, the worse candidate of all, but this fact cannot be seen directly from the graph. On the other hand, the candidate [s]u[s] is a local minimum: the arrows from all its neighbours point towards it. We also find valleys, such as the one formed by u[su] and u[us], or the one formed by [us]u and [su]u. These two valleys are formed by candidates of equal harmony, but situated lower than their surroundings. If we compare the local minima, [s]u[s], u[su], u[us], [us]u and [su]u, it turns out that all of them are global minima as well. However, the graph itself cannot help in determining which of the local minima is a global minimum.

By including additional constraints ranked below PARSE- σ , (such as alignment constraints, NONFINAL, FOOTBINARITY, IAMBIC, TROCHAIC or FOOT-NONFINAL, cf. Tesar and Smolensky, 2000, or other standard OT literature), we could differentiate between these local minima. Indeed, *all* constraints from the supposedly universal constraint set must appear somewhere in the hierarchy—even if very low. Similarly to the proposal in section 1.3.1, SA-OT could claim that two neighbours form a horizontal platform only if we knew all the constraints in the language. Nonetheless, if there is a constraint C such that each pair of neighbours differs for some constraint ranked not lower than C , then the constraints ranked below C do not play a role in SA-OT.

Although constraints have already been introduced earlier (definitions 5.1.1 and 5.1.2, as well as equation (4.8) in section 4.1.5), we repeat them here in order to accomplish the first part of step 3 of building an SA-OT model. If $\#A$ denotes cardinality of the set A (the number of elements in A), then

$$\begin{aligned} \text{COC}_{z,\sigma}(w) &= \sum_i \Delta(w_i, \sigma_i) + z \cdot \left| \|w\| - \|\sigma\| \right| \\ * \Sigma \Sigma(w) &= \# \left\{ \{\sigma_1, \sigma_2\} \mid \text{adjacent syllables of } w \text{ in different feet} \right\} \\ \text{PARSE} - \sigma(w) &= \# \left\{ \sigma \in \text{syllables of } w \mid \sigma \text{ is unparsed in } w \right\} \end{aligned} \quad (5.5)$$

In other words:

- OO-correspondence: compare to susu, syllable-by-syllable, assign one * to each difference. Plus: z times the difference in number of stresses.
- * $\Sigma\Sigma$: assign one * per “[“.
- PARSE- σ : assign one * per unparsed “u”.

The second half of step 3 is specifying the hierarchy. Earlier, we have argued that the hierarchy has to make the andante form the optimal candidate. Therefore, our starting hierarchy is:

$$\text{OOC} \gg * \Sigma \Sigma \gg \text{PARSE-}\sigma \quad (5.6)$$

The definition of temperature for SA-OT has been argued for very elaborately in chapter 2. What remains is performing the simulation for different cooling schedules.

5.4 Experimenting with the Schreuder-Gilbers model

To summarise and to focus first on Type 0 words (such as *fototoestel*), the search space includes all 43 possible parses of this 4-syllable input. The topology of the search space is defined by a *neighbourhood relation* such that a candidate w' is a neighbour of the candidate w if and only if w' can be constructed from w by applying exactly one of the five basic steps mentioned earlier: insertion of a monosyllabic foot, removal of a monosyllabic foot, enlarging a foot by moving its border, narrowing down a foot by moving its border, or changing the choice of the head syllable within a bisyllabic foot. The probability distribution over the set of the neighbours of a given candidate w is a constant function: each neighbour is chosen with equal probability. Furthermore, the Harmony function over the search space is defined by the constraint hierarchy $\text{OOC} \gg * \Sigma \Sigma \gg \text{PARSE-}\sigma$, using the definitions in (5.5).

We begin our experiments with the word *fototoestel* (Type 0), that is, the input is a four-syllable word, and OOC is calculated with respect to the input string *susu*, by postulating $z = 2$. The parameters used for the algorithm are: $T_{\max} = 3$, $T_{\min} = 0$, $K_{\max} = 3$ and $K_{\text{step}} = 1$. The parameter K_{\min} was chosen as a function of T_{step} :

$$\begin{aligned} K_{\min} &= -100, & \text{if } T_{\text{step}} > 0.5, \\ K_{\min} &= -30, & \text{if } 0.5 \geq T_{\text{step}} > 0.1, \\ K_{\min} &= -6, & \text{if } 0.1 \geq T_{\text{step}} > 0.05, \\ K_{\min} &= -3, & \text{if } T_{\text{step}} \leq 0.05, \end{aligned}$$

Furthermore, the algorithm will be launched from each of the 43 candidates as initial state. Similarly to the first experiments performed in section 2.3.1, our goal is to measure the probabilities of returning different outputs in the function of the parameters—first of all, as a function of T_{step} .

Thus, we launched the algorithm 300 times from each of the 43 candidates, which corresponds to $n = 12900$ runs. Supposing a Gaussian distribution of the outputs, this number of repetitions allows for a random fluctuation below 1%. As K_{\min} has been chosen low enough to ensure that the system reaches a local optimum, no other candidate is returned.

Now, this model includes three local minima. The candidate [s]u[su] is the global minimum—not surprisingly, as we have chosen the hierarchy so that the optimal one be this candidate, a possible parse of the andante form. The model includes two further local minima: [su]u[s] and u[us][s], which present possible

pitfalls for the algorithm. Candidate $[su]u[s]$ corresponds to the observed allegro form, whereas $u[us][s]$ is not observable in Dutch.

I do believe that the fact that such a simple hierarchy and a straightforward topology has immediately yielded the correct fast speech form as a local optimum is far from being trivial. It points towards the cognitive adequacy of the topology. Remember, the hierarchy has been chosen to account only for the andante form, whereas absolutely no empirical considerations have been given when defining the topology, besides metrical stress theory, based on very general cross-linguistic observations. It seems at first sight that the local optimum $[su]u[s]$ is an automatic consequence of $[s]u[su]$ being the global optimum. Yet, after this very first, unexpectedly easy success, we will have to struggle hard to get rid of the *artefact form* $u[us][s]$. The difficulty of this struggle will reinforce our conviction that the development of a successful model is far from being trivial.

Before going further, one can check that candidate $u[us][s]$ is indeed a local optimum (using OOC with respect to string $susu$, with $z = 2$). The following tableau contains the candidate $u[us][s]$ and all its neighbours. The \sim symbol refers to a local optimum in the present model, whereas the exclamation mark precedes the fatal violation with respect to $u[us][s]$.¹³

	OOC	* $\Sigma\Sigma$	PARSE- σ
\sim $u[us][s]$	**	*	*
$u[us]u$	**!*		**
$[s][us][s]$	**!*	**	
$uu[s][s]$	**	*	*!*
$u[su][s]$	**!*	*	*

(5.7)

Similarly, $[su]u[s]$ is also a local optimum, because it is more harmonic than all its neighbours:

	OOC	* $\Sigma\Sigma$	PARSE- σ
\sim $[su]u[s]$	**		*
$[s]uu[s]$	**		*!*
$[su][us]$	**	!*	
$[us]u[s]$	**!*		*
$[su][s][s]$	**!*	**	
$[su]uu$	**!*		**

(5.8)

Other candidates are not local optima. For instance, candidate $[s]u[s]u$ is less harmonic than its neighbour $[s]u[su]$, with PARSE- σ as fatal constraint. One could similarly demonstrate that all the remaining 39 candidates are not local optima. Yet, it is simpler to run a gradient descent (a simulated annealing launched with $K_{max} < 0$ if the lowest ranked constraint is in domain 0) from each of the candidates. As gradient descent is not able to escape from local optima, a simulation launched from such a state would return it necessarily. This experiment shows that the model, indeed, contains no other local optimum.¹⁴

¹³The reader is invited to check these tableaux by using the demo program available at <http://odur.let.rug.nl/~birot/sa-ot/Dutch-stress.php>.

¹⁴If the number of candidates is very large, checking by hand whether each candidate is a local optimum is not feasible. As a side remark, however, on philosophical grounds, one may

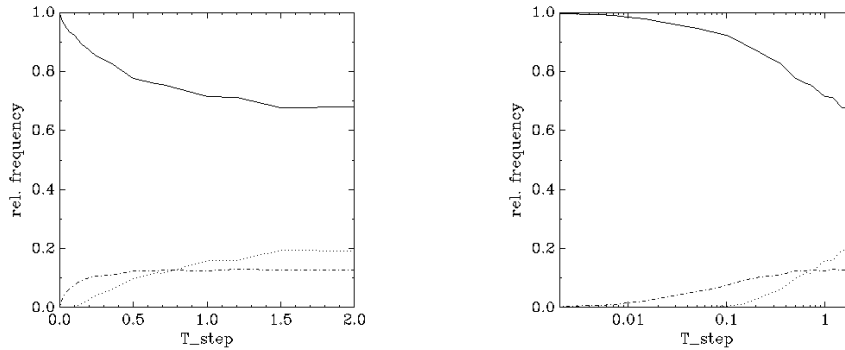


Figure 5.3: **Tuning T_{step} in the Schreuder-Gilbers model:** The proportions of producing the correct form according to competence ([s]u[su], solid line), the performance error form typical to fast speech ([su]u[s], dotted line), and the artefact form (u[us][s], dot-dashed line), as a function of the temperature step employed. The left box uses a linear scale, while the right box uses a logarithmic scale. Note that the exact values represented here originate from a different experiment from those appearing in Table 5.3, even if the shape of the curves are comparable.

Observe that although [su]u[s] and u[us][s] are local optima, they are far from being “quite good”. Indeed, several candidates ([su][su], [s]u[s]u, etc.) are much better than these, because they satisfy the highly ranked constraint OOC; and yet, they are not local optima. In this respect, metrical stress does not resemble spin glasses discussed in section 2.1.3: *the energy level in local optima is far from close to the level in the global optimum*. However, if SA-OT is successful in predicting speech errors, then the take-home message is that SA-OT is indeed a correct model of human speech, but the mistake was done by nature when implanting an optimisation algorithm in the brain that is not suitable for the given problem. In other words, the message could be that nature does compete with computational linguists in reaching ever higher precision, as long as communication is efficient enough.

After these preliminary observations, let us proceed to the results of the experiments (Fig. 5.3). As expected, the parameter T_{step} influences the results of the simulation the most. The other parameters being unchanged, it is inversely proportional to the number of iterations performed, that is, to the speed by which we reduce the temperature.

The global optimum form [s]u[su] is returned in some 68% of the cases when $T_{step} = 2$ and in more than 99% when $T_{step} = 0.01$. The proportion of the performance error form [su]u[s] drops below 10% at $T_{step} = 0.5$ and below 1%

refuse having the computer check it, similarly to the objection raised until recently against the 1977 proof of Appel and Haken of the Four-Colour Theorem. Fortunately, local optima in the present model are of much smaller significance than the Four-Colour Theorem.

approximately at $T_{step} = 0.1$. However, the second local optimum, $u[us][s]$, which corresponds to no observed form, is more persistent: with $T_{step} = 0.2$, it is still produced in almost 10% of the cases, and with $T_{step} = 0.01$, it appears in some 2% of the simulations. Interestingly, $u[us][s]$ is more stable than $[su]u[s]$ despite the fact that $u[us][s]$ is less harmonic than $[su]u[s]$ (compare tableaux (5.7) to (5.8)): as they are not neighbours, they actually never get compared. A moral of this observation is that, due to the complex structure of the random walk’s landscape, different local optima may behave in a very different manner as a function of the cooling schedule: some disappear quickly with a higher number of iterations, whereas other “traps” might be much more difficult to avoid for the simulation.

Consequently, the model successfully predicts that the form *susu* (to be more precise, the parse $[s]u[su]$) dominates normal speech, whereas the likelihood of the allegro form (*suus*, in the form of the parse $[su]u[s]$) increases significantly in fast speech. Notice, however, that the speed of the algorithm cannot be interpreted directly as speech pace, because the latter does not increase with one or two orders of magnitude as does the algorithm when contrasting $T_{step} = 2$ as to $T_{step} = 0.1$. So, we either see the present model only as a first approximation to how to model the fact that speech precision decreases due to increased speed; or we speculate about the speech organs failing to meet the increased pace set by the brain.

As regards the three types of words and their observed frequencies in the experiment of Schreuder and Gilbers (2004), we shall return to them after having refined our model.

What to do with the emergence of the absurd form $u[us][s]$? In section 5.6, an improved model will be proposed—including further constraints—that matches better the observed forms in Types 0-3, and which involves an unattested form only for Type 2. Nonetheless, a really good model, which also satisfies phonologists,¹⁵ is still under development.

Before changing the phonological model, however, let us analyse further the present system through a few more experiments. Not so much in order to find a better match between empirical results and the model, but rather in order to obtain a better understanding of Simulated Annealing Optimality Theory in general.

5.5 Further experiments

5.5.1 The role of T_{step}

The goal of this first simulation was to observe the role of T_{step} more closely, repeating some of the observations already advanced in the previous section. Again, the hierarchy was $OC_{z=2, \sigma=susu} \gg * \Sigma \Sigma \gg \text{PARSE-}\sigma$, corresponding to the Dutch word *fotoestel* (‘photo camera’). Furthermore, each of the 43 candidates acted 100 times as the starting point of the simulation, producing 4300 outputs in total. The parameters of the simulations were also the same as in the simulation presented in the previous section.

¹⁵For instance, ample phonological arguments have been brought in the literature for Dutch having no iambic foot (that is, $[us]$). Although the foot borders cannot be empirically tested, phonologists would like to see a model that parses the word *perfectionist* not as $[us]u[s]$, but as $u[s]u[s]$.

First, let us compare the outputs of the simulation with $T_{step} = 1$, corresponding to fast (allegro) speech, to those with $T_{step} = 0.1$, arguably modelling slow (andante) speech. With $T_{step} = 1$, the model has little chance to rove in the search space, and in practice, you just slide down the slope, and reach the bottom of the valley in which you are initially. With a number of iterations increased tenfold ($T_{step} = 0.1$), however, the search space could be walked through if there were no obstacles—as is the case in the initial phase of the simulation. The outputs of such an experiment are summarised in the following table:

Outputs	slow (andante) speech	fast (allegro) speech
[s]u[su]	3940	3051
[su]u[s]	17	660
u[us][s]	343	589

Table 5.2: Slow versus fast speech in the simplest Schreuder-Gilbers model

In both simulations, most of the 4300 runs produced [s]u[su], i.e. *fótotòdestel*. This is the output predicted by the underlying Optimality Theoretic model, and corresponds to the form seen as correct by the linguistic competence of the native speaker.

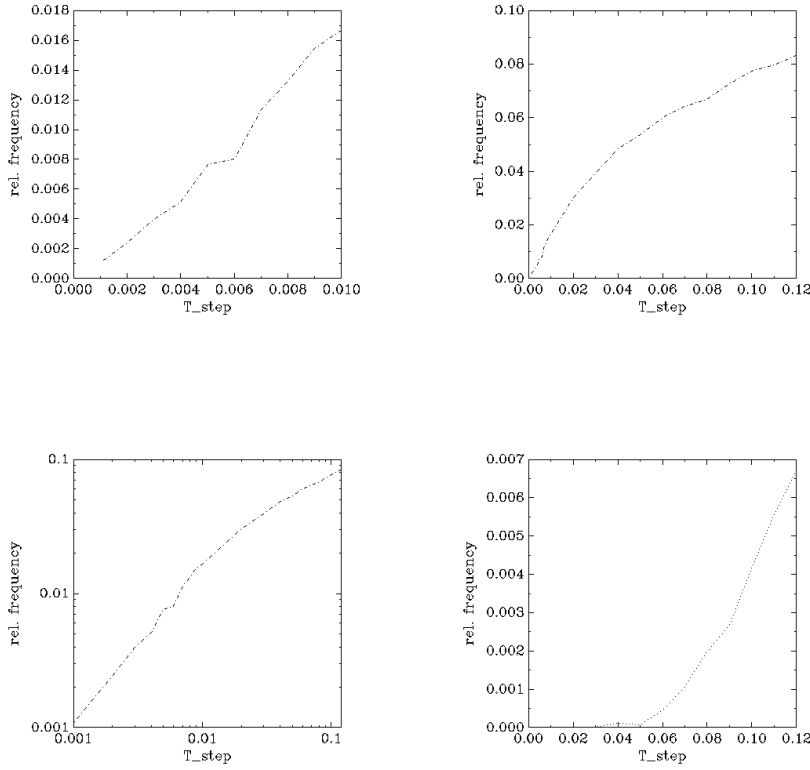
An encouraging point to note in the chart is, however, the drastic increase in the suboptimal form [su]u[s]. The 17 cases in slow speech is far below 1%. In fact, running the simulation again with the same parameters resulted sometimes in 0 or 1 cases only, among 4300. On the other hand, the parameter setting used for modelling fast speech produced [su]u[s] in approximately 15% of the cases. From repeating the experiment several times, the fluctuation of the output frequency of [su]u[s] in the fast speech model seems to be less than 2%. In sum, as we change the parameters of the simulation, the algorithm changes its behaviour, and produces more outputs that correspond to a local, but not global optimum. This result mirrors exactly the way in which changing speech rate results in a higher percentage of the fast speech forms. Unfortunately, but not surprisingly, the third parse, unattested in speech, u[us][s], is again present, and it is more stable as T_{step} diminishes than the more harmonic candidate [su]u[s].

Table 5.3 presents the number of outputs [s]u[su], [su]u[s] and u[us][s] produced under different speeds of simulation, in a repeated experiment, in which the simulation was run $n = 12900$ times (300 times for each candidate as starting point). The graph on Fig.5.3, presented earlier on page 135, reports on a distinct run of the experiment with the same settings: although the exact values are different, the shapes of the curves are similar.

A closer look both at Figure 5.3, as well as at Table 5.3 will reveal some interesting details. The higher T_{step} , the less probable it is that the correct output ([s]u[su]) is returned. However, the ratios of the two alternative outputs do not increase uniformly. The form u[us][s] is produced even with very low T_{step} values, while the form [su]u[s], the one observable in fast speech, is not produced at all for $T_{step} < 0.05$. For high T_{step} values, the two local optima have a comparable likelihood, meaning that the ratio of [su]u[s] grows more quickly. It even reaches a significantly higher frequency for the highest T_{step} -values.

Figure 5.4 relates the results of another repetition of the same experiment, which zoomed in on the domain $0.001 \leq T_{step} \leq 0.1$. In order to achieve a higher significance level, now 25800 outputs were generated for each T_{step} (i.e.,

T_{step}	nr. of [s]u[su]	nr. of [su]u[s]	nr. of u[us][s]
2	8825	2423	1652
1.5	8991	2316	1593
1	9110	2034	1756
0.75	9601	1741	1558
0.5	10150	1232	1518
0.25	10986	518	1396
0.2	11332	332	1236
0.15	11582	196	1122
0.1	11838	40	1022
0.05	12190	1	709
0.02	12565	0	335
0.01	12686	0	214

Table 5.3: Tuning T_{step} in the simplest Schreuder-Gilbers modelFigure 5.4: **Zooming in on the interval $0.001 \leq T_{step} \leq 0.1$:** The frequency of [su]u[s] (dotted line, lower right box), and of u[us][s] (dot-dashed line, three boxes), on different scales.

600 runs for each candidate as the starting points).

The upper two boxes present the relative frequency of producing the “artefact” form $u[us][s]$ as a function of T_{step} on two different scales. The lower left box combines these two figures within one loglog plot. The lower right box demonstrates the way the fast speech form $[su]u[s]$ appears with increasing T_{step} . Looking at the numerical values represented in the graph, one gets the impression that relative frequency first grows as an exponential function, then as a power law with an exponent higher than 1; later the exponent decreases to lower than 1, reaching finally an almost steady state function, such as the relative frequency of $u[us][s]$ with $T_{step} > 1$.

5.5.2 The role of T_{max} and T_{min} (1)

So far, we have varied the parameter T_{step} , while keeping the other parameters unchanged. One may wonder what role the other parameters play in the SA-OT algorithm (Fig. 2.8). Now, we are going to vary T_{max} and T_{min} , whereas the later part of Chapter 6 will examine the role of another interesting parameter, K_{max} .

In fact, the primary role of varying T_{step} was to change the number of iterations. We could have also introduced an additional parameter that specifies how many times we repeat the core of the loop before diminishing the temperature. Some readers would find it more elegant, though I think that the algorithm has already enough parameters without introducing an extra one. What would be the difference between varying T_{step} and introducing a parameter of repetition R ? The answer is related to the following question: how important is the exact value of the second component of the temperature $T = \langle K, t \rangle$? Indeed, if the second component t does not influence the outcome, we could have kept the temperature unchanged for a while and replaced the inner loop with the repetition of its core $R = \frac{T_{max} - T_{min}}{T_{step}}$ times.

Parameters T_{step} , T_{max} and T_{min} drive the inner loop of the SA-OT algorithm, which diminishes this second component t . Besides mere repetition, the role of t is to influence the transition probabilities, but only in the supposedly rare case when the fatal constraint’s index k is equal to the first component K of the temperature. How significant is this case? The following experiments demonstrate that it does influence the outcomes of the algorithm. Varying the upper and lower limits of the inner loop—that is, parameters T_{max} and T_{min} —results in a measurable, though small variation in the distribution of the outputs.

I conjecture that the effect would be larger in a model with fewer constraints, and smaller in the case of more constraints, because more constraints would diminish the chance of the fatal constraint’s index coinciding with K . If only few constraints are present, or if the differences in the violation levels of some constraints matter much, then t becomes more important. As an example, recall that the second component of the temperature is crucial in an extreme case such as the one appearing in tableau (2.18). In fact, this is the situation where SA-OT turns into traditional simulated annealing. In this case, moving from A to B increased the violation level of constraint $C1$ by $d = 1$, while moving from C to B increased it by $d = 2$. This difference in the differences d is exploited only by the probability $e^{-d/t}$ used when temperature is exactly in the domain of constraint $C1$. Thus, without this most complicated last case in the

Rules of moving (page 63) SA-OT would not display the behaviour described in subsection 2.3.1 (increasing precision with an increasing number of iterations). It is indeed worth bothering about this complicated definition.

There is a further motivation for asking about the role played by parameters T_{max} and T_{min} , and this is the inherent arbitrariness in the definition of the constraints as real valued functions. Indeed, a monotonic transformation of a constraint, such as $C'_i(w) := \alpha + C_i(w)$ or $C'_i(w) := \gamma \cdot C_i(w)$ ($\gamma > 0$), does not influence the underlying Optimality Theoretical grammar. The first transformation does not influence the output frequencies predicted by SA-OT either, for SA-OT requires only $d = C_i(w) - C_i(w')$, but the second one may influence SA-OT. As multiplying each constraint by γ is equivalent to dividing the algorithm's parameter t (hence, T_{max} , T_{min} and T_{step}) by γ , the experiments to be presented now may also demonstrate the role of the definition of the constraints in SA-OT.

Let us turn back to the experiments in which we varied T_{max} and T_{min} . The question is whether changing the half-closed interval $[T_{max}, T_{min})$ covered by t in the inner loop affects the outcome. Such an effect would demonstrate that t —hence, the last of the *Rules of moving*—does influence the algorithm. When varying T_{max} and T_{min} , however, we shall also adjust T_{step} , so that $R = \frac{T_{max} - T_{min}}{T_{step}}$, thus the number of iterations, be constant. The role of the number of iterations has already been demonstrated in the previous experiment, so we have to control for this factor.

Replacing the interval $[T_{max}, T_{min})$ by $[T_{max} + \tau, T_{min} + \tau)$ corresponds to increasing the value of t by τ in each iteration, which in turn means that $e^{-d/t}$ becomes $e^{-d/(t+\tau)}$, a probability closer to 1. Similarly, if the original interval $[T_{max}, 0)$ (that is, specifically $T_{min} = 0$), is replaced by $[\nu \cdot T_{max}, 0)$ (and T_{step} simultaneously becomes $\nu \cdot T_{step}$), then the transition probability grows to $[e^{-d/t}]^{1/\nu}$ at every time step. This is why varying $[T_{max}$ and $T_{min})$ can measure the importance of t 's exact value during the algorithm.

Based on a preliminary experiment, Table 5.4 compares the outputs produced under three different parameter-settings. Experiments A refer to the parameters used so far: in each domain, the temperature drops from $T_{max} = 3$ to $T_{min} = 0$, using some T_{step} . In experiments B, temperature drops from $T_{max} = 4$ to $T_{min} = 1$: for each T_{step} , we perform exactly the same experiment as in the corresponding experiment A, with the same number of steps, but the probabilities of moving upwards are increased a little bit. In experiments C, temperature drops in each domain from $T_{max} = 5$ to $T_{min} = 0$. This involves increasing the number of steps, so one must compare for instance the results of $T_{step} = 0.1$ in experiment settings C to the experiments A and B with $T_{step} = 0.06$, because these are the conditions under which temperature drops in 50 steps in each domain.

Let us have a closer look at this table. We can see that changing the conditions does not influence the phenomena drastically, the only major difference being that [su]u[s] begins appearing at different T_{step} values. The maximum of u[us][s] is around 12.9% (experiment A), 12.5% (experiment B) and 13.1% (experiment C). Further experiments show that the differences are not significant: running the simulation 43000 times (1000 times for each candidate as starting point; $T_{step} = 1.5$), we obtained 12.68% under conditions A, 12.87% under conditions B, and 12.96% under conditions C.

T_{step}	A	B	C	A	B	C	A	B	C
2	68	68	71	19	20	16	13	13	13
1.5	68	69	73	20	18	14	13	13	13
1.2	71	71	74	16	16	12	13	12	13
1	72	71	75	16	16	13	13	13	12
0.7	75	76	80	12	11	8	13	12	12
0.65	76	76	80	12	12	7	12	12	12
0.5	78	78	83	10	10	6	12	12	11
0.35	83	82	85	6	6	3	11	12	11
0.25	85	84	88	4	5	1.7	11	11	10
0.2	87	86	90	2.7	3.2	0.8	10	11	9
0.15	89	88	91	1.4	1.8	0.3	9	10	9
0.12	91	90	92	0.6	0.8	0.1	8	9	8
0.1	92	90	92	0.4	0.6	0.02	8	9	8
0.07	94	91	94	0.1	0.2	0	6	9	6
0.05	95	92	94	0.02	0	0	5	8	6
0.02	97	94	97	0	0	0	3	6	3
0.015	98	94	98	0	0	0	2	6	2
0.01	98	94	98	0	0	0	1.5	6	1.7
0.007	99.0	94	98.9	0	0	0	1.0	5	1.1
0.005	99.1	95	99.2	0	0	0	0.9	5	0.8
0.002	99.7	95	99.7	0	0	0	0.3	5	0.3

Table 5.4: **Varying T_{max} and T_{min} :** The percentages of producing $[s]u[su]$ (columns 2-4), $[su]u[s]$ (columns 5-7) and $u[us][s]$ (columns 8-10), as a function of T_{step} , out of 12900 runs (starting the simulation 300 times with each of the 43 candidates), under various circumstances (see text).

%	A	B	C
[s]u[su]	68.49 \pm 0.25	68.11 \pm 0.25	73.01 \pm 0.25
[su]u[s]	18.74 \pm 0.20	19.09 \pm 0.20	14.18 \pm 0.20
u[us][s]	12.76 \pm 0.15	12.80 \pm 0.15	12.81 \pm 0.15

Table 5.5: Varying T_{max} and T_{min} , with $T_{step} = 1.5$.

The ratio of [su]u[s] reaches a higher value under conditions A and B. Yet, the fact that for $T_{step} = 2$ we find only 16% in experiment C, whereas 19 – 20% under conditions A and B should not surprise us. This value for experiment C should be compared to the very similar values in experiments A and B with $T_{step} = 1.2$, for this refers to the case when three temperature values are used for each temperature-domain.¹⁶

When comparing the results under conditions A to those under conditions B, we do not see so much difference. Can we nonetheless find some significant differences? Summing up several experiments, altogether 129000 runs with $T_{step} = 1.5$, we obtained the results on Table 5.5, exhibiting weakly significant differences.¹⁷

5.5.3 The role of T_{max} and T_{min} (2)

In a next experiment, I ran a simulation producing five times 43000 outputs—43000 outputs corresponds to starting the simulation 1000 times from each of the 43 candidates—under four different conditions (see Table 5.5). Repeating the whole experiment five times helped in estimating the standard error of the experiment: I calculated not only the mean but also the dispersion of the five values obtained with the same parameter setting.

The first condition (3-0) corresponds to the original one: in each domain, the component t of temperature is decreased from $T_{max} = 3$ until $t > T_{min} = 0$. In the second condition (4-1), the same number of steps is achieved by decreasing the temperature from $T_{max} = 4$ until $T_{min} = 1$. In the third condition (6-0), the temperature decreases from $T_{max} = 6$ to $T_{min} = 0$, but the T_{step} values are doubled, in order to have the same number of steps. Similarly, under condition 1.5-0, temperature decreases from $T_{max} = 1.5$ to $T_{min} = 0$, by using the half of the original T_{step} values. Again, the reason of changing T_{step} in conditions 6-0 and 1.5-0 was to measure exclusively the influence of choosing the limits of the

¹⁶In the experiment mentioned earlier (running the simulation 43000 times, i.e. 1000 times with each candidate as starting point; $T_{step} = 1.5$), the proportions of [su]u[s] were: 18.82% for A, 19.23% for B, and 14.28% for C.

¹⁷I ran three times 43000 simulations (43000 simulations correspond to running 1000 times for each of the 43 candidates as starting point of the simulation), for both conditions A, B and C ($T_{step} = 1.5$).

The outputs under condition A: 68.50%, 68.68% and 68.30% for [s]u[su]; 18.82%, 18.57% and 18.84% for [su]u[s]; 12.68%, 12.74% and 12.85% for u[us][s]. The outputs under condition B: 67.90%, 68.15% and 68.28% for [s]u[su]; 19.23%, 18.97% and 19.06% for [su]u[s]; 12.87%, 12.87% and 12.66% for u[us][s]. Under condition C: 72.76%, 73.07% and 73.21% for [s]u[su]; 14.28%, 14.18% and 14.07% for [su]u[s]; and, finally, 12.96%, 12.75% and 12.72% for u[us][s].

Concerning the error of such an experiment, we can approximate the results with a binomial distribution. Thus, for instance, running the simulation A $N = 43000$ times, we obtained the output [s]u[su] with a relative frequency $p = 68.49\%$ and with a relative error of $\sigma = \sqrt{\frac{p(1-p)}{N}} \approx 0.25\%$.

T_{step}	output	3-0	4-1	6-0	1.5-0
1.5	[s]u[su]	68.46 ± 0.17	67.99 ± 0.23	67.70 ± 0.26	68.33 ± 0.26
	[su]u[s]	18.82 ± 0.04	19.16 ± 0.08	19.45 ± 0.21	18.76 ± 0.20
	u[us][s]	12.72 ± 0.15	12.85 ± 0.19	12.85 ± 0.11	12.90 ± 0.13
0.5	[s]u[su]	78.11 ± 0.25	77.60 ± 0.14	76.92 ± 0.19	77.48 ± 0.22
	[su]u[s]	9.79 ± 0.12	10.22 ± 0.12	10.49 ± 0.14	10.00 ± 0.12
	u[us][s]	12.11 ± 0.21	12.18 ± 0.11	12.59 ± 0.15	12.52 ± 0.17
0.06	[s]u[su]	94.01 ± 0.13	91.99 ± 0.11	91.94 ± 0.18	95.70 ± 0.10
	[su]u[s]	0.049 ± 0.008	0.073 ± 0.015	0.046 ± 0.014	0.034 ± 0.007
	u[us][s]	5.94 ± 0.14	7.94 ± 0.11	8.01 ± 0.19	4.27 ± 0.10
0.005	[s]u[su]	99.21 ± 0.06	94.53 ± 0.05	98.38 ± 0.07	99.64 ± 0.03
	[su]u[s]	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
	u[us][s]	0.79 ± 0.06	5.47 ± 0.05	1.62 ± 0.07	0.36 ± 0.03

Figure 5.5: **Four different T_{max} - T_{min} pairs:** Percentage of outputs (mean \pm standard deviation), obtained from running five sets of simulations, producing 43000 outputs each (43000 outputs corresponds to starting the simulation 1000 times from each of the 43 candidates). Note that the T_{step} values given in the left column were applied in the 3-0 and 4-1 cases, whereas the doubled values were applied in the 6-0 case, and the halved values in the 1.5-0 case, in order to keep the same number of steps.

inner loop of the algorithm, and to discard the effect of changing the number of iterations.

Both the 4-1 and the 6-0 cases differ from the 3-0 case in that the transition probabilities of moving uphill are slightly increased. Imagine, for instance, that the violation profiles of two neighbours differ fatally in that one of the candidates has one more violation of constraint C_k . Temperature T has just entered the domain of constraint C_k ($T = \langle k, T_{max} \rangle$). Then, the probability of moving from the better candidate to the worse one is $e^{-1/3} \approx 0.717$ in 3-0, and $e^{-1/4} \approx 0.779$ in 4-1—not a major difference. Before the temperature leaves the domain of that constraint, however, the same probability is almost 0 in 3-0, but more than $e^{-1} \approx 0.368$ in 4-1. Similarly, the transition probabilities in the 6-0 case are the square root of the corresponding probabilities in the 3-0 case at the same point of the simulation, since the value of t is the double of the value of t in the corresponding situation under 3-0. However, the same probabilities are diminished, reduced to their squares, under condition 1.5-0.

Based on the small but significant differences appearing on Table 5.5, we could already argue earlier that the system does make use of the counter-optimal moves that are slightly more probable under conditions 4-1 (called B in the previous experiment) than under conditions 3-0 (or A).

What are the significant differences introduced by changing the parameters? Decreasing the probabilities of moving uphill, as it happens when we change condition 3-0 to condition 1.5-0, has ambiguous effects. With the highest T_{step} values, no significant difference can be detected, with the middle high T_{step} values, more correct outputs ([s]u[su]) are produced under 3-0, whereas the opposite is true for the case with the two lowest T_{step} values. These significant differences confirm that whenever “real” simulated annealing is performed—namely, moving upwards in the search landscape is an option reasonably often

because temperature is decreased slowly enough—then moving upwards in the special case of $k = K$ indeed influences the output of the algorithm. Sometimes this influence is positive, and sometimes negative, but there is such an influence.

Increasing the probabilities (4-1 and 6-0, compared to 3-0), however, will reduce the probability of returning the globally optimal form [s]u[su]. Furthermore, the size of this difference increases as the number of steps performed within one domain grows, that is, as we decrease T_{step} .

Interestingly enough, the size of the difference in producing the fast speech form [su]u[s] decreases slightly with lower T_{step} values. However, it is the u[us][s] production that grows from a non significant difference ($T_{step} = 1.5$) into a highly significant difference ($T_{step} = 0.005$). Consequently, the size of the difference in the [s]u[su] production between the 3-0 condition on the one hand, and the increased probability conditions (4-1, 6-0) on the other hand has to be explained by different factors for different T_{step} values: for higher T_{step} values it is a consequence of the different [su]u[s] production, whereas for lower T_{step} values it originates from the different u[us][s] probabilities.

To sum up, we can say that the very complicated topology of the search space used is the key to understanding the behaviour of the simulation under different parameters. The search space can be vaguely seen as being composed of three major valleys. At the bottom of them we find [s]u[su], [su]u[s] and u[us][s], respectively, the first being the global optimum and the latter two being only local optima. The valley with [s]u[su] at its bottom is by far the widest, covering roughly 70% of the space, whereas the two other valleys cover approximately 15% each. At least, this is the picture obtained if we simply descend the slope when simulating fast speech. However, in the simulation of slow speech, we allow much more moving upwards, and therefore, we have more chance to find the global optimum. The fine structure of the search space would explain why it is much easier to get from the valley of [su]u[s] to the valley of [s]u[su] than from the valley of u[us][s]: either it depends on the distance of the valleys, or on the height of the intervening hills, or on the chances to enter each of the valleys from the region between the two valleys.

In the following section, we aim at improving our model and getting rid of the unattested form u[us][s] by introducing further constraints.

5.6 Improving the Schreuder-Gilbers model¹⁸

As we have seen, the candidate u[us][s] appears as the output of the simulation, even with very low T_{step} values, that is, when the number of iterations is high. Yet, this candidate does not correspond to anything observed in real speech. The goal of this section is to improve the underlying linguistic model in order to be able to account better for the observations.

In fact, having two stressed syllables adjacent to each other is surprising. This observation motivates introducing another constraint to the hierarchy: *CLASH, proposed originally by Kager (1994), assigns a violation mark to each stressed syllable that immediately follows another stressed syllable:

¹⁸The results of the present section have been presented as B    (2004) and published as B    (2005a). In this section, $T_{max} = 3$, $T_{min} = 0$, $K_{step} = 1$, while K_{max} is always one domain higher than the highest ranked constraint.

T_{step}	[s]u[su]	[su]u[s]
3	0.5883, 0.5824, 0.5855	0.4117, 0.4176, 0.4145
1	0.6266, 0.6268, 0.6405	0.3734, 0.3732, 0.3595
0.3	0.8182, 0.8172, 0.8142	0.1818, 0.1828, 0.1858
0.1	0.9771, 0.9775, 0.9769	0.0229, 0.0225, 0.0231
0.03	1.0000, 1.0000, 1.0000	0.0000, 0.0000, 0.0000
0.01	1.0000, 1.0000, 1.0000	0.0000, 0.0000, 0.0000

Table 5.6: ***Clash included:** The ratio of different outputs for a four-syllable word with the ranking $*CLASH \gg OOC_{\sigma=susu, z=2} \gg *ΣΣ \gg PARSE-σ$. As candidate u[us][s] is not a local optimum anymore, it is never returned. One simulation consists of running the algorithm 25800 times: that is, choosing 600 times each of the 43 candidates as initial candidate. The simulation was run three times with the same T_{step} value in order to assess the error of the algorithm.

Definition 5.6.1. *The number of violation marks assigned by constraint $*CLASH$ to candidate w is the number of substrings “s/[s]” within the candidate w .*

Now, the harmony function is defined by the hierarchy

$$*CLASH \gg OOC_{z=2} \gg *ΣΣ \gg PARSE-σ.$$

Under such circumstances, the candidate u[us][s] is not a local optimum anymore. Indeed, only two optima exist, those produced in real speech: the global optimum [s]u[su] and the fast speech form [su]u[s]. Experiments show that this model nicely fits empirical observations. The two local optima are returned with $T_{step} > 0.1$ (the parameter settings corresponding to fast speech), and the global optimum shows up alone for $T_{step} < 0.1$, similarly to slow speech (Table 5.6 and Fig. 5.6).

Ranking constraint $*CLASH$ high leads, however, to problems in the case of words such as *zúid.à.fri.kàans* (‘South African’) or *ùit.gè.ve.ríj* (‘publisher’) (Schreuder and Gilbers, 2004), which do include adjacent stresses in their grammatical form. If, on the other hand, constraint $*CLASH$ is ranked lower than OOC , then candidate u[us][s] remains a local optimum, as shown by tableau (5.7). Therefore, we replace $*CLASH$ with another constraint, $ALIGN(word, foot, left)$ (McCarthy and Prince, 1993a,b; McCarthy, 2002). This widely used constraint assigns one violation mark to each candidate whose left edge does not align with the left edge of some foot, and reflects the fact that the first syllable of most Dutch words bears some stress.¹⁹

Definition 5.6.2. $ALIGN(word, foot, left)(w) = 1$ if w begins with an unparsed syllable (u), and $ALIGN(word, foot, left)(w) = 0$ if w begins with the left bracket of some foot (l).

The results of the experiments with constraint $ALIGN(word, foot, left)$ appear in Table 5.7 and on Figure 5.7. The picture obtained by using the constraint $ALIGN(word, foot, left)$ is similar to the one that resulted from $*CLASH$, although

¹⁹McCarthy (2002)’s criticism against gradient constraints, corroborated by Bíró (2003), does not apply here: although $ALIGN(word, foot, left)$ is an alignment constraint, it is not gradient in any sense.

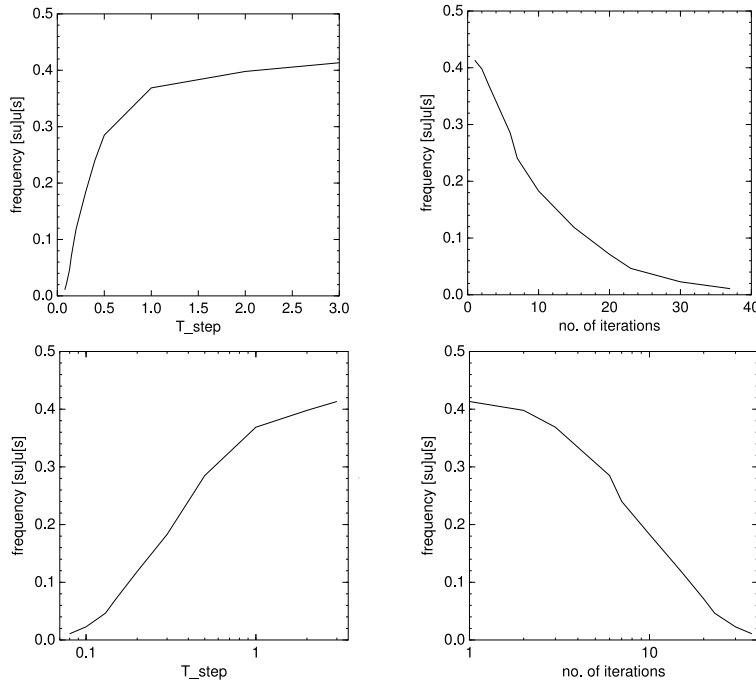


Figure 5.6: ***Clash included:** See caption of Table 5.6. The probabilities of the fast speech form [su]u[s] being returned are plotted as functions of T_{step} (left boxes) and of the number of iterations performed within one temperature domain ($3/T_{step}$, right boxes). The probabilities of obtaining the grammatical form [s]u[su] is 1 minus the plotted probabilities. The x axis is linear on the upper graphs, and logarithmic on the lower ones.

T_{step}	[s]u[su]	[su]u[s]
3	0.8036, 0.8048, 0.8005	0.1964, 0.1952, 0.1995
1	0.8520, 0.8519, 0.8522	0.1480, 0.1481, 0.1478
0.5	0.9031	0.0969
0.3	0.9518, 0.9548, 0.9540	0.0482, 0.0452, 0.0460
0.2	0.9706	0.0294
0.15	0.9848	0.0152
0.12	0.9912	0.0088
0.1	0.9953, 0.9956, 0.9959	0.0047, 0.0044, 0.0041
0.09	0.9965	0.0035
0.08	0.9980	0.0020
0.05	0.9997	0.0003
0.03	1.0000	0.0000

Table 5.7: **Align(word, foot, left)**: The ratio of obtaining different outputs for a four-syllable word with respect to the ranking $\text{ALIGN}(\text{word}, \text{foot}, \text{left}) \gg \text{OOC}_{\sigma=\text{susu}, z=2} \gg * \Sigma \Sigma \gg \text{PARSE-}\sigma$. One simulation consists of running the algorithm 25800 times. The simulation was run sometimes three times with the same T_{step} value in order to assess the error of the algorithm.

the ratio of the fast-speech form [su]u[s] is lower in the former case. Nevertheless, constraint $\text{ALIGN}(\text{word}, \text{foot}, \text{left})$ allows us to apply our model successfully to other inputs.

Indeed, so far, we have succeeded in accounting for the *andante* and *allegro* forms of Type 0 words, such as *fótotòdestel* ‘camera’ (output faithful by OOC to string susu). Can the new ranking $\text{ALIGN}(\text{word}, \text{foot}, \text{left}) \gg \text{OOC} \gg * \Sigma \Sigma \gg \text{PARSE-}\sigma$ also account for the fast speech phenomena corresponding to Types 1-3 on page 122)?

Type 1 included Dutch words such as *stu.die.toe.la.ge* (‘study grant’) or *weg.werp.aan.ste.ker* (‘disposable lighter’) that are five-syllable compounds, a two-syllable word followed by a three-syllable one. In these cases the stress pattern enforced by Output-Output Correspondence in slow, careful speech is susuu, the concatenation of su and of suu. However, a *right-shift* takes place in allegro speech, resulting in the suusu pattern. Simulated annealing with the previously used parameters imitates human-like behaviour: the percentage of producing [s]u[su]u grows from 49% to 96%, as T_{step} drops from 3 to 0.03.

The remainder of the outputs are, however, evenly distributed between the candidates [su]u[su], the empirically observed fast-speech form, and [su]u[us], an unattested form. The reason is clear: the latter two forms are assigned exactly the same violation profile and are neighbours in the search space. Consequently, they are situated at the bottom of an elongated valley. Whatever the probability of the system getting stuck in this valley, the two forms will be returned with equal probability. Yet, by adding a slight slope to this valley, we can favour one of the two forms. So we introduce an additional constraint, such as $\text{FOOTTYPE}(\text{trochaic})$, which will prefer the form [su]u[su] over its unattested neighbour, [su]u[us], and will drive the system towards the attested fast speech form. The constraint $\text{FOOTTYPE}(\text{trochaic})$, assigning one violation mark to each binary iambic foot [us], is a widely used constraint (Tesar and Smolensky, 2000), and its use is consonant with traditional analyses of Dutch arguing for

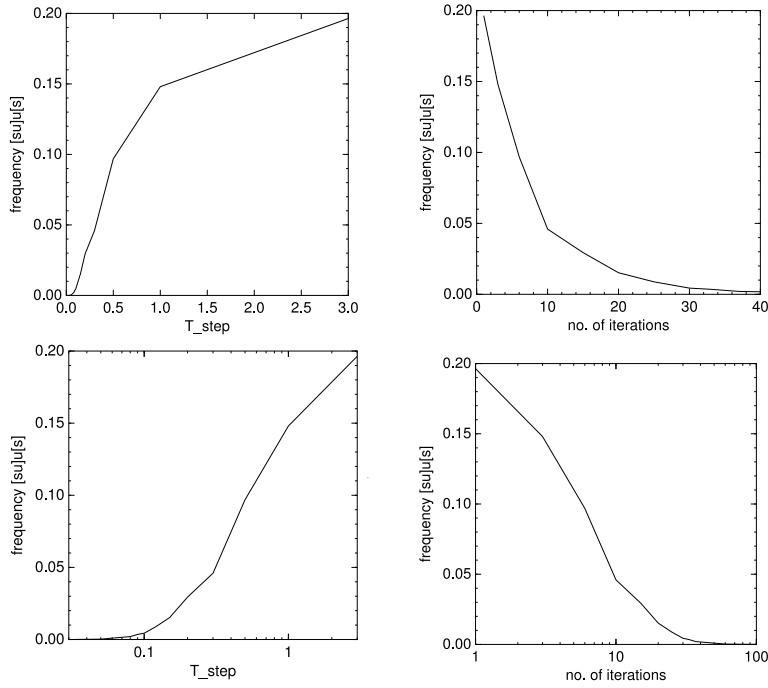


Figure 5.7: **Align(word, foot, left)**: See caption of Table 5.7. The probabilities of the fast speech form $[su]u[s]$ being returned are plotted as functions of T_{step} (left figures) and as functions of the number of iterations performed within one temperature domain ($3/T_{step}$, right figures). The probabilities of obtaining the grammatical form $[s]u[su]$ is 1 minus the plotted probabilities. The x axis is linear on the upper graphs, and logarithmic on the lower ones.

T_{step}	$[s]u[su]u$ $z = 2$	$[su]u[su]$ $z = 2$	$[s]u[su]u$ $z = 1$	$[su]u[su]$ $z = 1$
3	0.465	0.535	0.548	0.452
1	0.519	0.481	0.623	0.377
0.3	0.646	0.354	0.806	0.194
0.1	0.800	0.200	0.964	0.036
0.03	0.960	0.040	0.9998	0.0002

Table 5.8: Frequency of the outputs in a simulation for Type 1 words (such as *studietoelage*) with the hierarchy **ALIGN(word, foot, left)** \gg **OOC** $_{\sigma=susu}$ \gg *** $\Sigma\Sigma$** \gg **PARSE- σ** \gg **FOOTTYPE(trochaic)**. The observed andante form is *susuu*, whereas the allegro (fast speech) form is *suusu*. A first experiment (first two columns) used OOC with $z = 2$, whereas the last two columns report on a second experiment where $z = 1$. Both experiments included 71400 runs (600 times starting from each of the 119 candidates).

T_{step}	[us]u[s] $z = 2$	[s]u[us] $z = 2$	[s][su]u $z = 2$	[us]u[s] $z = 1$	[s]u[us] $z = 1$	[s][su]u $z = 1$
3	0.420	0.349	0.231	0.496	0.382	0.122
1	0.476	0.281	0.243	0.556	0.314	0.131
0.3	0.639	0.101	0.260	0.715	0.189	0.095
0.1	0.762	0.010	0.228	0.902	0.074	0.024
0.03	0.839	0.000	0.161	0.998	0.0017	0.0004

Table 5.9: Frequency of the outputs in a simulation for Type 2 words (such as *perfectionist*) with the hierarchy $\text{ALIGN}(\text{word}, \text{foot}, \text{left}) \gg \text{OOC}_{\sigma=\text{susu}} \gg * \Sigma \Sigma \gg \text{PARSE-}\sigma$. The observed andante form is *usus*, and the allegro form is *suus*. In the first experiment (first three columns) $z = 2$ for OOC , while $z = 1$ in the second experiment (last three columns). Both experiments included 25800 runs (600 times starting from each of the 43 candidates).

trochaic feet (e.g. Hayes, 1995, p. 305-306). Moreover, the fact that we have demoted this constraint to the bottom of the hierarchy, ensures that our previous results are not influenced by the introduction of this new constraint.²⁰

Definition 5.6.3. *The number of violation marks assigned by constraint $\text{FOOT-TYPE}(\text{trochaic})$ to candidate w is the number of binary iambic feet (the number of substrings “[us]”) within the candidate w .*

Table 5.8 presents the results of such an experiment. Observe that in “extreme fast speech” ($T_{step} = 3$), the percentage of the grammatical output *susu* is lower than 47%, significantly below 50%. Section 5.2 discussed the difficulties Boersma (1997)’s *Stochastic Optimality Theory* would face if required to produce the grammatical form in less than half of the cases using the reranking of constraints OOC and $* \Sigma \Sigma$. However an adequate model must be able to do so, since (Schreuder, 2006) observed a frequency as low as 38% for words in the type *studietoelage* (Table 5.1 on page 123).

Type 2 words (*per.fec.tio.nist*, ‘perfectionist’ or *a.me.ri.kaan* ‘American’) include a suffix that must bear a stress. Their careful pronunciation—determined by OO- Correspondence to the string *usus*—follows the stress pattern *usus*, whereas a *left-shift* occurs in fast speech yielding *suus*.

If we apply our model to this case (four-syllable input, $\text{OOC}_{\sigma=\text{usus}}$, the results are reported in Table 5.9), the grammatical form is again returned in the majority of the cases, and its proportion converges to 1 with decreasing T_{step}

²⁰The earlier landscape, the one corresponding to input *fotoetoestel* (four-syllable word, OOC calculated with reference string *susu*) and to hierarchy $\text{OOC}_{z=2} \gg * \Sigma \Sigma \gg \text{PARSE-}\sigma$, did not include any horizontal move, because each of the basic operations involved a change in the violation level of at least one of the constraints. Namely, inserting or deleting one monosyllabic foot certainly influenced the word’s behaviour with respect to $\text{PARSE-}\sigma$, as did moving one foot bracket; whereas changing the head syllable within a bisyllabic foot increased or decreased the violation level of $\text{OOC}_{\sigma=\text{susu}}$ by two, since the reference string contained a stressed and an unstressed syllable alternately. Consequently, the fatal constraint distinguishing between the violation profiles of neighbours was either $\text{PARSE-}\sigma$, or a higher ranked constraint, and this is why introducing new constraints would not influence the model if they were ranked lower than constraint $\text{PARSE-}\sigma$. If, however, the reference string of OOC includes a substring *ss* or *uu*, it is possible that altering the head syllable within a bisyllabic foot does not change the violation level of any of the previously mentioned constraints, and this is why a low ranked constraint $\text{FOOT-TYPE}(\text{trochaic})$ can become important.

values. Nonetheless, simulated annealing partially fails to predict the correct performance errors. Beside the empirically observed fast speech form [s]u[us], an additional local optimum emerges, namely, [s][su]u.

Interestingly enough, the candidate [s][su]u is even harmonically bounded ^{*}(see page 24) by [su]u[s]: both satisfy ALIGN(word, foot, left), both incur two marks by OO-CORRESPONDENCE _{$\sigma=usus$} (for any z), and one mark by PARSE- σ , while ^{*} $\Sigma\Sigma$ is satisfied only by [su]u[s]. And yet, these two candidates are not neighbours, so both can become local optima. Moreover, with $z = 2$, [s][su]u is a local optimum that is more stable with respect to decreasing T_{step} : even though [su]u[s] is returned slightly more frequently with $T_{step} = 3$ than [s][su]u, its frequency drops quickly as T_{step} is diminished, unlike the almost steady frequency of [s][su]u. The moral of this observation is that different local optima may have different stability with respect to variations of the parameters; and what is more, a local optimum may turn out to be relatively frequent despite the existence of a better local optimum that is easily avoided by careful annealing.²¹

This phenomenon also resists our attempts to introduce changes in the definition of Output-Output Correspondence, or to add new, low-ranked constraints, such as ALIGN(Word, Foot, right). For instance, if $z = 0$ in OOC, not only does [s][su]u disappear, but also the fast speech form [s]u[us], while a new (empirically unattested) local optimum, [s][su]u[s], shows up besides the global optimum [us]u[s]. For $z = 3$, the form [s][su]u is extremely persistent, its frequency even grows as T_{step} diminishes from 3 to approximately 0.4, and starts decreasing only not much before the likelihood of [s]u[us] drops drastically to but a few percent ($T_{step} < 0.3$). Table 5.10 presents the results of the experiment focusing on this surprisingly new type of behaviour of the system. This is the first time we observe that decreasing T_{step} does not automatically cause the probability of a local optimum to diminish.

The results obtained by using $z = 1$ (the right hand side of Table 5.9) will turn out to be one of the best experiments, when we compare the results of the same ranking using the other types of words. The ratio of [s][su]u is kept as low as 12–13% when $T_{step} = 3$, with a probability three times higher for [s]u[us]. In addition, none of the two local optima is more stable: the probability of [s]u[us] is steadily twice or three times higher than that of [s][su]u, while both converge to zero with diminishing T_{step} . A source of this relative success might be that for $z = 1$, [s][su]u and its neighbour, [s][su][s] violate equally OOC _{$\sigma=usus$} , only the lower ranked ^{*} $\Sigma\Sigma$ distinguish between them, and therefore, the random walker can more easily escape to [us]u[s].

A further idea has been to introduce constraint ALIGN(Word, Foot, right), which punishes [s][su]u, while favouring [us]u[s] and [s]u[us]:

Definition 5.6.4. *ALIGN(Word, Foot, right)(w) = 1 if w ends with an unparsed syllable (u), and ALIGN(Word, Foot, right)(w) = 0 if w ends with the right bracket of some foot ($()$).*

Nonetheless, candidate [s][su]u is still a local optimum besides [us]u[s] and [s]u[us] ($z = 2$, $z = 1$), unless ALIGN(Word, Foot, right) is ranked higher than

²¹Recall that Coetzee (2004) predicted more harmonic forms to be more frequent. We have already mentioned that SA-OT, contrary to Coetzee (2004)’s proposal, allows a candidate not to surface in the language, and yet to be better than an emerging form, if the first candidate is not a local optimum. Now, we can see an example for a worse candidate being more frequent even if the better, though less frequent candidate is also a local optimum.

T_{step}	[us]u[s]	[s]u[us]	[s][su]u
3	0.430	0.339	0.231
2	0.449	0.309	0.242
1	0.489	0.266	0.246
0.80	0.509	0.233	0.258
0.60	0.559	0.178	0.263
0.60	0.555	0.178	0.267
0.60	0.549	0.182	0.268
0.55	0.558	0.176	0.266
0.50	0.557	0.177	0.266
0.50	0.559	0.171	0.269
0.50	0.561	0.175	0.264
0.48	0.579	0.156	0.265
0.45	0.579	0.149	0.272
0.45	0.582	0.151	0.267
0.45	0.586	0.148	0.266
0.42	0.599	0.134	0.268
0.42	0.605	0.131	0.265
0.42	0.598	0.132	0.269
0.40	0.599	0.132	0.269
0.40	0.599	0.133	0.268
0.40	0.603	0.129	0.268

T_{step}	[us]u[s]	[s]u[us]	[s][su]u
0.38	0.607	0.130	0.263
0.38	0.601	0.126	0.272
0.38	0.600	0.131	0.269
0.35	0.618	0.111	0.270
0.35	0.619	0.113	0.268
0.35	0.618	0.116	0.267
0.32	0.636	0.099	0.266
0.30	0.644	0.084	0.271
0.30	0.652	0.086	0.262
0.30	0.649	0.086	0.264
0.25	0.665	0.075	0.261
0.25	0.661	0.075	0.264
0.25	0.665	0.075	0.260
0.20	0.693	0.048	0.259
0.18	0.702	0.039	0.258
0.15	0.723	0.026	0.251
0.12	0.755	0.013	0.232
0.10	0.769	0.009	0.222
0.08	0.794	0.0027	0.203
0.05	0.843	0.0003	0.156
0.03	0.889	0.000	0.111

Table 5.10: **Outputs when using $\text{OOC}_{\sigma=usus, z=3}$** : The hierarchy used was $\text{ALIGN}(\text{word}, \text{foot}, \text{left}) \gg \text{OOC}_{\sigma=usus, z=3} \gg * \Sigma \Sigma \gg \text{PARSE-}\sigma \gg \text{FOOT-TYPE}(\text{trochaic})$. Each experiment consisted of 25800 runs, that is, 600 runs launched from each of the 43 candidates. The experiment has been repeated for some T_{step} values in order to be able to assess the error. Many more runs would have been needed to determine exactly where the maximum of [s][su]u is located, but the figures clearly demonstrate that there is a maximum (or more maxima) around $T_{step} = 0.4$. This experiment demonstrates that decreasing T_{step} does not diminish the rate of some erroneous form automatically.

T_{step}	[us]u[s]	[s]u[us]
3	0.586	0.414
1	0.634	0.366
0.3	0.828	0.172
0.1	0.980	0.020
0.03	0.9999	0.0001

Table 5.11: **Introducing constraint Align(Word, Foot, right):** Frequencies for Type 2 words, using the hierarchy ALIGN(word, foot, left) \gg ALIGN(Word, Foot, right) \gg OOC $_{\sigma=usus, z=2}$ \gg * $\Sigma\Sigma$ \gg PARSE- σ \gg FOOT-TYPE(trochaic)

T_{step}	3	1	0.3	0.1	0.03
Type 0: susu					
[s]u[su]	0.818	0.906	0.992	1.0000	1.0000
[su]u[s]	0.182	0.094	0.008	0.0000	0.0000
Type 1: susu					
[s]u[su]u	0.549	0.628	0.824	0.970	0.9999
[su]u[su]	0.451	0.372	0.176	0.030	0.0001
Type 2: usus					
[us]u[s]	0.615	0.666	0.827	0.974	1.0000
[s]u[us]	0.385	0.334	0.173	0.026	0.0000
Type 3: ssus					
[s][s]u[s]	0.655	0.708	0.896	0.992	1.0000
[su]u[s]	0.345	0.292	0.104	0.008	0.0000

Table 5.12: Results using hierarchy (5.9). Each simulation was obtained by running the simulation 600 times from each candidate.

OOC. Seemingly, the hierarchy ALIGN(word, foot, left) \gg ALIGN(Word, Foot, right) \gg OOC $_{\sigma=usus, z=2}$ \gg * $\Sigma\Sigma$ \gg PARSE- σ \gg FOOT-TYPE(trochaic) produced only the required outputs (Table 5.11). The problem with this result is, however, that in the case of Type 1 words (*studietoelage*), this hierarchy (with OOC $_{\sigma=susu, z=2}$) does not have candidate [s]u[su]u (or any other parse of the andante form susu) optimal any more. Due to the highly ranked constraint ALIGN(Word, Foot, right), candidates parsing the ultimate syllable will become better. Hence, we have to reject this constraint ranking, even though it would solve our problem with respect to Type 2 words.

The only good combination of the constraints to be found was the following hierarchy:

$$\begin{aligned} \text{ALIGN(word, foot, left)} &\gg \text{OOC}_{z=1} \gg \text{ALIGN(Word, Foot, right)} \gg \\ &\gg * \Sigma \Sigma \gg \text{PARSE} - \sigma \gg \text{FOOT-TYPE(trochaic)} \end{aligned} \quad (5.9)$$

This hierarchy not only eliminates the unwanted local optimum in the case of Type 2 inputs, but also works for all the other types (Table 5.12). Its only drawback will be that the empirically observed frequencies cannot be reproduced adequately.

Finally, **Type 3** words incurred a *beat reduction* (deletion of a stress) in

T_{step}	[s][s]u[s]	[su]u[s]
3	0.557	0.443
1	0.680	0.320
0.3	0.835	0.165
0.1	0.970	0.030
0.03	1.000	0.000

Table 5.13: **Type 3 words**, $z = 0$: the hierarchy used was ALIGN(word, foot, left) \gg OOC $_{\sigma=ssus,z=0}$ \gg * $\Sigma\Sigma$ \gg PARSE- σ \gg FOOTTYPE(trochaic). 600 runs from each of the 43 candidates summed up to 25,800 runs for each T_{step} value.

T_{step}	[s][s]u[s]	[su]u[s]
3	0.652	0.348
1	0.682	0.318
0.3	0.836	0.164
0.1	0.972	0.028
0.03	0.99992	0.00007

Table 5.14: **Type 3 words**, $z = 1$: the hierarchy used was ALIGN(word, foot, left) \gg OOC $_{\sigma=ssus,z=1}$ \gg * $\Sigma\Sigma$ \gg PARSE- σ \gg FOOTTYPE(trochaic). 600 runs from each of the 43 candidates summed up to 25,800 runs for each T_{step} value.

fast speech, replacing the andante form *ssus* with the allegro form *suus*. The andante forms submit themselves once again to the requirements imposed by OO-Correspondence. This was the case in words such as *zuid.a.fri.kaans* (‘South African’) or *uit.ge.ve.rij* (‘publisher’).

For $z = 2$ in the definition of OOC, simulated annealing predicts (especially with higher $T_{step} = 3$ values) the emergence of an incorrect fast speech form, *viz.* [s]u[s][s], instead of [s]u[us] or [su]u[s]. In fact, [s]u[us] is not a local optimum, for its neighbour [s][s][us] is more harmonic with respect to the hierarchy. Similarly, [su][s][s], a neighbour of [su]u[s] incurs fewer violation marks by constraint OOC $_{\sigma=ssus,z=2}$: although one more violation mark originates from the mismatch in the third syllable, yet the difference in the number of stressed syllables adds two more violation marks to [s]u[us]. Candidate [s]uu[s] is even less harmonic than its neighbours [su]u[s] or [s]u[us]. In sum, none of the possible parses of the observed fast speech form *suus* is a local optimum.

For $z = 0$, however, [su]u[s] becomes a local optimum. Furthermore, it becomes the only local optimum besides the grammatical form [s][s]u[s], and the tuning of the parameter T_{step} nicely reproduces the difference between slow and fast speech, once again (Table 5.13).

As Table 5.14 shows, very similar results are returned for $z = 1$. The only significant difference is observed when $T_{step} = 3$. When cooling down the system quickly, one can check which valley a candidate belongs to. Thus, this observation shows that some of the candidates have moved from the valley of [su]u[s] to the valley of [s][s]u[s] when we changed the parameter z in the definition of constraint OOC.

To sum up, let us collect what the following hierarchy yields for the four types of words:

T_{step}	3	1	0.3	0.1	0.03
Type 2: usus					
[us]u[s]	0.394	0.398	0.431	0.495	0.581
[s][s]u[s]	0.606	0.602	0.569	0.505	0.419
Type 3: ssus					
[s][s]u[s]	0.661	0.680	0.835	0.969	1.00
[su]u[s]	0.339	0.320	0.165	0.031	0.00

Table 5.15: **Summary for hierarchy (5.10) with $z = 0$.** Each simulation was obtained by launching the simulation 600 times from each candidate.

T_{step}	Type 0	susu	Type 1	susu	Type 2	usus	Type 3	ssus
	[s]u[su]	[su]u[s]	[s]u[su]u	[su]u[su]	[us]u[s]	[s]u[us]	[s][s]u[s]	[su]u[s]
3	0.797	0.180	0.545	0.455	0.488	0.386	0.652	0.348
1	0.862	0.115	0.620	0.380	0.543	0.323	0.682	0.318
0.3	0.962	0.015	0.807	0.193	0.720	0.186	0.836	0.164
0.1	0.977	0.0001	0.964	0.036	0.907	0.070	0.972	0.028
0.03	0.977	0.0000	0.9998	0.0002	0.998	0.002	0.9999	0.0001

Table 5.16: **Summary for hierarchy (5.10) with $z = 1$.** Each simulation was obtained by launching the simulation 600 times from each candidate.

$$\begin{aligned}
& \text{ALIGN}(\text{word}, \text{foot}, \text{left}) \gg \text{OOC} \gg * \Sigma \Sigma \gg \\
& \gg \text{PARSE-}\sigma \gg \text{FOOTTYPE}(\text{trochaic})
\end{aligned} \tag{5.10}$$

For $z = 0$, we have trouble for Type 1 words. Namely, even though the global optimum is a parse of the andante form ([s]u[su]u), yet the model fails to predict the allegro form: the only other local optimum is [s]u[s][su]. Similar problem arises for Type 2 words ($\text{OOC}_{\text{sigma}=\text{usus}}$), where [s][s]u[s] is a local optimum, unlike any parse of the observed allegro form. Interestingly, however, this model correctly predicts that the grammatical form usus is extremely difficult to produce: even for low T_{step} values the error rate is higher than 50%, which is consonant with Schreuder (2006)’s experimental results (Table 5.15). For Type 0 words ($\text{OOC}_{\text{sigma}=\text{susu}}$), finally, the only local optimum is [s]u[su], that is, we again have trouble explaining why human fast speech can produce an “erroneous” form suus.

With weight $z = 1$ in the definition of OOC, the results shown on Table 5.16 were produced. Interestingly enough, the parse of the empirically observed fast speech form suus for Type 2 is predicted to be [s]u[us], and not [su]u[s]. These two parses behave similarly with respect to all constraints, except for the lowest ranked one, TROCHAIC. Indeed, [su]u[s] satisfies this constraint, unlike [s]u[us]. And yet, the less harmonic [s]u[us] is a local optimum, and therefore may and does emerge as an output; whereas [su]u[s] cannot be produced by the present model for it has a better neighbour, namely, [us]u[s].

Unfortunately, Type 2 keeps on producing a third output, [s][su]u. The rate of this empirically unattested form is between a third and a half (the latter for $T_{step} = 0.3$) of the rate of [s]u[us]. See the right part of Table 5.9, which reports on the same experiment but performed earlier (hence the minor differences in

<i>fo.to.toe.stel</i> 'camera'	<i>uit.ge.ve.ríj</i> 'publisher'	<i>stu.die.toe.la.ge</i> 'study grant'	<i>per.fec.tio.níst</i> 'perfectionist'
susu	ssus	susuu	usus
<i>fó.to.tòe.stel</i> fast: 0.82 slow: 1.00	<i>ùit.gè.ve.ríj</i> fast: 0.65 / 0.67 slow: 0.97 / 0.96	<i>stú.die.tòe.la.ge</i> fast: 0.55 / 0.38 slow: 0.96 / 0.81	<i>per.fèc.tio.níst</i> fast: 0.49 / 0.13 slow: 0.91 / 0.20
<i>fó.to.toe.stèl</i> fast: 0.18 slow: 0.00	<i>ùit.ge.ve.ríj</i> fast: 0.35 / 0.33 slow: 0.03 / 0.04	<i>stú.die.toe.là.ge</i> fast: 0.45 / 0.62 slow: 0.04 / 0.19	<i>pèr.fec.tio.níst</i> fast: 0.39 / 0.87 slow: 0.07 / 0.80

Table 5.17: **Observed vs. simulated frequencies:** Simulated frequencies Using hierarchy (5.10) with $z = 1$ are given in italics, while observed ones in bold (Schreuder, 2006). In the simulation, $T_{step} = 3$ was used for fast speech and $T_{step} = 0.1$ for slow speech.

the values), where the frequencies of [s][su]u are also mentioned explicitly.

Table 5.17 (from Bíró, 2005a) contrasts the empirical results of Schreuder (2006) with the simulation results. Using hierarchy (5.10) with $z = 1$ (Table 5.16), we model fast speech by using $T_{step} = 3$, and slow speech with $T_{step} = 0.1$. Then, the numerical match between Schreuder's experiment and our simulation is very nice for Type 3 words (*uitgeverij*), and reasonably good for Type 1 words (*studietoelage*). In the later case, the 81% of the slow speech could be easily reproduced by setting $T_{step} = 0.3$ (cf. Table 5.16) instead of $T_{step} = 0.1$, but then the reason of changing the value of T_{step} simulating *andante* speech would be unclear. The 38% of the correctness in fast speech remains a mismatch between the empirical data and the simulation, similarly to the values at both speech rates for Type 2 words (*perfectionist*).

These quantitative mismatches, as well as the presence of an alternative local optimum for Type 2 words should not discourage us. Modelling the *andante* forms as globally optimal candidates and the *allegro* forms as the only further local optima is in itself not a self evident task. Moreover, we could reproduce qualitatively the circumstance that different types display different frequencies—a result that other variations of Optimality Theory might have problems explaining. Even more encouraging is that we correctly predicted the relative rank of the three types: in both *andante* and *allegro* speech, Type 3 is realised more frequently by the *andante* form than Type 1 is, and Type 2 words have the least chance to be pronounced as their *andante* form. The fact that accounting for this relative ranking had not been set *a priori* as a goal, but was obtained as a surprise present, is an additional argument that SA-OT is “on the right track”.

Recall, finally, that the model with hierarchy (5.9) (Table 5.12) had the advantage that for all of the four Types, the global optima corresponded to the *andante* forms and the *allegro* forms were the only other local optima in the model. Hopefully, the reader will appreciate the non-triviality of this result by now, even if the frequencies produced by this model are far from the empirically observed ones. Another model (Table 5.15), which did not correctly reproduce the *allegro* forms, did however correctly predict that finding the global optimum is extremely difficult for Type 2 words, and produced frequencies for the *andante* form well below 0.5.

To sum up, I think I may be confident that one day somebody will find the winning hierarchy.

5.7 Getting rid of OOC: Biased initial state

OOC is clearly an awkward constraint. What it tries to capture—besides analogy effects—is the influence of the components on morphologically complex words. These phenomena used to be accounted for by cyclical rules in pre-OT Lexical Phonology (Kiparsky, 1982). As standard Optimality Theory sees phonology as one big box, there is no room for cyclicity. (Serial OT is a Lexical Phonology-type variation of standard OT.) As you have only input and output, and no intermediate stages of morphological derivation, only the surface forms of other lexical items can influence a given output form.

In SA-OT, however, there is an additional way of introducing the surface forms of the constituents into the computation of the compound form. Remember that the algorithm is launched from an initial candidate, that is, a potential surface form. So far, each element of the finite candidate set has been chosen with equal probability to be the initial state. However, we can also introduce a bias, and begin always with the candidate that is derived from the surface forms of the compound's components (that is, immediately from the reference string σ preferred by $\text{OOC}_{\sigma,z}$). Always launching the simulation from the candidate most faithful to the morphological components should favour the local optimum that is the closest. Thereby, we could get rid of the awkward OOC constraint (and especially of its arbitrary parameter z), and replace it by stipulating which candidate the algorithm should choose as initial candidate.

An additional argument could be brought in favour of such an approach. What should be the level in the language that is optimised during production: an utterance, a phrase, a word, a syllable? The search space (if not infinite) often grows exponentially in the length of the unit to be optimised—not a rosy perspective. Therefore, one could argue for a (parallel) optimisation of smaller units, which is followed by the optimisation of their combination. In this second phase, the optimisation process can concentrate on the way the smaller units are combined and on the phenomena occurring at their edges. This proposal seems to be sound, even if it might further reduce precision, the optimum of a combination not always being the optimal combination of the optimised building blocks. This proposal would correspond to the principle in Lexical Phonology (Kiparsky, 1982) according to which lexical rules are applied exclusively in derived context, that is, where some change has recently taken place, for instance on the boundaries of the *latest* morphological derivation. Phenomena supporting Kiparsky's principle would then be exactly the cases where serial local optimisation misses the global optimum.

In short, launching SA-OT from the candidate that is the concatenation of the previously optimised (*i.e.*, grammatical) forms of the compound's components introduces a bias that hopefully could account for phonological phenomena grounded in morphology, which have also motivated Lexical Phonology and OOC in the past. Additionally, such a “Serial SA-OT” would also simplify the computation by decomposing the problem.

Sadly, however, the aim of getting rid of constraint OOC has not been achieved yet. Still, the following experiments show at least what happens if

Outputs	slow (andante) speech	fast (allegro) speech
[s]u[su]	937	950
[su]u[s]	1	45
u[us][s]	62	5

Table 5.18: **Biased initial candidate:** The outputs of running SA-OT 1000 times, and always choosing [su][su] as the initial candidate. $T_{step} = 0.1$ for slow speech and $T_{step} = 1$ for fast speech.

we introduce such a bias into the choice of the initial candidate.

In the first experiment, I always took the candidate [su][su] as the starting point of the simulation. The reason for that is that the word *fototoestel* is obtained by compounding the words *fóto* and *tóestel*, that is, we join an [su] component to another [su] component. Therefore, the candidate [su][su] is arguably *the* candidate that should be taken as the starting point of the simulation.

The initial phase of the simulation, in which temperature is above the domain of the highest ranked constraint, was introduced precisely in order to diminish the role of the choice of the initial candidate. If the initial temperature is so high that the random walker can walk all over the whole search space several times before the decreasing temperature reaches the domain of the highest constraint, we obtain a situation practically equivalent to the previous experiments, where each candidate had equal chance to become the starting point of the walk. This is why in the present experiment K_{max} was not higher than the index associated with the highest ranked constraint.

Table 5.18 shows the number of outputs out of 1000 runs, each taking [su][su] as its starting point. The hierarchy was $OOC_{\sigma=susu, z=2} \gg * \Sigma \Sigma \gg \text{PARSE-}\sigma$. In each case, the starting value of the temperature was $K_{max} = 2$, which was the domain corresponding to the highest constraint (OOC). That is, no fully random walk preceded the simulation, the parameters of the simulation would have been otherwise similar to the ones described in the previous experiments. T_{step} used for decreasing the temperature was 0.1 for andante speech and 1 for fast speech.

Although the rate of the form [s]u[su]—the correct form for competence—is similar in slow and fast speech, the fast speech form [su]u[s] becomes much more frequent in the model of allegro speech. Here again, the “artefact” form u[us][s] spoils the beauty of our results.

In the next experiment, I tried to avoid using OOC. The same landscape is used for the three types with four syllables (0, 2 and 3), and the goal was to end up the simulation in different local optima in function of the initial candidate. Namely, the Type 0 word *fótotòestel* is a compound of [fóto] and [tóestel]. Hence, the initial candidate corresponding to this type is [su][su], and the closest local optimum should be some parse of susu, such as [su][su] or [s]u[su]. Type 2 word *perfèctioníst* is derived from *perf[éct]*, and the *íst* ending is stress bearing. Thus, SA-OT should choose u[s]u[s] as the initial candidate for this word, and should be stuck in this (or in a similar) local optimum. Last, the initial candidate corresponding to *ùitgèveríj* ought to be [s][su][s], which (or its neighbour [s][s]u[s]) must be also a local optimum.

An additional argument for choosing these initial candidates was merely pragmatic. Type 2 (usus) and Type 3 (ssus) words differ only in whether their

first syllable has a stress. As inserting or deleting a monosyllabic foot is one of the possible basic transformations, many of their parses are neighbours: $[s][su][s]$ and $u[su][s]$, as well as $[s][s]u[s]$ and $u[s]u[s]$. The goal, however, cannot be to have neighbouring candidates (for instance $[s][s]u[s]$ and $u[s]u[s]$) simultaneously local optima: if a valley with a horizontal bottom were created, then forms *ssus* and *usus* would be equally probable for both types. This is also why the non-neighbouring candidates $u[s]u[s]$ and $[s][su][s]$ were chosen as initial states, and why both have to be local optima where the simulation will be stuck.

Additionally, this same similarity renders finding an adequate hierarchy very difficult: how can we prefer saving the first monosyllabic foot in $[s][su][s]$, and simultaneously not insert that initial foot into $u[s]u[s]$? A solution might be the following model, which includes, among others, $[su][su]$, $u[s]u[s]$ and $[s][su][s]$ as local optima:

$$(\text{FTBIN} \bmod 2) \gg \text{FOOTTYPE}(\text{trochaic}) \gg \text{PARSE} - \sigma \quad (5.11)$$

Here, constraint $(\text{FTBIN} \bmod 2)$ is derived from the wide-spread constraint *FOOTBINARITY* (e.g. used by Tesar and Smolensky, 2000), which assigns one violation mark to each monosyllabic foot $[s]$. Yet, we replace the number of the monosyllabic feet by its remainder after dividing it by 2, in order to have candidates $u[s]u[s]$ and $[s][su][s]$ (with an even number of $[s]$) satisfy it, unlike their neighbours $uuu[s]$, $u[s]uu$, $u[s][us]$, $[us]u[s]$, $u[su][s]$ and $[s][su]u$ with an odd number of $[s]$.

By performing a gradient descent (a simulated annealing with K_{min} lower than constraint $\text{PARSE} - \sigma$, and thus not allowing any uphill moves) in this model, we can account for the three types of words without reference to OOC. The three types indeed correspond to launching the simulation from candidate $[su][su]$, $u[s]u[s]$ and $[s][su][s]$. As each of them is a local optimum where the simulation gets stuck, the parses corresponding to the morphological structure of the words— $[su][su]$ for Type 0 words, and so forth—are returned, and hence we have accounted for the andante forms.

The fast speech form *suus* remains to be explained, nevertheless. We would like the simulation to end up in some parse of *suus* whenever not ending up in the andante form. Thanks to constraint $(\text{FTBIN} \bmod 2)$, parse $[s]uu[s]$ is also a local optimum. There are, however, many more local optima in this model: $[s][s][s][s]$, $[s][s][s]u$, $uu[s][s]$, etc. In turn, if K_{max} is higher, many different candidates may be returned, not exclusively the initial candidate of the walk (the andante form) and $[s]uu[s]$ (the allegro form).

Further research may come up with a hierarchy that is able to do that. If some parse of the fast speech form, for instance $[s]uu[s]$, were the global optimum, and separated the local optima from each other,²² then faster speech would correspond to more steps at higher temperature domains. In such a model, gradient descent would be stuck in the local optimum that is the closest to the initial candidate, thereby accounting for OOC-like phenomena in andante speech. At the same time, fast speech would correspond to a “hothead speaking”, to more iterations spent at higher temperature, which enhances the

²²That is, paths connecting two local optima require climbing very high summits—which is prohibited if K_{max} is chosen adequately—unless the path passes by the global optimum. In such a “star-shaped” neighbourhood structure, the algorithm returns either the local optimum in the branch of the star from where the walk is launched, or the global optimum situated at the heart of the star.

chance of finding the globally optimal parse of *suus*. As this model employs only markedness constraints describing what is “easy to pronounce”, the globally optimal candidate would be indeed universal, independent of the input, optimising for pronunciation—as opposed to the morphologically informed *andante* form which depends on the morphologically biased choice of the initial candidate.

Observe that this last model already moves away from the original idea in this chapter to have higher T_{step} values (fewer iterations) for slower speech, and lower T_{step} values (more iterations) for fast speech. In this respect, this last proposal brings us to the models introduced in the next chapter where K_{max} will play an important role.

5.8 Discussion

Chapter 2 has introduced a version of *simulated annealing* as an algorithm for finding the optimal candidate within some Optimality Theoretical grammar. Although simulated annealing, like other heuristic techniques for combinatorial problems, does not guarantee that one finds *the* optimal candidate, there are quite a few arguments for using it. First, generation in Optimality Theory can be of a very high complexity if the candidate set is huge. Heuristic techniques have been actually developed to approximate the solution of these hard problems within a reasonable time framework.

Second, if Optimality Theory is supposedly an adequate model for linguistic competence, then Simulated Annealing for Optimality Theory models linguistic performance. It is an “approximately good” algorithm that returns *some* form within limited time, using a restricted computational capacity, and the chance is not bad for the returned form to be the grammatically correct form predicted by competence / OT. Furthermore, if circumstances require, the performance model is able to work faster, although the probability of returning the grammatical form diminishes. Not just anything can be returned as erroneous outputs, but only local optima. This latter phenomenon may be used as a model for the observed typical performance errors produced in fast speech.

In this chapter, we have demonstrated this argument by showing how SA-OT can account for Dutch stress assignment. After having formulated the building blocks of SA-OT, such as the candidate set, the neighbourhood structure and the constraints, we played around with the model. We have demonstrated how fast speech errors can be reproduced by decreasing the number of iterations, that is, by increasing the parameter T_{step} . Then, we varied in addition the parameters T_{min} and T_{max} , and concluded that they also influence the success of the algorithm slightly. This observation is relevant for instance if one asks about the effect on SA-OT of monotonic transformations of the constraints, such as $C'_i(w) := \gamma \cdot C_i(w)$ ($\gamma > 0$).

Subsequently, we tried to include further constraints in order to better account for the behaviour of the three types of words dealt with by Schreuder and Gilbers (2004) and Schreuder (2006). Different word types produced different landscapes due to constraint *Output-Output Correspondence*, and therefore they returned the *andante* form and the *allegro* form with different frequencies for the same parameter setting, even though the same process lies always in the background. It is the topology and the landscape, formed by *all* candidates, even by those not appearing on the surface, that explain the different frequencies

for different word types.

The results presented might not be fully convincing, and yet, they should encourage further work. The demo at <http://odur.let.rug.nl/~birot/sa-ot/index.php> might help the interested to experiment further. During this random walk in the search space of different constraints and rankings, we have also encountered a number of interesting phenomena, such as the increasing ratio of a non-global local optimum with decreasing T_{step} in Table 5.10. Finally, we tried to get rid of morphologically motivated constraints, such $OOC_{\sigma,z}$, by introducing a morphological bias into the choice of the initial candidate of the algorithm, and argued for further research.

An objection by several readers was that stress shift in fast speech can be simply explained by the tendency of languages to keep stress apart. While the length of one intervening unstressed syllable is enough in normal speech, this suggested time span requires two syllables at a faster rate. This explanation, also found in Schreuder’s work, does not account, however, for the differences found between word types.

A second objection argued that fast speech phenomena might be caused not necessarily by increased speed *per se*, but by reduced degree of effort (cf. eg., Kirchner, 1998). Notice, however, that the central factor in the simulations presented was T_{step} , that is, the number of iterations, which can be readily translated to milliseconds only if we suppose that one iteration steps is performed in a constant time. This, however, may be not so simple, because the speed of our simulations changed sometimes with a factor of 10 or 100. So the claim that the speed of the algorithm models the speed of speech is true only qualitatively, and the relationship between the two is not always one-to-one. But, on the other hand, the number of iterations can be interpreted as the degree of effort at least as much as real time.

How consonant is our model with related work on stress assignment? Both computational (Eisner, 2000a) and psycholinguistic studies (Schiller, 2003) have argued for an incremental—left-to-right directional—assignment of stress. Our performance model, however, computes the stress of the whole word at once. Yet, this apparent contradiction can be resolved easily. On the one hand, Eisner’s *directional evaluation* of some constraints, different from those used in the present study, is a theoretical construct on the level of the competence model, determining which form is grammatical. Whether this idea can have any influence on performance models remains unclear. On the other hand, Schiller et al. (2004) advances an alternative explanation of the outcome of their psycholinguistic findings. The fact that the stress in the first syllable of bisyllabic words is identified 60-70 ms earlier than the stress in the last syllable (Schiller, 2003) does not need to be explained by an incremental stress assignment; it can also be due to the “sequential nature of a perceptual mechanism used to monitor lexical stress”. In other words, it may be the case that the position of the stress is calculated in the same amount of time for words with initial and with final stress; and yet, word initial stress is recognised much faster because when looking for the position of the stress in some experimental settings, the word, whose stress has been already determined, is checked left to right.

The following chapter tackles a new linguistic phenomenon, employing a new type of search space, and investigating new aspects of SA-OT.

Chapter 6

Dutch Voice Assimilation with SA-OT

6.1 The magic square

In this chapter, we are discussing a set of linguistic variations that one could call the *magic square*. Their common characteristic is that two related features vary in a synchronous way. The basic structure is represented in Figure 6.1, where + and – are the possible values of the two consecutive features.

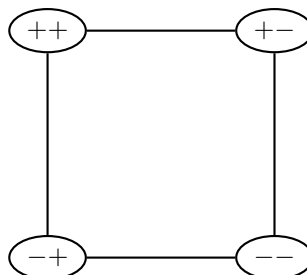


Figure 6.1: The “magic square”

The prototypical example, which we shall presently use in the discussion, is *voice assimilation*: if two neighbouring stops have different [voice] features, either regressive or progressive assimilation takes place, yielding a homogeneous sequence with respect to this articulative feature. Hence, candidates ++ and -- are favoured over candidates +- and -+. Steriade, Lombardi, Joe Pater, Eric Bakovic and others have argued that it is very uncommon (impossible) for a language to follow a third strategy, such as inserting an epenthetic vowel in order to avoid clash in voicing (e.g. Lombardi, 2001), even though epenthesis is a frequently used strategy to avoid prohibited consonant clusters, in general.¹

¹Thus, in Brazilian Portuguese, *football* translates to *futebol* and English *handball* to *handebol*. For schwa epenthesis in Modern Hebrew, see for instance Bíró and Hamp (2002). But epenthesis is claimed not to be a repair strategy for syllable codas having an unwanted [voice] feature. Not only can an epenthetic vowel never intervene between two stops with dissimilar [voice] features, but suffixing a final schwa is also not an option in languages prohibiting voiced consonants in a word-final position.

If + in Fig. 6.1 stands for [+voice], and – for [-voice], then out of the four possibilities, only ++ and – – may be grammatical in many languages. Furthermore, although for some input usually only one of ++ and – – is grammatical, yet sometimes the other may surface as an alternate form. Our paradigmatic example will be the Dutch word pair *op die* (‘in this...’), where the clash in the voice feature can be solved by assimilation in either of the two possible directions. Dutch phonology requires *regressive* voice assimilation, that is, *o[bd]ie*, and yet, often *o[pt]ie* emerges as the result of *progressive* voice assimilation.

Further examples can be also found that exhibit a similar *magic square*. The Dutch word *partij* (‘(political) party’) is sometimes pronounced as [ptɛi],² forming an otherwise prohibited open consonant cluster with the deletion of two segments. Now + represents the presence of the segment and – its absence in the rhyme, and again the same diagonally opposed forms alternate: the grammatical form ++ with the alternative form – –. The two other forms involving only partial deletion in the rhyme, +- and -+, are not allowed.

An analogous situation is used by Bíró and Gervain (2006): the *resyllabification* of the [z] in the Hungarian definite article *a* / *az*. The choice between the two allomorphs depends on whether the next word begins with a consonant or with a vowel:

$$\begin{array}{ll} az \text{ énekesnő} & \text{‘the soprano’}, \\ a \text{ kopasz énekesnő} & \text{‘the bald soprano’}. \end{array} \quad (6.1)$$

The definite article is also prone to undergo resyllabification turning the [z] into the onset of the subsequent syllable. In other words, the pause between the article and the subsequent word can drop, and therefore the segment [z] can be perceived as belonging to the next word, sometimes leading to misunderstanding in the case of minimal pairs, and sometimes to language games.

Judit Gervain performed a controlled psycholinguistic experiment measuring the frequency of this phenomenon. Hungarian distinguishes at least five speech levels on the basis of the rate/speed of speech, and she tested two of them: (i) motherese or infant-directed speech, characterised by a rather slow pace and emphatic, exaggerated prosody, (ii) and fluent, casual, conversational style with a medium speech rate. The hypothesis, which had been already described theoretically but never measured empirically (Kiefer, 1994), claims that more resyllabification occurs with the acceleration of the speech rate. The experiments confirmed this hypothesis by measuring the overall length and the presence of pauses in critical minimal pairs (e.g. *az ár* ‘the price’ vs. *a zár* ‘the lock’) excised from test sentences pronounced by three female native speakers

I am thankful to everybody who answered my question on Linguist List in February 2005. That is where I was referred, among others, to the following urls:

<http://roa.rutgers.edu/view.php3?id=29>,
http://www.linguistics.ucla.edu/people/steriade/papers/P-map_for_phonology.doc,
<http://people.umass.edu/pater/pater-balantak.pdf>,
http://camba.ucsd.edu/bakovic/work/bakovic_wilson_lar.pdf.

²At least, it was pronounced so by the late Dutch prime minister Joop den Uyl within the context *Partij van de Arbeid*, ‘Labour Party’. Otherwise, the native speaker would not judge [ptɛi] necessarily better than [patei] or [prtɛi], or would also allow the insertion of a schwa ([p@tɛi]). Note, however, that, similarly to Boersma (2004b), in SA-OT grammaticality judgements need not correlate with production: a form might be produced by the performance model (SA-OT) even if it is judged absolutely out of question by the underlying competence model (OT).

in both conditions. In the production of the slower infant-directed speech, resyllabification (*az ár* pronounced as *a zár*) happens about in 40% of the cases, which raises to about 80% in conversational style. On the perception side, she tested whether a naive group of native speakers could identify which of the minimal pairs were pronounced by the speakers. When the segmentation cue—the pause—was present, subjects identified the words with a 85% accuracy, while they were at chance when the words were pronounced without a pause.

Now, the + and – values of the *magic square* describes the presence and the absence of the segment [z] on either side of the syllable boundary. Candidate +- corresponds to the input obtained by concatenating the lexical items (*az.énekesnő*), whereas -+ is the resyllabified form (*a.zénekesnő*). According to Gervain, the -- form (*a.énekesnő*) appears in children's speech.

In an alternative analysis, + corresponds to the preferred syllable structures (an empty coda and an onset filled with [z]), and – to the disfavoured ones (a coda filled with [z] and an empty onset). Now, the original form *az.énekesnő* is -- and the resyllabified form *a.zénekesnő* is ++, whereas +- (*a.énekesnő*) and -+ (*az.zénekesnő*) could be but are non attested—similarly to the +- and -+ candidates of all of the previous examples. Observe that the resyllabified form is the best with respect to syllable structure, whereas the original form is the worst one among the four theoretical possibilities. And yet, if with respect to some other factor forms +- and -+ are worse, candidate -- becomes a local optimum thanks to the two candidates separating it from the global optimum in Fig. 6.1. We shall return to this phenomenon in section 7.1, and present a detailed analysis based on a subsequent experiment of Judit Gervain.

As an example from syntax, I take the favourite one of Modern Hebrew linguist purists. The “correct” form for ‘three shekels’ would be *šloša škalim*, with agreement both in gender (morphologically visible on the cardinal number) and in number (the [-im] plural suffix on the noun). Nonetheless, most speakers use *šaloš šekel*, omitting both agreement features.³ This magic square is formed by the presence (+) or absence (–) of agreement in number (first position) and in gender (second position). The grammar, in general (for noun+adjective pairs), requires again candidate ++ to win, but in some special cases (namely, with numerals) speaker might also produce --, but not candidates +- and -+.

Notice that in many of these examples, form --, the alternative one, occurs in frequent (“semi-lexicalised”) constructions. Progressive voice assimilation in Dutch would be unconceivable in nouns such as *zakdoek* or *duikboot*, only in bi-grams of unstressed function words. The pronunciation [ptɛi] of the word *partij* was characteristic to Joop Den Uyl, the late prime minister of the Netherlands (D. Gilbers and M. Schreuder, personal communication) in the expression *Partij van de Arbeid* (‘Labour Party’). The lack of double agreement in Modern Hebrew occurs only in frequently used expressions of quantity.

An Optimality Theoretical account of these phenomena should include at least two constraints. The first constraint C1 requires surface homogeneity, punishing the heterogeneous forms +- and -+. The lower ranked constraint

³Some speakers in colloquial Hebrew omit the agreement in gender for all cardinal + noun constructs. In their case, however, one may argue that the masculine forms of the numerals have been removed from the language altogether for the sake of paradigm uniformity, due to their counter-intuitiveness. Namely, in semitic languages the gender morphemes on numerals are the opposite of the gender morphemes generally found in the grammar.

C2 prefers ++ to --, whereas -+ and +- may either satisfy or violate C2. In the following tableau, the well-known \mathbb{E} points to the grammatical form, whereas the \sim symbol shows the alternative form:

	C1	C2
\mathbb{E} ++		
\sim --		*
+-	*!	?
-+	*!	?

(6.2)

Indeed, tableau (6.2) together with the candidate space topology in Fig. 6.1 will turn ++ into the global optimum, and -- into the only alternative local optimum (see Fig. 6.4 on page 172 for a concrete example).

To tell the truth, each story is a little bit more complex if we want to stay linguistically correct. Still, the basic structure of the tableaux remains similar. A crucial property of these tableaux is that +- and -+ are defeated in an earlier stratum, while the difference between ++ and -- appears only lower in the hierarchy.

In case of the word *partij* pronounced [ptɛi], one may propose using two faithfulness constraints. Constraint C1 is FAITHFULNESS[RHYME]: each syllable rhyme in the input is identical to its image in the output form, *if* there is such an image. The rhyme /ar/ in the input /par.tei/ is identical to its image [ar] in the candidate [partɛi], but different from [a] and [r] in candidates [patɛi] and [prtɛi] respectively. That rhyme, however, has no correspondent in the string [ptɛi] for it has been deleted altogether, therefore the constraint is satisfied vacuously.⁴ Subsequently, constraint C2 is a faithfulness constraint on segments: deleting each segment increases the number of violation marks by one. This constraint then favours [partɛi] with zero violation marks to [ptɛi] with two violation marks. Candidates [prtɛi] and [patɛi] are assigned only one violation mark each, but they have been already put out of the game previously by constraint C1.

In Hungarian resyllabification, C1 is a constraint which requires a strictly alternating vowel-consonant sequence, which is satisfied both by *a.zénekesnő* and by *az.énekesnő*, but not by *a.énekesnő* or by *az.zénekesnő*. C2 can be derived from constraints ONS and NOCODA known from Basic Syllable Structure Theory (Prince and Smolensky, 2004): by their sum, each empty syllable onset and each filled coda incurs one violation mark.

Using a “pseudo-minimalist” approach in the case of syntactic agreement in Modern Hebrew, constraint C1 in (6.2) can be said to require agreement features

⁴Similarly to what will be said on agreement in Hebrew, we could differentiate between a candidate [p.tei] in which the correspondence relation is not defined on the underlying rhyme /ar/ (thus, no image in the candidate), and a candidate [p∅∅.tei] in which the correspondence relation maps the rhyme /ar/ to an empty string, incurring two violation marks. Introducing the second candidate does not have any effect, for it is always a loser, because it is harmonically bounded by other candidates. A different, maybe more convincing but less elegant solution is to use two constraints to eliminate candidates [prtɛi] and [patɛi]. The first candidate can be easily eliminated by using highly ranked syllable structure constraints that do not allow a complex onset [prt] (being too complex and violating sonority requirements), and do not allow for the syllabification of [r] as a nucleus either. The second candidate may be eliminated by using a simpler and more convincing version of FAITHFULNESS[RHYME]: a rhyme in the surface form has to correspond to a rhyme in the underlying form. This second version of FAITHFULNESS[RHYME] is satisfied by [par.tei], [ptɛi], [prtɛi] and [pr.tei], but not by [pa.tei], for the rhyme [a] does not correspond to the rhyme /ar/ in the input form.

to be either checked or unchecked: *in the case* checking does take place overtly, then no feature may be left unchecked (one violation mark per feature left unchecked). The form — (*šaloš šekel*) satisfies this constraint automatically because no feature checking occurs. (An alternative candidate, identical on the surface, would violate this constraint twice if it involves feature checking, but then both gender and number are left unchecked.) Forms +— and —+ (*šloša šekel* and *šaloš škalim*) do involve feature checking, but not all features are checked, which leads to violating constraint C1. Subsequently, constraint C2 requires features to be checked, so any unchecked (not agreeing) feature incurs one violation mark. Hence *šaloš šekel* with two unchecked features is worse than *šloša škalim* (both features checked) for C2. The two constraints are almost identical, the only difference being that the lower ranked constraint requires features be checked always, whereas the higher ranked constraint requires it only if feature checking is performed in general.

Finally, in the voice assimilation example, phonology would propose a markedness constraint [αvoice][αvoice], requiring a homogeneous sequence with respect to the [voice] feature; as well as a faithfulness constraint that punishes any change of the value in the [voice] feature compared to the input form. These two constraints would not distinguish however between *o[bd]ie* and *o[pt]ie*, for both satisfy markedness and both violate faithfulness once. Therefore, the regressiveness of the assimilation should be also incorporated into the analysis. In addition, we also would like to consider forms with epenthesis in a subsequent approach, thus the constraint DEP is required to punish epenthetic forms. In the following section, we work out the details of this analysis.

6.2 Voice assimilation in Dutch

Voice assimilation in general, and regressive voice assimilation of neighbouring stops in particular, is an extremely widespread phenomenon across languages. Not surprisingly, we can also observe it in Dutch, a language that tends to neutralise the [voice] feature of obstruents in other contexts, as well, such as in the word-final position.

The middle consonant cluster in words such as *duikboot* (‘submarine’) or *zakdoek* (‘handkerchief’) exemplifies *regressive voice assimilation*: in these cases we obtain [gb] and [gd] respectively. The coda of the previous syllable assimilates to the onset of the subsequent syllable. The traditional way to account for this phenomenon in Optimality Theory is to assume two constraints, namely a faithfulness constraint overranked by a markedness constraint. The constraint FAITH[VOICE] requires the value of the [voice] feature be kept unchanged in the output, whereas ASSIMILATE[VOICE] punishes adjacent stops not sharing their [voice] feature in the surface form. The need for the faithfulness constraint is supported by hypercorrect (or extremely careful) pronunciation yielding *za[kd]oek* and *dui[kb]oot*: in this register FAITH[VOICE] is promoted above the markedness constraint ASSIMILATE[VOICE], due to which assimilation may not take place.⁵

⁵As Paul Boersma pointed out, the word *handboek* (‘hand book’) is pronounced as *han[tb]oek* in equally careful speech, violating both FAITH[VOICE] and ASSIMILATE[VOICE]. This case might be influenced by another factor, such as an Output-Output Correspondence to the word *han[t]*.

Actually, careful vs. careless speech is often seen as a parameter orthogonal to speech rate (e.g., Kiefer, 1994), the first factor being dependent upon the social context, while the second being determined by time pressure on the individual speaker. Fast careful speech may have different characteristics from careless speech, which itself can also have different speech rates. Hence, they might have to be modelled separately. This is why employing constraint reranking to account for these extremely careful or hypercorrect forms does not contradict our agenda of using SA-OT for speech rate dependent phenomena. In turn, constraint reranking—either performed categorically, or in a Stochastic OT-style—reflects the intuitive view that extremely careful or hypercorrect speech is indeed about faithfulness to the underlying form; or, more precisely, to the written form in literate languages with a prescriptive tradition. Hypercorrectness could be seen as a separate register (or language), thus stipulating a separate hierarchy is not in conflict with our previous criticism about supposing separate hierarchies for different speech rates.

After these considerations, we can focus on phenomena that are typical to speech rate (or other factors), and we may ignore the hypercorrect forms. Dutch features an additional variation: the preposition *op* followed by *die* (only as a demonstrative pronoun or an article, such as in *op die manier* ‘in that way’) may sometimes involve *progressive* voice assimilation, and result in the consonant cluster [pt], besides the form [bd] yielded by regressive assimilation.⁶

Progressive voice assimilation between stops seems to contradict our belief in a homogeneous Dutch phonological system, because exclusively regressive assimilation is allowed everywhere else. In order to save the uniform phonology, as part of the supposed linguistic competence of the native speaker, we shall try explaining the form *o[pt]ie* as a performance phenomenon and reproducing it with simulated annealing. Our strategy here is to exile exceptions from competence (the static mental representation of the language), and to use the performance (or computational-production) model to account for them. If it works, we can keep the competence model simple and still account for all observed data.

Two models will be introduced, and these two models will demonstrate the capabilities and restrictions of SA-OT. Indeed, the real goal of this chapter is to further analyse what SA-OT is able to do, rather than to account for Dutch progressive assimilation in particular. The latter is taken as a mere example out of the analogous phenomena listed in the previous section, and ongoing and future work (such as Bíró and Gervain, 2006) should collect more empirical data that our simulations ought to reproduce.

The first model uses a finite (actually, quite restricted) search space, and is only able to account for a 50%-50% distribution of the forms *o[pt]ie* ([pt] in short) vs. *o[bd]ie* ([bd], henceforth), independently of the parameter settings.

⁶As I have been informed by my readers, for further references on the subject see for instance: Wim Zonneveld: Lexical and phonological properties of Dutch voice assimilation, in: Van der Broecke *et al.* (eds.): *Sound Structures, Studies for Antonie Cohen*, Floris, Dordrecht, 1983:297-312; or Mirjam Ernestus: *Voice Assimilation and Segment Reduction in casual Dutch: A corpus-based study of the phonology-phonetics interface*, PhD thesis, Vrije Universiteit, Amsterdam, Amsterdam, 2000. Wim Zonneveld claims that the double forms are limited to clitic-like non-lexical categories, so *op deze lijst* ‘on this list’ can be realised both as [bd] and [pt], but *op dikke boeken* ‘on thick books’ must be [bd].

Adam Albright pointed out that Northeastern Yiddish displays a similar progressive voice assimilation that is basically limited again to a function word, namely, to the reflexive pronoun *zikh*. For instance, *golt zikh* ‘shave-3sg’ becomes *gol[ts]ikh*.

The model will be a slightly more complicated version of the toy example presented in section 2.3.2. The lesson is that SA-OT does not necessarily converge towards maximal precision. Instead of interpreting this observation as a failure of SA-OT, I propose to see it as a source of hope for the frequent cases where simple and elegant linguistic models have had to be turned into very complex ones just because of some few annoying exceptions.

Based on this model, I argue that linguistic data might be reproduced by keeping the competence model simple and by leaving the dirty job to such performance models. And since this family of performance models always predicts errors, independently of the parameter settings, one cannot distinguish *a priori* between phenomena related to competence in its narrow sense and between phenomena constantly introduced by the second level on Table 2.1 (page 43). This case is contrasted to the situation presented in Chapter 5, where the output frequencies depended on the parameters, and therefore the *allegro* form, whose frequency increased at higher speech rate, could be identified as the performance error form.

In contrast to the first one, the second model will allow tuning the frequencies of candidates [bd] (*o*[bd]*ie*) and [pt] (*o*[pt]*ie*) by varying the parameters. True, this second model necessitates a constraint which may not meet the expectations of all phonologists, for it is a markedness constraint referring also to the underlying form. However, the model turns to be illuminating about the possibilities of Simulated Annealing Optimality Theory, whereas further research may replace the problematic constraint with a less controversial one.

6.3 The building blocks of Simulated Annealing

First, we have to define the candidate set with respect to a given underlying form. Let the underlying form be a pair of stops $\sigma_1\sigma_2$. Now, σ'_1 denotes the stop that has the same features as σ_1 , but the [voice] feature is different; similarly for σ'_2 . The candidate set will then be a set of strings beginning with either σ_1 or σ'_1 , ending with either σ_2 or σ'_2 , and having zero or more epenthetic vowels (say, schwas) in-between—the simple regular language $\{\sigma_1, \sigma'_1\} \times @^* \times \{\sigma_2, \sigma'_2\}$.

We have not really argued for the need to epenthesise yet, but we advance it here for the sake of the second model to be presented in this chapter. After all, epenthesis is always a possibility for a phonologist, who would never prevent GEN from producing candidates including epenthetic segments, footnote 1 on page 161 notwithstanding.

To simplify notation, let us replace σ_1 and σ_2 by [p] and [d] (from the input *op die*). We write then the underlying (input) form as /pd/, and the output forms (candidates) as [pd], [bd], [pt], [bt], [p@d], [b@@t], etc. The @ symbol will refer to the epenthetic vowel, and a superscript may refer to its repetition n times (e.g. [p@ ^{n} d]), zero or more times (Kleene-star: [p@*d]), or one or more times (Kleene-plus: [p@⁺d]).

As we follow the usual steps of introducing the building blocks of SA-OT (see page 45 or page 129), we have to define next the neighbourhood structure on this set. We shall regard two candidates as neighbours if and only if one candidate can be reached from the other by performing exactly one of the following *basic steps*:

- Insert or delete exactly one epenthetic vowel (from $\sigma_1@^n\sigma_2$ to $\sigma_1@^{n\pm 1}\sigma_2$).

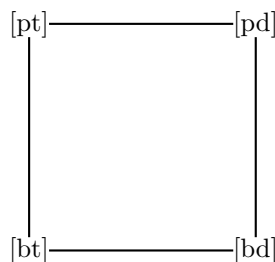


Figure 6.2: **Search space used in the first model for voice assimilation:** neighbours are connected by a line.

- Change the value of the [voice] feature of exactly one of the bordering stops (from $\sigma_1 @^n \sigma_2$ either to $\sigma'_1 @^n \sigma_2$ or to $\sigma_1 @^n \sigma'_2$).

In our first model the candidate set will be restricted to the four pairs of stops without allowing any epenthesis ($n = 0$) (Fig. 6.2). By adding the possibility of iterative epenthesis, we arrive at the structured candidate set appearing in Fig. 6.3, the one to be used in the second model.

As for the *a priori* probabilities, that is, the second part of the definition of a topology on the search space, we simply give equal probability to each neighbour of a candidate (Eq. (2.5) on page 49).

Now, let us move forward, defining the constraints. Suppose \mathcal{C}_w is the *correspondence relation* (see also section 4.1.5), that is a partial bijection⁷ mapping (some of) the segments (tokens) of the input string onto (some of) the segments of candidate w fulfilling *contiguity*,⁸ as defined by *Correspondence Theory* (cf. McCarthy and Prince (1993b) p. 67).

In our case, \mathcal{C}_w maps the first and the second underlying stops onto the first and the last stops in the candidate w , respectively. The epenthetic vowels are not contained in the range of \mathcal{C}_w .

We shall use the following constraints (the definition provided is more general than needed for the present case):⁹

- DEP (DEPENDENCY, “don’t epenthesise!”): one violation mark assigned to each segment in the candidate that does not correspond to a segment in the input string. In other words, a candidate w is assigned as many violation marks as the number of its segments, minus the cardinality of the range of \mathcal{C}_w
- ASSIMILATE[VOICE]: one violation mark to each pair of segments (σ_1, σ_2) in the candidate such that σ_1 immediately precedes σ_2 (in the *candidate*

⁷I shall call a relation $\mathcal{R} \subset A \times B$ a *partial bijection* if and only if \mathcal{R} is a bijection—a one-to-one mapping—between its domain and its range, even if its domain and its range may be a proper subset of A and B respectively.

⁸In the present model, *contiguity* requires that for all segments σ_1 and $\sigma_2 \in \text{Domain}(\mathcal{C}_w)$, segment $\mathcal{C}_w(\sigma_1)$ is left of $\mathcal{C}_w(\sigma_2)$ in the candidate string if and only if σ_1 is left of σ_2 in the input string.

⁹For historical reasons, constraints are typically defined in terms of what criteria must be met: what is the structure that does not incur any violation mark. However, Optimality Theory constraints are functions on the candidates that have not necessarily Boolean (true / false) values. Very often, the *number* of violation marks assigned plays a crucial role. This is why I repeatedly argue that constraints should be defined positively, by giving the *number* of violation marks they assign to a given candidate.

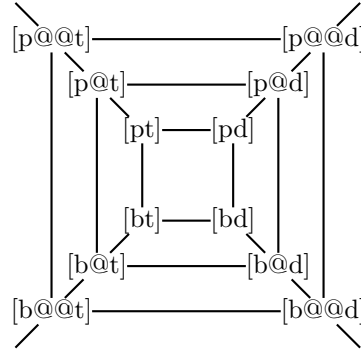


Figure 6.3: **Search space used in the second model for voice assimilation:** the character @ stands for the epenthetic schwas.

string), both have a [voice] feature, and yet, these features have a different value.

- STOPASSIMILATION=REGRESSIVE: one violation mark to each pair of segments (σ_1, σ_2) in the underlying representation such that σ_1 immediately precedes σ_2 (in the *underlying form*), both are elements of the domain of \mathcal{C}_w , furthermore σ_2 , $\mathcal{C}_w(\sigma_1)$ and $\mathcal{C}_w(\sigma_2)$ are stops with a [voice] feature, and yet these three voice features do not have the same value.
- FAITH[VOICE]: one violation mark is assigned to each segment σ in the underlying form such that $\mathcal{C}_w(\sigma)$ exists, both σ and $\mathcal{C}(\sigma)$ have a [voice] feature, and yet these features have a different value.

Constraint DEP, originally FILL in Prince and Smolensky (1993), is *the* standard constraint in Optimality Theory that prohibits inserting elements into a candidate that were not present in the input form. Constraint ASSIMILATE[VOICE], a straightforward way to compel stops to agree in voicing, is called AGREE by Lombardi (1995). She refers to FAITH[VOICE] as IDENT(laryngeal). Yet, her constraint IDENTONSET(laryngeal), which causes voice assimilation to be regressive by punishing unfaithful onsets but not codas, is not going to be useful in our analysis. Indeed, our fourth constraint, STOPASSIMILATION=REGRESSIVE might be claimed to be the most questionable.¹⁰

Notice that constraint STOPASSIMILATION=REGRESSIVE punishes all underlyingly adjacent pairs of stops that do not assimilate regressively (either do not assimilate at all or assimilate progressively, if they are different underlyingly), even if they are not adjacent on the surface. This constraint sounds quite weird to the ears of a phonologist, for markedness constraints should refer to properties of the surface form alone, without touching upon the underlying form. Nevertheless, we shall need it for our second approach.¹¹ In the first approach, the following simpler alternative may replace it:

¹⁰My impression was that finding the constraint corresponding to STOPASSIMILATION=REGRESSIVE is also the most difficult task when applying this model to the analogous phenomena mentioned earlier.

¹¹What the model requires is tableau (6.5). Alternative formulations of this constraint might be possible, which will assign violation marks with no significant change. In the footsteps of Lombardi (1995)'s constraint IDENTONSET(laryngeal), we could for instance give the following definition: one violation mark is assigned to each stop in an onset position that (i) is not

- STOPASSIMILATION=REGRESSIVE2: one violation mark goes to each pair of segments (σ_1, σ_2) in the candidate string iff σ_1 immediately precedes σ_2 (in the *candidate string*), both are elements of the range of \mathcal{C}_w , furthermore $\sigma_1, \sigma_2, \mathcal{C}_w^{-1}(\sigma_1)$ and $\mathcal{C}_w^{-1}(\sigma_2)$ are stops with a [voice] feature, assimilation has taken place (σ_1 and σ_2 share the same [voice] feature), and yet σ_1 has the same [voice] feature as $\mathcal{C}_w^{-1}(\sigma_1)$, whereas σ_2 differs from $\mathcal{C}_w^{-1}(\sigma_2)$ in this feature. (In short, progressive assimilation has occurred.)


As we have to check whether assimilation has been progressive or regressive, we probably cannot avoid referring to the underlying form, at least in some hidden way. Yet, this second formulation differs in two respects from the first one. Firstly, it does not punish progressive assimilation anymore if epenthetic vowels intervene between the stops in the surface form: this is exactly the point making the first definition unattractive to a phonologist but necessary for the second model to be presented. Secondly, the second formulation does not assign any violation mark to candidates where no assimilation has taken place (vacuous application), whereas the first formulation punished them for missing the occasion of assimilating (in a regressive way). Consequently, candidates [pd] and [bt] vacuously fulfil STOPASSIMILATION=REGRESSIVE2, whereas they violate STOPASSIMILATION=REGRESSIVE. This difference will not have any effect in the models, since these two candidates are already defeated by their neighbours due to a higher ranked constraint, namely ASSIMILATE[VOICE]. The second model will, nevertheless, crucially exploit the fact that [b@⁺d] are the only candidates with epenthesis satisfying this constraint, consequently that model necessitates that [p@⁺d] and [b@⁺t] violate it, too.

The last step in constructing our Simulated Annealing Optimality Theory model is to define the hierarchy. As explained in the introduction, the faithfulness constraint has to be demoted below the markedness constraints, otherwise no assimilation will take place. Constraint DEP, which will play a role only in the second model, should be ranked high in order to avoid forms with epenthesis becoming successful. Similarly, the relative ranking of ASSIMILATE[VOICE] and STOPASSIMILATION=REGRESSIVE is determined by the fact that *o*[pt]*ie* should emerge as an alternative form, and not *o*[pd]*ie*. In summary, the following ranking is the most likely to help us:

$$\begin{aligned} \text{DEP} &\gg \text{ASSIMILATE[VOICE]} \gg \\ &\gg \text{STOPASSIMILATION=REGRESSIVE} \gg \text{FAITH[VOICE]} \end{aligned} \quad (6.3)$$


Before going on with the analysis of Simulated Annealing, let us review the tableaux produced by this constraint hierarchy. The well-known $\mathbf{13}$ symbol refers to the optimal candidate, whereas \sim will refer again to the alternative form. In the first model, we may use STOPASSIMILATION=REGRESSIVE2, yielding the following chart (*vac* meaning that the constraint is satisfied vacuously):

faithful in its [voice] feature if the following syllable nucleus is original; (ii) is faithful to the input form in its [voice] feature if the following syllable nucleus is epenthetic.

/pd/	DEP	ASSIM[VOICE]	STASS=RGR[VC]2	FAITH[VOICE]
 [bd]				*
~ [pt]			*!	*
[pd]		*!	vac	
[bt]		*!	vac	**

(6.4)

For the second model, we need to use the original formulation of the constraint STOPASSIMILATION=REGRESSIVE and to enlarge our candidate set. The @ symbol refers to the epenthetic vowel (for instance, a schwa), and the exponent n multiplies the preceding character (in the tableau $n > 0$).

/pd/	DEP	ASSIM[VOICE]	STASS=RGR[VC]	FAITH[VOICE]
 [bd]				*
~ [pt]			*!	*
[pd]		*!	*	
[bt]		*!	*	**
[b@d]	*!			*
[p@t]	*!		*	*
[p@d]	*!		*	
[b@t]	*!		*	**
...
[b@ ⁿ d]	* ⁿ !			*
[p@ ⁿ t]	* ⁿ !		*	*
[p@ ⁿ d]	* ⁿ !		*	
[b@ ⁿ t]	* ⁿ !		*	**
...

(6.5)

6.4 Model 1: Finite search space

In the first approach, the search space (the structured candidate set) is restricted to the four candidates appearing in Fig. 6.2.

What does the landscape of the search look like? The landscape—represented in three dimensions in Fig. 6.4—is determined by the difference in the violation profiles of the *neighbouring* candidates. As this difference depends only on the highest ranked constraint distinguishing between the two profiles, phonologists can replace constraint STOPASSIMILATION=REGRESSIVE with STOPASSIMILATION=REGRESSIVE2, and both tableau (6.4) and the first rows of tableau (6.5) may be used. The global optimum is above [bd], and another local optimum, diagonally opposed to it, above [pt]. At the two ends of the other diagonal, [pd] and [bt] represent peaks.

Let us run simulations under the usual conditions. Temperature drops from above the highest constraint to much below the lowest constraint, so that enough time is given both to walk freely around the search space initially, and to find the local optimum (relax) finally. The domains containing the constraints follow

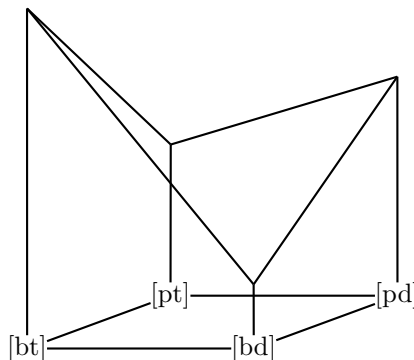


Figure 6.4: **3-D landscape of the first model for voice assimilation**: more harmonic candidates are drawn lower than the less harmonic ones (searching for the optimum corresponds to searching for the minimum). Candidate [bd] is the global optimum, [pt] is another local optimum, whereas [bt] is the least harmonic candidate.

each other, $K_{step} = 1$, and the real part of the temperature drops from $T_{max} = 3$ to $T_{min} = 0$ in equal steps T_{step} .

Based on our previous experience, we would predict Simulated Annealing Optimality Theory to produce the following behaviour: with slow simulation, the global optimum [bd] is easily found, whereas accelerated simulation may be stuck in the erroneous local optimum [pt]. The faster the simulation, the more frequently [pt] is expected to be returned—according to our intuition.

How very wrong we are! Implementing the model shows that [bd] and [pt] are returned with equal probability, 50% each, with some random dispersion. This happens so independently of the parameter setting!

The reason for this surprising result is easy to understand, and lies in the symmetry of the landscape. We face a case similar to those discussed in section 2.3.2. In fact, there is a greater symmetry than one would initially suppose. On the one hand, the chance of leaving the two local optima are equal at every moment of the simulation. To leave either of them, temperature has to allow violating ASSIMILATE[VOICE] once, which is possible at high temperatures, impossible at low temperatures, and has a chance between 0 and 1, when temperature is exactly in the domain of this constraint. The fact that [pt] violates a lower ranked constraint not violated by [bd] does not influence anything. On the other hand, from [pd] and [bt] both local optima are chosen with an *a priori* probability of 50%. Once chosen, the random walker moves always, both to [bd] and [pt], with a transition probability of 1. In turn, nothing guarantees that the random walker will prefer moving to [bd] to moving to [pt] from candidates [pd] and [bt]. In brief, both local optima are reached with equal probability and are left with equal probability—independently of the parameters of the simulation.

How could we break the symmetry of the search space just described, which results in the two local optima being found with equal probability? A first idea might be to increase the probability for the random walker to move from [pd] and [bt] to [bd], and to decrease the chance of moving to [pt]. The second idea will be to enlarge the search space in an asymmetric way, as will be demonstrated in the second model.

As both [bd] and [pt] are more harmonic than [pd] and [bt], it would con-

tradict the idea of simulated annealing not to have the transition probability $P(w \rightarrow w') = 1$, once a neighbour w' of $w = [\text{pd}]$ or $w = [\text{bt}]$ has been chosen. So, one may alter rather the *a priori* probabilities determining the choice of the neighbours. So far, each neighbour has been chosen with an equal probability, but this symmetry can be deformed *ad hoc*. One may also reconsider simulated annealing, and connect the horizontal structure of the landscape to the vertical one, although it is quite unclear to me how this could be done in the general case. This direction would involve taking the possible gain in the harmony function (the vertical structure) into consideration when determining the *a priori* chance to pick a neighbour (the horizontal structure): the more you can gain in harmony, the more it is probable that you will consider the possibility to move to this neighbour.¹²

In any case, experiments (for $p = 0.67$ and $p = 0.8$) have shown that this direction is still fruitless. If the *a priori* chance of picking $[\text{bd}]$ (as opposed to choosing $[\text{pt}]$) when the random walker is in $[\text{pd}]$ or $[\text{bt}]$ is increased, say, to $p = 2/3$, then the random walker will indeed prefer moving to $[\text{bd}]$. In the next step, however, leaving $[\text{bd}]$ has still the same probability as leaving $[\text{pt}]$. In general, if the probability of moving from $[\text{pd}]$ or $[\text{bt}]$ to $[\text{bd}]$ is p , and to $[\text{pt}]$ is q (with $p + q = 1$), then the simulation will return $[\text{bd}]$ with probability p and $[\text{pt}]$ with probability q —independently of the cooling schedule! Such a model can describe empirical data with a distribution different from 50% - 50%, and yet the interpretation of why p and q have some specific values would still be missing. Likewise missing is the interpretation explaining how and why these parameters of the horizontal structure are tuned by different speech situations.¹³

What would lead to success is a model in which the random walker is less likely to leave $[\text{bd}]$ than to leave $[\text{pt}]$ —at least, in some phase of the simulation. Once in $[\text{bd}]$, it is captured there, while leaving $[\text{pt}]$ is still possible. In order to end up in $[\text{pt}]$, the system has to choose to move always back to $[\text{pt}]$ —and never to $[\text{bd}]$ —each time the system has escaped from $[\text{pt}]$. The slower the cooling schedule, the more often such a decision has to be made. If $[\text{pt}]$ is chosen with probability q , then a cooling schedule offering n such decisions would return $[\text{pt}]$ with a probability of q^n , and $[\text{bd}]$ with a probability of $1 - q^n$. A slower cooling schedule means a higher n , resulting in a lower q^n with a higher $1 - q^n$. As discussed in section 2.3.2, however, it is unclear how the difference of two violation profiles could be defined in order to have the system escape from $[\text{pt}]$ more easily than from $[\text{bd}]$. The last resort is obviously the introduction of a new, highly ranked constraint satisfied by $[\text{bd}]$ and violated by the other three candidates, so that once temperature has dropped below this new constraint, escaping from $[\text{bd}]$ is not possible any more, but escaping from $[\text{pt}]$ still has some chance.

¹²In the present case, for instance, moving from $[\text{pd}]$ or $[\text{bt}]$ to $[\text{pt}]$ can be seen as an “improvement of one constraint level”, because the highest violation mark incurred is assigned by constraint `STOPASSIMILATION=REGRESSIVE[VOICE]` instead of `ASSIMILATE[VOICE]`. Similarly, moving from $[\text{pd}]$ or $[\text{bt}]$ to $[\text{bd}]$ is an “improvement of two constraint levels”, for $[\text{bd}]$ ’s highest violation mark originates from `FAITH[VOICE]`. The *a priori* chance to pick a neighbour with an improvement of two constraint levels may, in turn, be assigned double weight, as opposed to the chance of picking a neighbour with an improvement of only one constraint level.

¹³In general, how should we interpret the tuning of the probabilities related to the *horizontal* structure of the landscape? Nevertheless, by supposing that the horizontal structure may be slightly different for each individual (yet constant within a person), we can account for variations among speakers, dialects or sociolects.

To summarise, this model is analogous to the situations described in subsection 2.3.2 discussing the cases where SA-OT does not work. The present model with our four candidates is a variant of that basic situation. If traditional simulated annealing converges asymptotically, but SA-OT does not, should we conclude that simulated annealing has not been applied properly? The discussion in Chapter 3 aimed at demonstrating that peculiarities of SA-OT follow directly from the core of Optimality Theory. Thus, it is only to be hoped that the divergence between simulated annealing and SA-OT can be interpreted within linguistics and OT.

Consequently, I have a good and a bad bit of news: which do you want to hear first? The bad news is that this approach can only produce an equal distribution of the two forms, which is too strong a prediction. It is quite unlikely that empirical research would report on an exactly 50%-50% distribution. Stochastic approaches—and simulated annealing is one of them—aim at reproducing quantitative phenomena; why shall we content ourselves, then, with the qualitative result that both candidates can be reproduced? Just as in good-news-bad-news jokes, however, the good news will also resolve the bad news.

Now, the good news. Well, this sounds initially also as a bad news: we have to give up our expectations about the precision of SA-OT converging to 1, as simulated annealing is performed slower. And yet, this is a good piece of news. Optimality Theory Simulated Annealing is claimed to be a performance model on top of OT as a competence model (Table 2.1), and we know that performance is indeed always full of errors. Why do we actually expect it to be precise asymptotically, then?

Being even more radical, I suggest reformulating some basic ideas in linguistics. So far, phenomena independent of external factors (such as speech rate) were supposed to belong to competence, to the core of linguistic knowledge deeply encoded in the brain (or, at least, in the physiology of the speech production-perception system). However, many phenomena may steadily persist in language, even if they “contradict” the (static) mental representation of the given language, because they are necessarily introduced by the (dynamic) computational production process. In the present case, even though competence in its narrow sense would require regressive assimilation (hence, its model, the OT grammar, yields exclusively [bd] as optimal); and yet, the computational production process, modelled by SA-OT, cannot help but also return the [pt] form displaying progressive assimilation.

The fact that the ratio of the “erroneous” form is constant and does not depend on speech rate makes it impossible to argue *a priori* for a certain form to be the performance error. Earlier, namely, we could identify the form whose frequency increases in fast speech as the performance error, based on the assumption that fast speech cannot be more correct than normal speech. Our aim was to reproduce this behaviour using SA-OT. Now, however, it is only the model that turns a certain form into the grammatical form (by having it as the globally optimal candidate), and other forms as performance errors (local optima), and not pre-theoretical observations. The only hint was that the form with progressive assimilation seemed to be an exception from the general trend displaying only regressive assimilation.

I am convinced that models of the mental representation of languages could be kept simpler if many “ugly cases” were exiled to the production-computation process. The present example has shown us the way: without reformulating

Dutch phonology, we could reproduce the exceptional progressive assimilation in *o[pt]ie*.

Obviously, the question arises why the same pressures do not apply to *zakdoek* ('handkerchief') or to *duikboot* ('submarine'). This brings us back to the bad news: if we were able to modulate frequencies by tuning the parameters, we could simply argue that the unaccented frequent function words constituting *op die* are produced much more quickly—that is, with a different parameter setting—than relatively infrequent nouns such as *zakdoek* and *duikboot*. This is why we have to confront the second model, whose parameters will influence again the output frequencies.

The last good piece of news then is that the second model and the mathematical challenges posed by its formal analysis (which can be safely skipped by the reader less interested in math) will turn out to be illuminating about the techniques offered by SA-OT.

6.5 Model 2: Infinite search space

6.5.1 Enlarging the search space

The second model involves enriching the search space with new candidates, and in this way breaking its symmetry. The candidate set becomes huge—actually infinite. The four candidates of the previous model (Fig. 6.2) form but the central zone of the new search space (already advanced in Fig. 6.3). As the periphery of the latter does not exhibit the same symmetry as the centre, the two local optima may be returned with probabilities different from 50% each. The more use the system makes of the periphery, the more significant the difference from the 50%-50% distribution will be.

Importantly, the periphery will be less optimal than the central valley, and therefore we can get farther in the periphery only in the *first phase* of the simulation. This is, when temperature is still higher than the highest ranked constraint. In other words, to distance the system from the 50%-each distribution, we have to allow many iterations in the first phase. Hence the novelty of this model: unlike in the different uses of SA-OT so far, all of which included but a finite search space, the parameter K_{max} is assigned now a leading role.

Parameter K_{max} is starring in the present model also for a second reason, which is similarly related to the fact that the candidate set is infinite. Due to this fact, we cannot launch the simulation from each of the candidates with equal probability, as we have done before. One option would be to define a probability distribution on the candidate set; but we leave this option open to future research, and we rather launch the simulation always from one of the four candidates in the central basin. This is why K_{max} will determine how far from the central basin the random walker can get, and thereby, how much of the asymmetry of the search space's external regions we can make use of.

After this introduction, the question is raised: how can we enrich the search space? A straightforward direction, copying the classical paradigms in OT, is to allow epenthesis: let us insert an epenthetic vowel (a schwa) between the two consonants. Indeed, vowel epenthesis is frequently employed by natural languages to break up unwanted consonant sequences, even if not necessarily to

resolve clashes in voice.¹⁴

The possibility of inserting only one schwa has not proven to be fruitful. Inserting any number of schwas recursively is more interesting (Fig. 6.3 depicting the structured candidate set is repeated here as Fig. 6.5). This is so even if forms with more than one epenthetical vowel are—most probably—not attested in any language.

This paradox, namely the fact that forms not attested in natural languages render the model fruitful, is worth emphasising here. Following B     and Gervain (2006), we could call this phenomenon the “*Bald Soprano*” effect, or even the *Godot effect*: there are characters in the play who never appear openly on the scene, and yet, they influence importantly the whole story line. In fact, the present study refutes a possible criticism to Optimality Theory in general: why should a model include an infinite set of candidates, if not for the sake of simplicity and of mathematical beauty? OT’s main goal is to account for linguistic typology and typologies include only a very restricted number of types, whence one would expect a very restricted finite candidate set. Do the candidates that can never win (the *losers* according to Samek-Lodovici and Prince, 1999) play any role in Optimality Theory at all? We shall see presently that they do, at least in SA-OT.

6.5.2 The landscape

After such a long introduction, let us enter the linguistic details of the new model. The infinite candidate set and its topology have already been defined in section 6.3, so we turn our attention to the constraints.

Now the constraint STOPASSIMILATION=REGRESSIVE in its first formulation will play a role. Recall tableau (6.5) in section 6.3, repeated here below. All forms with an epenthesis violate DEP (as many times as the number of epenthetical vowels included), and satisfy ASSIMILATE[VOICE]. The third most important constraint is STOPASSIMILATION=REGRESSIVE, which is satisfied by [b@+d] (a positive number of epenthetical vowels surrounded by [b] and [d]) and, crucially, violated by the other candidates with epenthesis.

¹⁴See footnote 1 on page 161. For a concrete example of schwa insertion, consider Modern Hebrew. (Notice that the word “schwa” originates from the concept of *schwa mobile* coined by the Biblical Hebrew grammarians.) A clash occurs when the past tense singular 2nd masculine suffix [-ta] is added to a verb ending in a [d], such as *lamad* ‘learn, study’. In such cases, two forms may emerge: the first one involves regressive assimilation ([lamatta]), while the second one inserts an epenthetic schwa ([lamad@ta]). In fact, this case serves as an example for the prohibition of homorganic consonant clusters in general [Schwarzwald (2001, pp. 11-12.); for further examples, see B     and Hamp (2002)]. Still, the behaviour of Modern Hebrew with respect to homorganic consonant clusters can be only described by using a candidate set that includes forms with epenthetical vowels, as well as by constraint DEP overruled by a markedness constraint *[ PLACE][ PLACE]. In sum, Modern Hebrew—among, most probably, a huge number of further languages—does support the need to include the new candidates into the candidate set. If one requires such a support at all, as most phonologists view GEN as a black box generating literally “everything”.

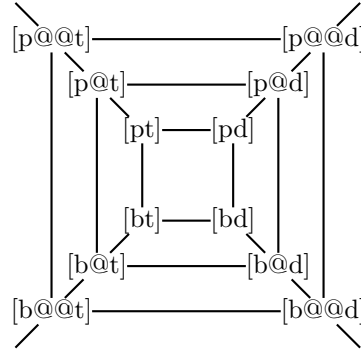


Figure 6.5: The second search space used for *op die*. The character @ stands for the epenthetic schwas.

/pd/	DEP	ASSIM[VOICE]	STASS=RGR[VC]	FAITH[VOICE]
[bd]				*
~ [pt]			*!	*
[pd]		*!	*	
[bt]		*!	*	**
[b@d]	*!			*
[p@t]	*!		*	*
[p@d]	*!		*	
[b@t]	*!		*	**
...
[b@ ⁿ d]	* ⁿ !			*
[p@ ⁿ t]	* ⁿ !		*	*
[p@ ⁿ d]	* ⁿ !		*	
[b@ ⁿ t]	* ⁿ !		*	**
...

(6.6)

Consequently, the landscape looks as follows: in the “middle” we find a *central basin*, formed by the four candidates without epenthesis and having the form already discussed (in section 6.4, and especially in Fig. 6.4), which is in turn surrounded by ever rising hills (the higher, the less optimal). This picture is the result of promoting DEP, the constraint penalising recursion, to the highest position. Furthermore, this “radial” structure of the landscape is modulated by a “tangential” structure, to be presented soon in Fig. 6.6. In each concentric circle outside the central basin, [b@ⁿd] ($n > 0$) is lower (more harmonic) than [p@ⁿd], [p@ⁿt] and [b@ⁿt], due to constraint STOPASSIMILATION=REGRESSIVE. This search space can thus be visualised as a circular crater with a smaller radial valley formed by a river that runs down in a centripetal direction towards the central basin.

Our goal has been exactly to create this channel [b@ⁿd], and this is why we require the first definition of constraint STOPASSIMILATION=REGRESSIVE. Imagine the water falling on such a landscape, which sooner or later reaches some deepest valley in the landscape by flowing down the slope. The valley

collects most of the water and streams it to the basin in the form of a river or channel. Even though an initial rain has spread the water, say, uniformly in a larger region, still water will concentrate more and more in the river, and later in the central basin, as time passes. Remembering this metaphor might help better understand the behaviour of our SA-OT system.

In the first, unhindered stage of the simulation, the freely roaming random walker may be found, more or less, everywhere in the landscape. The likelihood $P_0(w)$ of the random walker being at a certain point w of the search space by the end of this first phase (we will be calculated exactly later) resembles the quantity of water in w after the initial rain. The dispersion is not necessarily even (that is, $P_0(w_1) = P_0(w_2)$ does not necessarily hold for any w_1 and w_2), but “smooth”. Additionally, the total amount of water is unity, corresponding to the fact that $\sum_{w \in \text{Gen}(UR)} P_0(w) = 1$ must hold.

Now the water starts flowing; that is, the probability distribution $P_t(w)$ of the random walker being in w changes as time t advances in each time step of the simulation. Obviously, the total amount of water ($\sum_{w \in \text{Gen}(UR)} P_t(w)$) remains the same over time. As temperature reaches the domain of the highest ranked constraint, DEP, not all moves are equally likely anymore. In particular, centrifugal moves increasing the number of the epenthetical vowels become blocked in this stage. Once moving upwards in the landscape becomes difficult for the random walker, the water—the probability $P_t(w)$ —will be collected and streamed to the central basin by the structure of the landscape, and especially by channel $[b@^+d]$. By the end, $P_\infty(w)$ (the “amount of water collected in w ”) gives you the probability of the algorithm returning candidate w : usually 0, unless w is a local optimum.

How does channel $[b@^+d]$ work? Suppose the random walker is “out in the hills”, that is, not in the central basin, when temperature drops to the domain of the constraint STOPASSIMILATION=REGRESSIVE. At this moment, some tangential moves—moves changing the [voice] feature of the stops—are not free anymore either: the transition probability of stepping from $[b@^nd]$ to either $[p@^nd]$ or $[b@^nt]$ becomes less than 1—and this probability quickly diminishes to zero—because such steps would require incurring a violation mark by this constraint. In turn, $[b@^nd]$ serves as a trap for the tangential component of the random walk. The “water” is collected by channel $[b@^+d]$ during the tangential steps, and the channelled water has no other option but to flow towards the central basin through a series of centripetal steps (deleting the epenthetic vowels, but not altering the voiced feature of the consonants).

Now, the clue to this model is the fact that this channel enters the central basin at $[bd]$. This is crucial, since all the “water” (probability of the random walker being there) channelled by the river or channel will be stuck in $[bd]$, cannot end up in $[pt]$, for $[bd]$ is a local optimum. The *channelling effect* starts when temperature falls to the domain of STOPASSIMILATION=REGRESSIVE. At this stage of the simulation, escaping from the two local optima is not possible anymore, because escaping would require incurring a violation mark by ASSIMILATE[VOICE], which is higher than the actual temperature.

On the other hand, the water that has not been collected by the channel may end up in $[pt]$, which is also a trap, a local optimum, due to tableau (6.6). The water reaching the basin in $[pt]$ from $[p@t]$ gets caught there; whereas the water arriving into $[pd]$ and $[bt]$ (from $[p@d]$ and $[b@t]$) is equally divided between

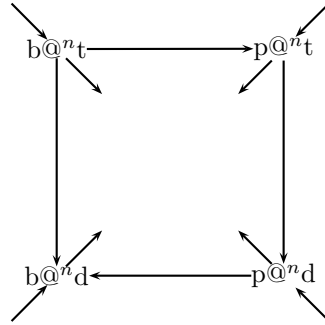


Figure 6.6: “**Channelling**” effect in the infinite search space, in the n th ($n > 0$) concentric layer, which is formed by the candidates with n epenthetical vowels. The arrows point to the more harmonic candidate, based on tableau (6.5). If temperature is below the lowest ranked constraint, the random walker can move exclusively towards the centre, or towards candidate $b@^n_d$, the most harmonic candidate among the candidates with n epenthetical vowels. In this sense, we can speak of the $b@^n_d$ valler or channel.

the two local optima, as explained in the context of the first model.

6.5.3 Tuning the output of the model

Consequently, moving away from the 50%-50% distribution between the two local optima of the four-candidate model is possible by channelling as much “water” as possible into channel $[b@^+d]$, and thereby increasing the probability of our simulation returning $[bd]$. Notice that decreasing it is not feasible in our model. Nonetheless, it is more likely that accounting for empirical data would require increasing the probability of $[bd]$ rather than decreasing it.¹⁵

What we need to do then is to disperse initially the constant amount of water on a region as large as possible, by providing the walker the possibility to move much without any obstacles (i.e., a long initial phase in the simulation). Why is this so? In short, once the temperature has reached the domain of DEP, centrifugal moves become prohibited, and a competition starts between centripetal and tangential moves.

The competition is about the water reaching the channel first or the basin first. If the channel is reached, $[bd]$ will certainly be the output, otherwise $[bd]$ and $[pt]$ have equal chance. The more “water” reaches the channel, the higher the likelihood of $[bd]$. Additionally, observe that from a larger distance, more centripetal steps are required to arrive at the central basin, which increases the chance to reach the channel first by performing a few tangential steps.

In sum, the farther the random walker is from the central basin at the end of the unhindered phase, the more likely it is for $[bd]$ to be returned by the algorithm. This is the technique we can have our SA-OT model returning the two outputs with different probabilities.

¹⁵Decreasing the probability of $[bd]$ is possible by using constraints that define a very similar landscape but with a channel $[p@^+t]$. Then, the more water that is channelled, the higher the frequency of output $[pt]$. Observe that such a model is possible even if $[bd]$ is the global optimum, and not $[pt]$. In other words, here we see an example of an SA-OT model with the global optimum being returned in less than or equal to half of the cases.

At this point an additional issue arises, the choice of the initial candidate. We must remember that the search space is infinite, unlike in our previous models. So far, we could choose each candidate with equal probability to be the starting point of the random walk, but now, we have to come up with a different solution. For instance, a Gaussian-style distribution could be defined so that the likelihood of a candidate w_0 with n epenthetical vowels ($[C_1 @^n C_2]$) being the initial candidate of the walk be proportional to $e^{-n^2/2\sigma^2}$. Then, a larger σ will disperse the “rain” over a wider region, resulting in a higher frequency of output [bd], due to the channelling effect.

Instead of introducing an additional parameter σ to the model, however, we rather leave this idea to future research, and prefer exploiting the already existing parameters of the SA-OT Algorithm. Probably the most natural choice is to employ exclusively the four basic candidates of the central basin ([pd], [pt], [bd] and [bt]) as initial candidates, with 25% chance each. In turn, having “the initial rain covering a wide region” corresponds to allowing the random walker to get really far away from the initial candidate in the first, unhindered phase of the simulation. Using the water metaphor, the water is poured in the four central candidates, before it splashes to the initially unhindering mountains.

Then, a last point remains to be clarified in our train of thought: the way we lengthen the initial phase of the simulated annealing, that is, the phase in which the random walker moves freely even to worse neighbours. From the parameters of the algorithm (see Fig. 2.8 on page 64), two are the most straightforward candidates: K_{max} and T_{step} . In other words, we either add extra upper domains that the temperature has to traverse before reaching the domain of the highest ranked constraint (increase K_{max}); or increase the number of steps to be performed within each domain in order to have more steps also in the domain(s) superior to the domain of the highest ranked constraint. The second strategy can be realised in the simplest way by decreasing T_{step} , and this is the technique we have used the most often so far. So, should we increase K_{max} or decrease T_{step} , if we would like to have more iterations in the first, unhindered phase of the simulation?

In contrast to our previous models, it turns out that simply decreasing T_{step} does not work now.¹⁶ The reason, in short, is that increasing the number of steps within one domain will also increase the number of steps while temperature is between the domains of DEP and STOPASSIMILATION=REGRESSIVE. The reason, in detail, requires some mathematical discussion, which can be skipped without losing the general train of thought of my dissertation.

6.5.4 The interaction of K_{max} with T_{step}

The present subsection aims at presenting a formal analysis of how the parameters of the model influence the probabilities of returning [pt] and [bd]. First,

¹⁶Note this major difference between the present case and stress assignment in fast speech. For stress assignment, increasing K_{max} alone would not have any effect: the candidate set is finite, and not only can the random walker rove around the whole search space in the initial phase of the simulation, but also each point has an equal chance to be the starting point of the random walker. The phenomenon arises from changing the number of steps in the *second* phase of the simulation, that is, when temperature has already reached the domains of the constraints. In the present case, however, the search space is infinite, we start the simulation from a small subset of it (the four central candidates), and the goal is to have the random walker also visit candidates as remote as possible.

let us introduce a few notations:

$$\tau = \left\lfloor \frac{T_{max} - T_{min}}{T_{step}} \right\rfloor + 1 \approx \frac{T_{max} - T_{min}}{T_{step}} = \frac{3}{T_{step}} \quad (6.7)$$

$$k = K_{max} - K_{highest} = K_{max} - 3 \quad (6.8)$$

where $K_{highest}$ is the index associated with the highest ranked constraint, 3 in the present case. Further, τ stands for the number of repetitions performed by the inner loop of the SA-OT algorithm, that is, the number of iterations while temperature decreases one domain. Here, we employ our standard values, $T_{max} = 3$ and $T_{min} = 0$, and $\lfloor x \rfloor$ represents the integer part of x .

K_{max} is located k domains above constraint DEP, so temperature traverses k domains in the first phase of the simulation. In the period when temperature is exactly in the domain of DEP, centrifugal moves become banned only gradually: in the beginning of this period, they are almost free, and later almost impossible. Let us approximate this gradual effect by supposing that the random walker can freely move away from the centre as long as the temperature crosses the first $k + 0.5$ domains, and this direction becomes maximally prohibited immediately afterwards, from that point onwards when temperature enters the lower part of the domain of DEP. Consequently, the number of steps performed by the random walker in the first (unhindered) phase of the simulation is:

$$N = (k + 0.5) \cdot \tau \quad (6.9)$$

Remember that a candidate $[C_1 @^n C_2]$ (with $n > 0$) has four neighbours, two in a tangential direction (that is, also including n epenthetical vowels: $[C_1' @^n C_2]$ and $[C_1 @^n C_2']$), and two in a radial direction ($[C_1 @^{n+1} C_2]$ and $[C_1 @^{n-1} C_2]$). As each neighbour has an equal *a priori* probability of 0.25, the number of radial steps among these first N steps can be approximated by

$$N_{radial} = N_r \approx \frac{N}{2} = (k + 0.5) \cdot \frac{\tau}{2} \quad (6.10)$$

Estimating $\pi_N(n)$

Now, we calculate the probability $\pi_N(n)$ of being exactly at a distance n from the central valley by the end of the first phase, that is, of starting the “competition” from some candidate with exactly n epenthetical vowels ($[C_1 @^n C_2]$). The radial component of this first phase is a one-dimensional *Brownian motion* with equal probability of moving in both directions (centripetal and centrifugal). One flips a symmetrical coin N_r times, with head corresponding to the insertion of a @, and tail to the deletion of a @. Ending up with n epenthetical vowels requires exactly $\frac{N_r + n}{2}$ heads and $\frac{N_r - n}{2}$ tails, supposing that N_r and n have the same parity. Consequently, $\pi_N(n)$ can be approximated with a binomial distribution.¹⁷

¹⁷Observe that in each prefix of the insertion-deletion (head-tail) sequence, the number of deletions must not exceed the number of insertions, as we launch our algorithm from a candidate with no epenthetical vowel. Yet, we can overcome this problem by employing a trick. Observe that when the number of epenthetical vowels is zero, we still may flip our coin, but both head and tail should correspond to insertion, with deletion having zero probability. So, if flipping the coin returns then tail, we reverse the roles of heads and tails: from now on

$$\pi_N(n) = \begin{cases} 0 & \text{if } n \not\equiv N_r \pmod{2} \\ \binom{N_r}{N_r/2} \cdot 0.5^{N_r} & \text{else if } n = 0 \\ 2 \binom{N_r}{(N_r+n)/2} \cdot 0.5^{N_r} & \text{else} \end{cases} \quad (6.11)$$

Using the basic properties of the binomial coefficients, one can quickly check that $\sum_n \pi_N(n) = 1$. What we shall need is actually the sum of $\pi_N(n)$ over a large range of its argument n in function of N (i.e., of N_r), so we can render our life simpler by “smoothing” $\pi_N(n)$ (dividing the probabilities among $\pi_N(2k)$ and $\pi_N(2k \pm 1)$):

$$\begin{aligned} \pi_N(n) &= \begin{cases} \binom{N_r}{(N_r+n)/2} \cdot 0.5^{N_r} & \text{if } n \equiv N_r \pmod{2} \\ \binom{N_r}{(N_r+n+1)/2} \cdot 0.5^{N_r} & \text{if } n \not\equiv N_r \pmod{2} \end{cases} \\ &\approx \frac{1}{\sqrt{2\pi}} \frac{2}{\sqrt{N_r}} e^{-\frac{n^2}{2N_r}} \end{aligned} \quad (6.12)$$

Here, we have employed the well-known fact that a binomial distribution ($p = q = 0.5$ in our case) can be approximated with a normal distribution, namely¹⁸

$$\binom{n}{k} p^k q^{n-k} \approx \frac{1}{\sqrt{npq}} \varphi\left(\frac{k - np}{\sqrt{npq}}\right) \quad (6.13)$$

for large n , where $\varphi(x)$ is the standard normal distribution:

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \quad (6.14)$$

Equation (6.12) makes clear that $\pi_N(n)$, the position of the random walker by the end of the first phase of the simulation, follows approximately the positive part of a Gaussian distribution, centred around the origin, with a standard deviation¹⁹

head will correspond to deletion and tail to insertion. Therefore, each of the 2^{N_r} different head-tail series can be made a legitimate one, and each of them has an equal probability.

This more complicated interpretation of heads and tails can be visualised if head is still seen as always moving one unit to the positive direction of the scale, and tail as moving one to the negative one; but we allow moving to both directions of the origin, with both positions $+n$ and $-n$ corresponding to n insertions in the candidate string. Indeed, once the coin returns tail when no deletion is possible, we move from 0 to -1 on the scale, which corresponds to an insertion (-1 also meaning one insertion) together with the reversion of the roles of heads and tails from the viewpoint of the number of epenthetical vowels (e.g. a second tail would bring to -2 , i.e. to a second insertion).

As arriving at both $-n$ and $+n$ corresponds to n insertions, $\pi_N(n)$ has to be multiplied by 2 if $n \neq 0$, as compared to the standard binomial distribution. In short, the two halves of the scale are folded around the origin.

Note finally that this train of thought would work correctly only if the *a priori* probabilities corresponding to the elements of the central valley had been defined in a slightly different way. Namely, by assigning a 50% chance to insertion, and 25% chance to changing the [voice] feature of one of the stops, similarly to the *a priori* probabilities of the other candidates. Now that each of the three neighbours has an equal probability of 1/3, the following formula is but an approximation.

¹⁸See e.g. <http://mathworld.wolfram.com/NormalDistribution.html>.

¹⁹It would have sufficed to refer to the fact that in a *Brownian motion* the expected value of the squared displacement is proportional to the number of steps performed. Let us take a


$$\sigma = \sqrt{N_r} = \sqrt{(k + 0.5) \cdot \frac{\tau}{2}} \quad (6.15)$$

Estimating n_0

In what follows, we estimate the distance n_0 beyond which channelling is expected to take place. If the random walker has reached this distance by the end of the first phase, then it will most probably end up in candidate [bd]; otherwise, it has an equal chance to return us [pt] or [bd]. Thus, once n_0 has been estimated, we predict candidate [pt] to be returned with probability

$$\mathcal{P}([\text{pt}]) = \frac{1}{2} \sum_{i=0}^{n_0} \pi_N(i) \approx \int_0^{n_0} \frac{1}{\sqrt{N_r 2\pi}} e^{-\frac{t^2}{2N_r}} dt \quad (6.16)$$

Let us repeat here tableau (6.6):

/pd/	DEP	ASSIM[VOICE]	STASS=RGR[VC]	FAITH[VOICE]
 [bd]				*
~ [pt]			*!	*
[pd]		*!	*	
[bt]		*!	*	**
[b@d]	*!			*
[p@t]	*!		*	*
[p@d]	*!		*	
[b@t]	*!		*	**
...
[b@ ⁿ d]	* ⁿ !			*
[p@ ⁿ t]	* ⁿ !		*	*
[p@ ⁿ d]	* ⁿ !		*	
[b@ ⁿ t]	* ⁿ !		*	**
...

(6.17)

Imagine that temperature is just crossing the domain of ASSIMILATE[VOICE], and the random walker is located somewhere in the epenthetical hills. Say, at $[C_1@^nC_2]$. The four neighbours ($[C'_1@^nC_2]$, $[C_1@^nC'_2]$, $[C_1@^{n+1}C_2]$ and $[C_1@^{n-1}C_2]$) are chosen with equal probability. The centrifugal move (inserting an extra @) is prohibited for $T \ll \text{DEP}$, whereas choosing the centripetal step (deleting one @, which is always possible) results in bringing you towards the central basin with transition probability 1. The two other neighbours are chosen *a priori* with 0.5 chance, and moving in this tangential direction is still fully free. Since temperature is high, channelling does not occur. A race with time starts: the centripetal steps performed in the approximately 1/4 of the

one-dimensional random walk (*Brownian motion*) starting from $x_0 = 0$, with p and q being the probabilities of stepping one unit to the right ($x_{i+1} = x_i + 1$) and to the left ($x_{i+1} = x_i - 1$) respectively. The expected value of the location of the walker after N steps is $\overline{x_N} = N(p - q)$, whereas the dispersion is $(x_N - \overline{x_N})^2 = 4Npq$. See for instance Hubbey (1999, p. 229) and references therein, or <http://scienceworld.wolfram.com/physics/BrownianMotion.html> and references there. A very creative derivation is found in Reif (1965, pp. 13-16). In our case, $p = q = 0.5$.

iterations should not bring you back to the central basin until temperature has reached constraint $\text{STOPASSIMILATION}=\text{REGRESSIVE}$ so that channelling can be effective.

As an approximation, let us say that there are 2τ iterations—while temperature drops from the middle of the domain of DEP to the middle of the domain of $\text{STASS}=\text{RGR}[\text{VC}]$ —during which centrifugal moves are prohibited, but tangential and centripetal moves are free. On average, a quarter of these time steps are used to bring us closer to the central basin. This is why the random walker has to reach a distance larger than $n_1 = \tau/2$ by the end of the first phase, if it should not be probable for the random walker to reach the central basin until channelling is effective (until temperature reaches constraint $\text{STASS}=\text{RGR}[\text{VC}]$).

Once the $[\text{b}@\text{d}]$ channel becomes visible, a few more steps are still needed for the random walker to reach it, and not the central basin. The expected number of tangential steps for the random walker to reach the channel is $\kappa = 2.5$, which is slightly altered whenever constraint $\text{FAITH}[\text{VOICE}]$ becomes also active.²⁰

While the system performs κ tangential steps, it also tries—on average— $\frac{\kappa}{2}$ centrifugal steps (in vain), and performs $\frac{\kappa}{2}$ centripetal steps. In other words, if the random walker has been not farther than $n_2 = \frac{\kappa}{2}$ from the central valley, there is a chance of reaching the central valley at a random point (that is, yielding outputs $[\text{bd}]$ and $[\text{pt}]$ with equal chance), and not through channelling. If, however, the random walker has been farther away, the random walker will have reached the $[\text{b}@\text{d}]$ channel, before entering the central valley in $[\text{bd}]$ *due to* the channelling effect.

In sum, at the end of the first phase the random walker has to reach at least a distance of

$$n_0 = n_1 + n_2 = \frac{\tau}{2} + \frac{\kappa}{2} + 1 \quad (6.19)$$

for the channelling effect to take place significantly (remember $\tau = \frac{3}{T_{\text{step}}}$ and $\kappa = 2.5$). Even after having deleted n_1 epenthetical vowels while $T \approx \text{ASSIM}[\text{VOICE}]$, and having lost subsequently n_2 epenthetical vowels while trying to reach the already visible channel, there must be at least one $@$ left.

²⁰Suppose that temperature is such that constraint $\text{STASS}=\text{RGR}[\text{VC}]$ already prohibits leaving $[\text{b}@\text{d}]$ (the channel acts as a trap), but constraint $\text{FAITH}[\text{VOICE}]$ is not yet active to block some of the other tangential moves.

Let us focus now on the tangential component of the moves, by projecting the search space onto a circle of four candidates, $[\text{bd}]$, $[\text{pd}]$, $[\text{pt}]$ and $[\text{bt}]$ (or $[\text{b}@\text{d}]$, $[\text{p}@\text{d}]$, $[\text{p}@\text{t}]$ and $[\text{b}@\text{t}]$). Supposing that the random walk in this small space is free, but $[\text{bd}]$ is a trap, how many steps are required on average for the walker to get stuck in $[\text{bd}]$?

Let k_w be the expected number of tangential steps that is required to reach $[\text{bd}]$ from candidate w . From $[\text{b}@\text{t}]$ we either move to $[\text{b}@\text{d}]$ (1 step required to reach the channel; with probability 0.5), or we move to $[\text{p}@\text{t}]$ ($1 + k_{[\text{pt}]}$ steps required to reach the channel; with probability 0.5). Similarly for the other candidates, which yields the following equations:

$$\begin{aligned} k_{[\text{bd}]} &= 0 \\ k_{[\text{bt}]} &= 0.5 \cdot 1 + 0.5 \cdot (1 + k_{[\text{pt}]}) \\ k_{[\text{pd}]} &= 0.5 \cdot 1 + 0.5 \cdot (1 + k_{[\text{pt}]}) \\ k_{[\text{pt}]} &= 0.5 \cdot (1 + k_{[\text{pd}]} + 0.5 \cdot (1 + k_{[\text{bt}]}) \end{aligned} \quad (6.18)$$

By solving these equations, we obtain $k_{[\text{bd}]} = 0$, $k_{[\text{bt}]} = 3$, $k_{[\text{pd}]} = 3$ and $k_{[\text{pt}]} = 4$. This is why $\kappa = 0.25k_{[\text{bd}]} + 0.25k_{[\text{bt}]} + 0.25k_{[\text{pd}]} + 0.25k_{[\text{pt}]} = 2.5$.

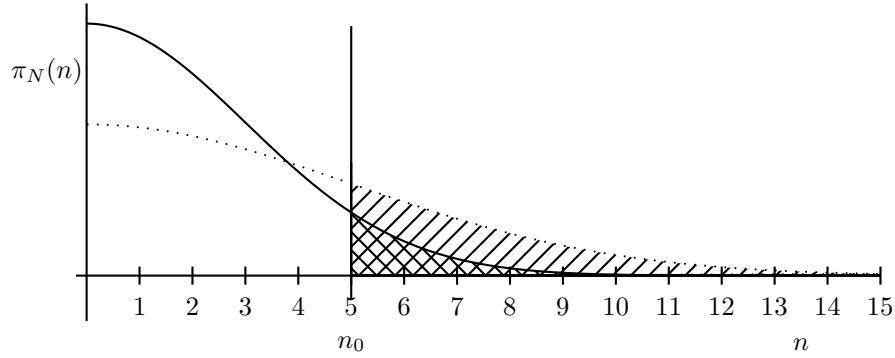


Figure 6.7: **Distribution** $\pi_N(n)$ of the random walker's position at the end of the first phase. The standard deviation of a distribution is $\sqrt{N_r}$. Here, the dotted distribution corresponds to a larger number of radial steps N_r (to a larger N) than the solid one. For a given n_0 , the chance of the random walker getting to a distance of at least n_0 increases as N_r grows larger.

If the random walker has reached this distance by the end of the first phase, the output will be most probably [bd]. If the random walker has not reached this distance by the end of the first phase, both [bd] and [pt] have equal chance to be returned. In brief, equation (6.19) defines the n_0 to be used in equation (6.16), which we repeat here:

$$\mathcal{P}([\text{pt}]) = \frac{1}{2} \sum_{i=0}^{n_0} \pi_N(i) \approx \int_0^{n_0} \frac{1}{\sqrt{N_r 2\pi}} e^{-\frac{t^2}{2N_r}} dt \quad (6.20)$$

Understanding the role of the parameters

Figure 6.7 helps us summarising what we have so far. We are interested in the impact of two parameters, namely K_{max} (or k , see equation (6.8)) and T_{step} (or τ , see equation (6.7)), on the output frequencies estimated by equation (6.20). This estimation includes two derived parameters, N_r and n_0 . According to equation (6.10), N_r depends on both K_{max} and T_{step} ; whereas n_0 depends exclusively on T_{step} by equation (6.19).

Thus, let us first fix T_{step} (hence, n_0), and consider the influence of K_{max} on the outputs. As Fig. 6.7 illustrates, a larger K_{max} (a larger k , a larger N_r) increases the chance of the random walker finishing up beyond the fixed n_0 (the curve has a thicker tail), thereby decreasing the probability of returning [pt] by equation (6.20). Experiments performed will support this prediction in the next subsection.

What happens if K_{max} is fixed and T_{step} varies? Our experience in earlier chapters has been that a larger T_{step} increases the probability of returning the suboptimal alternating form, [pt] in the present case. Will the same happen now, as well?

Let us transform equation (6.20) into the integral of the standard normal distribution by employing a replacement $u = t/\sqrt{N_r}$:

$$\mathcal{P}([\text{pt}]) \approx \int_0^{n_0} \frac{1}{\sqrt{N_r} 2\pi} e^{-\frac{t^2}{2N_r}} dt = \int_0^{n_0/\sqrt{N_r}} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} du = \Phi\left(\frac{n_0}{\sqrt{N_r}}\right) \quad (6.21)$$

It becomes clear that the frequency of [bd] increases, that is, the frequency of [pt] decreases, if the argument of Φ —the integral of the standard normal distribution, a monotone increasing function—decreases; that is, if n_0 decreases and N_r increases. Increasing K_{max} and keeping T_{step} fixed is increasing N_r with n_0 kept unchanged. The influence of varying T_{step} (with K_{max} being constant), on the other hand, depends on the influence of T_{step} on the argument G of Φ :

$$G := \frac{n_0}{\sqrt{N_r}} = \sqrt{\frac{\tau}{2(k+0.5)}} + \sqrt{\frac{(\kappa+2)^2}{2\tau(k+0.5)}} \quad (6.22)$$

For small τ values, the second addend dominates, whereas for large τ , the first one does. As we increase τ (decrease T_{step}), the value of G will first decrease, and then, as the first addend turns dominant, G grows larger again. Decreasing G corresponds to decreasing the frequency of [pt] by equation (6.21), and increasing G brings the frequency of [pt] closer to 0.5.

By employing the fact that the geometrical mean is always less or equal to the arithmetic mean, we obtain ($\kappa = 2.5$):²¹

$$G \geq \sqrt{2 \frac{\kappa+2}{k+0.5}} = \frac{3}{\sqrt{k+0.5}} \quad (6.23)$$

and G is minimal iff $\tau = \kappa + 2 = 4.5$. That is, iff $T_{step} \approx \frac{3}{\kappa+2} = \frac{2}{3}$.

Notice, however, that by its definition, τ must be an integer (the number of steps in a domain), so in the case of our standard T_{max} and T_{min} values, we expect the turning point to be around $1 > T_{step} \geq 0.75$ (corresponding to $\tau = 4$). It will turn out that on the other side of the turning point—for T_{step} values corresponding to $\tau = 5$ ($0.75 > T_{step} \geq 0.6$)— G grows faster, so these parameters will produce more [pt] outputs than parameters corresponding to $\tau = 4$.

Another prediction is that for $T_{step} \ll 1$, when the second addend in (6.22) becomes negligible, different parameter settings will produce the same frequencies if $\frac{\tau}{k+0.5}$ (that is, if $T_{step} \cdot (k+0.5)$) is kept constant.

After such a long mathematical discussion, let us probe the pudding now!

6.5.5 Experiments

The results of a few experiments are summarised in Tables 6.1 and 6.2, as well as in Fig. 6.8. In each of the cases, one of the two parameters K_{max} and T_{step} is kept constant, while the other varies. For each parameter setting, an experiment consisted of running 100 000 simulations, that is launching the simulation 25 000 times from each of the four central candidates, and of calculating the frequencies of the outputs. By repeating this experiment two more times, we could also determine the mean and the $\sigma(n-1)$ error of the measured frequency. Finally, these tables also show the estimated frequencies based on equation (6.21).

²¹ $a + b \geq 2\sqrt{ab}$. Furthermore, $a + b = 2\sqrt{ab}$ if and only if $a = b$. For our purpose, take the two addends in (6.22) as a and b . This trick saves us calculating $\frac{\partial G}{\partial \tau}$.

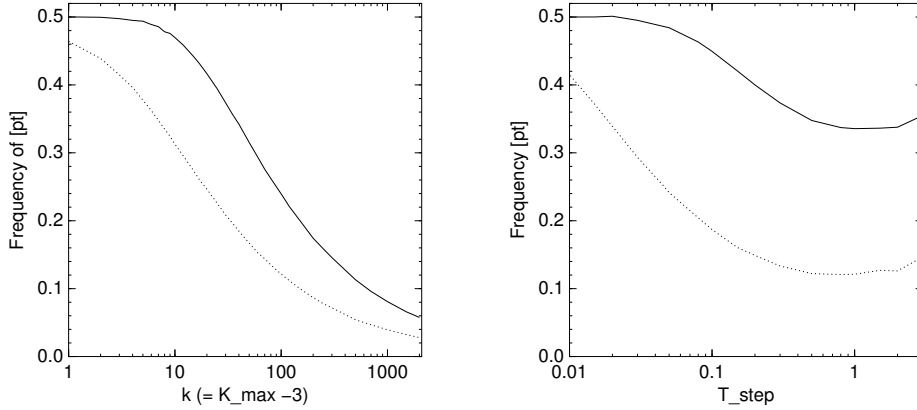


Figure 6.8: **Frequency of [pt] when varying either K_{max} or T_{step} .** The parameter varied is represented on a logarithmic scale. **Left box:** K_{max} changes, while $T_{step} = 0.05$ (solid line) or $T_{step} = 0.5$ (dotted line). Compare it to Table 6.1. **Right box:** T_{step} changes, while $K_{max} = 10$ (solid line) or $K_{max} = 100$ (dotted line). The frequencies are the same as those in Table 6.2.

The frequencies represented on the left panel of Figure 6.8, that is, in Table 6.1, confirm our prediction that the frequency of [pt] decreases as K_{max} (or k) grows larger, both for $T_{step} = 0.5$ and for $T_{step} = 0.05$. The curve corresponding to $T_{step} = 0.05$ runs higher than the one corresponding to $T_{step} = 0.5$. Not surprisingly, since the right panel of Figure 6.8 (that is, Table 6.2) demonstrates that the frequency of [pt] grows as T_{step} diminishes, supposing that $T_{step} < 0.8$. The turning point predicted by equation (6.23) can also be observed around $T_{step} \approx 1$, but we return to this point soon.

Subsequently, Fig. 6.9 presents the two dimensional *phase space*, that is, the behaviour of the system in function of both parameters. The radii of the circles are proportional to the difference of the frequencies of the two outputs. That is to say, the dots in the lower left corner correspond to the system returning [bd] and [pt] with (practically speaking) the same probability, whereas the large circles in the upper right corner visualise how [bd] becomes dominant. The largest circle is at $k = 27$ and $T_{step} = 0.8$, where the probability of [bd] reaches 79%.

One can also observe in Fig. 6.9 that circles of the same size are located, roughly speaking, on a diagonal straight line. As the figure uses logarithmic scales on both axes, such a diagonal straight line corresponds to a hyperbola on linear scales, and confirms our earlier prediction that for small T_{step} values keeping $T_{step} \cdot (k + 0.5)$ constant yields similar output frequencies.

In order to verify this observation in a more precise way, Table 6.3 presents a few parameter combinations that have been proven to yield [pt] with a chance of 0.25—that is, when channelling is effective in exactly half of the runs. This case corresponds to $G \approx 0.67$ by equation (6.21). If $T_{step} \ll 1$, the second addend in equation (6.22) becomes negligible:

$K_{max} - 3$	[pt] ($T_{step} = 0.05$)	$Pred.$	[pt] ($T_{step} = 0.5$)	$Pred.$
1	0.5001 ± 0.0014	0.5000	0.4641 ± 0.0031	0.4933
2	0.4996 ± 0.0018	0.4999	0.4389 ± 0.0013	0.4724
3	0.4976 ± 0.0013	0.4992	0.4149 ± 0.0016	0.4474
4	0.4950 ± 0.0030	0.4972	0.3963 ± 0.0019	0.4235
5	0.4938 ± 0.0004	0.4940	0.3776 ± 0.0003	0.4019
6	0.4889 ± 0.0022	0.4895	0.3621 ± 0.0020	0.3828
7	0.4858 ± 0.0003	0.4842	0.3478 ± 0.0015	0.3658
8	0.4782 ± 0.0024	0.4783	0.3348 ± 0.0016	0.3508
9	0.4758 ± 0.0008	0.4720	0.3244 ± 0.0030	0.3373
10	0.4697 ± 0.0014	0.4654	0.3128 ± 0.0003	0.3252
12	0.4585 ± 0.0006	0.4521	0.2955 ± 0.0001	0.3044
15	0.4417 ± 0.0010	0.4326	0.2737 ± 0.0004	0.2793
17	0.4314 ± 0.0019	0.4204	0.2605 ± 0.0011	0.2656
20	0.4165 ± 0.0009	0.4033	0.2464 ± 0.0008	0.2484
25	0.3943 ± 0.0011	0.3782	0.2258 ± 0.0022	0.2258
30	0.3737 ± 0.0014	0.3568	0.2084 ± 0.0006	0.2084
35	0.3562 ± 0.0024	0.3385	0.1960 ± 0.0009	0.1945
40	0.3423 ± 0.0021	0.3226	0.1848 ± 0.0019	0.1831
50	0.3152 ± 0.0014	0.2963	0.1671 ± 0.0001	0.1651
60	0.2940 ± 0.0013	0.2755	0.1529 ± 0.0017	0.1516
70	0.2758 ± 0.0013	0.2584	0.1440 ± 0.0003	0.1409
80	0.2621 ± 0.0010	0.2442	0.1349 ± 0.0003	0.1323
100	0.2395 ± 0.0012	0.2215	0.1214 ± 0.0007	0.1188
120	0.2203 ± 0.0006	0.2042	0.1115 ± 0.0012	0.1088
150	0.2003 ± 0.0013	0.1844	0.1000 ± 0.0004	0.0976
200	0.1742 ± 0.0012	0.1612	0.0867 ± 0.0003	0.0848
250	0.1583 ± 0.0001	0.1451	0.0778 ± 0.0001	0.0759
300	0.1458 ± 0.0011	0.1329	0.0718 ± 0.0006	0.0694
500	0.1132 ± 0.0004	0.1038	0.0542 ± 0.0001	0.0539
700	0.0960 ± 0.0010	0.0880	0.0468 ± 0.0004	0.0456
1000	0.0809 ± 0.0008	0.0738	0.0395 ± 0.0005	0.0382
1500	0.0660 ± 0.0006	0.0604	0.0324 ± 0.0001	0.0312
2000	0.0577 ± 0.0004	0.0524	0.0279 ± 0.0005	0.0270

Table 6.1: **Frequency of [pt] as a function of K_{max}** , while $T_{step} = 0.05$ (second column) and $T_{step} = 0.5$ (fourth column). Each frequency has been calculated by running 100 000 simulations trice. Error is $\sigma(n - 1)$. The figures in the first column are $k = K_{max} - 3$, that is, the number of strata above the highest ranked constraint. The third and the fifth columns show the estimations based on equation (6.21): the correspondence with the results of the experiments is often very good.

T_{step}	[pt] $K_{max} = 10$	$Pred.$	[pt] $K_{max} = 100$	$Pred.$
3	0.3549 ± 0.0025	0.4222	0.1466 ± 0.0007	0.1532
2	0.3377 ± 0.0023	0.3970	0.1262 ± 0.0009	0.1371
1.5	0.3363 ± 0.0018	0.3823	0.1269 ± 0.0005	0.1290
1	0.3356 ± 0.0013	0.3682	0.1212 ± 0.0008	0.1218
0.8	0.3372 ± 0.0022	0.3643	0.1207 ± 0.0006	0.1198
0.5	0.3476 ± 0.0008	0.3658	0.1222 ± 0.0015	0.1206
0.3	0.3735 ± 0.0023	0.3818	0.1334 ± 0.0012	0.1287
0.2	0.4000 ± 0.0009	0.4032	0.1489 ± 0.0020	0.1408
0.15	0.4208 ± 0.0011	0.4214	0.1608 ± 0.0018	0.1526
0.1	0.4495 ± 0.0018	0.4481	0.1870 ± 0.0013	0.1740
0.08	0.4632 ± 0.0005	0.4617	0.2043 ± 0.0010	0.1883
0.05	0.4842 ± 0.0023	0.4842	0.2415 ± 0.0012	0.2245
0.03	0.4951 ± 0.0008	0.4965	0.2934 ± 0.0010	0.2729
0.02	0.5011 ± 0.0007	0.4994	0.3390 ± 0.0018	0.3168
0.015	0.5001 ± 0.0008	0.4999	0.3716 ± 0.0014	0.3498
0.01	0.5000 ± 0.0008	0.5000	0.4152 ± 0.0015	0.3960

Table 6.2: **Frequency of [pt] as a function of T_{step}** ($[pt] \pm \sigma(n-1)$), while $K_{max} = 10$ and $K_{max} = 100$. Each frequency has been calculated by running 100 000 simulations trice. The third and the fifth columns show the estimations based on equation (6.21), not rarely matching the observed frequencies.

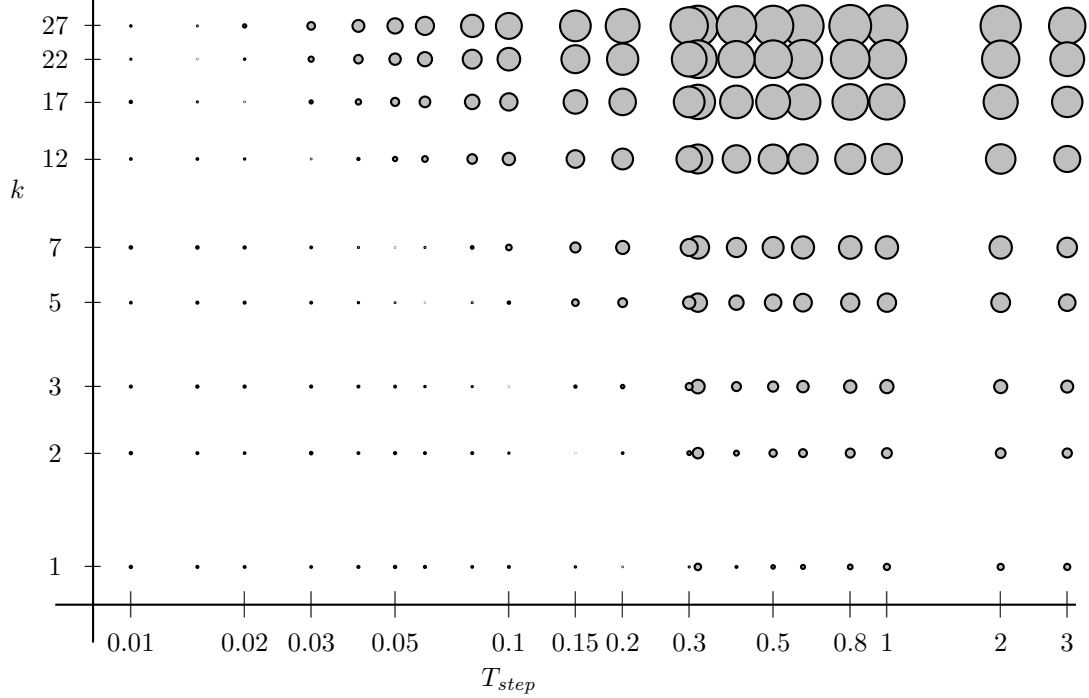


Figure 6.9: **The phase space:** the behaviour of the system in the function of the two parameters k (K_{max} minus the rank of the highest ranked constraint) and T_{step} , on a log-log scale. The radius of each circle is proportional to the difference of the probability of the two forms. Small dots represent (almost) 50%-50% distribution, whereas large circles correspond to [bd] dominating.

T_{step}	$K_{max} = k + 3$	Frequency [pt]	$T_{step} \cdot (k + 0.5)$	$\frac{T_{step}(k+0.5)}{(1+1.5T_{step})^2}$
0.0105	400	0.24996 ± 0.00379	4.17	4.05
0.0210	200	0.2493 ± 0.00725	4.15	3.89
0.0270	160	0.2501 ± 0.0029	4.25	3.93
0.045	100	0.2505 ± 0.0044	4.39	3.85
0.060	80	0.2495 ± 0.0029	4.65	3.91
0.105	50	0.2490 ± 0.0057	4.99	3.72
0.145	40	0.2506 ± 0.0043	5.44	3.67
0.220	32	0.2503 ± 0.0040	6.49	3.67
0.425	24	0.2503 ± 0.0038	9.14	3.408

Table 6.3: **Parameter settings producing [pt] with 25% chance:** (T_{step}, K_{max}) combinations that return candidate [pt] in 25% of the cases. We have noted that $T_{step} \cdot (k + 0.5)$ should be approximately constant for such parameter combinations. This prediction turns out to be correct only in a first approximation, and further factors (the second addend in equation (6.22) first of all) become gradually more important as T_{step} grows larger.

$$\begin{aligned}
G &= \sqrt{\frac{\tau}{2(k+0.5)}} + \sqrt{\frac{(\kappa+2)^2}{2\tau(k+0.5)}} = \\
&= \left(1 + \frac{\kappa+2}{\tau}\right) \cdot \sqrt{\frac{\tau}{2(k+0.5)}} = \\
&= (1 + 1.5T_{step}) \cdot \sqrt{\frac{\tau}{2(k+0.5)}} \approx \sqrt{\frac{\tau}{2(k+0.5)}} \quad (6.24)
\end{aligned}$$

If $G = 0.67$ and $\tau = 3/T_{step}$, then we predict $T_{step}(k + 0.5) \approx 3.34$. As Table 6.3 shows, the larger the value of T_{step} , the larger this product, for the second addend in (6.22) also contributes to G . Indeed, a better approximation following from equation (6.24) is that $\frac{T_{step}(k+0.5)}{(1+1.5T_{step})^2}$ must be constant.

Finally, let us check the prediction according to which there is a turning point in the frequencies around $T_{step} = 0.75$ if K_{max} is kept constant. Equation (6.23) gives a lower bound on G . Yet, the corresponding frequencies (that can be calculated by integrating equation (6.21) until this G value) cannot be reproduced, because G is predicted to be minimal for $\tau = \kappa + 2 = 4.5$, but τ is an integer. Consequently, we tried to find the minima in the frequencies of [pt] by varying T_{step} , for different K_{max} (Table 6.4). Since we required very accurate values, the number of iterations was very large: each piece of data in Table 6.4 originates from running 500 000 simulations trice in order to estimate also the error $\sigma(n-1)$ of the frequencies.

The experiment confirms our predictions for $K_{max} \geq 16$: [pt] is produced the least frequently for $\tau = 4$ ($T_{step} = 0.8$ in our experiment²²), and these frequencies are only slightly larger than the minimal that would correspond to

²²Further results not reported here for lack of space show that different T_{step} values corresponding to the same τ (such as 2 and 1.5, or 1.2 and 1) have not produced significantly different frequencies. Probably many more runs are required in order to be able to demonstrate the role of the t values in the inner loop of the algorithm, as we have done in subsections 5.5.2 and 5.5.3.

K_{max}	$T_{step} = 1.5$	$T_{step} = 1.0$	$T_{step} = 0.8$	$T_{step} = 0.6$	$T_{step} = 0.4$	Pred.
150	1046 \pm 0005	0994 \pm 0001	0987 \pm 0003	0990 \pm 0006	1052 \pm 0002	0975
100	1269 \pm 0003	1213 \pm 0002	1201 \pm 0001	1212 \pm 0004	1277 \pm 0001	1194
70	1503 \pm 0003	1439 \pm 0002	1425 \pm 0001	1433 \pm 0006	1518 \pm 0004	1425
50	1754 \pm 0006	1689 \pm 0005	1679 \pm 0001	1691 \pm 0003	1778 \pm 0005	1683
40	1939 \pm 0006	1870 \pm 0007	1857 \pm 0002	1877 \pm 0005	1986 \pm 0002	1879
35	2058 \pm 0004	1984 \pm 0008	1981 \pm 0007	1997 \pm 0010	2110 \pm 0005	2006
30	2196 \pm 0002	2124 \pm 0010	2119 \pm 0001	2141 \pm 0011	2261 \pm 0005	2163
25	2380 \pm 0002	2307 \pm 0009	2304 \pm 0004	2329 \pm 0003	2457 \pm 0004	2364
22	2506 \pm 0003	2444 \pm 0004	2436 \pm 0004	2466 \pm 0004	2604 \pm 0008	2515
20	2609 \pm 0002	2544 \pm 0003	2542 \pm 0005	2576 \pm 0002	2709 \pm 0009	2633
18	2713 \pm 0003	2660 \pm 0003	2657 \pm 0007	2699 \pm 0004	2839 \pm 0008	2769
16	2849 \pm 0012	2786 \pm 0003	2794 \pm 0004	2824 \pm 0008	2984 \pm 0003	2929
14	2994 \pm 0007	2947 \pm 0009	2962 \pm 0004	2994 \pm 0004	3156 \pm 0009	3118
12	3170 \pm 0007	3127 \pm 0009	3153 \pm 0004	3186 \pm 0002	3363 \pm 0005	3348
10	3378 \pm 0004	3353 \pm 0003	3374 \pm 0005	3423 \pm 0004	3602 \pm 0002	3633
8	3638 \pm 0004	3629 \pm 0004	3656 \pm 0005	3718 \pm 0006	3910 \pm 0010	3996
7	3786 \pm 0003	3792 \pm 0006	3845 \pm 0004	3888 \pm 0003	4091 \pm 0003	4213
6	3967 \pm 0001	3979 \pm 0002	4039 \pm 0009	4097 \pm 0008	4288 \pm 0005	4456
5	4168 \pm 0001	4202 \pm 0003	4266 \pm 0002	4327 \pm 0003	4500 \pm 0007	4711
4	4433 \pm 0005	4480 \pm 0008	4539 \pm 0006	4590 \pm 0007	4729 \pm 0007	4928

Table 6.4: **The turning point around $\tau = \kappa + 2$:** the frequency of [pt] for different parameters, with the initial “0.” truncated due to lack of space. The turning point (a local minimum in the frequency of [pt]) predicted to be at $T_{step} = 0.8$ ($\tau = 4$) can be observed for larger k , even though $T_{step} = 1$ ($\tau = 3$) often produces frequencies that are not significantly different. Nonetheless, further factors become important for lower k values, and the turning point slowly shifts towards larger T_{step} : to $T_{step} = 1$ for $16 \geq K_{max} \geq 7$, and to $T_{step} \geq 1.5$ for $7 \geq K_{max} \geq 4$. Finally observe that wherever these further factors are not yet observable, our expectations on the lower bound predicted by equations (6.21) and (6.23) are met: the values in the fourth column ($T_{step} = 0.8$, $\tau = 4$) are but slightly larger than those in the last one (corresponding to $\tau = 4.5$).

$\tau = 4.5$ according to equation (6.21). However, our approximations are not good enough any more if $K_{max} \leq 16$. Most probably further factors have to be taken into considerations, or our approximations must be refined, in order to explain why the turning point shifts towards larger T_{step} , and why the observed frequencies are much lower than the predicted lower bound.

6.6 What have we learnt from [voice] assimilation?

The starting problem of the present chapter was Dutch voice assimilation in linguistic forms such as *op die* and *zakdoek*. The first model presented included four candidates with a topology of the form that we called a “magic square”. Similarly to some tableaux in the three-candidate search space of subsection 2.3.2, this magic square, together with the hierarchy we employed, demonstrates that SA-OT does not necessarily converge towards maximal precision as the number of iterations increases. The proposed version of simulated annealing for OT cannot avoid getting caught in a local optimum with a constant probability—independently of the cooling schedule.

As argued above, however, this phenomenon may help accounting for certain irregularities. Instead of making the model more complicated in order to include them, the model of the static mental representation (competence) can be kept simple, and irregularities are quarantined in the dynamic computational process. For instance, the model of Dutch phonology will include only regressive voice assimilation, that is, the only global optimum is *o[bd]ie*. Nevertheless, the local optimum *o[pt]ie* is also returned by the dynamic computational process as an irregular form. As the *o[bd]ie* vs. *o[pt]ie* alternation is most probably not a fast speech phenomenon, there is no need to tune the frequencies through the cooling schedule, as we have done in Chapter 5: the frequencies have to be the same under different speech conditions (for different speech rates).²³

From a methodological point of view, the difference between this variation and fast speech was that the observation that the frequency of the *andante* form diminishes at higher speech rate makes possible to point immediately to the *allegro* form as performance error. Whereas in the present case, only the theoretical model will decide which is the form that can be easily described (for instance, by having it globally optimal in OT), and which form has to be exiled to the dynamic computational process. Such an approach may prove to be advantageous even for language acquisition: a learning algorithm robust enough to deal with the inevitable noise will learn the simpler grammar faster, in which case the “performance effects” are realised for free.

But the situation is not so simple. The 50-50% distribution might turn to be incorrect empirically for *op die*; and is certainly false for other words such as *zakdoek* where only regressive assimilation may occur. We have, therefore, introduced a second model that involved an infinite search space.

However, the parameters influenced this model in a surprising way. Unlike earlier, decreasing T_{step} decreased the chance of returning the grammatical form *o[bd]ie*, while K_{max} also played an important role. And yet, this divergent be-

²³According to Paul Boersma, the voiceless variant might be more common in fast speech, which situation could be modelled using T_{step} values larger than the turning point.

haviour can be nicely interpreted. Fast speech phenomena were analysed using the parameter T_{step} in the previous chapter, whereas the present phenomenon is different, and is consequently analysed employing another parameter, namely K_{max} . If the $o[bd]ie \sim o[pt]ie$ variation is dialect dependent, speaker dependent or register dependent, then this variation should be driven by a parameter different from the one driving speech-rate dependent variations.

Furthermore, this observation also helps in explaining the difference between *op die* as opposed to *zakdoek*. An important point about Simulated Annealing Optimality Theory is that it is not only able to account for the presence and the absence of variation by tuning its parameters, but that the parameters can also be interpreted. Namely, parameter settings yielding variation often correspond to a faster production process than parameter settings yielding almost exclusively a given form. Now, production speed may not depend only on speech rate, but also on word frequency: frequent words, such as the unstressed function words in *op die*, are probably more quickly processed than relatively less frequent nouns. That is, processing *zakdoek* involves a much greater K_{max} , which causes the computation to take longer *even for the same speech rate*, and consequently the frequency of the regressive assimilation form to converge to 1. (Note, however, that our argument will be different for the Hungarian definite article, which is a phenomenon related to speech rate: there, we employ the fact that a larger K_{max} requires a longer computational time, and thus may be viewed as also corresponding to a slower speech rate.)

The second model also shows the usefulness of an infinite candidate set. Candidates that can never win are not only necessary for the mathematical consistency and beauty of the model, but they may also influence the search algorithm. In traditional OT, loser candidates (candidates winning for no constraint ranking) could be already excluded from GEN, but in SA-OT they play a role behind the scenes. Even if they are never returned as outputs, the system may rove through them, and it is exactly because the search space is infinite that the output frequencies can be tuned by varying K_{max} .

The analysis of this second model for voice assimilation with an infinite search space reveals an additional peculiarity. Let us alter slightly the definition of the constraints so that [pt] is the global optimum:

/pd/	DEP	ASSIM[VOICE]	C ₃	FAITH[VOICE]
~ [bd]			*!	*
☞ [pt]				*
[pd]		*!	*	
[bt]		*!	*	**
[b@d]	*!			*
[p@t]	*!		*	*
[p@d]	*!		*	
[b@t]	*!		*	**
...
[b@ ⁿ d]	* ⁿ !			*
[p@ ⁿ t]	* ⁿ !		*	*
[p@ ⁿ d]	* ⁿ !		*	
[b@ ⁿ t]	* ⁿ !		*	**
...

(6.25)

This model is expected to display the same behaviour as the original one. At small K_{max} values, both local optima, [pt] and [bd], have 50% chance to be returned; but channelling through the [b@ⁿd] river becomes significant as K_{max} increases, making [bd] more probable. Consequently, we have a model in which the global optimum, now [pt], can never be returned in more than half of the cases, and its frequency can even converge to zero.

Chapter 7

Word Structure and Syllable Structure with SA-OT

This chapter enlarges further the scope of techniques that the Simulated Annealing Optimality Theory Algorithm offers us. Two phenomena are dealt with, both related to syllabification and syllable structure. First, the results of Judit Gervain’s psycholinguistic experiments on the cliticisation of the definite article in Hungarian are modelled, and then Prince and Smolensky (1993)’s well-known basic syllable structure theory is implemented using SA-OT.

The topologies of the candidate sets in both models are similar to that presented in section 6.5. Recall the infinite search space used there (Fig. 6.3 on page 169), and the fact that we distinguished between radial and tangential moves. In fact, the search space had a centre, an origin in which candidates had no epenthetic segments. Centrifugal moves involved applying epenthesis recursively, centripetal moves undid epenthesis, whereas tangential moves corresponded to other operations on the candidate string, which did not change the number of epenthetic elements, that is, the “distance” from the origin. As a typical candidate had four neighbours, and each of them was assigned the same *a priori* probability, therefore performing a tangential step had an *a priori* probability of 50%, whereas centripetal and centrifugal steps had 25% each (with the remark at the end of footnote 17 on page 181).

In the case of both phenomena to be introduced, we shall start with a similar topology, but then make them more complex. Some of the conclusions to be drawn can also be applied to the description of phenomena discussed earlier.

7.1 Th’ article in Hungarian

7.1.1 The behaviour of the definite article

Let us now review a model accounting for the behaviour of the definite article in Hungarian, which is similar to the model we have just discussed (section 6.5),

but which displays further features of SA-OT.¹ The definite article has two allomorphs, and the choice between them depends on whether the next word begins with a consonant or with a vowel. For instance:

$$\begin{array}{ll} az\ alma & \text{'the apple'}, \\ a\ szalma & \text{'the straw'}. \end{array} \quad (7.1)$$

The pause between the *az* allomorph of the article and the subsequent word is prone to omission, resulting in the cliticising of the article. The hypothesis has been that cliticising is more frequent with an acceleration of the speech rate (Kiefer, 1994). In order to support the hypothesis, Judit Gervain has performed a series of controlled psycholinguistic experiment measuring the frequency of the phenomenon (so far unpublished, reported by Bíró and Gervain, 2006). Her experiments confirmed this hypothesis by measuring the presence and the overall length of pauses in critical minimal pairs (e.g. *az ár* 'the price' as opposed to *a zár* 'the lock') excised from test sentences pronounced by four female native speakers in three conditions.

She reports that for the *a* allomorph, cliticisation is the default, irrespective of speed. However, *az* cliticises (i.e., the length of pause is less than 3 msec) only in 1 case out of 12 (8.3%) at a slow speaking rate. This proportion grows to 4 cases out of 12 (33.3%) at a medium rate, and to 8 cases out of 12 (66.7%) at a fast rate. Furthermore, detailed results show that the length of the pauses correlates inversely with speed, the average length of the pause being significantly shorter at a medium speech rate than in slow speech. A first explanation based on the intuition of the native speakers could be that if the *a* allomorph cliticises, the syllable boundaries still align with the morpheme boundaries; if, however, the *az* allomorph cliticises, the segment [z] is resyllabified into the onset of the subsequent word, resulting in a violation of the relevant alignment constraint, which should be avoided at slower speech rate.

7.1.2 Constructing a model

The model to be presented resembles the one employed to account for the magic square-type phenomena, such as Dutch voice assimilation. This model is also based on the infinity of the search space, even if its structure is slightly different. Moreover, we shall tune the frequencies again by having the random walker rove away from the origin due to large K_{max} values.

The candidates to be considered for an input such as *az ár* will have the form $[az^n \#^m E]$: between the [a] segment of the article and the arbitrary initial vowel E of the subsequent word, segment [z] of length n is followed by a pause of length m ($n, m \geq 0$). Exponents n and m can also be thought of as time units, for instance given in msec. The initial candidate, from which the simulations are launched, will always be $[az \# E]$, that is $n = m = 1$, and basic steps alter the values of n and m . Thus, candidates of the form $[a \#^m z^n E]$ ($n, m > 0$) are never reached, even though these would come into play if the input were something like *a zebra* ('the zebra') or *a zár* ('the lock'). The pause between the article and the subsequent word is considered to be omitted if the exponent m

¹The present section builds upon Bíró and Gervain (2006), but presents a more elaborate model for the same phenomenon.

of the pause symbol # is less than three, corresponding to Gervain's definition of cliticising (measuring a pause shorter than 3 msec).²

A basic step consists of changing both n and m by 1—one may not change only n or m . This stipulation may sound *ad hoc*, but the success of the model will depend on it, and has the advantage of creating a simple topology.³ A general candidate $[az^n \#^m E]$ has four neighbours only: $[az^{n-1} \#^{m-1} E]$, $[az^{n+1} \#^{m+1} E]$, $[az^{n-1} \#^{m+1} E]$ and $[az^{n+1} \#^{m-1} E]$. Obviously, if n or m is 0, the candidate has fewer neighbours. In other words, one may shorten the candidate by 2, one may lengthen the candidate by 2, and one may change the proportions of $[z]$ and the pause without changing the overall length.

The resulting topology, presented in Fig. 7.1, has a similar structure to the topology dealt with in section 6.5 (Fig. 6.3 on page 169). There are radial steps (shortening the candidate is a centripetal step, while lengthening the candidate is taking a centrifugal step), as well as tangential steps, perpendicular to the radial ones, which change the difference of the number of z 's and $\#$'s. If changing only n without changing m , and changing only m without changing n were also allowed, then diagonal steps could be also possible. In some sense, radial moves change the “quantity”, and tangential steps change the “quality” of the candidate, and combining the two is not possible within one *basic step*. In the present case, permitting n or m not to change in a basic step, that is, having for instance $[az^{n\pm 1} \#^m E]$ as a neighbour of $[az^n \#^m E]$, might be viewed as a diagonal move in Fig. 7.1, which would be a combination of both qualitative and quantitative changes in the candidate. Such diagonal moves would, however, prevent the candidates that should be returned by the algorithm to become local optima.

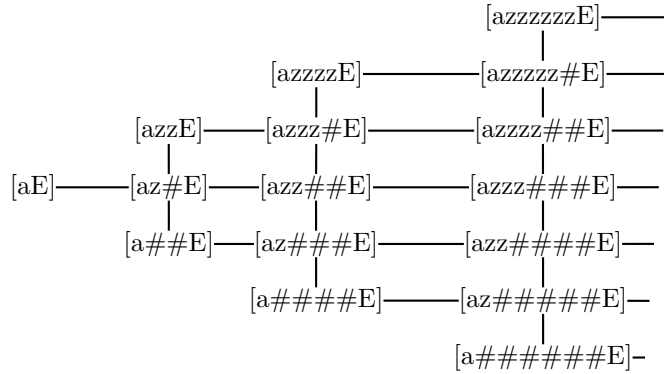
The markedness constraints to be used are very simple:

$$\begin{aligned} C_1(w) &= \text{KEEPSHORT}([az^n \#^m E]) = n + m \\ C_0(w) &= \text{KEEPSEGMENTSHORT}([az^n \#^m E]) = n \end{aligned} \quad (7.2)$$

Both reflect some principles of economy: KEEPSHORT punishes long strings in general, whereas KEEPSEGMENTSHORT disprefers long segments. Importantly, “pronouncing” the pause requires negligible energy, so no separate constraint KEEPPAUSESHORT—whose value on candidate $[az^n \#^m E]$ would be m —is needed; alternatively, such a constraint should be ranked (universally) lower for the same rationale. Observe, furthermore, that these constraints follow also

²As correctly remarked by Paul Boersma, the exponents of $[z]$ and of $[\#]$ cannot have both the interpretation of measuring the respective durations in msec, because the outputs are going to be candidates with $n = 1$, while the duration of a segment $[z]$ in reality is about 100 milliseconds. Hence, this asymmetry has to be accounted for, either in the definition of the exponents, or otherwise.

³Readers criticised this stipulation, similarly to the decision in Chapter 5 of not allowing the insertion and deletion of a bisyllabic foot as a basic step. As proposed there, too, the general principle might be that the set of basic steps should be minimal in the sense that leaving out one of them would create a topology in which not all candidates can be reached from any other candidate. Now, as both n and m should change, the candidates with an odd $n + m$ value cannot be reached from candidates with an even $n + m$ value. And yet, I have the impression (which will not be shared by most readers) that leaving out the “odd half” of the candidate set does not really influence the general structure, while including further basic steps breaks the logic of the structure to be explained soon. Changing n only (or m only) corresponds to the idea of a “diagonal” step that could be decomposed into a radial and a tangential one. Nonetheless, I am also open to alternative proposals.

Figure 7.1: Search space for the Hungarian article *az*.

the logic of the search space: **KEEPSHORT** is the constraint that bridles recursive insertions of $[z\#]$, while **KEEPSEGMENTSHORT** influences the tangential movements. In Fig. 7.1, the first constraint forces the system to stay left, and the second constraint to stay as close to the bottom as possible.

Each of the four possible basic steps involves a well-defined change in the violation level of each constraint, so there is no need to re-evaluate the candidates at every iteration of the algorithm. The difference of the violation profiles follows directly from the basic step chosen by the algorithm. This strong connection between the structure of the candidates, the topology and the (markedness) constraints improves the speed of the algorithm, and is an illustration of what I refer to as the SA-OT implementation being built organically upon the underlying traditional OT model.

Our goal is to have the system return candidates $[az\#^{2k+1}E]$: the consonant of the article is kept always short, while the pause might have different lengths. The special case $k = 0$ corresponds to cliticisation, because the pause is so short that it is unperceivable (our system is unable to return candidate $[azE]$), whereas a larger $m = 2k + 1$ corresponds to a (shorter or longer) audible pause. In order to make these candidates local optima, we still need to disqualify the bottom left-most candidates ($[a\#^{2k}E]$), which are more harmonic than their neighbours for the constraints introduced so far. That is a simple task, once we observe that the candidates to be disqualified miss the $[z]$ segment of the input. The following constraint will do the work:

$$C_2(w) = \text{FAITHFULNESS}([az^n\#^mE]) = \begin{cases} 1 & \text{if } n = 0 \\ 0 & \text{if } n \geq 1 \end{cases} \quad (7.3)$$

It is only by ranking this latter constraint above both markedness constraints that candidates $[az\#^{2k+1}E]$ become local optima in our topology:

$$\text{FAITHFULNESS} \gg \text{KEEPSHORT} \gg \text{KEEPSEGMENTSHORT} \quad (7.4)$$

The expected behaviour of this system is similar to that of the one analysed in section 6.5. After being launched from candidate $[az\#E]$, the random walker may freely rove in the initial phase of the simulation. The larger the K_{max} ,

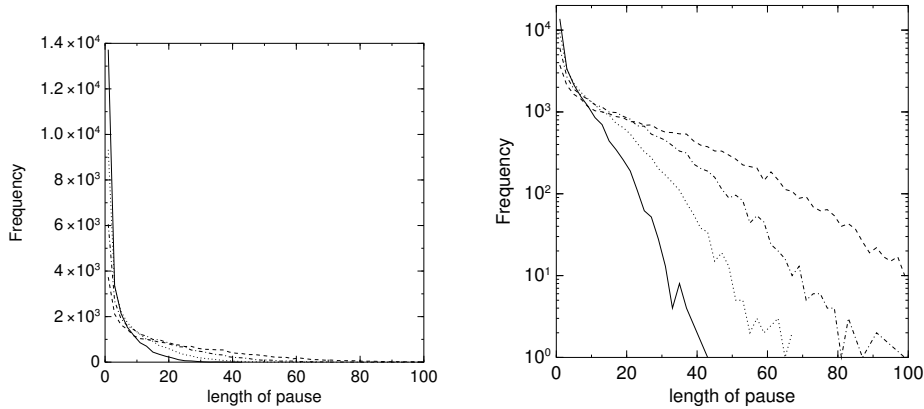


Figure 7.2: **Frequencies of $[az\#^m E]$ employing different K_{max} values**, as a function of m , with a linear (left box) and a logarithmic (right box) frequency axis. The different graphs correspond to $K_{max} = 1$ (solid line), $K_{max} = 3$ (dotted line), $K_{max} = 5$ (dotdashed line) and $K_{max} = 7$ (dashed line) respectively. $T_{step} = 0.1$.

the farther it gets. Once temperature reaches the domain of constraint KEEP-SHORT, insertions are prohibited, and the random walker slowly gravitates back towards the origin. In this second phase, approximately half of the time steps are employed to perform “tangential” moves, that is, vertical ones in Fig. 7.1.

As soon as the temperature cools down to the domain of constraint KEEP-SEGMENTSHORT, candidates $[az\#^{2k+1}E]$ become traps wherein the random walker may get stuck. The “competition” analysed in the previous section translates here to a competition between the centripetal moves leading back to candidate $[az\#E]$ and the tangential moves trapping the system to another candidate $[az\#^{2k+1}E]$. The difference between the two models is that the $[b@^+d]$ channel was only a trap to the tangential moves and “channelled the water” towards $[bd]$ in a centripetal direction, whereas now candidates $[az\#^{2k+1}E]$ are local optima, for centripetal steps are also prohibited by the high-ranked constraint FAITHFULNESS. The “water channel” is replaced here by a series of “water reservoirs”. Even without working out the formal analysis that would be similar to the one presented in the previous section, we expect the chance of returning candidate $[az\#^{2k+1}E]$ ($k > 0$) to increase as parameter K_{max} grows larger.

Running the simulation under the “usual” conditions results in Fig. 7.2. Constraints were assigned the indices 2, 1 and 0. The parameters were $T_{max} = 3$, $T_{min} = 0$, $T_{step} = 0.1$ and $K_{step} = 1$. Furthermore, instead of employing K_{min} , the algorithm was run each time until the random walker had not moved for 30 iterations: in the case of four neighbours, the likelihood of having a more harmonic neighbour but not finding it in 30 trials is $0.75^{30} < 0.0002$. This technique corresponds to measuring the “specific heat” in standard simulated annealing: there, the stopping condition is that the specific heat—the decrease in the target function divided by the decrease in temperature—drops below a certain value, that is, not much is expected to be gained if going further. Finally, the simulation was run 25000 times using each of four different K_{max} values,

and Fig. 7.2 presents the frequency of being returned candidate $[az\#^mE]$ as a function of m (only odd m 's appear on the graphs).

The graphs confirm our prediction. While the probability of $m \leq 25$ is around 99.5% if $K_{max} = 5$ and around 95% if $K_{max} = 10$, increasing the length of the initial phase results in only 84% for $K_{max} = 20$, and 68% for $K_{max} = 40$. Candidates with a probability above 1% are the candidates $[az\#^mE]$ with $m < 21$ for $K_{max} = 5$, $m < 29$ for $K_{max} = 10$, $m < 39$ for $K_{max} = 20$, and $m < 53$ for $K_{max} = 40$. In brief, the larger K_{max} , the broader the distribution of the outputs. As a larger K_{max} requires a longer running time, our model correctly predicts longer pauses at slower speech rates.

7.1.3 Refining the model by changing the topology

This result does not satisfy us, however. These distributions are, namely, centred around the origin, and the candidate with $m = 1$ (omission of the pause) has always the highest chance; whereas experimental results observed a longer gap in most of the cases for slower speech rates, with the pause being almost never omitted. Therefore, we need a model that produces a distribution similar to those in Fig. 7.3: in fast speech, the length of the pauses has a distribution around zero, but at a slower rate, the distribution is centred around some larger m . The slower the rate, the further is the distribution shifted to the right. How can we produce such a distribution?

Observe that so far the random walker has had an equal chance to move to the centripetal and to the centrifugal direction in the initial phase. Hence, both in the section on voice assimilation and in the present model, the distribution of the random walker's position at the end of the initial phase has been centred around the origin. It has been true that larger K_{max} values increase the chance of reaching a more remote region of the search space, but the region with the highest probability has always remained the centre.

It is by introducing a bias into the random walk that one can force the random walker to leave the central region. As pointed out in footnote 19 on page 183, the expected position of the random walker in an asymmetric, one-dimensional Brownian motion is proportional to the difference of the probability of moving left and moving right. Consequently, if the *a priori* probability of lengthening the candidate is increased, and the *a priori* probability of shortening the candidate is decreased, a drift is introduced into the system, and the random walker's position by the end of the initial phase will be some distribution centred around a more remote point in the search space. Then, even if lengthening the candidate becomes impossible after the temperature has reached constraint KEEPSHORT, and the system starts gravitating backwards, the final distribution is not necessarily centred around the origin. If the most probable position of the random walker at the end of the initial phase is far enough, then the chance is very low for the random walker to get back to the origin, and most probably it will be stuck in some farther local optimum.

Figure 7.3 has been obtained by introducing a simple change into the *a priori* probabilities. Earlier, the exponent of $[z]$ was increased or decreased by one with a probability of 0.5 each, and the same applied, independently, to the exponent of $[\#]$ (whenever possible). Now, we first toss a coin, and with a chance of 0.5, we lengthen the candidate (both exponents are increased by 1), and with a chance of 0.5, we apply the earlier algorithm. Thereby, the

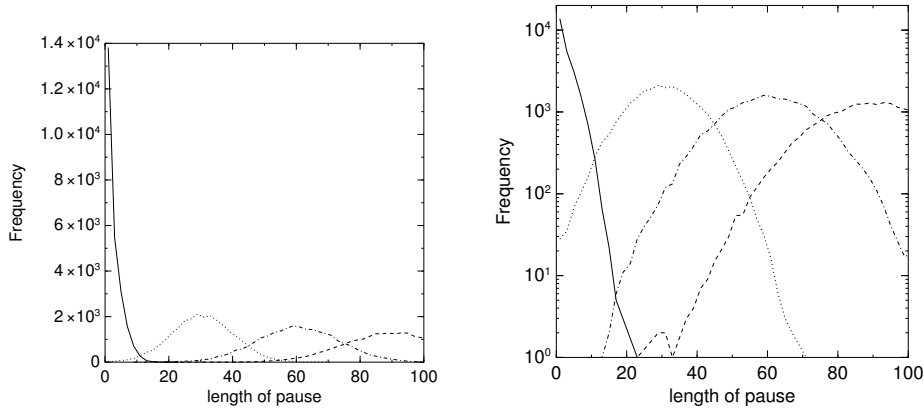


Figure 7.3: **Frequencies of $[az\#^m E]$ employing different K_{max} values**, as a function of m , with the altered *a priori* probabilities. The frequency axis is linear on the left box, and logarithmic on the right one. The different graphs correspond to $K_{max} = 5$ (solid line), $K_{max} = 10$ (dotted line), $K_{max} = 20$ (dotdashed line) and $K_{max} = 40$ (dashed line) respectively. $T_{step} = 0.1$.

earlier *a priori* probabilities of 0.25 each (in the general case) have been altered: $P_{choice}([az^{n+1}\#^{m+1}E] \mid [az^n\#^m E]) = 5/8$, whereas the *a priori* probabilities of the three other neighbours have been reduced to $1/8$ (in the case of $n, m > 0$). This technique enables us to have the model fit the empirical observations better. But it also demonstrates the important role that a further component of the SA-OT algorithm, namely, the *a priori* probabilities, play in determining the output frequencies.

7.1.4 Refining the model by demoting constraints

Judit Gervain's psycholinguistic experiment has also confirmed the different behaviour of the *a* allomorph from that of the *az* allomorph. So far, our model accounts for the speech-rate dependent cliticisation of the *az* allomorph, but are we also able to include the fact that the *a* allomorph almost always cliticises (the case of *a zebra* 'the zebra')? The solution demonstrates a further dimension of SA-OT models.

Figure 7.4 shows the search space analogous to the previous one (Fig. 7.1), but for the *a+zebra* case. As our constraints have been insensible to the order of the pause and the segment $[z]$ within a candidate, the two models should display exactly the same behaviour. (The definitions in (7.2) and in (7.3) can be easily generalised to a candidate of the form $[a\#^m z^n E]$.) Which is not what we aim at, because a certain parameter setting is supposed to be characteristic for a particular speaker, a particular speech situation or a particular speech rate, so our goal is to predict significantly different frequencies for the two types (*a+zár* 'the lock', as opposed to *az+ár* 'the price') using the same parameter setting.

It has been already mentioned that in the *a+zár* case, a support or permission to cliticisation might be that the concatenation creates a sequence in which the well-formed syllable structure preserves the morphological structure. In the *az+ár* case, however, either the syllables are suboptimal (by including a

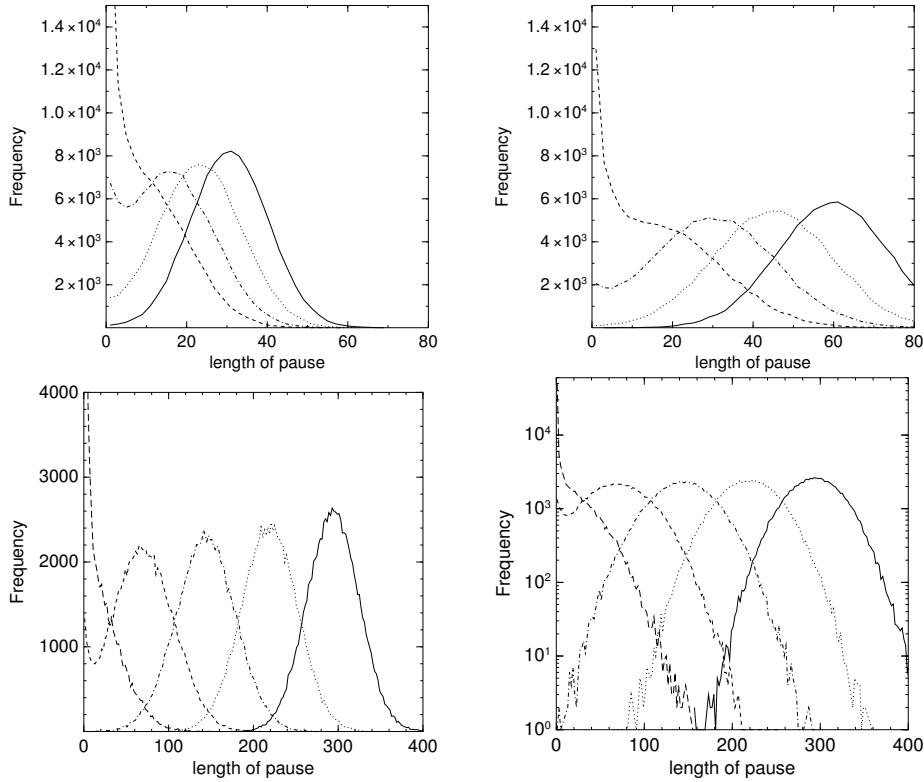


Figure 7.5: **The effect of demoting constraint `KeepSegmentShort`:** The frequency of an output with a pause of length m among 10^5 outputs is displayed as a function of m , for $T_{step} = 0.1$ (upper left box), for $T_{step} = 0.05$ (upper right box), and for $T_{step} = 0.01$ (lower boxes, with linear and logarithmic frequency axes respectively; notice the change of the horizontal scale, as well). In each box, constraint `KEEPSEGMENTSHORT` is associated either with domain 0 (solid line, rightmost), or with domain -2 (dotted line), -4 (dotdashed line), -6 (shortdashed line), or -8 (longdashed line, leftmost, only in the lower boxes). $K_{max} = 3$.

0, while the index of `*COMPLEXONSET` should be -1 or even lower (if further constraints intervene).

In the *az+ár* case, the vacuously fulfilled, low-ranked `*COMPLEXONSET` does not interfere with our previously presented simulation results. Similarly, constraint `*COMPLEXCODA` is vacuously fulfilled in the *a+zár* case, but it does influence the frequencies. Namely, it lengthens the phase during which the random walker gravitates back to the origin, and candidates $[a\#^{2k+1}zE]$ do not act yet as a trap. This idea corresponds to increasing n_1 in (6.19) on page 184: the random walker has a larger probability of reaching the origin $[a\#zE]$, or at least of getting stuck in a local optimum $[a\#^mzE]$ with a small m .

The plots in Fig. 7.5 presents experimental results on demoting constraint `KEEPSEGMENTSHORT` to lower domains, while constraint `FAITHFULNESS` is kept associated with index (domain) 2, and `KEEPSHORT` with 1. In each box, the solid line represents what we have had so far (cf. Figure 7.3, but $K_{max} = 3$),

that is, constraint `KEEPSEGMENTSHORT` is associated with domain 0. This situation corresponds to the case when `KEEPSEGMENTSHORT` is in fact constraint `*COMPLEXCODA` for input $[az+E]$. As discussed in the previous subsection, the distribution of the outputs is shifted to the right due to the bias introduced into the topology. If, however, constraint `KEEPSEGMENTSHORT` is demoted and associated with a lower domain (that is, the lower ranked `*COMPLEXONSET` is acting on input $[a+zE]$), while no effective constraint is associated with domain 0, then this distribution is shifted back to the left. In the extreme case, if `KEEPSEGMENTSHORT` is ranked very low, then the normal-like distribution centred around some m turns into a quickly decreasing distribution. For $T_{step} = 0.1$ and constraint `KEEPSEGMENTSHORT` demoted to domain -6 , output $m = 1$ is returned in 20% of the cases, and $m = 3$ in another 11%. The same values are 50% and 5.5%, if $T_{step} = 0.01$ and constraint `KEEPSEGMENTSHORT` is demoted to domain -8 . Besides, this figure also demonstrates the influence of T_{step} on this model: lower T_{step} involves more steps in the initial phase, therefore the distribution is shifted further to the right, and is broader.

Here, we make crucial use of the way temperature decreases through the domains of the constraints. The length of the pause in $a+zár$ or in $a+zebra$ can be tuned (or an observable pause can be practically eliminated) by introducing further domains between `*COMPLEXCODA` and `*COMPLEXONSET`—either by inserting other constraints in between, or by simply associating constraint `*COMPLEXONSET` with a lower index. Hence, the temperature has to cross empty domains (domains not associated with any constraint or with any relevant constraint) before it reaches `*COMPLEXONSET`.

Here, we touch upon the role of parameter K_{step} , so far absolutely neglected, since by varying this parameter it is also possible to tune the model. Without changing the indices of constraints `*COMPLEXCODA` and `*COMPLEXONSET`, but by decreasing parameter K_{step} , you can also introduce empty domains between these constraints. If, for instance, $K_{step} = 0.5$, then a phase is added to the simulation during which the first component of the temperature is -0.5 , and which phase does not differ from the case when the first component of the temperature is -1 , but `*COMPLEXONSET` is associated with index -2 .

Finally, this solution also has a consequence for traditional OT. In a traditional approach, based on the data presented, one would not be able to rank constraints `*COMPLEXCODA` and `*COMPLEXONSET` relative to each other, because for some inputs `*COMPLEXCODA` is vacuously satisfied, whereas in other cases `*COMPLEXONSET` does not influence the computation. In SA-OT, however, the hierarchy must be such as in (7.6), because this is the hierarchy that guarantees that allomorph a cliticises much more often than allomorph az . The reversed ranking would reverse the frequency of cliticisation.

7.1.5 Conclusion

In the present section, a model has been presented to account for the resyllabification or cliticisation of the Hungarian article. Similarly to the extended magic square model advanced in section 6.5, this model also exploits the infinity of the search space, but has an infinite number of local optima. The major difference is therefore the fact that the local optima, the candidates $[az\#^nE]$, are not neighbours of each other, so they do not form a “channel” any more.

Explaining the different behaviour of the two allomorphs was possible by

employing two different constraints, ranked into different domains. Both constraints measured the number (or length) of the segment $[z]$, but temperature reached that domain later in the $[a+zE]$ case. In other words, an empty domain (the domain of a vacuously satisfied constraint) is introduced for $[a+zE]$ inputs. Consequently, not only does this model demonstrate the proper role of domains to be traversed by temperature above the highest ranked constraint, it also shows that empty domains *between* the constraints can also influence the model. Moreover, introducing empty domains can be replaced by employing $K_{step} \leq 0.5$, too, thus even this parameter of the algorithm might influence the output.

Finally, if lengthening the candidate is assigned a higher *a priori* probability than shortening it, we introduce a bias in the random walk, that is, a drift, which helps us better reproducing experimental observations. This alternation of the topology already leads us to discussing the next model, the goal of which is to see what happens if the *a priori* probabilities vary in a given interval.

7.2 Syllabification (CVT) theory

In the following section, we implement the basic CV Theory for syllabification using simulated annealing. There are a few reasons for doing that. Besides being the most known example of an OT grammar since Prince and Smolensky (1993), it is also a model that has been implemented using different technologies. As already mentioned in section 1.2, Tesar and Smolensky (2000, Chapter 8) employ dynamic programming (chart parsing) for this task, whereas Gerdemann and van Noord (2000) demonstrate the *matching approach* to Finite-State OT exactly on syllabification. Furthermore, it offers us the possibility to discuss further the role of the topology in SA-OT.

To begin with, the search space generated is infinite due to the possibility of inserting epenthetic vowels and consonants in a recursive way. Not only that, but the number of neighbours will grow with the length of the candidate, since a longer candidate can be altered at more places. The topology will, therefore, turn more complex than any other topology discussed in my dissertation, even though its basic structure consisting of radial and tangential moves is the same as the structure of the topologies discussed so far. Additionally, the constraints also operate on each substring of a candidate independently, which results in a high number of local optima and, hence, in a drop in the precision of SA-OT.

By enlarging the set of basic transformations in an *ad hoc* way, the performance of the model can be improved; nevertheless, further work is required to create a really elegant model. As it is the case for SA-OT in general, the precision of SA-OT lags behind the precision of other techniques (dynamic programming, finite state OT); but I conjecture that SA-OT will be able to account for typical performance phenomena, such as the drop (underparsing) of syllable parts in fast speech (*partij* 'political party' becoming *p'tij* in Dutch, *azt hiszem* 'I think' shortening to *asszem* in Hungarian, as well as numerous examples from other languages). Indeed, a major claim of my thesis is that human performance lags also behind the precision of dynamic programming and finite state OT.

7.2.1 Basic CV Theory

First, we repeat Prince and Smolensky (1993)'s syllable theory, which follows Jakobson's typology.

The input is a string of segments, such as **abab**. What GEN does is to parse its segments into a syllable structure. A candidate is a series of syllables, each syllable containing a nucleus, preceded by an optional onset, and followed by an optional coda. Additionally, the string may contain *underparsed* (deleted) segments, which we shall return to soon. For the sake of simplicity, we assume that the phonemes of the language can be divided into two distinct sets, vowels may appear only as nuclei, and consonants only as onsets or codas.⁵

Hence, an underlying vowel can be either underparsed or parsed as a nucleus, and an underlying consonant can be either underparsed or parsed as an onset or as a coda. Additionally, *overparsing* may insert onsets, nuclei and codas that do not contain underlying segments, but epenthetic (default) material, such as a schwa or a [t]. Ignoring the underparsed segments, which are not pronounced, the result must be a well-formed word, that is, a sequence of well-formed syllables (exactly one nucleus, preceded optionally by an onset and followed optionally by a coda). The candidate set corresponding to an input is the set of all possible candidates whose underlying (*i.e.*, not epenthetic) material forms that input, by also keeping the linear order of the segments.⁶

For the sake of convenience, we do not allow complex (branching) onsets and codas, that is, a syllable may contain at most one consonant as onset, and at most one consonant as coda. This constraint applies only to certain languages, but now our only aim is to keep the model as simple as possible.

Following the notation of Gerdemann and van Noord (2000), let $N[a]$ denote an underlying element (the phoneme /a/ in this case) parsed as a nucleus; moreover, $O[b]$ and $D[b]$ refer to the consonant /b/ parsed as onset and coda respectively. By $X[a]$ or $X[b]$ we represent the underparsing (deletion) of some underlying material, whereas $N[_]$, $O[_]$ and $D[_]$ shows the insertion of the default epenthetical vowel or consonant in the position of nucleus, onset or coda. Not surprisingly, $X[_]$ is avoided: deleting a previously inserted element is realised by simply removing it from the string.

For example, if the underlying representation is **ba**, possible candidates include $O[b]N[_]O[_]N[a]$, $O[_]N[_]D[b]N[a]$, and $X[b]N[a]N[_]$. However, $O[b]O[_]N[a]$ or $O[b]D[_]N[a]$ are not valid candidates, since the first one includes a branching onset, while the first syllable of the second one lacks a nucleus. The otherwise well-formed candidates $N[a]D[b]$, $O[b]N[a]D[b]$ or $O[d]N[e]$ are not element of the candidate set either, for they correspond to different inputs.

Following Prince and Smolensky (1993), and most work in their footsteps, we use the following five constraints:

- **ONSET (ons)**: the number of nuclei in the candidate that are not preceded immediately by an onset.⁷

⁵Syllabic consonants are therefore seen as vowels. What we ignore is the possibility of changing the syllabicity of a segment, such as turning a vowel into a glide or a plain consonant into a syllabic consonant. In that respect, we follow in the footsteps of the earlier work referred to, as our first goal is to illustrate SA-OT, and not to account exhaustively for specific linguistic phenomena.

⁶Consequently, this model does not allow metathesis and reduplication.

⁷Immediate precedence is understood in the surface form. That is, underparsed segments

- NoCODA (noc): the number of codas (Ds) in the candidate.
- PARSE (prs): the number of underlying segments that is underparsed in the candidate (the number of Xs).
- FILLONSET (fio): the number of onsets in the candidate that are not parses of an underlying segment (the number of O[]s).
- FILLNUCLEUS (fin): the number of nuclei in the candidate that are not parses of an underlying segment (the number of N[]s).

Motivated by Jakobson’s typology, and found in Prince and Smolensky (2004, p. 106), constraint ONSET requires each syllable to have an onset, and constraint NoCODA prefers each syllable not to have a coda. Unlike these *markedness constraints*, the last three are *faithfulness constraints*: they punish any difference between the input and the output.

7.2.2 Syllabification with simulated annealing I.

From the building blocks of the SA-OT Algorithm, we have introduced the constraints, so let us now concentrate on the definition of the topology. Through what basic step shall we construct a neighbour from the actual candidate, the present position of the random walker? Unlike so far, instead of defining formally the set of neighbours $Neighb(w)$ and the *a priori* probability distribution on it, now we rather advance a procedure that constructs some neighbour w' which the present candidate w will be compared to.

The proposed algorithm first checks if there is any intervocalic consonant that can be reparsed (turn an onset into a coda or a coda into an onset). If there is at least one, the algorithm generates a random number r between 0 and 1 with an equal distribution, and if $r < P_{reparse}$, the basic step to be performed is reparsing. In this case, one of the possible loci for reparsing is chosen with equal probability, where subsequently reparsing takes place.

If no such locus exists, or if $r \geq P_{reparse}$, then the word is lengthened or shortened. The next decision to be made is whether to lengthen or to shorten the candidate. If shortening is not possible (the random walker is located in the centre of the search space), then lengthening takes place. Otherwise, each has a chance ($P_{centrifugal}$ and $P_{centripetal}$) of 50%, because if shortening would be preferred over lengthening, then the random walker would stay around the origin (the candidate with no epenthetical position and all underlying segments underparsed). Increasing the probability $P_{centrifugal}$ of lengthening might be an option for future research, analogous to the model producing Fig. 7.3 (on page 201, in subsection 7.1.3), but does not seem to be very promising: here—unlike there—falling into distant local optima is not attested in speech. In fact, the 50%-chance-each model parallels the earlier model described in section 7.1, as well as the model in section 6.5: moving away from the centre has the same *a priori* probability as moving backwards ($P_{centrifugal} = P_{centripetal}$). This stipulation allows us to concentrate on further parameters of the topology, whereas the role of parameter $P_{centrifugal}$ has already been analysed in subsection 7.1.3.

intervening between an onset and a nucleus do not cause the candidate to violate this constraint.

Subsequently, once it has been decided that the candidate will be shortened, then all possibilities of shortening the candidate are listed, and one of the possibilities is chosen with equal chance. A possibility involves performing one of the following operations at a given locus of the candidate string:

1. Underparse an underlying consonant (parsed as an onset or coda).
2. Delete an epenthetic onset ($O[_]$) or coda ($D[_]$).
3. Underparse an underlying vowel (parsed as a nucleus), supposing that what remains is a well-formed candidate.⁸
4. Delete an epenthetic nucleus ($N[_]$), supposing that what remains is a well-formed candidate.

For instance, candidate $O[b]N[a]N[_]D[c]N[a]$ can be shortened at five different points, so each possibility has a chance of 20%.

Similarly, if the candidate's fate is to be lengthened, then one is picked from the possibilities, where a possibility is rewriting a single locus using one of the following operations:

1. Insert an epenthetic nucleus ($N[_]$), which is possible everywhere (between any two parsed, underparsed or overparsed elements, at the beginning and at the end of the string).
2. Insert an epenthetic onset ($O[_]$), supposing that the previous parsed (or overparsed; if there is one) element is not an onset, and the next one is a nucleus.
3. Insert an epenthetic coda ($D[_]$), supposing that the previous parsed (or overparsed) element is a nucleus, and the next one (if there is one) is not a coda.
4. Turn an underparsed vowel into a nucleus.
5. Turn an underparsed consonant into an onset, supposing that the previous parsed (or overparsed; if there is one) element is not an onset, and the next one is a nucleus.
6. Turn an underparsed consonant into a coda, supposing that the previous parsed (or overparsed) element is a nucleus, and the next one (if there is one) is not a coda.

So far, we have a topology similar to those in sections 6.5 and 7.1, but more complex. The similarity is that the search space has a centre, the candidate whose length is minimal in pronunciation. Furthermore, possible moves are either radial or tangential with respect to this centre. The present model is more complex, however, for a candidate string can be lengthened at any point. Not only that, but we have also introduced a parameter, $P_{reparse}$, that determines the probability of the tangential moves. In the earlier models, the probability

⁸A nucleus can be deleted if it is the first parsed element of the candidate and the next parsed element is not a coda; if the previous or the next parsed element is a nucleus; or if the previous parsed element is a coda and the next one is not a coda.

of considering a tangential move was 50%, because each neighbour had equal *a priori* probabilities, and most often two neighbours out of four represented a tangential move.

Additionally, we render our model even more complex by introducing further neighbours. The reason for that is that if you run a simulation with the present model, you will most often be stuck in some local optimum, similarly to the search spaces in section 7.1. But unlike that case, one is unhappy now if this happens, because local optima are non-attested in speech. These local optima include cases such as epenthetical syllables: if the substring $O[_]N[_]$ is followed by a consonant or by the end of the word, then deleting the nucleus is impossible, while deleting the onset makes the candidate worse if $ONSET \gg FILLONSET$. Another type of local optima is formed by candidates with a substring such as $N[_]X[a]$ if $ONSET \gg PARSE$: deleting the epenthetical nucleus might bring to an ill-formed string, whereas reparsing the vowel [a] first increases the violations of constraint $ONSET$.

Therefore, an additional parameter $P_{postproc}$ is introduced in the definition of the topology. After having performed exactly one basic operation (resyllabification, lengthening or shortening), some post-processing may also occur. Each substring $O[_]N[_]$ is considered, and if the next parsed element is not a coda, then this substring is deleted with probability $P_{postproc}$. Similarly, each substring $N[_]X[v]$, $X[v]N[_]$, $O[_]X[c]$ and $X[c]O[_]$ (where v stands for any vowel and c for any consonant) is collapsed into $N[v]$ or into $O[c]$ with the same probability. These operations help to avoid certain traps, but the analysis of the experiments performed demonstrate that further operations should also be allowed in the future.⁹ Note finally that due to the irreversibility of these post-processing operations, the neighbourhood relation is not symmetric anymore.

In sum, two parameters determine the *a priori* probabilities of the topology, $P_{reparse}$ and $P_{postproc}$. How do they influence the precision of the algorithm? The experiments display huge differences in function of the input string and of the hierarchy employed, but also of T_{step} . The role of the latter here is similar to its role in section 6.5: lower values allow for the random walker to move farther away from the origin in the initial stage of the simulation.

Here, I report on some short experiments performed with the hierarchy

$$noc \gg prs \gg ons \gg fin \gg fio$$

and with initial form $O[l]N[a]D[b]O[d]N[a]D[k]$ (from input /labdak/). The optimal candidate is $O[l]N[a]O[b]N[_]O[d]N[a]O[k]N[_]$ with two epenthetical nuclei, as epenthesis is preferred to deletion and to having codas. The constraints were associated with ranks 0 to 4, and the parameters of the algorithm were: $T_{max} = 3$, $T_{min} = 0$, $T_{step} = 0.01$, $K_{max} = 5$, $K_{min} = -2$ and $K_{step} = 1$. The simulations were run 750 times for a certain $(P_{reparse}, P_{postproc})$ pair, so that the mean and the standard deviation ($\sigma(N-1)$) of the frequencies

⁹These include deleting sequences of $N[_]O[_]$, as well as reparsing whole syllables: turning $X[c]X[v]$ into $O[c]N[v]$, and $X[c]$ into $O[c]N[_]$ in one go. The gradual inclusion of such *ad hoc* operations lessens the simplicity of the model, and future research will hopefully propose a more elegant solution. Another direction has also been advanced, namely, to prevent GEN from generating candidates with similar redundant or verbose substructures.

$P_{reparse}$	0[1]N[a]0[b]N[_]0[d]N[a]0[k]N[_]	0[1]N[a]0[b]N[_]0[d]N[a]X[k]
0.1	0.57 ± 0.05	0.10
0.3	0.50 ± 0.02	0.11
0.5	0.38 ± 0.05	0.17
0.7	0.24 ± 0.04	0.17
0.9	0.11 ± 0.03	0.18

$P_{postproc}$	0[1]N[a]0[b]N[_]0[d]N[a]0[k]N[_]	0[1]N[a]0[b]N[_]0[d]N[a]X[k]
0.1	0.23 ± 0.02	0.11
0.3	0.41 ± 0.01	0.11
0.5	0.50 ± 0.02	0.11
0.7	0.53 ± 0.01	0.12
0.9	0.55 ± 0.04	0.13

Table 7.1: **Varying the parameters of the *a priori* probabilities:** The frequencies of the optimal form and of the most frequently returned non-global local optimum are reported in function of the parameters $P_{reparse}$ and $P_{postproc}$. See text for more details. In the upper table $P_{postproc} = 0.5$, while in the lower table $P_{reparse} = 0.3$.

could be calculated based on the values measured in three groups of 250 runs each.

The results appear in Table 7.1 and Fig. 7.6. The differences in the precision across different parameter combinations are significant in most of the cases, demonstrating how important role the *a priori* probabilities play in the SA-OT algorithm. Moreover, an interesting observation has been that the second most frequent candidate, 0[1]N[a]0[b]N[_]0[d]N[a]X[k], has a much more stable chance to be returned, even though there seems to be a major jump between $P_{reparse} = 0.3$ and $P_{reparse} = 0.5$. For $P_{reparse} = 0.9$, this candidate is the most frequent one, but there are also further non-optimal candidates that emerge more frequently than the optimal one.

However, precision varies enormously with hierarchy and input. The next subsection (Tables 7.3 and 7.4) exemplifies the precision's dependence on the constraint ranking, and preliminary experiments not reported here demonstrated the dependence upon the input. Even for the same hierarchy and input, a different T_{step} value results in a very different behaviour of the system. For instance, if $T_{step} = 0.1$, the random walker in the model just discussed is unable to get far enough from the origin in the initial stage of the simulation, therefore heavily underparsed candidates (such as X[1]X[a]X[b]X[d]X[a]X[k]) are returned most often. By introducing further post-processing steps, more local optima can be avoided, but the model becomes more complex. It is only to be hoped that a more elegant model will emerge from future research.

7.2.3 Syllabification with simulated annealing II.

Now, let us turn to a few further interesting lessons that might be learnt from early, preliminary experiments. These experiments were performed using a

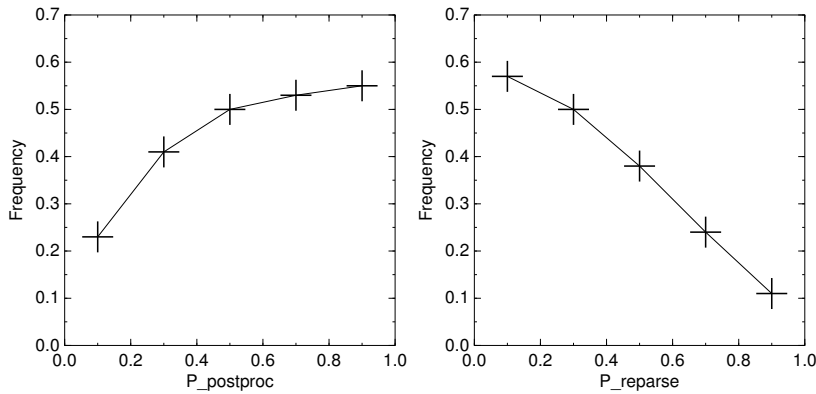


Figure 7.6: **Varying the parameters of the *a priori* probabilities:** The frequency of the optimal form is plotted as a function of parameters $P_{reparse}$ and $P_{postproc}$ (cf. Table 7.1). On the left panel $P_{postproc} = 0.5$, while on the right panel $P_{reparse} = 0.3$.

slightly different topology, so first let us describe it.¹⁰

Similarly to the previous model, we first have to decide whether we want to change the length of the candidate (by inserting, deleting or underparsing segments), or to reparse the present structure, that is, to move the syllable borders. Let $P_{reparse}$ denote again the probability of the latter, whenever possible; whereas $1 - P_{reparse}$ is the probability of changing the word's length.

If we have chosen to reparse, we move one syllable border, that is, we reparse randomly one of the intervocalic consonants: if it has been an onset, we reparse it as a coda, and vice versa (for instance, we turn $O[b]N[a]O[b]N[a]D[_]$ into $O[b]N[a]D[b]N[a]D[_]$).

Alternatively to reparsing, a basic step can either insert an epenthetical vowel or consonant, or delete a phoneme.

Both insertion and deletion of a segment are chosen with a probability of 50%, similarly to an unbiased, one-dimensional random walk. Hence, after t time steps, the expected value of the distance of a random walker's position from its origin is proportional to $t^{1/2}$. Therefore, if we want our simulated annealing algorithm to reach even remote regions of the search space that correspond to k insertions, we must allow a number of steps proportional to k^2 in the first phase of the simulation, when temperature is high and any move is allowed with transition probability 1.

The positions of insertion or deletion are chosen with equal probability again. But from this point onwards come a few differences compared to the topology presented in the previous subsection.

Insertion means only overparsing randomly either an onset, or a nucleus, or a coda, that is, to insert an $O[_]$, a $N[_]$ or a $D[_]$. The three options are chosen with a probability of 40%, 40% and 20%, respectively, making the chance

¹⁰This topology involves too many decision points, involving too many (hidden) parameters. Furthermore, it may run into an infinite loop, because it checks the possibility of an operation only after having performed it. These are the reasons why this topology was revised. However, I did not have the time to reproduce all experiments reported here, using the topology described earlier.

$P_{reparse}$	%	$P_{reparse}$	%	$P_{postproc}$	%	$P_{postproc}$	%	$P_{postproc}$	%
0.00	15	0.60	20	0.00	19	0.35	19	0.70	14
0.10	15	0.70	15	0.05	11	0.40	15	0.75	15
0.20	15	0.80	14	0.10	8	0.45	12	0.80	16
0.30	16	0.90	9	0.15	10	0.50	13	0.85	14
0.40	14	1.00	3	0.20	14	0.55	11	0.90	16
0.50	17			0.25	18	0.60	11	0.95	21
				0.30	14	0.65	14	1.00	25

Table 7.2: *The role of ($P_{reparse}$ and $P_{postproc}$) in CV Theory:* Percentage of simulations returning the correct parse 0[_]N[a]D[n]0[t]N[a], for UR **anta** and hierarchy ONSET \gg FILLNUCLEUS \gg PARSE \gg FILLONSET \gg NoCODA. Each ($P_{reparse}, P_{postproc}$) parameter combination was run 10 times. The left panel shows the results for different $P_{reparse}$ values (average over all possible $P_{postproc}$ values), whereas the right panel presents the role of $P_{postproc}$ (averaged over all possible values of $P_{reparse}$).

of inserting a consonant slightly higher than inserting a vowel. I acknowledge that these values are fully arbitrary, and further research may investigate the role of that choice. Once it has been decided that we want to overparse an onset (similarly in the case of overparsing a nucleus or a coda), a random position is chosen, and the insertion takes place (unless the result is an ill-formed string). Sometimes, forced reparsing takes place in order to obtain a valid candidate.

If deletion is the operation to perform, then a random element of the candidate is chosen, and depending on its present parse, we alter its status. If it has been a parsed, underlyingly existing element, then it gets underparsed (turned into X[.]). If it is an epenthetical element, then it gets deleted. Lastly, and differently from the previous model, if it has been underparsed so far, then it gets reparsed (vowels as nuclei, consonants as onsets or codas). This last operation belonged to insertion (lengthening the candidate) in the previous subsection.

Finally, we can “cheat” again by allowing some *post-processing*, that is some greater changes in the string, larger steps in the search space that improve the word. Let $P_{postproc}$ denote again the probability of these changes, whenever possible. These steps help the system to escape from trivial local minima. In our case, we have considered deleting 0[_]N[_] substrings (whole epenthetical syllables), furthermore, contracting adjacent epenthetical nuclei with underparsed vowels, as well as epenthetical onsets with underparsed consonants. The right panel of Table 7.2 shows how increasing $P_{postproc}$ has improved our results.

After having defined the *basic step*, applying simulated annealing is straightforward. The algorithm presented in Figure 2.8 requires some values for T_{max} , T_{min} and T_{step} . Now, we used 4, 0 and 0.1 respectively.

Note that $P_{reparse}$ and $P_{postproc}$ are parameters of the model that determine the *a priori* probabilities of the topology. The left panel on Table 7.2 reports on some experiments on the role of parameter $P_{reparse}$. Significant difference was found only for the highest $P_{reparse}$ values. Seemingly, a low value for this parameter does not affect the results too much, because a number of insertions and deletions, involving also some forced reparses (in order to make the string a valid word), can replace reparsing. However, increasing $P_{reparse}$ too much will prevent the model from performing the insertions and deletions required in the optimal candidate.

Finally, one may ask whether precision depends on the hierarchy, besides—as we have seen—parameters of the topology (such as $P_{reparse}$ or $P_{postproc}$) and of the cooling schedule (such as T_{step} or K_{max}).

Therefore, simulated annealing was run for the input **anta**, with parameters $P_{reparse} = 0.60$ and $P_{postproc} = 0.95$, and with all the possible 120 rankings. By comparing the outputs to the correct output produced by a finite state technique (Gerdemann and van Noord, 2000), 17 out of the 20 outputs for ranking $\text{PARSE} \gg \text{FILLONSET} \gg \text{NoCODA} \gg \text{FILLNUCLEUS} \gg \text{ONSET}$ were found correct, whereas only 3 for $\text{PARSE} \gg \text{FILLONSET} \gg \text{FILLNUCLEUS} \gg \text{ONSET} \gg \text{NoCODA}$. For more results, see Tables 7.3 and 7.4.

Ranking	%
prs fio fin noc ons	31
prs fio fin ons noc	26
prs fio noc fin ons	78
prs fio noc ons fin	84
prs fio ons fin noc	14
prs fio ons noc fin	72
prs fin fio noc ons	38
prs fin fio ons noc	25
prs fin noc fio ons	30
prs fin noc ons fio	23

Table 7.3: Percentage of correct outputs for different rankings, out of 100 simulations for each.

7.2.4 Conclusion

The present chapter aimed at realising the potential in the model introduced in section 6.5. This model had assigned equal *a priori* probabilities to tangential and radial moves, and within the second, to centripetal and to centrifugal moves ($P_{tangential} = 0.5$, $P_{centripetal} = 0.25$, $P_{centrifugal} = 0.25$).¹¹ Subsection 7.1.3 increased the chance of centrifugal moves, whereas the present section demonstrated the role of $P_{tangential}$, called here $P_{reparse}$. As mentioned, not all possibilities have been tried out yet, for only $P_{centripetal} = P_{centrifugal}$ has been considered in the present section.

A further connection between the models in sections 7.1 and 7.2 is the high number of local optima. However, it was exactly our goal to have the algorithm be stuck in them in section 7.1, and thereby to account for the observed pause following the definite article in Hungarian; whereas these local optima are not attested in speech in the case of CV-Theory. Thus, we had to introduce some post-processing steps, that is, to enlarge the set of neighbours of some candidates, in order not to make them local optima. Interestingly, these post-processing steps are not reversible, so the new neighbourhood relation is not symmetric any more. The role of parameter $P_{postproc}$, determining the chance of applying these post-processing steps, has also been analysed.

¹¹As mentioned in footnote 17 on page 181, one should reconsider the *a priori* probabilities for unepenthesised candidates in order to fully meet this definition.

Ranking	Correct output	%	%	Ranking	Correct output	%	%
prs fio fin noc ons	N[a]D[n]O[t]N[a]	31	33	fin noc prs ons fio	O[]N[a]X[n]O[t]N[a]	56	47
prs fio fin ons noc	N[a]D[n]O[t]N[a]	26	24	fin noc fio prs ons	N[a]X[n]O[t]N[a]	64	60
prs fio noc fin ons	N[a]O[n]N[]O[t]N[a]	78	76	fin noc fio ons prs	X[a]X[n]O[t]N[a]	46	48
prs fio noc ons fin	N[a]O[n]N[]O[t]N[a]	84	77	fin noc ons prs fio	O[]N[a]X[n]O[t]N[a]	19	19
prs fio ons fin noc	N[a]D[n]O[t]N[a]	14	18	fin noc ons fio prs	X[a]X[n]O[t]N[a]	35	22
prs fio ons noc fin	N[a]O[n]N[]O[t]N[a]	72	75	fin ons prs fio noc	O[]N[a]D[n]O[t]N[a]	38	39
prs fin fio noc ons	N[a]D[n]O[t]N[a]	38	32	fin ons prs noc fio	O[]N[a]D[n]O[t]N[a]	38	30
prs fin fio ons noc	N[a]D[n]O[t]N[a]	25	28	fin ons fio prs noc	X[a]X[n]O[t]N[a]	13	11
prs fin noc fio ons	N[a]D[n]O[t]N[a]	30	27	fin ons fio noc prs	X[a]X[n]O[t]N[a]	15	25
prs fin noc ons fio	O[]N[a]D[n]O[t]N[a]	23	22	fin ons noc prs fio	O[]N[a]X[n]O[t]N[a]	22	19
prs fin ons fio noc	O[]N[a]D[n]O[t]N[a]	22	19	fin ons noc fio prs	X[a]X[n]O[t]N[a]	32	23
prs fin ons noc fio	O[]N[a]D[n]O[t]N[a]	27	12	noc prs fio fin ons	N[a]O[n]N[]O[t]N[a]	39	38
prs noc fio fin ons	N[a]O[n]N[]O[t]N[a]	77	84	noc prs fio ons fin	N[a]O[n]N[]O[t]N[a]	37	42
prs noc fio ons fin	N[a]O[n]N[]O[t]N[a]	80	78	noc prs fin fio ons	N[a]O[n]N[]O[t]N[a]	31	29
prs noc fin fio ons	N[a]O[n]N[]O[t]N[a]	76	76	noc prs fin ons fio	O[]N[a]O[n]N[]O[t]N[a]	29	32
prs noc fin ons fio	O[]N[a]O[n]N[]O[t]N[a]	59	47	noc prs ons fio fin	O[]N[a]O[n]N[]O[t]N[a]	29	28
prs noc ons fio fin	O[]N[a]O[n]N[]O[t]N[a]	57	48	noc prs ons fin fio	O[]N[a]O[n]N[]O[t]N[a]	19	28
prs noc ons fin fio	O[]N[a]O[n]N[]O[t]N[a]	65	45	noc fio prs fin ons	N[a]O[n]N[]O[t]N[a]	34	33
prs ons fio fin noc	O[]N[a]D[n]O[t]N[a]	19	16	noc fio prs ons fin	N[a]O[n]N[]O[t]N[a]	34	36
prs ons fio noc fin	O[]N[a]O[n]N[]O[t]N[a]	65	46	noc fio fin prs ons	N[a]X[n]O[t]N[a]	44	38
prs ons fin fio noc	O[]N[a]D[n]O[t]N[a]	19	15	noc fio fin ons prs	X[a]X[n]O[t]N[a]	34	35
prs ons fin noc fio	O[]N[a]D[n]O[t]N[a]	21	18	noc fio ons prs fin	X[a]O[n]N[]O[t]N[a]	33	31
prs ons noc prs fin	O[]N[a]O[n]N[]O[t]N[a]	56	42	noc fio ons fin prs	X[a]X[n]O[t]N[a]	34	20
prs ons noc fin fio	O[]N[a]O[n]N[]O[t]N[a]	61	38	noc fin prs fio ons	N[a]X[n]O[t]N[a]	38	34
fio prs fin noc ons	N[a]D[n]O[t]N[a]	39	35	noc fin prs ons fio	O[]N[a]X[n]O[t]N[a]	36	22
fio prs fin ons noc	N[a]D[n]O[t]N[a]	32	39	noc fin fio prs ons	N[a]X[n]O[t]N[a]	44	40
fio prs noc fin ons	N[a]O[n]N[]O[t]N[a]	46	57	noc fin fio ons prs	X[a]X[n]O[t]N[a]	34	35
fio prs noc ons fin	N[a]O[n]N[]O[t]N[a]	47	47	noc fin ons prs fio	O[]N[a]X[n]O[t]N[a]	11	14
fio prs ons fin noc	N[a]D[n]O[t]N[a]	32	19	noc fin ons fio prs	X[a]X[n]O[t]N[a]	13	18
fio prs ons noc fin	N[a]O[n]N[]O[t]N[a]	45	45	noc ons prs fio fin	O[]N[a]O[n]N[]O[t]N[a]	14	5
fio fin prs noc ons	N[a]D[n]O[t]N[a]	32	39	noc ons prs fin fio	O[]N[a]O[n]N[]O[t]N[a]	10	9
fio fin prs ons noc	N[a]D[n]O[t]N[a]	34	37	noc ons fio prs fin	X[a]O[n]N[]O[t]N[a]	14	17
fio fin noc prs ons	N[a]X[n]O[t]N[a]	48	35	noc ons fio fin prs	X[a]X[n]O[t]N[a]	17	13
fio fin noc ons prs	X[a]X[n]O[t]N[a]	30	28	noc ons fin prs fio	O[]N[a]X[n]O[t]N[a]	14	9
fio fin ons prs noc	X[a]X[n]O[t]N[a]	15	20	noc ons fin fio prs	X[a]X[n]O[t]N[a]	13	11
fio fin ons noc prs	X[a]X[n]O[t]N[a]	33	23	ons prs fio fin noc	O[]N[a]D[n]O[t]N[a]	17	16
fio noc prs fin ons	N[a]O[n]N[]O[t]N[a]	34	38	ons prs fio noc fin	O[]N[a]O[n]N[]O[t]N[a]	11	14
fio noc prs ons fin	N[a]O[n]N[]O[t]N[a]	34	27	ons prs fin fio noc	O[]N[a]D[n]O[t]N[a]	20	12
fio noc fin prs ons	N[a]X[n]O[t]N[a]	42	29	ons prs fin noc fio	O[]N[a]D[n]O[t]N[a]	13	27
fio noc fin ons prs	X[a]X[n]O[t]N[a]	29	24	ons prs noc fio fin	O[]N[a]O[n]N[]O[t]N[a]	7	4
fio noc ons prs fin	X[a]O[n]N[]O[t]N[a]	24	25	ons prs noc fin fio	O[]N[a]O[n]N[]O[t]N[a]	11	8
fio noc ons fin prs	X[a]X[n]O[t]N[a]	34	34	ons fio prs fin noc	X[a]O[n]N[]O[t]N[a]	9	8
fio ons prs fin noc	X[a]O[n]N[]O[t]N[a]	35	35	ons fio prs noc fin	X[a]O[n]N[]O[t]N[a]	16	10
fio ons prs noc fin	X[a]O[n]N[]O[t]N[a]	27	26	ons fio fin prs noc	X[a]X[n]O[t]N[a]	17	7
fio ons fin prs noc	X[a]X[n]O[t]N[a]	17	10	ons fio fin noc prs	X[a]X[n]O[t]N[a]	12	11
fio ons fin noc prs	X[a]X[n]O[t]N[a]	38	38	ons fio noc prs fin	X[a]O[n]N[]O[t]N[a]	16	8
fio ons noc prs fin	X[a]O[n]N[]O[t]N[a]	26	23	ons fio noc fin prs	X[a]X[n]O[t]N[a]	8	8
fio ons noc fin prs	X[a]X[n]O[t]N[a]	31	25	ons fin prs fio noc	O[]N[a]D[n]O[t]N[a]	22	22
fin prs fio noc ons	N[a]D[n]O[t]N[a]	57	56	ons fin prs noc fin	O[]N[a]D[n]O[t]N[a]	20	17
fin prs fio ons noc	N[a]D[n]O[t]N[a]	58	56	ons fin fio prs noc	X[a]X[n]O[t]N[a]	15	8
fin prs noc fio ons	N[a]D[n]O[t]N[a]	64	58	ons fin fio noc prs	X[a]X[n]O[t]N[a]	16	8
fin prs noc ons fio	O[]N[a]D[n]O[t]N[a]	67	55	ons fin noc prs fio	O[]N[a]X[n]O[t]N[a]	14	15
fin prs ons fio noc	O[]N[a]D[n]O[t]N[a]	52	50	ons fin noc fio prs	X[a]X[n]O[t]N[a]	10	8
fin prs ons noc fio	O[]N[a]D[n]O[t]N[a]	57	65	ons noc prs fio fin	O[]N[a]O[n]N[]O[t]N[a]	19	7
fin fio prs noc ons	N[a]D[n]O[t]N[a]	58	57	ons noc prs fin fio	O[]N[a]O[n]N[]O[t]N[a]	10	10
fin fio prs ons noc	N[a]D[n]O[t]N[a]	54	65	ons noc fio prs fin	X[a]O[n]N[]O[t]N[a]	12	16
fin fio noc prs ons	N[a]X[n]O[t]N[a]	63	53	ons noc fio fin prs	X[a]X[n]O[t]N[a]	5	7
fin fio noc ons prs	X[a]X[n]O[t]N[a]	54	37	ons noc fin prs fio	O[]N[a]X[n]O[t]N[a]	14	16
fin fio ons prs noc	X[a]X[n]O[t]N[a]	23	21	ons noc fin fio prs	X[a]X[n]O[t]N[a]	10	8
fin fio ons noc prs	X[a]X[n]O[t]N[a]	48	37				
fin noc prs fio ons	N[a]X[n]O[t]N[a]	63	55				

Table 7.4: For the input **anta**, the number of correct outputs for different rankings, out of 100 runs. In each hierarchy, the highest ranked constraint appears first. The third column shows the results with parameters $P_{\text{reparse}} = 0.60$ and $P_{\text{postproc}} = 0.95$, while the results in the fourth column was obtained with parameters $P_{\text{reparse}} = 0.60$ and $P_{\text{postproc}} = 0.30$. The correct outputs were calculated using finite state techniques. The precision does not depend on the output either, for there are significant differences between rankings that yield the same optimal candidate.

It has been mentioned that implementing the CV Theory for syllabification allows us to compare the SA-OT Algorithm to other computational approaches to Optimality Theory. It should not be surprising by now that SA-OT has a lower computational complexity than most of its competitors, but in exchange, it has a lower precision. It is to be hoped that a new and more elegant topology for CV Theory will be able to increase precision and account for fast speech phenomena, such as dropping of syllables or syllable parts.

Indeed, SA-OT is polynomial in time: quadratic in the number of constraints, and approximately cubic in the length of the word. Evaluating candidates is linear in the number of constraints, and linear in the length of the input word (however, as candidates blow up with inserted materials, the evaluation becomes longer). The more constraints, the higher K_{max} , i.e. the interval traversed by T . Furthermore, the longer the word, the larger the region to walk through, i.e. the more insertions and deletions to try out. A good estimate for the number of steps required is proportional to the square of the distance we want to walk.

How can one make use of simulated annealing, if it sometimes returns the correct output only in 20% of the cases? One can run more simulations in parallel, and then choose the output returned the most often. This solution would work if the erroneous outputs have an even lower probability, which is true only in part of the SA-OT models.

Another possibility is to compare the few different outputs obtained by parallel simulations, using the classical OT evaluation methods. Although this solution seems to returning to the original methods, it is not the case, for now we only need to compare a few candidates, and not the entire boundless candidate set. In fact, if our algorithm returns the optimal candidate only with a probability of 20%, but we run it ten times, then the optimal candidate will be returned at least once with a chance of almost 90%. However, once the optimal candidate appears in the output set, it will win using traditional tableau comparison of the ten outputs. Running 20 parallel simulations will increase the precision to almost 99%.

It is also possible to combine the previous two solutions: run many simulations in parallel, choose the most frequent outputs, and compare them using a tableau.

If SA-OT is indeed an adequate model of language production, then the observation about the differences in precision across different rankings also has an important consequence for OT's claim on *factorial typology*. The traditional approach in a Chomskyan style is that attested and non attested language types result from what the human brain (the Universal Grammar) is able or is unable to realise—that is, the idea of *factorial typology* in Optimality Theory. Jäger (2003a) has proposed however that some gaps in factorial typology (i.e., a language type predicted by some constraint ranking, but not attested among the languages of the world) may be explained by it being unstable during evolution (across generations). Boersma (2004a) has raised a second option: some constraint rankings may turn out to be not learnable. Now, we may add a third, even more trivial possibility: some constraint rankings might not appear in attested languages, just because they are not producible. That is, SA-OT has only a very low chance to find the correct output. From the observation that some constraint rankings are much more advantageous than other ones, we predict that hard-to-compute rankings are less likely to be attested in the

languages of the world.

What happens if the global optimum is relatively hard to reach (because it is located in a narrow valley), whereas some other local optima are returned relatively often? If this is the case for most words, we expect this language not to be stable: the next generation will learn another ranking. But if this phenomenon happens only to a few words, we may predict the surfacing of the sub-harmonic form, seen as either irregularity, or free alternation, or slip of the tongue.

Chapter 8

Conclusion: Is Simulated Annealing Optimality Theory, thus, Better?

8.1 Summary

This dissertation aimed at introducing a new variant of *Optimality Theory*, namely, the *Simulated Annealing for Optimality Theory Algorithm* (SA-OT).

After having overviewed existing variants and the “philosophical” background of OT in Chapter 1, Chapter 2 motivates the use of heuristic optimisation algorithms—such as simulated annealing—and then introduces the SA-OT Algorithm (Fig. 2.8, on page 64). The main argument for simulated annealing was that it is a plausible model for the “implementation” of language in the brain: it is fast, efficient, does not require large computational power, but makes *certain* mistakes, the ratio of which increases with production rate. Therefore, SA-OT could be used as a model of some aspects of performance in phonology.

Subsequently, Chapter 3 introduced some formal approaches to OT in order to underpin the SA-OT Algorithm. Both polynomials and ordinal numbers were introduced for that purpose. The following chapter, a set of open questions more than full-fledged proposals, points to related linguistic issues—such as the role of the lexicon and learnability—that should be elaborated in the future. It also introduced a new definition for Output-Output Correspondence.

The remaining chapters present several applications. The goal of these chapters is not so much linguistic, but methodological. Even though I tried to argue for the linguistic well-foundedness of the models, more cooperation with fellow linguists might have been useful here and there to reach an analysis which might withstand linguistic criticism. It is only to be hoped that the model will arouse enough interest among general linguists to help improve these models. More importantly, I urge experimental linguists to provide quantitative experimental data so that the predicted frequencies of new models can be tested in the future against empirical results. Nevertheless, these chapters have hopefully illustrated the methodological issues arising if one decides to use SA-OT as the framework for a linguistic (performance) model.

Indeed, SA-OT is a complex model involving many parameters and many decisions to be made, such as the definition of the topology, the choice of the constraints and their indices (their association with the domains of temperature), and so on. Consequently, it offers the possibility for tuning at many different points. Some may even argue that there are *too many* such points. To this criticism my answer is threefold.

First, SA-OT's aim is to account for a complex quantitative data set (the frequencies of different forms in different conditions), hence the complexity of the model. If the model is simpler than the data to be described, then the model has explained something from the observable complexity. There is an indication that such a reduction in complexity happens when the model correctly accounts for something that has not been aimed at originally: for instance, when the model in Chapter 5 was tuned to return the correct *andante* and *allegro* forms, but it also turned out to correctly predict which word type is more likely to change in fast speech. Second, practice has demonstrated that the high number of parameters does not trivialise the task of tuning the model, and finding a correct model is far from being a sinecure. It is not the case that just “anything” can be reproduced simply using SA-OT. Third, the parameters are restricted by further guidelines. The topology and the constraints should be cross-linguistically universal and well-founded, so they cannot be defined in an *ad hoc* way.¹ Varying certain parameters (typically T_{step}) can and should be interpreted as varying the run time of the algorithm (the speed of speech production), whereas varying other parameters (e.g., those related to the *a priori* probabilities) may not have such an interpretation. If the variation depends on the frequency of the word (rare content words being pronounced more carefully than frequent function words), one may tune the parameters of the cooling schedule again, while the *a priori* probabilities of the topology, I hypothesise, accounts for differences among speakers only, since a certain speaker does not alter his or her topology.

In particular, Chapter 5 works out a model accounting for Dutch stress assimilation in normal and fast speech, and thereby analyses the role of parameters T_{step} , T_{max} and T_{min} . Varying the former is the simplest and probably the most straightforward tool for reproducing fast speech, whereas the later two also have a slight influence on the output frequencies. Besides, this chapter also analysed the role of the definition of the constraint Output-Output Correspondence, employing what had been introduced in the previous chapter.

If Chapter 5 focuses on parameters T_{step} , T_{max} and T_{min} , then Chapter 6 and section 7.1 add parameters K_{step} , K_{max} and K_{min} to the analysis. Here, unlike in traditional OT, candidates that can never win and constraints that are vacuously satisfied may significantly interfere with a model's output frequencies. In particular, section 6.5 presents a model—followed by a mathematical discussion and experiments—that relies heavily on parameter K_{max} in addition to T_{step} . Due to the infinity of the search space, a larger initial stage in the simulation enhances the “channelling effect”. A similar model appears in section 7.1,

¹Many readers have not been convinced by my arguments for certain topologies being a natural choice in the particular case. Future work should therefore either proliferate the number of phenomena that require a certain topology, so that the choice becomes an unquestionable necessity; or some general principles should determine the topology. For instance, Gerhard Jäger has proposed to connect the neighbourhood structure to a psycholinguistic notion of similarity, whereas Adam Albright has suggested Steriade's P-Map.

which ends by remarking that the different behaviour of the two allomorphs of the Hungarian article can be tuned by varying the indices (domains of temperature) with which the relevant constraints are associated. Introducing empty domains (or inactive constraints) between these two constraints lengthens the period in which temperature is located between these constraints, thereby making the divergence in the behaviour of the two allomorphs more pronounced. This technique is related to the way the first component of the temperature is diminished in the outer loop of the SA-OT Algorithm, hence also an observation on the role of K_{step} . Finally, the same model makes parameter K_{min} dispensable by measuring the “specific heat”: the algorithm runs until the random walker has not moved for 30 consecutive iterations.

An additional morale of the first sections in Chapter 5 is that—unlike standard simulated annealing and the SA-OT models presented in earlier chapters—some SA-OT models do not converge to maximal precision if the number of iterations is increased. This remark has also opened some speculation about how to account for linguistic irregularities by using a simple grammar together with an algorithm that is not always correct but which makes predictable errors.

Finally, Chapter 7 (especially subsection 7.1.3 and section 7.2) brings another parameter to our attention, namely, the definition of the topology (the neighbourhood structure). We demonstrate how changing the parameters of the *a priori* probabilities influences the output frequencies. The issue turns more important as the candidate set becomes larger (infinite), and as candidates are assigned a larger number of neighbours.

In what follows, we return to the assessment of the OT variants in section 1.3, and ask the question: is SA-OT any better?

8.2 Advantages (and disadvantages) of SA-OT

8.2.1 SA-OT and specific linguistic phenomena

Arguments for some approach and against other ones can be of three different sorts. People often present cases where a given model is unable to account for some phenomenon. Second, one may show that an approach is *in general* unable to come to grips with some aspects of the explanandum. Finally, one might formulate “philosophical” preferences and theoretical expectations not matched by that approach. As an example, the reader is referred to Keller and Asudeh (2002)’s criticism of Boersma and Hayes (2001)’s Stochastic OT, replied to by Boersma (2004b).

Concerning the first sort of criticism, I refer to tableau (5.4) on page 128. It shows that for both Types 0 and 2 in Dutch stress assignment, all possible parses of the observed fast speech forms are harmonically bounded, so that therefore, the loser forms cannot win for any hierarchy. This fact could be a counter-argument for all approaches that wish to generate the alternative forms with constraint reranking (such as an *ad hoc* reranking, Anttila’s proposal or Boersma’s Stochastic OT). The same tableau demonstrates why Coetzee’s approach would fail: an attested alternative form violates the highest ranked constraint, hence the critical cut-off point must be above this highest ranked constraint, therefore all candidates should be attested. At the same time, I could argue that Simulated Annealing Optimality Theory does the job nicely.

Although this type of argument may support a certain approach, and can help it become more popular, it is certainly not decisive. Namely, nothing shows that using different constraints would not work within the alternative approaches (cf. e.g., Boersma, 1998a). Add a new constraint to the top of the hierarchy, and the train of thought holds no longer. Keeping in mind that the set of constraints is supposed to be universal across languages, even while it varies across linguists and papers, we have been shown only that we have not been creative enough.

Similar remarks apply to any argument demonstrating that a certain SA-OT model is not able to reproduce a certain phenomenon: in which the output frequencies do not match the experimentally observed frequencies, or in which the local optima are not exactly the attested alternating forms. An example for both was actually the case of Type 2 words—such as *perfectionist*—in Table 5.17 on page 155. But such a fact is not an argument against SA-OT in general, for nothing proves that different rankings, different constraints and different neighbourhood relations would have no chance to work either. This failure only points to the need for future work. Only the repetitive failure of a model and the lack of success might slowly motivate the linguistic community to drop it and to adopt different approaches.

Therefore, we now turn to the second type of arguments. A number of observations show that several linguistic phenomena—such as Dutch stress assignment or the behaviour of the Hungarian article—are about a gradual shift of frequencies in function of certain parameters (speech rate, sociolinguistic parameters, and so forth). This fact is a serious argument against Coetzee’s approach, who refuses to predict the frequencies quantitatively, as well as against Anttila’s proposal, which requires a very different grammar in order to approximate a slightly different frequency distribution (as mentioned by Boersma and Hayes, 2001). Nonetheless, Stochastic OT, MaxEnt OT and SA-OT make it possible to vary the output frequencies as a function of external parameters.

Our SA-OT experiments on the definite article in Hungarian showed how simply the frequency of the most harmonic candidate can be fine-tuned between 0% and 100%. For instance, in Fig. 7.3 (page 201), $K_{max} = 20$ or 40 never returned the globally optimal form [az#E], whereas $K_{max} = 3$ would have done it with a frequency close to 100%. On the other hand, remember that in section 6.5 the channelling effect could also enforce the alternate form using a slightly modified tableau: then the global optimum can never be produced in more than half of the cases. Hence, the framework of SA-OT does not restrict the possibilities in a purely categorical fashion. As language data do not seem to be restricted too strongly, either, a strong prediction would be an argument against a certain approach. Indeed, Stochastic Optimality Theory requires the interplay of at least three, almost equally ranked constraints, otherwise it predicts that the grammatical form must have a probability exceeding 50%. While this would seem to count against it, only time can decide which of the two approaches fits better all kinds of linguistic phenomena.

Additionally, the parameter that determines the output frequencies can often be simply interpreted in SA-OT, because it is directly related to the algorithm’s run time. Therefore, fast speech phenomena indeed emerge in a speeded up algorithm. In other cases, nonetheless, similarly to the parameters determining the output frequencies in Stochastic OT and MaxEnt OT, the connection is not so obvious: why would for instance sociolinguistic factors or word frequency

(familiarity) influence K_{max} ?²

A further argument can be brought in favour of SA-OT based on the Dutch fast speech data, which, again, is not decisive, for different constraints might do the job within Stochastic OT, as well. As mentioned, the empirical data display a huge difference in the behaviour of different word types. Output-Output Correspondence (Output-Output Faithfulness) is able to account for the different winning *forms*, but is it also able to account for the different *frequencies*? In Stochastic OT, the frequency of the alternating form was derived from the probability of reranking the constraints at evaluation time. Thus, we have to postulate that different morphological types either employ a different evaluation noise (why?); or associate Output-Output Correspondence with a slightly different rank. The latter possibility is not absurd, for OOC is an odd constraint anyway, and its rank might depend on an argument, the reference string.³ But how? On the other hand, SA-OT predicted correctly which form is more likely to be mispronounced—a surprising result, since we had not created our model with this goal in mind. The reason why different word types result in different frequencies in SA-OT is that OOC alters the landscape, due to which the local optima have a catchment area (the basin from which rain flows into a particular river) of different size for different inputs. Again, the candidates not appearing on the surface heavily influence the output frequencies.

In brief, while Coetzee’s proposal explicitly rejects accounting for quantitative phenomena that SA-OT can explain, Anttila’s approach is unable to cope with them, but Boersma’s Stochastic OT, similarly to MaxEnt OT, is theoretically able to face them. Nonetheless, there are certainly cases where competitors of SA-OT could turn out to be more convincing.

For instance, it seems to be a coincidence for SA-OT that fast speech prefers the forms that are phonologically less marked, and slower speech is more faithful to morphology. In the constraint reranking approaches, however, these intuitive observations become *the* explanation of the phenomenon. SA-OT’s replies that it is exactly the neighbourhood structure and the “landscape” that explain *why* faithfulness becomes less important and markedness more significant in fast speech: if faithfulness is ranked higher than markedness, then the faithful global optimum seems to be difficult to find in fast speech, and less faithful but unmarked local optima may be returned more easily.

In general, however, a major disadvantage of Simulated Annealing Optimality Theory—compared to its competitors—is that it is hard to understand exactly why it works in certain cases. Developing an exact analysis of SA-OT’s behaviour is difficult even for relatively simple landscapes. The interactions between SA-OT’s components (the neighbourhood structure, the constraint hierarchy and the algorithm’s parameters) are so complex that the success or

² K_{max} is not necessarily to be interpreted as being connected to the speed of the algorithm. Its role is to determine the length of the initial phase of the simulation, in case the simulation is launched from the same one or few candidates. If the initial candidate is chosen from a wider pool, however, then the initial phase can be omitted. Therefore, different observed frequencies can be reproduced by changing the way the initial candidate is chosen, while K_{max} (hence, run time) is kept constant. In other words, the random walk in the initial phase can be viewed as not belonging to the SA-OT algorithm, but as a way to chose the initial candidate from this wider pool. Then, K_{max} is a parameter that determines the choice of the candidate from where (the most interesting part of) the algorithm is launched.

³In other words, a whole family of OOC constraints should exist, with each member being associated with a slightly different rank, and each member acting upon a different word type.

failure of a model cannot be simply predicted using paper-and-pen linguistics, without implementing the simulation on computers.⁴ (For an extreme case, recall Table 5.10 on page 151 where the frequency of a local optimum surprisingly increased as we diminished T_{step} .) This is the reason why the reader is welcome to try out the SA-OT demo page at <http://www.let.rug.nl/~birot/sa-ot/>.

8.2.2 SA-OT, competence and performance

Finally, let us turn to the third level of criticism against various approaches, that is, to “more philosophical issues”. In particular, we shall ask how SA-OT, as well as its competitor models reflect the traditional dichotomy of competence and performance.

Often, both forms A and B are equally grammatical. In other cases, however, the use of the two forms is not symmetrical, and one may argue for form A to be the “grammatical form”, while B is regarded as a “performance error”—without any value judgement. For instance, in Chapter 5 on Dutch stress assignment, we identified the grammatical form with the *andante* pattern, that is, whose frequency diminishes at higher speech rate. (Otherwise, one should claim that fast speech is more grammatical, which would be odd.) Therefore, we expect a linguistic model to predict which form is grammatical (whatever that means), as well as what other forms may also emerge under certain conditions.

Here I refer to the idea sketched on Fig. 2.1 (page 43), which replaces the competence-performance dichotomy with a three-level picture. Between the competence in-the-narrow-sense (the static linguistic knowledge encoded in one’s brain) and the performance in-the-narrow-sense (including all the extralinguistic factors influencing linguistic products, speech), one also finds the dynamic language production process. Phenomena that are traditionally reckoned to belong to performance, but are determined by linguistic factors, might be analysed on this intermediate level. Consequently, let us ask the various approaches how they distinguish the static knowledge of the language from the dynamic language production, and whether they see a difference between theoretically grammatical forms and forms observable, say, in a corpus.

MaxEnt OT assigns a positive probability to all candidates generated by GEN. Consequently, there is no chance to differentiate in a principled way between forms that are so agrammatical (even absurd) that they cannot be attested for sure, and forms attested, though only rarely—unless one restricts GEN in a language specific manner, or the assigned probabilities drop drastically at a certain point. Such a model might be most welcome in cases where all candidates are attested in a corpus, and only their frequencies require explanation, such as was the case in Jäger and Rosenbach (2006)’s model for English genitive constructions.

Assigning the same violation profile to several candidates, creating a new grammar by reranking some constraints by hand, and the unranked hierarchies of Anttila form the next group of models. They all account for linguistic variation on the level of the competence model, at the core of the OT architecture. These approaches are adequate when the alternating forms are not differentiated with respect to their grammaticality, which definitely is the case in certain forms of variation.

⁴Note that Lauri Karttunen has made the same remark on Optimality Theory in general in his FSMNLP talk in Helsinki in September 2005 (cf. also Karttunen, 2006).

Coetzee's model is different. It distinguishes sharply between the form that is grammatical, by making it the optimal output, on the one hand, and alternate forms, the second, third best candidates still emerging, on the other. His model can be, thus, used for phenomena displaying alternation of an arguably grammatical and most frequently occurring form with ungrammatical ones.

Even if it is not necessary, Boersma's Stochastic Optimality Theory may also be seen as reflecting the distinction between competence and performance. Namely, the unperturbed hierarchy can represent competence, in conformance with standard OT; whereas the noisy evaluation cycles are the model for the dynamic language production process. The form optimal for the unperturbed hierarchy is the grammatical form, whereas perturbations explain why other forms are also attested. Postulating, for instance, a larger σ in the evaluation noise at a higher speech rate will account for the increased frequency of the allegro form. The decoupling of the model's two levels is appealing, but it is unclear why noise should increase in fast speech.

I argue that competence and performance are most radically separated in Simulated Annealing OT. Similarly to the unperturbed hierarchy in Stochastic OT, the underlying traditional OT model accounts again for linguistic competence: the grammatical form is the (global) optimum. On top of that, however, we have introduced a separate search algorithm, which models the functioning brain during speech production. The search algorithm is computationally simple, arguably plausible: each time only one form has to be stored, which is then altered in an elementary way, supposing that this basic change does not incur too many extra violations. Not only in memory requirements is the algorithm a plausible model of the brain's functioning, but also in run time, which can be kept constant. Moreover, the precision of the algorithm (the probability of returning the grammatical form) depends also upon parameters that have a direct connection to speech rate: T_{step} can easily be argued to change in function of speech rate, since it directly influences the algorithm's run time.

SA-OT also competes with different implementations of OT whose goal is to find the optimal candidate in the candidate set. Even if SA-OT does not guarantee that one always finds the optimal candidate, it may still be more adequate in a cognitive sense than its competitors.

Indeed, I argue that simulated annealing is an adequate model for the computations in the human mind for several reasons. First of all, no severe restrictions must be made on GEN and on the constraints as in Finite State Optimality Theory (Eisner, 1997; Frank and Satta, 1998; Karttunen, 1998; Gerdemann and van Noord, 2000; Jäger, 2002; Bíró, 2003, 2005c; Karttunen, 2006). As described elsewhere (Bíró, 2005c), two approaches exist within Finite State Optimality Theory: either a new automaton must be built for each input, requiring huge computational power, or a finite state transducer maps any input to its optimal output, but this latter approach works in very restricted cases only.

Simulated annealing is an algorithm that may find the optimal candidate of a combinatorial problem in a reasonable time with a reasonable precision, even in the case of NP-complete problems, which Optimality Theory may pose (Eisner, 2000b). It does not require computational capacities as large as the ones finite-state approaches need, or even those genetic algorithms or chart-parsing may ask for. Simulated annealing produces *some* output within constant time, similarly to speech, where *something* must be produced within a specific time span—conversation partners are not computer users who are willing to watch

the hourglass! Furthermore, simulated annealing can be speeded up, just as human speech: in both cases the price to pay for shortening the time span is precision. In fast speech, indeed, we accept some ungrammatical forms for the sake of expressing ourselves more quickly. Hence, I argue again, a linguist may gain an explanation of fast speech phenomena, alternation forms or performance errors by adding some topology to the candidate set.

Paul Smolensky's implementation of Harmony Grammar (closely related to Optimality Theory) within a connectionist framework has already long included a notion of neighbourhood, whence follows the possibility of local optima existing in the system. Yet, his goal, which he successfully reaches, is to avoid allowing the system to find them.⁵ My approach, however, differs from his not only in that I turn the errors made by the algorithm (when the system gets stuck in local optima) into my advantage; but also in that mine is not confined to neural networks. So non-connectionist scholars—linguists, but also other cognitive scientists—may employ it.

8.3 SA-OT as a general cognitive model

More than a decade before 1993, the year when Optimality Theory appeared in linguistics, Seymour Papert (1980) proposed a remarkable cognitive model.

Imagine a child is shown the following experiment: the water is poured from a broader vessel into a narrower one, so the level of the water will be higher than it was in the original vessel. According to observations, children below the age of six or seven will tell you that the amount of liquid has increased, whereas above this age children suddenly change their mind and give an answer in conformance with the principle requiring that the amount of liquid be conserved. How to explain both answers and the switch between them?

Papert suggested the following model (Papert, 1980, p. 166f). Suppose there are (at least) three homunculi present in the brain of a person who has to compare quantities. Each of them works very simply, and the answer of the child is derived from the answers given by these homunculi.

The first of them, Papert proposes, judges the amount of anything according to its height. As objects in the world usually have more or less constant dimensional ratios, judging from the height should be reliable. After all, many important judgements in the world—for instance, the age, role, power and might of an unknown fellow human—can be made based on the other's height. According to Papert, this homunculus serves even the youngest children very well, when they have to distribute Coca Cola or hot chocolate equally among glasses.

The second homunculus relies on the horizontal dimensions. Papert writes that this homunculus is usually not as skilled as the first one, so he influences the judgements of the child much less frequently. He comes to a role in statements such as “there might be really much water in the sea”.

The third homunculus is called History, and teaches that “if two amounts were equal, then they remain equal”. It is a “folk” version of the principle of conservation of matter in physics. Even if we increased the amount of the water, this homunculus of Papert would come to this conclusion.

⁵Chapter 20, sections 3.7.4, 3.7.5 of an October 2004 print out of Smolensky and Legendre (2006), which was made available before the KNAW Masterclass “Cognitive Foundations of Interpretation” in Amsterdam.

Now, how to account for the answers of the younger children, and for those of the older ones? When the younger child is asked the question whether the amount of water has changed while pouring it from one vessel to the other, the first homunculus, the vertically minded one, will give the answer.

Concerning the older child, Papert offers three explanations. Either the first two homunculi become more “sophisticated”, so that they interfere only if everything else remains unchanged: for instance, the first homunculus learns to have an opinion only if the width of the object is the same. The second explanation—the most interesting one from the point of view of the developments in linguistics ten years after Papert’s book was published—proposes a reordering in the relative prominence (he calls it the “seniority”) of the homunculi: homunculus History suddenly jumps forward, and becomes the “dominant voice”. Does not this idea remind you of OT learning algorithms? Papert’s third possible answer introduces a fourth homunculus from the critical age onwards that combines the answers given by the first two homunculi (the geometrical ones), so this fourth homunculus will cancel their contradictory opinions.

We can summarise and reinterpret Papert’s model in the following way. In order to solve a cognitive task, the human brain invites several of its “modules” to give some answer. (Modularity of the brain was probably in the air already in 1980.⁶) Out of the pool of possibilities, each module picks one, and returns it as the solution. These modules work in very simple ways, and would quite often mislead the brain if they had to work alone. However, the interaction of them (unspecified by Papert, although he already alludes to some hierarchy in importance among them) results in a cognitive capacity leading to an evolutionarily successful behaviour. The brain does not necessarily return the mathematically exact solution always, and yet, even with such an “imperfect human logic”, humanity has been reasonably successful.

Indeed, this kind of Optimality Theory as a general cognitive strategy, together with simple heuristics, such as “take the higher”, “take the wider”, “quantities do not change”, form the building blocks of the *Heuristic-and-Biases Program* launched by Tversky and Kahneman (1974), and of the *ABC Research Project* (Gigerenzer et al., 1999).

Gigerenzer et al. (1999, p. 24-25) summarise the *ABC Research Project* with the following words:

The research program [...] is designed to elucidate three distinct but interconnected aspects of rationality [...]:

1. *Bounded rationality.* Decision-making agents in the real world must arrive at their inferences using realistic amounts of time, information, and computational resources. We look for inference mechanisms exhibiting bounded rationality by designing and testing computational models of fast and frugal heuristics and their psychological building blocks. The building blocks include heuristic principles for guiding search for information or alternatives, stopping the search, and making decisions.

⁶The reader who would like to argue against the strong modularity of the brain in the sense of Jerry Fodor, is welcome to replace the term “module” used here with something like “basic computational unit”, probably smaller ones than those argued for by the proponents of the modularity of the brain.

2. *Ecological rationality.* Decision-making mechanisms can exploit the structure of information in the environment to arrive at more adaptively useful outcomes. To understand how different heuristics can be ecologically rational, we characterize the ways information can be structured in different decision environments and how heuristics can tap that structure to be fast, frugal, accurate, and otherwise adaptive at the same time.
3. *Social rationality.* The most important aspects of an agent's environment are often created by the other agents it interacts with. [...] Social rationality is a special form of models of fast and frugal heuristics that exploit the information structure of the social environment to enable adaptive interactions with other agents. [...]

These three aspects of rationality look toward the same central goal: to understand human behavior and cognition as it is adapted to specific environments (ecological and social), and to discover the heuristics that guide adaptive behavior.

Typical examples for the “fast and frugal heuristics” used in the *ABC Research Program* include for instance: “if one of two objects is recognized and the other is not, then infer that the recognized object has the higher value” (which of the two cities mentioned is larger?, Gigerenzer et al., 1999, p. 41) or “feed your children from youngest to oldest” (Gigerenzer et al., 1999, p. 314). The claim is that modelling decision-making using such heuristics is a cognitively adequate description of the human mind, on the one hand; and that such heuristics are computationally simple, and yet efficient techniques, on the other.

The key to the success of such heuristics is the structure of the world: the structure of the information, of the society, of communication, and so on. In Todd's words: “[i]n our program, we see heuristics as the way the human mind can take advantage of the structure of information in the environment to arrive at reasonable decisions, and so we focus on the inferences” (p. 28).

Still, nothing guarantees avoiding errors. If human mind makes decision based on such heuristics, then... *errare humanum est!* But the interpretation of these errors had changed much in 25 years: what was seen by the heuristics-and-biases program (Tversky and Kahneman, 1974) as a hindrance to sound reasoning (“rendering *Homo sapiens* not so sapient”, Gigerenzer et al., 1999, p. 29), is perceived by the *ABC Research Group* rather as “enabling us to make reasonable decisions and behave adaptively in our environment—*Homo sapiens* would be lost without them” (*ibid*).

In the context of the cognitive research lines just described, Prince and Smolensky (1993)'s *Optimality Theory* can be seen as a concrete case for the specific cognitive subfield of language. OT constraints are similar heuristics, that is, simple rules to evaluate the possibilities (candidates): “take the one with the least codas”, “take the one with the most onsets”, “take the one with the least epenthetic segments”, and so forth.⁷ Finding the most harmonic candidate (“take the best” for the ABC Research Group) is performed in OT with

⁷Here I leave open the question whether “simplicity” is also meant in computational terms, in the form of requirements that, for example, constraints must be “primitive”, finite-state friendly (Eisner, 1997; B    , 2003). Even constraints that do not meet these requirements are immensely simpler than grammars themselves.

respect to the *lexicographic strategy*, one of the three possibilities besides the linear model and the classification trees (Gigerenzer et al., 1999, p. 136-139). Cognitive research on decision making in general, therefore, corroborates the use of Optimality Theoretical grammars, and the emergence of OT-style linguistic systems during the evolution of general cognitive skills becomes more plausible.

A major difference still remains, however. In Optimality Theory, the candidate optimising the constraints is *by definition* the solution sought, whereas “fast and frugal heuristics” aim only at *approximating* the solution of the complex problem posed to the cognitive faculties (by seeking “good (i.e. near-optimal) solutions at a reasonable computational cost”, Reeves, 1995, p. 6). Indeed, “fast and frugal heuristics” are employed to answer problems under time pressure and when information may be incomplete. In a different study, Gary Klein reports that fireground commanders make around 80 percent of their decisions in less than one minute, sometimes within a few seconds, whereas chess players under blitz conditions make a move in average in six seconds Klein (1999, p. 4).

This difference between Optimality Theory and its cognitive background, “fast and frugal heuristics” can be explained, though. Suppose that (OT-style) language evolved from the heuristic inference system, and used its architecture to define the rules of linguistic communication. Unlike in the case of a question such as “which amount of water / which city is larger”, however, there is no *a priori* uniquely good solution to the problem how to encode a thought into a utterance. Hence, the system that had had only limited precision when solving cognitive problems, could now perfectly encode the rules of language—just because these rules⁸ were formulated in terms of this system.

And yet, this new system of communication did not work so perfectly—maybe due to the proliferation of meanings to be expressed, leading to the proliferation of lexical items and possible structures. Here came a second level of heuristics into play, at least according to the main claim of my thesis. Even though the best candidate sought after is well defined in terms of the constraints, still, it is not always possible to find it at production time. Now we move from the first level to the second level on Table 2.1. Once the language community has accepted a form as the grammatical (the optimal) one, locating it becomes an analogous task to finding the *a priori* correct answer to any other question faced by the cognitive system: the individual is expected to make her utmost effort to approximate the correct solution as closely as possible, within a reasonable time, and by using a limited computational capacity.

Therefore the individual will utilise heuristic techniques again. Even though simulated annealing as a “heuristic optimisation algorithm” is “heuristic” in a very different sense from the way the ABC Research Group employs the word “heuristic”, some crucial similarities still point to a possibly deeper connection. Namely, the tolerance of errors, as well as the use of the information’s structure. Errors often emerge from the trade-off between the precision required by the situation, on the one hand, and the computational resources and time available to the system, on the other. Moreover, the structure of the wor(l)d—*i.e.*, of the search space—is taken into account. That is to say, the “fast and frugal heuristics” of the *ABC Research Group* have developed during evolution so that they reflect the structure of the world, and thereby help increasing precision;

⁸The word “rules” does *not* refer here to traditional generative rewrite rules, whose dismissal was exactly OT’s main agenda. By “rules” I simply mean the laws governing language.

while in SA-OT, the neighbourhood structure mirrors what GEN does, due to which SA-OT can exploit features of the candidate set.

Additionally, it can be argued that the larger human cognitive system also employs some “heuristic optimisation algorithm” that has features remarkably reminding us of simulated annealing. Klein (1999, p. 30) observes that “[d]ecision makers usually look for the first workable option they can find, not the best option”, that “they do not have to generate a large set of options to be sure they get a good one”, and that “[t]hey generate and evaluate options one at a time and do not bother comparing the advantages and disadvantages of alternatives”.⁹ All that because “[t]he emphasis is on being poised to act rather than being paralyzed until all the evaluations have been completed”. Later, Klein (p. 287) summarises his model for human decision making as “analytical”: “... generative, channeling the decision making from opportunity to opportunity rather than exhaustively filtering through all the permutations”. He even adds that human decision making mechanisms “trade accuracy for speed and *therefore* allow errors” [emphasis added—T. B.].

To turn to a very different domain, to believe systems, Bainbridge (2006) introduces simple connectionist networks to model agents in a society. He concludes then that “[l]ocal minima are actually very interesting, because they represent a very human quality: a reluctance to give up beliefs that function pretty well at the cost of never finding the real truth. One way of expressing the thesis of this book is to say: *Religious faith is a local minimum*” (p. 83, italics in the original text). Hence, unlike Klein’s firefighters, but like the random walker in SA-OT, Bainbridge’s believer agents are happy with being stuck in a local optimum.

In sum, I propose to view linguistic competence in its narrow sense (the first level on Table 2.1) as a by-product of the heuristics used by the human cognitive capacities. Then, on a second level, when the individual comes to produce the grammatical form defined by the first level and accepted by the language community, then these (or similar) heuristics ought to be used again, in order to solve an otherwise computationally challenging task. As a result, *errare humanum est*, even in matters of language, which is a domain created by the human mind. For is it not surprising that the system of communication developed by the human mind poses to the same mind problems whose difficulty is similar to the difficulty of the problems posed by external factors? Why is it so hard to find the right words?

⁹This last observation does not contradict the argument that Klein’s decision makers follow an algorithm similar to simulated annealing. Namely, he does not describe here *how* the decision maker comes up with particular options on a micro-level, but this process could be imagined as the random walk in gradient ascent or simulated annealing. My only point here is that neither Klein’s model, nor simulated annealing perform global comparisons or comparisons of distant options.

An important difference is indeed that Klein’s model, unlike simulated annealing, is able to decide whether a particular option (a local optimum) is in itself “good enough” or the search should be continued further. On the other hand, if the experienced speaker could somehow judge whether the local optimum returned by SA-OT is “good enough”, then the precision of SA-OT could be improved. So, similarly to Klein’s firefighter who decides to search further if the solution arrived at involves too much risk, the speaker would also run another simulation if the locally optimal output still incurs too many violation marks.

Bibliography

- Arto Anttila. Morphologically conditioned phonological alternations. *Natural Language and Linguistic Theory*, 20:1–42, 2002. Also: ROA-425.
- Arto Anttila. Deriving variation from grammar. In Frans Hinskens, Roeland van Hout, and W. Leo Wetzels, editors, *Variation, Change and Phonological Theory*, pages 35–68. Benjamins, Amsterdam – Philadelphia, 1997a. Also: ROA-63.
- Arto Anttila. *Variation in Finnish Phonology and Morphology*. Doctoral dissertation, Stanford University, Stanford, California, 1997b.
- Arto Anttila and Young-mee Yu Cho. Variation and change in Optimality Theory. *Lingua*, 104(1-2):31–56, 1998.
- Arto Anttila and Vivienne Fong. The partitive constraint in Optimality Theory. *Journal of Semantics*, 17:281–314, 2000. Also: ROA-416.
- William Sims Bainbridge. *God from the Machine: Artificial Intelligence Models of Religious Cognition*. Cognitive Science of Religion Series. Altamira Press, Lanham, etc., 2006.
- Eric Baković. Unbounded stress and factorial typology. In R. Artstein and M. Holler, editors, *RuLing Papers 1 (Working Papers from Rutgers University)*. Rutgers University Department of Linguistics, New Brunswick, NJ, 1998. Also: ROA-244.
- Tamás B    . Quadratic alignment constraints and finite state Optimality Theory. In *Proceedings of the Workshop on Finite-State Methods in Natural Language Processing (FSMNLP), held within EACL-03, Budapest*, pages 119–126, 2003. Also: ROA-600¹⁰
- Tam  s B    . When the hothead speaks: Simulated Annealing Optimality Theory for Dutch fast speech. presented at CLIN 2004, Leiden, 2004.
- Tam  s B    . When the hothead speaks: Simulated Annealing Optimality Theory for Dutch fast speech. In Crit Cremers, Hilke Reckman, Michaela Poss, and Ton van der Wouden, editors, *Proceedings of the 15th Meeting of Computational Linguistics in the Netherlands (CLIN 2004)*, Leiden, 2005a.
- Tam  s B    . How to define Simulated Annealing for Optimality Theory? In *Proceedings of the 10th Conference on Formal Grammar and the 9th Meeting on Mathematics of Language*, Edinburgh, August 2005b.

¹⁰ROA stands for *Rutgers Optimality Archive* at <http://roa.rutgers.edu/>.

- Tamás Bíró. Squeezing the infinite into the finite: Handling the OT candidate set with finite state technology. In *Proceedings of the Workshop on Finite-State Methods in Natural Language Processing (FSMNLP)*, Helsinki, August 2005c.
- Tamás Bíró. Az optimalitáselméleti modell megvalósítása szimulált hőkezeléssel [Realising the Optimality Theoretical Model with Simulated Annealing]. seminar paper, Eötvös Loránd University, 1997.
- Tamás Bíró and Judit Gervain. L' acantatrice chauve: Loser candidates in SA-OT and speech rate. paper presented at OCP 3, 2006.
- Tamás Bíró and Anna Hamp. Schwa and roots: A non-concatenative lexical morpho-phonology. In *Selected Papers of Docsymp 6, the Graduate Students' Sixth Linguistics Symposium*, pages 9–22, Budapest, 2002.
- Reinhard Blutner. Some aspects of optimality in natural language interpretation. *Journal of Semantics*, 17:189–216, 2000.
- Rens Bod, Jennifer Hay, and Stefanie Jannedy, editors. *Probabilistic Linguistics*. The MIT Press, Cambridge, Mass., [etc.], 2003.
- Paul Boersma. Review of *B. Tesar & P. Smolensky* (2000): Learnability in Optimality Theory. ROA-638, 2004a.
- Paul Boersma. Prototypicality judgments as inverted perception. m.s., ROA-742, 2005.
- Paul Boersma. The odds of eternal optimization in OT. ms., ROA-429, 2000.
- Paul Boersma. A stochastic OT account of paralinguistic tasks such as grammaticality and prototypicality judgments. ms., ROA-648, 2004b.
- Paul Boersma. How we learn variation, optionality, and probability. *Proceedings of the Institute of Phonetic Sciences, Amsterdam (IFA)*, 21:43–58, 1997.
- Paul Boersma. *Functional Phonology: Formalizing the interactions between articulatory and perceptual drives*. PhD thesis, Amsterdam University, The Hague, 1998a.
- Paul Boersma. Review of Arto Anttila: Variation in Finnish phonology and morphology. *GLOT International*, 5(1):33–40, 2001. URL <http://www.fon.hum.uva.nl/paul/papers/anttila.review.pdf>.
- Paul Boersma. Typology and acquisition in functional and arbitrary phonology. Presented at Utrecht Phonology Workshop, June 22, 1998, 1998b. URL http://www.fon.hum.uva.nl/paul/papers/typ_acq.pdf.
- Paul Boersma and Bruce Hayes. Empirical tests of the Gradual Learning Algorithm. *Linguistic Inquiry*, 32:45–86, 2001. Also: ROA-348.
- Ellen Broselow and Zheng Xu. Differential difficulty in the acquisition of second language phonology. *International Journal of English Studies*, 4(2):135–163, 2004. Also: ROA-743.

- Roger Brown. *A First Language: The Early Stages*. Harvard University Press, Cambridge, MA, 1973.
- Luigi Burzio. Missing players: Phonology and the past-tense debate. *Lingua*, 112:157–199, 2002.
- Luigi Burzio. Missing players: Phonology and the past-tense debate. unpublished manuscript, August 1999. Similar to ROA-341.
- Patrik Bye. Coda maximisation in Northwest Saamic. *Nordic Journal of Linguistics*, 28(2):189–221, 2005. Also: ROA-757.
- V. Černý. Thermodynamical approach to the travelling salesman problem: an efficient simulation algorithm. *Journal of Optimization Theory and Applications*, 45:41–51, 1985.
- Noam Chomsky. *Syntactic Structures*. Mouton, The Hague and Paris, 1957.
- Noam Chomsky. *Aspects of the Theory of Syntax*. The M.I.T. Press, Cambridge, Mass., 1965.
- Noam Chomsky and Morris Halle. *The Sound Pattern of English*. Harper & Row, New York, N.Y. [etc.], 1968.
- Brady Clark. On stochastic grammar. *Language*, 81(1):207–217, 2005.
- Andries W. Coetzee. *What it Means to be a Loser: Non-optimal Candidates in Optimality Theory*. PhD thesis, University of Massachusetts, Amherst, 2004. Also: ROA-687.
- Andries W. Coetzee. Grammar is both categorical and gradient. In Stephen Parker, editor, *Phonological Argumentation*. Equinox Publishers, London, to appear. Also: ROA-864.
- John G. Daugman. Brain metaphor and brain theory. In E. L. Schwartz, editor, *Computational Neuroscience*, pages 9–18. MIT Press, Cambridge, MA, 1990.
- Ferdinand de Saussure. *Course in General Linguistics*. Peter Owen, London, 1974.
- A. E. Eiben and J. E. Smith. *Introduction to Evolutionary Computing*. Springer, Berlin, etc., 2003.
- Jason Eisner. Directional constraint evaluation in Optimality Theory. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000)*, Saarbrücken, August 2000a.
- Jason Eisner. Easy and hard constraint ranking in Optimality Theory: Algorithms and complexity. In J. Eisner, L. Karttunen, and A. Thériault, editors, *Finite-State Phonology: Proceedings of the 5th Workshop of the ACL Special Interest Group in Computational Phonology (SIGPHON)*, pages 57–67, Luxembourg, 2000b.

- Jason Eisner. Efficient generation in Primitive Optimality Theory. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL-1997) and 8th EACL*, pages 313–320, Madrid, 1997. Also: ROA-206.
- T. Mark Ellison. Phonological derivation in Optimality Theory. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING)*, Kyoto, pages 1007–1013, 1994. Also: ROA-75.
- Robert Frank and Giorgio Satta. Optimality Theory and the generative complexity of constraint violability. *Computational Linguistics*, 24(2):307–315, 1998.
- Dale Gerdemann and Gertjan van Noord. Approximation and exactness in finite state Optimality Theory. In Jason Eisner, Lauri Karttunen, and Alain Thériault, editors, *SIGPHON 2000, Finite State Phonology*, 2000.
- Gerd Gigerenzer, Peter M. Todd, and the ABC Research Group. *Simple Heuristics That Make Us Smart*. Oxford University Press, Oxford, 1999.
- Dicky Gilbers and Helen de Hoop. Conflicting constraints: An introduction to Optimality Theory. *Lingua*, 104:1–12, 1998.
- Dicky Gilbers and Wouter Jansen. Klemtoon en ritme in Optimality Theory. *TABU*, 26(2):53–101, 1996.
- Dicky Gilbers and Maartje Schreuder. Taal en muziek in Optimaliteitstheorie [Language and Music in Optimality Theory]. *TABU*, 30(1-2), 2000. English version: ROA-571.
- Sharon Goldwater and Mark Johnson. Learning OT constraint rankings using a maximum entropy model. In Jennifer Spenader, Anders Eriksson, and Östen Dahl, editors, *Proceedings of the Stockholm Workshop on Variation within Optimality Theory*, pages 111–120, Stockholm, 2003. Stockholm University.
- Matthew Gordon. A factorial typology of quantity-insensitive stress. *Natural Language and Linguistic Theory*, 20:491–552, 2002.
- Trevor A. Harley. *The Psychology of Language: From Data to Theory*. Psychology Press and Taylor & Francis Inc., Hove, UK and New York, NY, 2nd edition, 2001.
- Bruce Hayes. *Metrical Stress Theory: Principles and Case Studies*. University of Chicago Press, Chicago, 1995.
- Jack Hoeksema. Corpus study of negative polarity items. In *IV-V Jornades de corpus lingüístics 1996-1997*, pages 67–86. IULA, Universitat Pompeu Fabra, Barcelona, 1998. URL <http://www.let.rug.nl/~hoeksema/docs/barcelona.html>.
- M. Holz, K. Steffens, and E. Weitz. *Introduction to Cardinal Arithmetic*. Birkhäuser, Basel and Boston and Berlin, 1999.

- T. Howells. VITAL: a connectionist parser. In *Proceedings of 10th Annual Meeting of the Cognitive Science Society*, pages 18–25. Lawrence Erlbaum, 1988.
- H. Mark Hubbey. *Mathematical Foundations of Linguistics*. Lincom Europa, München, 1999.
- Henrietta J. Hung. *The Rhythmic and Prosodic Organization of Edge Constituents*. PhD thesis, Brandeis University, Amherst, 1994. Also: ROA-24.
- William J. Idsardi. A simple proof that Optimality Theory is computationally intractable. *Linguistic Inquiry*, 37(2):271–275, 2006a.
- William J. Idsardi. Misplaced optimism. ROA-840, 2006b.
- Gerhard Jäger. Gradient constraints in finite state OT: The unidirectional and the bidirectional case. ms., ROA-479, 2002.
- Gerhard Jäger. Simulating language change with functional OT. In Simon Kirby, editor, *Language Evolution and Computation, Proceedings of the Workshop at ESSLLI, Vienna*, pages 52–61, 2003a.
- Gerhard Jäger. Maximum entropy models and Stochastic Optimality Theory. m.s., ROA-625, 2003b.
- Gerhard Jäger and Anette Rosenbach. The winner takes it all – almost: Cumulativity in grammatical variation. *Linguistics*, 44(5):937–971, 2006. URL <http://www.uni-bielefeld.de/lili/personen/gjaeger/jaegerRosenbach.pdf>.
- Douglas C. Johnson. *Formal Aspects of Phonological Description*. Mouton, The Hague [etc.], 1972.
- René Kager. Ternary rhythm in Alignment Theory. manuscript, Research Institute for Language and Speech, Utrecht University, 1994.
- Lauri Karttunen. The proper treatment of Optimality Theory in computational phonology. In *Finite-state Methods in Natural Language Processing*, pages 1–12, Ankara, 1998.
- Lauri Karttunen. The insufficiency of paper-and-pencil linguistics: the case of finnish prosody. ROA-818, 2006.
- Frank Keller and Ash Asudeh. Probabilistic learning algorithms and Optimality Theory. *Linguistic Inquiry*, 33(2):225–244, 2002.
- Gerard Kempen and Theo Vosse. Incremental syntactic tree formation in human sentence processing: a cognitive architecture based on activation decay and simulated annealing. *Connection Science*, 1:273–290, 1989.
- Michael Kenstowicz. Base-Identity and Uniform Exponence: Alternatives to cyclicity. In Jacques Durand and Bernard Laks, editors, *Current Trends in Phonology: Models and Methods*. CNRS, Paris-X and University of Salford Publications, Paris, 1995. Also: ROA-103.

- Ferenc Kiefer, editor. *Strukturális Magyar Nyelvtan [A Structural Hungarian Grammar]*, volume 2, Fonológia. Akadémiai Kiadó, Budapest, 1994.
- Paul Kiparsky. From cyclic phonology to lexical phonology. In H. van der Hulst and N. Smith, editors, *The Structure of Phonological Representations*, volume 1, pages 131–175. Floris, Dordrecht, 1982.
- Robert Martin Kirchner. *An Efford-Based Approach to Consonant Lenition*. PhD thesis, UCLA, Los Angeles, 1998. Also: ROA-276.
- S. Kirkpatrick, C. D. Gelatt Jr., and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983.
- Gary Klein. *Sources of Power: How People Make Decisions*. MIT Press, Cambridge, Massachusetts–London, England, 1999.
- András Kornai. Is OT NP-hard? ROA-838, 2006a.
- András Kornai. Guarded optimalism. ROA-841, 2006b.
- Kimmo Koskeniemi. Two-level morphology: A general computational model for word-form recognition and production. *Publication No. 11, Department of General Linguistics, University of Helsinki*, 1983.
- Stan A. Kuczaj. The acquisition of regular and irregular past tense forms. *Journal of Verbal Learning and Verbal Behavior*, 16:589–600, 1977.
- Jonas Kuhn. Processing Optimality-theoretic syntax by interleaved chart parsing and generation. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL-2000)*, Hongkong, pages 360–367, 2000.
- Géraldine Legendre, Yoshiro Miyata, and Paul Smolensky. Harmonic grammar – a formal multi-level connectionist theory of linguistic well-formedness: Theoretical foundations. In *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society*, pages 388–395, Cambridge, MA., 1990a.
- Géraldine Legendre, Yoshiro Miyata, and Paul Smolensky. Harmonic Grammar – a formal multi-level connectionist theory of linguistic well-formedness: An application. In *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society*, pages 884–891, Cambridge, MA., 1990b.
- Géraldine Legendre, Yoshiro Miyata, and Paul Smolensky. Can connectionism contribute to syntax? Harmonic Grammar, with an application. In *Proceedings of the 26th Meeting of the Chicago Linguistic Society*, Chicago, IL, 1990c.
- Piroska Lendvai. *Extracting Information from Spoken User Input: A Machine Learning Approach*. PhD thesis, Tilburg University, Tilburg, 2004.
- Linda Lombardi. Restrictions on Direction of Voicing Assimilation: an OT account. *University of Maryland Working Papers in Linguistics*, 3:89–115, 1995. Also: ROA-246.

- Linda Lombardi. Why Place and Voice are different: Constraint-specific alternations in Optimality Theory. In Linda Lombardi, editor, *Segmental phonology in Optimality Theory: Constraints and Representations*. Cambridge University Press, Cambridge, 2001. Also: ROA-105.
- Susanna C. Manrubia, Alexander S. Mikhailov, and Damián H. Zanette. *Emergence of Dynamical Order: Synchronization Phenomena in Complex Systems*. World Scientific, New Jersey, etc., 2004.
- John J. McCarthy. Against gradience. ROA-510, 2002.
- John J. McCarthy and Alan Prince. Generalized alignment. In *Yearbook of Morphology*, pages 79–153. Kluwer, Dordrecht, 1993a. Also: ROA-7.
- John J. McCarthy and Alan Prince. Prosodic morphology: Constraint interaction and satisfaction. Technical Report nr. 3. of the Rutgers University Center for Cognitive Science (RuCCS-TR-3), ROA Version: ROA-482 (2001), 1993b.
- John J. McCarthy and Alan Prince. The emergence of the unmarked: Optimality in prosodic morphology. In *Proceedings of the North East Linguistics Society 24*, pages 333–379, Amherst, MA., 1994. Graduate Linguistic Student Association.
- James L. McClelland and K. Patterson. Rules or connections in past-tense inflections: what does the evidence rule out? *Trends in Cognitive Sciences*, 6(11):465–472, 2002.
- Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of state calculation by fast computing machines. *Journal of Chemical Physics*, 21(6):1087–1092, 1953.
- Tom M. Mitchell. *Machine learning*. WCB/McGraw-Hill, New York [etc.], 1997.
- Anthony James Mullen. *An Investigation into Compositional Features and Feature Merging for Maximum Entropy-Based Parse Selection*. PhD thesis, Groningen University, Groningen, 2002.
- Naomi Nagy and William T. Reynolds. Optimality Theory and variable word-final deletion in Faetar. *Language Variation and Change*, 9:37–55, 1997.
- Rafael E. Núñez. Creating mathematical infinities: Metaphor, blending, and the beauty of transfinite cardinals. *Journal of Pragmatics*, 37:1717–1741, 2005.
- Mitsuhiko Ota. The learnability of the stratified phonological lexicon. *Journal of Japanese Linguistics*, 20, 2004. Also: ROA-668.
- Seymour Papert. *Mindstorms: Children, Computers, and Powerful Ideas*, volume 14 of *Harvester Studies in Cognitive Science*. The Harvester Press Ltd., Brighton, Sussex, 1980.
- Joe Pater. Non-convergence in the GLA and variation in the CDA. ms., ROA-780, 2005a.

- Joe Pater. Learning a stratified grammar. In Alejna Brugos, Manuella R. Clark-Cotton, and Seungwan Ha, editors, *Proceedings of the 29th Boston University Conference on Language Development*, pages 482–492. Cascadilla Press, Somerville, MA, 2005b. Also: ROA-739.
- Steven Pinker and Alan Prince. On language and connectionism: analysis of a parallel distributed processing model of language acquisition. *Cognition*, 28: 73–193, 1988.
- Steven Pinker and Michael T. Ullman. The past-tense debate: The past and future of the past tense. *Trends in Cognitive Sciences*, 6(11):456–463, 2002.
- Alan Prince. Anything goes. ms., ROA-536, 2002.
- Alan Prince and Paul Smolensky. Optimality Theory: Constraint Interaction in Generative Grammar. ROA Version: 537-0802, <http://roa.rutgers.edu>, August 2002.
- Alan Prince and Paul Smolensky. *Optimality Theory: Constraint Interaction in Generative Grammar*. Blackwell, Malden, MA, etc., 2004.
- Alan Prince and Paul Smolensky. Optimality Theory: Constraint Interaction in Generative Grammar. Technical Report nr. 2. of the Rutgers University Center for Cognitive Science (RuCCS-TR-2), 1993.
- Alan Prince and Bruce Tesar. Learning phonotactic distributions. In R. Kager, J. Pater, and W. Zonneveld, editors, *Constraints in Phonological Acquisition*, pages 245–291. CUP, 2004.
- Douglas Pulleyblank and William J. Turkel. Learning phonology: Genetic algorithms and Yoruba tongue-root harmony. In Joost Dekkers, Frank van der Leeuw, and Jeroen van De Weijer, editors, *Optimality Theory: Phonology, Syntax, and Acquisition*, pages 554–591. Oxford University Press, Oxford, 2000.
- Colin R. Reeves, editor. *Modern Heuristic Techniques for Combinatorial Problems*. McGraw-Hill, London, etc., 1995.
- Frederick Reif. *Fundamentals of Statistical and Thermal Physics, International Student Edition*. McGraw-Hill, New York, etc., 1965.
- William Thomas Reynolds. *Variation and Phonological Theory*. PhD thesis, University of Pennsylvania, Philadelphia, 1994.
- H. Robbins and S. Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407, 1951.
- D. E. Rumelhart and James L. McClelland. On learning the past tenses of english verbs. In *McClelland, Rumelhart et al.: Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, volume 2, pages 216–271. Bradford, MIT Press, Cambridge, MA, London, England, 1986.
- Vieri Samek-Lodovici and Alan Prince. Optima. ROA-363, 1999.

- Niels O. Schiller. Metrical stress in speech production: A time course study. In M. J. Solé, D. Recasens, and J. Romero, editors, *Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS 2003)*, pages 451–454, Barcelona, August 2003.
- Niels O. Schiller, B. M. Jansma, J. Peters, and W. J. M. Levelt. Monitoring metrical stress in polysyllabic words. *Language and Cognitive Processes*, 2004. URL http://www.psychology.unimaas.nl/images/pdf_files/NielsSchiller/Schiller_et_al_LCP_accepted.pdf.
- Maartje Schreuder. *Prosodic Processes in Language and Music*. PhD thesis, Rijksuniversiteit Groningen, Groningen, Netherlands, 2006.
- Maartje Schreuder and Dicky Gilbers. The influence of speech rate on rhythm patterns. In Dicky Gilbers, Maartje Schreuder, and Nienke Knevel, editors, *On the Boundaries of Phonology and Phonetics*, pages 183–201. University of Groningen, Groningen, 2004.
- Ora R. Schwarzwald. *Modern Hebrew*. Languages of the World/Materials 127. Lincom Europa, München, 2001.
- Bart Selman and Graeme Hirst. A rule-based connectionist parsing system. In *Proceedings of the Seventh Annual Meeting of the Cognitive Science Society*, Irvine, pages 212–221. Lawrence Erlbaum, Hillsdale, NJ, 1985.
- Bart Selman and Graeme Hirst. Parsing as an energy minimization problem. In Geert Adriaens and Udo Hahn, editors, *Parallel natural language processing*, pages 238–254. Ablex Publishing, Norwood, NJ, 1994.
- Paul Smolensky. Information processing in dynamical systems: Foundations of Harmony Theory. In *Rumelhart et al.: Parallel Distributed Processing*, volume 1, pages 194–281. Bradford, MIT Press, Cambridge, MA, London, England, 1986.
- Paul Smolensky and Géraldine Legendre, editors. *The Harmonic Mind: From Neural Computation to Optimality-Theoretic Grammar*. MIT Press, 2006.
- James C. Spall. *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control*. Wiley-Interscience, Hoboken, New Jersey, 2003.
- Emmanuel A. Stamatakis, William Marslen-Wilson, Lorraine K. Tyler, and Paul Fletcher. Inter-regional covariances in a fronto-temporal system for speech and language. In *10th International conference on Functional Mapping of the Human Brain*, page Poster Number MO 104, Budapest, 2004. URL http://www.meetingassistant.com/ohbm/meeting_plan/ohbm_mtg_index_search.php.
- Patrick Suppes. *Axiomatic Set Theory*. Dover, New York, 1972.
- Niels A. Taatgen and Mariëtta Dijkstra. Constraints on generalization: Why are past-tense irregularization errors so rare? In *Proceedings of the 25th annual conference of the Cognitive Science Society*, pages 1146–1151, Mahwah, NJ, 2003. Erlbaum.

- Bruce Tesar. Robust interpretive parsing in metrical stress theory. *The Proceedings of WCCFL*, 17:625–639, 1999. Also: ROA-262.
- Bruce Tesar and Alan Prince. Using phonotactics to learn phonological alternations. m.s., ROA-620, 2003.
- Bruce Tesar and Paul Smolensky. *Learnability in Optimality Theory*. The MIT Press, Cambridge, MA - London, England, 2000.
- Bill Turkel. The acquisition of Optimality Theoretic systems. m.s., ROA-11, 1994.
- A. Tversky and D. Kahneman. Judgment under uncertainty: Heuristics and biases. *Science*, 185:1124–1131, 1974.
- Lorraine K. Tyler, Paul DeMornay-Davies, Rebekah Anokhina, Catherine Longworth, Billi Randall, and William D. Marslen-Wilson. Dissociations in processing past tense morphology: Neuropathology and behavioral studies. *Journal of Cognitive Neuroscience*, 14(1):79–94, 2002.
- Leonoor van der Beek and Gerlof Bouma. The role of the lexicon in Optimality Theoretic syntax. In Miriam Butt and Tracy Holloway King, editors, *Proceedings of the LFG04 Conference*, University of Canterbury, 2004. CSLI Publications. Also: ROA-690.
- Peter J. M. van Laarhoven. *Theoretical and Computational Aspects of Simulated Annealing*. PhD thesis, Erasmus University Rotterdam, Rotterdam, 1987.
- Theo Vosse and Gerard Kempen. Syntactic structure assembly in human parsing: a computational model based on competitive inhibition and a lexicalist grammar. *Cognition*, 75:105–143, 2000.
- Harold Todd Wareham. *Systematic Parameterized Complexity Analysis in Computational Phonology*. PhD thesis, University of Victoria, April 1999. Also: ROA-318.
- Max Wheeler. Cluster reduction: Deletion or coalescence? ROA-718, 2005.

Summary

This dissertation presents an implementation of *Optimality Theory* (OT, Prince and Smolensky, 1993) that also aims at accounting for certain variations in speech. The *Simulated Annealing for Optimality Theory Algorithm* (SA-OT, Fig. 2.8, on page 64) combines OT with *simulated annealing*, a widespread heuristic optimisation technique. After a general introduction to Optimality Theory and the discussion of certain “philosophical background questions” (especially on the role of probabilities in linguistics; Chapter 1), the SA-OT Algorithm is introduced (informally in section 2.2, mathematically in sections 3.3 and 3.4), put into a broader context (section 2.1, Chapter 4, and sections 8.2 and 8.3), and experimented with (section 2.3, Chapters 5-7).

Reeves (1995) defines *heuristic* as “a technique which seeks good (i.e. near-optimal) solutions at a reasonable computational cost without being able to guarantee either feasibility or optimality, or even in many cases to state how close to optimality a particular feasible solution is.” Even if they are not exact, these algorithms are very useful in solving efficiently hard computational problems, similar to the task of finding the optimal candidate in an OT candidate set. A *good* solution suffices in many applications, and there is no need to allocate huge computational resources to find the *best* solution. As section 2.1 argues, heuristic algorithms—such as SA-OT—may serve as adequate models of the computations performed by the human brain for at least three reasons: (1) many of these algorithms are simple, (relatively) efficient and produce *some* output within a predefined time span, even if (2) they may make errors, and finally (3) the algorithm can be speeded up with a price to be paid in reduced precision. A faster computation is possible, but more prone to make errors. The adequacy of such a model is corroborated if besides the grammatical forms it also reproduces the empirically observable error patterns under different conditions. Importantly, these predictions are quantitative, and the algorithm’s parameters can “fine-tune” the output frequencies of the erroneous or alternating forms.

Table 2.1 (page 43) formulates this idea: by distinguishing between a linguistic model and its implementation, one can account for both linguistic competence and certain types of linguistically motivated performance phenomena. Thus an adequate linguistic model (a *grammar*, such as a well-founded OT grammar) predicts correctly which forms are judged as grammatical by the native speaker. This layer refers to the *static knowledge of the language* in the native speaker’s brain. On top of that is built the *implementation of the grammar* as a model of the *dynamic language production process*. Similarly to human speech, the implementation of the grammar need not be exact, but the errors made by the implementation should correspond to the observed performance errors.

In particular, SA-OT requires a *topology* (a *neighbourhood structure*) on the OT candidate set. Consequently, the notion of a *local optimum* is introduced: a candidate that is more harmonic than all its neighbours is a local optimum, independently of whether it is the most harmonic element of the entire candidate set. Local optima are the candidates that can emerge as outputs in SA-OT. The *global optimum* predicts the grammatical form, whereas all other outputs should model performance errors.

How does the SA-OT Algorithm work? A random walk is performed on the candidate set. In each iteration, the random walker chooses randomly a candidate w' among the neighbours of its present position w . For instance, a minimal *basic operation* transforms w into w' . The definition of the topology on the candidate set also includes the *a priori probabilities* $P_{\text{choice}}(w'|w)$, which determine the chance of choosing w' with the condition that the random walker is in w . Then, the random walker truly moves to w' with some other *transition probability* $P(w \rightarrow w'|T)$ —which depends on the harmonies (violation profiles) of candidates w and w' —else it stays in w .

The *a priori probabilities* $P_{\text{choice}}(w'|w)$ do not depend on the violation profiles of the candidates (hence, on the constraint ranking) and are constant during the simulation. The *transition probabilities* $P(w \rightarrow w'|T)$, however, diminish as a function of the parameter T (called *temperature*), gradually, from 1 at the beginning of the simulation to 0 at its end—if w' is less harmonic than w . Otherwise, the *transition probability* from w to w' remains 1, so the random walker is always allowed to move to a better neighbour. (See equation (2.2) on page 39 for this idea in traditional simulated annealing, and equation (2.17) on page 63 and the subsequent “Rules of moving” for SA-OT.) Thus, the random walker will be stuck in some local optimum by the end. The output of the algorithm is the final position of the random walker, and its precision is the likelihood of this local optimum being also globally optimal. Frequently—but certainly not always, as Chapter 6 demonstrates—a slower pace of diminishing the parameter T (that is, a larger number of iterations performed, a slower *cooling schedule*) results in a reduced chance of being stuck in a local optimum that is not globally optimal.

Combining simulated annealing with Optimality Theory has been far from trivial. Traditional simulated annealing optimises a real-valued target function, which is different from the Harmony function employed in OT. In order to introduce the *transition probabilities*, first the difference of two violation profiles has to be defined, and then temperature T is introduced as a pair $\langle K, t \rangle$. The idea is first presented in section 2.2 in a relatively informal way, whereas Chapter 3 argues for the same algorithm by making use of several mathematical formalisms. Chapter 3 starts with the mathematical definition of OT—also in order to show which assumptions are needed in OT and what assumptions can be generalised in future research—followed by a discussion on how to realise the Harmony function using polynomials, on the one hand, and using ordinal numbers, on the other. Both approaches lead to the same way of combining OT with simulated annealing.

The following chapter speculates about two “hot topics” in linguistics: the lexicon and learnability. Apart from trains of thought that are left open for future research and do not play a central role in this dissertation, it introduces a new and formal definition of *Output-Output Correspondence*, or rather *Constituent-Output Correspondence*, which is used in the subsequent chapter.

The remaining three chapters before the conclusion present how concrete linguistic phenomena could be tackled within the framework of SA-OT. In these chapters the emphasis lies not so much on the phonological details of the specific analyses, rather on the methodological problems raised by SA-OT. The experiments with the models demonstrate the role of certain parameters of the algorithm, as well as the importance of certain decisions that have to be made when one decides to use SA-OT. Additionally, different techniques and tricks, different ways of experimenting with the models and different ways of understanding their behaviour are presented throughout these chapters. Section 8.1 summarises in details the various methodological issues dealt with in these chapters.

Using a few observations on Dutch metrical stress assignment, Chapter 5 demonstrates how fast speech phenomena can be reproduced by varying the speed of the algorithm. If the algorithm is run more quickly (with fewer iterations, with a faster *cooling schedule*), then the frequencies of the alternating forms change similarly to the way they change when moving from slow (normal, careful) speech to fast speech, as reported by laboratory experiments.

The toy models in section 2.3 advanced some of the problems that are further dealt with in Chapter 6, using the example of Dutch regressive and progressive voice assimilation. Two models are presented, the first one involving a very simple and restricted candidate set, and the second one displaying an infinite topology. The landscapes, that is, the topologies of the OT candidate set with the Harmony function, were simpler in these models than in the previous chapter, and therefore an analytical discussion of the behaviour of the models could also be included, besides the experimentation with the parameters. Furthermore, these models demonstrated that in the case of SA-OT—unlike in the case of traditional simulated annealing—increasing the number of iterations (having a slower cooling schedule) does not always necessarily lead to an increased probability of returning the globally optimal candidate. Supposing that SA-OT is indeed an adequate model for speech production, this observation opened the floor to speculations about how to keep a grammar simple while still accounting for “irregularities”. Namely, the “irregular” forms are conjectured to be the erroneous outputs (the non-global local optima) that the dynamic language production process cannot avoid producing under any condition.

Chapter 7 discusses two phenomena, both related to word or syllable structure. First, the cliticisation of the article in Hungarian is accounted for, as a function of the speech rate and of the allomorph chosen. The topology of the model has an overall structure that makes it similar to the second topology analysed in Chapter 6. The same type of topology is enriched in the last model presented in this dissertation, which is a first and preliminary attempt to implement the classical OT paradigm for syllabification (basic CV Theory) using SA-OT. So the simulations presented in Chapters 6 and 7 also demonstrate how a certain type of model—which, I conjecture, might become important in future research—can be made gradually more complex. In the same chapters we see how never winning (loser) candidates can still influence the output frequencies and should therefore be included in the candidate set (the *Godot effect*). Moreover, section 7.1 demonstrates how SA-OT can supply arguments for ranking constraints that could not be ranked based on the arguments used in traditional OT.

The concluding Chapter 8 summarises the results of the present thesis and

compares SA-OT to other OT approaches to linguistic variation. Future research should decide whether SA-OT or its competitors, the already existing stochastic OT models are more fruitful, but I believe that they may complement each other. Moreover, future research should also work out certain details, which have been judged so far by many readers as *ad hoc*, in a more persuasive manner. Finally, in section 8.3, SA-OT is put into the context in which Optimality Theory was born more than a decade ago, namely, the cognitive sciences.

Samenvatting

Het doel van dit proefschrift is *Optimality Theory* (Prince and Smolensky, 1993) zo te implementeren dat we de variatie in spraak kunnen beginnen te verklaren. Het *Simulated Annealing Optimality Theory* algoritme (SA-OT, Fig. 2.8, pagina 64) combineert Optimaliteitstheorie met de zogenaamde *Simulated Annealing*, een wijd verspreid heuristisch optimalisatie-algoritme waarmee we de frequentie van verschillende alternatieven nauwkeurig kunnen modelleren. Verder onderzoek zal in de toekomst moeten uitwijzen of SA-OT op termijn succesvoller is dan de bestaande, stochastische optimaliteitstheoretische modellen.

In Hoofdstuk 1 worden deze concurrerende modellen op een rijtje gezet, en wordt de “filosofische” achtergrond van het proefschrift geschetst. In Hoofdstuk 1 wordt het gebruik van een heuristisch optimalisatiealgoritme – zoals *Simulated Annealing* – verantwoord, en het SA-OT algoritme geïntroduceerd. Tabel 2.1 (pagina 43) formuleert een van de centrale ideeën van dit proefschrift: door het scheiden van het taalkundige model en de implementatie ervan, kunnen taalkundige competentie en performancefouten verklaard worden. Een adequaat model, zoals een optimaliteitstheoretische grammatica, kan voorspellen welke vormen door de moedertaalspreker als grammaticaal worden beoordeeld. Dit deel van het model komt overeen met de *statische kennis van de taal* in de hersenen van de moedertaalspreker. Daarnaast is er de implementatie van de grammatica, die gezien kan worden als model van het *dynamische taalproductieproces*. Net als de menselijke spraak hoeft de implementatie van de grammatica niet exact (juist) te zijn, maar de door de implementatie gemaakte fouten dienen wel overeen te komen met geobserveerde performancefouten.

SA-OT introduceert een topologie van de verzameling optimaliteitstheoretische kandidaten, namelijk een *nabijheidsstructuur* (Engels: ‘neighbourhood structure’). Op deze manier kan het begrip *lokaal optimum* gedefinieerd worden: een lokaal optimum is een kandidaat die een hogere harmonie heeft dan zijn directe omgeving. Het zijn deze lokale optima die, ookal zijn ze niet globale optima, als ‘onjuiste’ uitvoer geselecteerd kunnen worden, en zo performancefouten modelleren. Het globale optimum komt overeen met de grammaticale vorm.

Hoe gaat dit in zijn werk? SA-OT voert een *random walk* uit op de kandidatenverzameling. Bij elke iteratie kiest de random walker willekeurig vanuit de huidige positie w een naburige kandidaat w' (er is bijvoorbeeld een basisoperatie die kandidaat w in w' verandert), op basis van de *a priori probabilities*. Vervolgens verplaatst de random walker zich naar w' met een waarschijnlijkheid die we de *transition probability* zullen noemen. De *a priori probabilities* blijven constant tijdens de simulatie, maar de *transition probability* neemt af,

als functie van de parameter T ('temperatuur'), tenzij w' een harmonieuzere (of een even zo harmonieuze) kandidaat is dan w , in welk geval de *transition probability* altijd 1 is. Aan het eind van een simulatie staat de random walker altijd in een lokaal optimum dat niet noodzakelijk het globale optimum is. De uitvoer van het algoritme is de eindpositie van de random walker. Vaak – maar, zoals is gebleken, lang niet altijd – is de kans dat het juiste, globale optimum wordt bereikt groter wanneer de temperatuur T langzamer afneemt, en er meer iteraties worden uitgevoerd.

Het combineren van Simulated Annealing en Optimaliteitstheorie is niet triviaal. De *targetfunctions* die geoptimaliseerd worden in de traditionele Simulated Annealing hebben reële getallen als functiewaarden, maar de Harmoniefunctie uit de Optimaliteitstheorie niet. De oplossing voor dit probleem wordt informeel geïntroduceerd in sectie 2.2, en wiskundig uitgewerkt in hoofdstuk 3. Dit hoofdstuk begint met een formele definitie van Optimaliteitstheorie, gevolgd door een bespreking van twee manieren om de Harmoniefunctie te formuleren: door middel van polynomen, of met ordinale getallen. Beide benaderingen leiden tot dezelfde combinatie van Optimaliteitstheorie met Simulated Annealing.

In het daaropvolgende hoofdstuk speculeer ik over twee taalkundige 'hot topics': het mentale lexicon (de woordenschat), en het leren van een grammatica. Naast vragen die ik in dit hoofdstuk en deze dissertatie verder onbeantwoord zal laten, introduceer ik ook een nieuwe formele definitie van de zogenaamde *Output-Output Correspondence*, of liever *Constituent-Output Correspondence*, die in het volgende hoofdstuk gebruikt zal worden.

De laatste drie hoofdstukken laten zien hoe een aantal concrete, taalkundige verschijnselen aangepakt kunnen worden met behulp van SA-OT. Hier ligt de nadruk dan ook niet op de fonologische details van de analyses, maar op de methodologie van het toepassen van SA-OT. De besproken modellen laten het belang van de modelparameters zien in het algoritme. Daarnaast worden er enkele technieken en trucs geïntroduceerd en verschillende experimenten die men kan uitvoeren met de modellen, en wordt besproken hoe de uitkomsten van deze experimenten begrepen moeten worden. Sectie 8.1 vat de verschillende methodologische observaties nog eens uitvoerig samen.

Metrische klemtoon in het Nederlands dient in hoofdstuk 5 als voorbeeld van hoe spreeknelheid met behulp van het algoritme gemodelleerd kan worden. Bij een langzamere simulatie (met meer iteraties, met een geleidelijker *cooling schedule*) veranderen de relatieve frequenties van de door het model geproduceerde vormen op dezelfde wijze als de door een mens geproduceerde vormen bij het veranderen van de spreeknelheid in een laboratoriumexperiment.

Hoofdstuk 6 gaat over stemassimilatie, en de methodologische kwesties uit dit hoofdstuk gaan terug op enkele problemen uit sectie 2.3. De 'landschappen', dat wil zeggen de topologieën met de Harmoniefunctie, zijn eenvoudiger in deze modellen dan die van de modellen in hoofdstuk 5, waardoor een analytische bespreking van het gedrag van de modellen mogelijk is. Verder bewijzen deze modellen dat – in tegenstelling tot traditionele Simulated Annealing – een hoger aantal iteraties niet altijd leidt tot een grotere kans het globale optimum, de beste kandidaat, te vinden. Als SA-OT daadwerkelijk een adequaat model van de spraakproductie is, dan suggereert dit dat we onregelmatige vormen ook met een eenvoudige grammatica kunnen uitleggen. Namelijk, de onregelmatige vormen zijn dan de 'foutjes' die het dynamische taalproductieproces onherroepelijk maakt.

In hoofdstuk 7 passeren twee verschijnselen de revue, die met woord- en lettergreepstructuur te maken hebben. Eerst bestudeer ik de clitisering van het Hongaarse lidwoord, afhankelijk van het spreektempo en de gekozen allomorf. Daarna implementeer ik met SA-OT het klassieke paradigma van lettergreepstructuur van Prince en Smolensky: *Basic CV Theory*. De topologie in het model van het Hongaarse lidwoord heeft een algemene structuur die overeenkomt met een van de topologieën die in hoofdstuk 6 gebruikt zijn. De topologie die gebruikt is bij het implementeren van Basic CV Theory is een verdere ontwikkeling hiervan. Samengenomen hebben we in hoofdstukken 6 en 7 verschillende aspecten van modellen met een vergelijkbare topologie bestudeerd. Ik vermoed dat deze ‘familie’ van modellen in toekomstig werk ook prominent aanwezig zal zijn.

In het achtste en laatste hoofdstuk worden de bevindingen samengevat, en SA-OT vergeleken met andere modellen van variatie in taal. Tot slot wordt SA-OT geplaatst in de context waarin de Optimaliteitstheorie meer dan een decennium geleden ontstond: de cognitiewetenschap.

Összefoglalás

Az Optimalitáselmélet (*Optimality Theory*, OT Prince and Smolensky, 1993) az elmúlt évtized egyik legnépszerűbb elmélete, elsősorban a fonológiában (a hangtanban), de más nyelvészeti területeken is. A jelen disszertáció célja a modell számítógépes implementációja, vagyis olyan algoritmus kidolgozása, amely megtalálja az OT jelöltek halmazán az optimális jelöltet, amely az elmélet szerint az adott nyelv grammatikus alakjának felel meg.

Az optimális jelölt megkeresésére a *szimulált hőkezelés* (*szimulált lehűtés*, *simulated annealing*) nevű algoritmust használtam, amely a statisztikus fizikából a számítástudományba átvett, elterjedt heurisztikus optimalizálási algoritmus. Az *SA-OT Algoritmust* (*Simulated Annealing for Optimality Theory Algorithm*) a 2.8. ábra mutatja be, a 64. oldalon. Mint a legtöbb heurisztikus algoritmus, az SA-OT sem garantálja azt, hogy mindig megtaláljuk az alaphalmaz legjobb elemét, pontossága (a keresett elem megtalálásának a valószínűsége) általában kisebb 100%-nál. Mégis azt állítom, hogy az emberi beszédprodukciónak adekvát modellje, mivel (1) egyszerű, (2) tetszőleges időtartam alatt produkál egy outputot, és (3) ez az időtartam lerövidíthető, az algoritmus felgyorsítható a pontosság rovására. A beszédpartnerünknek nem kell várnia egy komplexebb mentális számítás esetén sem, és ha gyorsan kell beszélnünk, legfeljebb bevállaljuk a hibázás nagyobb esélyét. Ezért az SA-OT Algoritmus a performanciahibák modellezésére is alkalmas.

A nyelvi kompetencia és a nyelvészetileg motivált performancijelenségek egységes kezelésére tett javaslatomat mutatja be a 2.1 táblázat a 43. oldalon. Ha különválasztjuk a nyelvészeti modellt, például egy OT-nyelvtant, annak az implementációjától, akkor az előbbi a nyelvi kompetenciát (a nyelv statikus ismeretét az agyban), utóbbi pedig a performanciát (a dinamikus nyelvprodukciónak) adhatja vissza. A nyelvten jóslatot tesz arra nézve, hogy mely alakokat tartja az anyanyelvi beszélő grammatikusnak, míg a nyelvten implementációja kvantitatív előrejelzéseket tehet a grammatikus, ill. a kevésbé grammatikus alakok előbukkanási gyakoriságára. Az SA-OT Algoritmus, paramétereinek finombeállítása révén, éppen ezen valószínűségek reprodukciójára alkalmas.

Az SA-OT egy *topológiát* (egy *szomszédsági struktúrát*) igényel az OT jelölthalmazon. Így *lokális optimumokról* is beszélhetünk, vagyis olyan jelöltekről, amelyek harmonikusabbak a szomszédaiknál, függetlenül attól, hogy az egész halmaznak *globális optimumjai*-e. Az SA-OT outputjai épp ezek a lokális optimumok lesznek, vagyis azok a jelöltek felelnek meg a performanciahibáknak, amelyek globálisan nem optimálisak, de amelyeket, lokális optimumok lévén, az algoritmus kiadhat.

Hogyan történik ez? Az SA-OT Algoritmus egy véletlen bolyongást valósít

meg a jelöltek halmazán. Ha a véletlen bolyongó épp a w pozícióban található, a szomszéd jelöltek közül kiválaszt egyet, w' -t, $P_{choice}(w'|w)$ a priori valószínűséggel. Ezek a valószínűségek állandóak, függetlenek a jelöltek harmóniájától, a constraint-ek rendezésétől, és a topológia definíciójának a részei. Majd összehasonlítja w -t és w' -t, és egy másik $P(w \rightarrow w'|T)$ valószínűséggel átlép w' -be. Utóbbi valószínűségek függenek a jelöltek constraint-sértéseitől, és változnak az algoritmus T paramétere (a „hőmérséklet”) függvényében. Ha a w' jelölt nem rosszabb, mint w , akkor ez a valószínűség mindig 1 (harmonikusabb jelöltre mindig át szabad lépni), ellenkező esetben pedig a szimuláció elején 1, majd fokozatosan lecsökken 0-ra. Ezért a véletlen bolyongó a szimuláció elején még ki tud szabadulni a lokális optimumokból, míg a végén belejük ragad, és ez a végállapot válik az algoritmus outputjává. Ha a T „hőmérsékletet” lassabban csökkentjük, vagyis több lépést, több iterációt engedünk meg, akkor sok esetben (de nem mindig, lásd a 6. fejezetet) megnő annak a valószínűsége, hogy a véletlen bolyongó az algoritmus végére megtalálja a globális optimumot, amelyet a hagyományos Optimalitáselmélet a grammatikus alaknak feleltet meg.

Az 1. fejezet bemutatja az Optimalitáselméletet és néhány változatát, valamint a sztochasztikus módszerek nyelvészeti relevanciáját veti fel. A 2. fejezet a heurisztikus módszerek mellett érvel, majd bevezeti az SA-OT Algoritmust. A 3. fejezet az Optimalitáselmélet matematikai megalapozását nyújtja, annak érdekében, hogy az SA-OT Algoritmust formális eszközökkel is bevezethesse. Bemutatja azt, hogy miképp lehet a jelöltek harmóniáját polinomokkal, valamint (transzformáció) rendszámokkal ábrázolni. A 4. fejezet spekulációi az SA-OT és a lexikon, illetve a tanulhatóság viszonyát feszegetik.

Az 5. fejezettől kezdve konkrét nyelvészeti példákon teszteljük az SA-OT-t. Az 5. fejezet a holland hangsúlyok eltolódását szimulálja gyorsbeszédben. A lassan lefuttatott szimuláció a két lehetséges alak normális tempójú beszédben előforduló gyakoriságait, míg a gyorsan lefuttatott szimuláció a gyorsbeszédbeli gyakoriságait hivatott visszaadni. A 6. fejezet a regresszív és progresszív zöngésségi harmóniát, míg a 7. fejezet a magyar névelő tapadását és a Prince és Smolensky-féle szótagolás-modellt tárgyalja.

Ezekben a fejezetekben a sok ponton támadható fonológiai modelleknél fontosabb az, ahogyan az SA-OT lehetőségeit fokozatosan kiaknázzuk. Az említett jelenségek ürügyén az algoritmus paramétereit és a különféle topológiákat teszteljük, különböző trükköket alkalmazunk. A rendszer viselkedését, a kísérletek mellett, a 6. fejezetben analitikus eszközökkel is igyekszünk megérteni. A 6. és a 7. fejezetben egy topológia-típust fokozatosan teszünk egyre összetettebbé. A 6. fejezetben arra is példát látunk, hogy egy nyelvtant meg lehet örizni egyszerűnek, ha a „szabálytalan” alakokat a performanciamodell területére száműzzük: a vizsgált példában az algoritmus mindig elő fogja állítani a globálisan nem optimális lokális minimumot, vagyis azt jósoljuk, hogy a nyelvtannak ellentmondó alak minden körülmények közt elő fog fordulni. Érveket láttunk amellett is, hogy a felszínen soha meg nem jelenő alakokat is bevegünk a jelöltek halmazába, és az SA-OT olyan constraintek rendezésében is segít, amelyeket a hagyományos OT nem tudna rendezni.

Az 5-7. fejezet tanulságait részletesebben a 8.1 alfejezet foglalja össze. A 8.2 alfejezet az Optimalitáselmélet korábban már tárgyalt változataival veti össze az SA-OT Algoritmust. Végezetül pedig az SA-OT-t visszahelyezzük abba a környezetbe, amelyben az Optimalitáselmélet eredetileg született, a kognitív tudományok közé.

Groningen dissertations in linguistics (GRODIL)

1. Henriëtte de Swart (1991). *Adverbs of Quantification: A Generalized Quantifier Approach.*
2. Eric Hoekstra (1991). *Licensing Conditions on Phrase Structure.*
3. Dicky Gilbers (1992). *Phonological Networks. A Theory of Segment Representation.*
4. Helen de Hoop (1992). *Case Configuration and Noun Phrase Interpretation.*
5. Gosse Bouma (1993). *Nonmonotonicity and Categorical Unification Grammar.*
6. Peter I. Blok (1993). *The Interpretation of Focus.*
7. Roelien Bastiaanse (1993). *Studies in Aphasia.*
8. Bert Bos (1993). *Rapid User Interface Development with the Script Language Gist.*
9. Wim Kosmeijer (1993). *Barriers and Licensing.*
10. Jan-Wouter Zwart (1993). *Dutch Syntax: A Minimalist Approach.*
11. Mark Kas (1993). *Essays on Boolean Functions and Negative Polarity.*
12. Ton van der Wouden (1994). *Negative Contexts.*
13. Joop Houtman (1994). *Coordination and Constituency: A Study in Categorical Grammar.*
14. Petra Hendriks (1995). *Comparatives and Categorical Grammar.*
15. Maarten de Wind (1995). *Inversion in French.*
16. Jelly Julia de Jong (1996). *The Case of Bound Pronouns in Peripheral Romance.*
17. Sjoukje van der Wal (1996). *Negative Polarity Items and Negation: Tandem Acquisition.*
18. Anastasia Giannakidou (1997). *The Landscape of Polarity Items.*
19. Karen Lattewitz (1997). *Adjacency in Dutch and German.*
20. Edith Kaan (1997). *Processing Subject-Object Ambiguities in Dutch.*
21. Henny Klein (1997). *Adverbs of Degree in Dutch.*
22. Leonie Bosveld-de Smet (1998). *On Mass and Plural Quantification: The case of French 'des'/'du'-NPs.*
23. Rita Landeweerd (1998). *Discourse semantics of perspective and temporal structure.*
24. Mettina Veenstra (1998). *Formalizing the Minimalist Program.*
25. Roel Jonkers (1998). *Comprehension and Production of Verbs in aphasic Speakers.*
26. Erik F. Tjong Kim Sang (1998). *Machine Learning of Phonotactics.*
27. Paulien Rijkhoek (1998). *On Degree Phrases and Result Clauses.*
28. Jan de Jong (1999). *Specific Language Impairment in Dutch: Inflectional Morphology and Argument Structure.*
29. H. Wee (1999). *Definite Focus.*
30. Eun-Hee Lee (2000). *Dynamic and Stative Information in Temporal Reasoning: Korean tense and aspect in discourse.*
31. Ivilin P. Stoianov (2001). *Connectionist Lexical Processing.*
32. Klarien van der Linde (2001). *Sonority substitutions.*
33. Monique Lamers (2001). *Sentence processing: using syntactic, semantic, and thematic information.*
34. Shalom Zuckerman (2001). *The Acquisition of "Optional" Movement.*

35. Rob Koeling (2001). *Dialogue-Based Disambiguation: Using Dialogue Status to Improve Speech Understanding.*
36. Esther Ruigendijk (2002). *Case assignment in Agrammatism: a cross-linguistic study.*
37. Tony Mullen (2002). *An Investigation into Compositional Features and Feature Merging for Maximum Entropy-Based Parse Selection.*
38. Nanette Bienfait (2002). *Grammatica-onderwijs aan allochtone jongeren.*
39. Dirk-Bart den Ouden (2002). *Phonology in Aphasia: Syllables and segments in level-specific deficits.*
40. Rien Withaar (2002). *The Role of the Phonological Loop in Sentence Comprehension.*
41. Kim Sauter (2002). *Transfer and Access to Universal Grammar in Adult Second Language Acquisition.*
42. Laura Sabourin (2003). *Grammatical Gender and Second Language Processing: An ERP Study.*
43. Hein van Schie (2003). *Visual Semantics.*
44. Lilia Schürcks-Grozeva (2003). *Binding and Bulgarian.*
45. Stasinios Konstantopoulos (2003). *Using ILP to Learn Local Linguistic Structures.*
46. Wilbert Heeringa (2004). *Measuring Dialect Pronunciation Differences using Levenshtein Distance.*
47. Wouter Jansen (2004). *Laryngeal Contrast and Phonetic Voicing: A Laboratory Phonology.*
48. Judith Rispens (2004). *Syntactic and phonological processing in developmental dyslexia.*
49. Danielle Bougairé (2004). *L'approche communicative des campagnes de sensibilisation en santé publique au Burkina Faso: Les cas de la planification familiale, du sida et de l'excision.*
50. Tanja Gaustad (2004). *Linguistic Knowledge and Word Sense Disambiguation.*
51. Susanne Schoof (2004). *An HPSG Account of Nonfinite Verbal Complements in Latin.*
52. M. Begña Villada Moirón (2005). *Data-driven identification of fixed expressions and their modifiability.*
53. Robbert Prins (2005). *Finite-State Pre-Processing for Natural Language Analysis.*
54. Leonoor van der Beek (2005) *Topics in Corpus-Based Dutch Syntax.*
55. Keiko Yoshioka (2005). *Linguistic and gestural introduction and tracking of referents in L1 and L2 discourse.*
56. Sible Andringa (2005) *Form-focused instruction and the development of second language proficiency.*
57. Joanneke Prenger (2005) *Taal telt! Een onderzoek naar de rol van taalvaardigheid en tekstbegrip in het realistisch wiskundeonderwijs.*
58. Neslihan Kansu-Yetkiner (2006) *Blood, Shame and Fear: Self-Presentation Strategies of Turkish Women's Talk about their Health and Sexuality.*
59. Mónika Z. Zempléni (2006) *Functional imaging of the hemispheric contribution to language processing.*
60. Maartje Schreuder (2006) *Prosodic Processes in Language and Music.*
61. Hidetoshi Shiraishi (2006) *Topics in Nivkh Phonology.*
62. Tamás Biró (2006) *Finding the Right Words: Implementing Optimality Theory with Simulated Annealing.*



GRODIL
 Secretary of the Department of General Linguistics
 P.O. Box 716
 9700 AS Groningen, The Netherlands