

*The emergence of French phonology* presents a computational framework for modeling the acquisition of sound patterns. It argues that in order to model this acquisition, not just phonological but also extra-phonological levels are necessary. To build on previous literature of phonological acquisition, especially in Optimality Theory, the framework is constraint-based. The thesis then presents a number of studies that utilize one such multi-level constraint-based framework, Bidirectional Phonetics and Phonology, to simulate a number of scenarios in both first and second language acquisition. In this way, it demonstrates the viability of a multi-level and data-driven approach to phonology.

JAN-WILLEM VAN LEUSSEN

# THE EMERGENCE OF FRENCH PHONOLOGY

## UITNODIGING

voor het bijwonen van  
de openbare verdediging van  
mijn proefschrift getiteld

## THE EMERGENCE OF FRENCH PHONOLOGY

op vrijdag 19 juni 2020  
om 12.00 uur

*Receptie na afloop*

JAN-WILLEM VAN LEUSSEN  
[jwvanleussen@gmail.com](mailto:jwvanleussen@gmail.com)

*Paranimfen*

MARGARITA GULIAN  
[margarita.gulian@gmail.com](mailto:margarita.gulian@gmail.com)

HANNAH VISCHER  
[hannahvischer@gmail.com](mailto:hannahvischer@gmail.com)

# The emergence of French phonology

Printed by Ipskamp Printing  
Typeset in L<sup>A</sup>T<sub>E</sub>X, based on a template by Alexis Dimitriadis

Cover artwork by Hannah Vischer

ISBN: 978-94-028-2069-0  
NUR: 616

Copyright © 2020 Johannes Willem van Leussen. All rights reserved.

# The emergence of French phonology

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor  
aan de Universiteit van Amsterdam  
op gezag van de Rector Magnificus  
prof. dr. ir. K.I.J. Maex  
ten overstaan van een door het College voor Promoties ingestelde  
commissie, in het openbaar te verdedigen  
op vrijdag 19 juni 2020, te 12.00 uur

door

Johannes Willem van Leussen

geboren te Gorinchem

**Promotiecommissie:**

Promotor:	Prof. Dr. P.P.G. Boersma	Universiteit van Amsterdam
Copromotor:	Dr. T. S. Biró	Eötvös Loránd Tudományegyetem
Overige leden:	prof. dr. A.P. Versloot	Universiteit van Amsterdam
	prof. dr. E.O. Aboh	Universiteit van Amsterdam
	dr. T.O. Lentz	Universiteit van Amsterdam
	dr. R. Bermúdez-Otero	The University of Manchester
	prof. dr. C. Lyche	Universitetet i Oslo
	prof. dr. C.C. Levelt	Universiteit Leiden

Faculteit der Geesteswetenschappen



The research reported in this thesis was funded by the Netherlands Organization for Scientific Research (NWO) grant number 277.70.008 awarded to Paul Boersma.

*“Science! Curse thee, thou vain toy; and cursed be all the things that cast man’s eyes  
aloft to that heaven, whose live vividness but scorches him, as these old eyes are even  
now scorched with thy light, O sun!”*

– Captain Ahab in Herman Melville’s *Moby-Dick*, chapter 118



---

## Contents

---

Acknowledgements . . . . .	xi
Author contributions . . . . .	xiii
<b>1 General introduction</b>	<b>1</b>
1.1 Introduction . . . . .	1
1.2 Theoretical and computational foundation . . . . .	2
1.3 L2 perception as a window onto L1 phonology . . . . .	2
1.4 A scalable multi-level learning model . . . . .	2
1.5 A corpus-based analysis of liaison . . . . .	2
1.6 Conclusion and discussion . . . . .	3
<b>2 A framework for multi-level constraint grammars</b>	<b>5</b>
2.1 Introduction . . . . .	5
2.2 Introduction to Optimality Theory . . . . .	5
2.2.1 Basics of OT evaluation . . . . .	6
2.2.2 OT as a linguistic model . . . . .	10
2.3 Learning and parsing . . . . .	14
2.3.1 Parsing hidden structure . . . . .	15
2.4 Stochastic ranking and gradual learning . . . . .	16
2.4.1 Stochastic ranking . . . . .	16
2.4.2 Gradual learning . . . . .	17
2.5 Multiple levels of representation . . . . .	19
2.5.1 Beyond two-level OT . . . . .	19
2.6 Bidirectional Phonetics and Phonology . . . . .	24
2.6.1 On the BiPhon model . . . . .	24
2.6.2 Why multiple levels of representation? . . . . .	25
2.7 Description of the simulation framework . . . . .	27
2.8 Conclusion . . . . .	29



<b>3</b>	<b>Learning to perceive and recognize a second language: the L2LP model revised</b>	<b>31</b>
3.1	Introduction	32
3.2	The L2LP model revised	36
3.2.1	Architecture of the L2LP-revised: levels and connections	36
3.2.2	Evaluating optimal paths	37
3.2.3	Sequential vs. interactive processing	40
3.2.4	Meaning-driven learning	41
3.3	Computational modeling with the L2LP-revised model	42
3.3.1	Acoustic input data for the simulated learners	43
3.3.2	Training and testing procedures	43
3.4	Modeling results	45
3.4.1	L1 learning	46
3.4.2	L2 learning	46
3.5	Discussion	49
<b>4</b>	<b>Efficient evaluation &amp; learning in multi-level parallel constraint grammars</b>	<b>51</b>
4.1	Multi-level parallel constraint grammars	52
4.2	Visualization of candidates with multi-level OT tableaux	54
4.2.1	The serial analysis of gender selection and liaison	54
4.2.2	The parallel analysis of gender selection and liaison	60
4.2.3	Candidate graphs	62
4.3	Efficient evaluation and learning	62
4.3.1	Constraints in the graph	62
4.3.2	Efficient evaluation: the elimination version	64
4.3.3	Efficient learning from meaning–sound pairs	71
4.4	Learning French gender allomorphy	73
4.4.1	The constraints	74
4.4.2	Grammars that work	74
4.4.3	Error-driven learning	79
4.4.4	The random baseline learner	80
4.4.5	Incremental learning procedures	81
4.4.6	A grammar with a constraint against phonetic schwa	86
4.4.7	Harmonic Grammar	86
4.4.8	Conclusion	87
4.5	The relation to complexity reductions for other parameters than number of levels	88
4.6	Conclusion	89
<b>5</b>	<b>Learning from corpus data in multi-level constraint grammars</b>	<b>91</b>
5.1	Introduction	91
5.1.1	A multi-level model of liaison	92
5.1.2	Using data from the PFC corpus	93
5.1.3	Aims and limitations of this study	93

5.1.4	Outline . . . . .	94
5.2	Model and data . . . . .	94
5.2.1	Reducing teleological bias . . . . .	94
5.2.2	Formalizing levels of representation and forms . . . . .	96
5.2.3	A dynamic and local GEN . . . . .	97
5.2.4	Level-by-level walkthrough of the extended model . . . . .	98
5.2.5	Constraining GEN and generating CON . . . . .	109
5.3	Simulation 1: toy French revisited . . . . .	113
5.3.1	Data set 1A: three-noun toy grammar . . . . .	114
5.3.2	Data set 1B: variation in schwa . . . . .	117
5.3.3	Data set 1C: Gender-allomorphic forms . . . . .	121
5.3.4	Data set 1D: Plural forms . . . . .	123
5.4	Simulation 2: PFC dataset . . . . .	128
5.4.1	Preprocessing of input data . . . . .	128
5.4.2	Method . . . . .	129
5.4.3	Simulation 2A: Learnability of serial and parallel gram- mars . . . . .	130
5.4.4	Simulation 2B: liaison consonants as abstract segments . . . . .	132
5.4.5	Simulation 2C: merely-lexical grammars . . . . .	134
5.4.6	Simulation 2D: lexically limited grammars . . . . .	135
5.5	Conclusion and discussion . . . . .	137
5.5.1	Interpretation of simulation results . . . . .	137
5.5.2	Suggestions for future research . . . . .	138
<b>6</b>	<b>Conclusion</b> . . . . .	<b>141</b>
6.1	Summary . . . . .	141
6.2	Implications and limitations . . . . .	142
6.2.1	Implications . . . . .	143
6.2.2	Limitations of the presented studies . . . . .	144
6.3	Suggestions for future research . . . . .	145
6.4	Conclusion . . . . .	146
<b>A</b>	<b>List of minimal pairs used as target lexical items in Chapter 3</b> . . . . .	<b>147</b>
	Bibliography . . . . .	149
	Summary . . . . .	163
	Samenvatting . . . . .	165



---

## Acknowledgements

---

When Paul Boersma suggested I apply for a PhD position in his NWO VICI project, I was honored. Paul's name strikes awe in the hearts of phonologists and phoneticians, and his eye for detail and commitment to scientific rigor are unparalleled. Equally inspiring is the amount of time he dedicates to helping others, be they a first-year student or a world-class researcher. His influence on the ideas presented in this book will be obvious.

My co-promotor Tamás Biró has likewise been instrumental in the completion of this thesis. I fondly remember our discussions in his Amsterdam office, where he approached phonological questions with the mathematician's poetic playfulness. After he moved to Budapest, there were many occasions where Tamás patiently spurred me on to keep working when my progress had all but ground to a halt.

Although it ultimately did not result in a thesis chapter, working with Gideon Borensztajn was one of the most rewarding periods of my years in the Bungehuis. Gideon's willingness to question fundamental assumptions was a much-needed reminder of what had once attracted me to the study of speech sounds.

Paola Escudero, co-author of Chapter 3, taught me much about the process of efficiently producing academic writing. Compared to the other chapters, our collaboration and its publication seemed to take place in the blink of an eye. Since then it has garnered a few dozen citations, a testament to Paola's keen insight.

For providing stimulating discussions and an enjoyable atmosphere, I am grateful to my office mates Kateřina Chládková and Karin Wanrooij, as well as other members of our research group not mentioned above: Titia Benders, Bart de Boer, Jeroen Breteler, Silke Hamann, Mirjam de Jonge, Wolfgang Kehrein, Sophie ter Schure, Klaas Seinhorst, Jasmin Pfeifer, Tessa Verhoef, Dirk Jan Vet and David Weenink. I also wish to thank the members of my doctorate committee for taking the time to read this work and act as opponents at the defense.

So far I have restricted these acknowledgements to people who contributed to the research in this thesis. Those close to me know that completing it was a struggle, which at times cast a shadow over my well-being. Like Ahab in *Moby-Dick*, I have found myself cursing science, or rather my decision to pursue an academic career. But the comparison is in many respects not apt. For one, the voyage of the *Pequod* lasted only four years, and resulted in a very entertaining book. More importantly, my story does not end in shipwreck. The leviathan is vanquished.

I could not have succeeded without a supporting crew of friends, family and colleagues who stood with me for more than a decade. Chief among these are my paranympths Margarita and Hannah. Many others played a role, and to rattle off your names in order of appearance would not do your contribution justice. I am profoundly in your debt, and I look forward to repaying you and celebrating with you in freedom.

Amsterdam, May 2020

---

## Author contributions

---

Chapter 3, titled *Learning to perceive and recognize a second language: the L2LP model revised*, was co-authored by Jan-Willem van Leussen (JL) and Paola Escudero (PE) and published in *Frontiers in Psychology* in August 2015. PE was chiefly responsible for section 1 (Introduction), while JL wrote sections 2, 3 and 4, and the simulations described therein. Section 5 (Conclusion) was written by both authors together.

Chapter 4, titled *Efficient evaluation and learning in multi-level parallel constraint grammars*, was co-authored by Paul Boersma (PB) and JL and published in *Linguistic Inquiry*, volume 48, issue 3, in July 2017. Most sections were written by PB and JL together, with PB taking the lead and JL making occasional suggestions or improvements. The simulations described in this chapter were implemented by both authors independently.

JL is the sole author of all other chapters.



# CHAPTER 1

---

## General introduction

---

### 1.1 Introduction

This thesis presents a computational framework for phonological acquisition – that is, the process by which people learn the sounds of a spoken language, together with rules about how these can combine to form meaningful words. It builds on a rich tradition of using computer simulations as a tool to investigate how learning shapes sound structure in language. In particular, the simulations conducted in this thesis are informed by Bidirectional Phonetics and Phonology (Boersma 1998 et seq; henceforth BiPhon). The stated aim of that model is to increase explanatory adequacy by being able to do *whole-language simulations*, capable of handling not just phonology but also its interface with other aspects of the grammars.

This work aims to get closer to this ambition, by expanding realism of the simulations in a number of ways. First, the input data presented to simulated learners is taken from empirical studies and corpora, to more closely mirror the input available to real learners. Second, by increasing the efficiency of learning and evaluation in the model, many different assumptions over the model and parameter settings can be tested systematically. The feasibility of these expansions is tested in a number of case studies.



## 1.2 Theoretical and computational foundation

Chapter 2 lays out the general properties of the framework that informs the three simulation chapters at the heart of the thesis. This framework models phonology as a constraint-based, stochastic, multi-level grammar. Phonological acquisition is modeled using learning algorithms that alter the prominence of constraints on the basis of input data. I argue that these properties, in particular the use of multiple levels of representation outside of phonology proper, are useful for arriving at a more complete model of phonological acquisition. The chapter introduces Bidirectional Phonetics and Phonology as the particular theory that informs the *representational* choices made in each chapter. However, it should be noted that the broader simulation framework is also compatible with other multi-level, constraint-based theories of linguistic structure.

## 1.3 L2 perception as a window onto L1 phonology

In Chapter 3, the simulation framework is used to test and refine the L2LP model, which builds on BiPhon to form a theory of the mechanisms that shape phonological perception in a second language, at the level of the individual learner. The L2 acquisition of Spanish front vowels by native speakers of Dutch is simulated, using data from vowel production studies in both languages. The predicted outcome is shaped not just by these input data but also by representational choices within a learner's grammar. By comparing different outcomes with data from perception studies, the relative merit of these choices can be scrutinized, and the predictive power of the model improved.

## 1.4 A scalable multi-level learning model

This chapter illustrates multi-level evaluation and learning on a minimal "toy" grammar of French liaison. It is shown that even for a small number of input data, the size of the candidate set generated for each input can grow exponentially in the number of levels. An efficient method for evaluation and learning is therefore proposed, whose complexity grows only linearly in the number of levels. The method is tested on the toy liaison grammar, illustrating again how the multi-level, data-driven approach can be used to evaluate the learnability of different analyses that yield the same surface data. Various approaches to learning in constraint grammars are also compared.

## 1.5 A corpus-based analysis of liaison

Chapter 5 presents a final simulation, which scales up the liaison learning model of Chapter 4 in two respects. First, a large number of input data with

and without liaison is obtained from a corpus of spoken French. Second, the number of candidates generated for each input datum is increased considerably, by implementing a candidate and constraint generation function with relatively few assumptions about possible morphological and phonological structures and their relations. Although the resulting grammars are many orders of magnitude larger, the efficient evaluation and learning algorithm of Chapter 4 keeps learning computationally tractable. After a walkthrough of the candidate and constraint generation mechanism, the results of two series of simulations are presented. The first series generates small “toy” languages similar to those of Chapter 4, to demonstrate the viability of the generation mechanism. In the second series, virtual learners are presented with an input data set generated on the basis of the corpus *Phonologie du Français Contemporain* (PFC, a corpus of spoken French). As in Chapter 3, variations of the base model are compared for their ability to successfully reproduce the patterns espoused by real speakers. For successful learners, the particular analyses chosen are inspected and compared with theoretical analyses from the literature, and with findings from studies of phonological acquisition.

## 1.6 Conclusion and discussion

A brief final chapter summarizes the results from Chapters 3 to 5 and evaluates them in light of the main theme of whole-language simulation. Implications and limitations of these results are discussed, leading to some suggestions for future research. The thesis concludes with some general methodological recommendations and remarks.



## CHAPTER 2

---

### A framework for multi-level constraint grammars

---

#### **2.1 Introduction**

This chapter establishes the framework for multi-level constraint grammars (MLCGs) that is used in the analyses and simulations presented in the rest of this work. Sections 2.2 and 2.3 introduce the central concepts of standard two-level Optimality Theory, learning through reranking, and stochastic evaluation. Sections 2.5.1 and 2.6 introduce MLCGs and the particular framework that informs the simulations described in subsequent chapters, Bidirectional Phonetics and Phonology. Finally, section 2.7 sets out the computational framework that underlies the simulations of Chapters 3, 4 and 5. While more specific implementational details are sometimes given in the relevant chapters, all simulations described within this work follow the general framework described in this chapter, and the reader may want to refer back to it in order to fully understand the details of the simulations in later chapters.

#### **2.2 Introduction to Optimality Theory**

This section introduces the basics of Optimality Theory (OT), the constraint framework underpinning the simulations in this thesis. Readers already familiar with OT may safely skip to Section 2.3.

### 2.2.1 Basics of OT evaluation

Optimality Theory was introduced by Prince and Smolensky (1993). It is chiefly used in the discipline of phonology, where it became the dominant paradigm in a matter of years after its introduction. However, it has also seen use in other subdisciplines of linguistics such as syntax (e.g. Bresnan, 2000; Legendre et al., 2001), and semantics, (e.g. Blutner et al., 2003, De Swart, 2009) and even outside of linguistics proper (for instance Jones, 2003 on kinship terms, Biró, 2011 on religious rituals; see Biró and Gervain, 2011 for more examples).

The central concept of OT is that the outcome of a given cognitive process is decided through an ordered set of **constraints**. These constraints express a (dis)preference for various properties of **candidates**, the possible outcomes of the process. Constraints are ordered by importance, and this **hierarchy** determines the relative desirability of a candidate (its **harmony**). The **evaluation** process, given a hierarchy and a candidate set, appoints the most harmonious candidate(s) as **optimal**: informally put, the candidate satisfying the most important constraints in the hierarchy emerges as the winner.

The idea that several conflicting constraints, laws or principles shape a decision is a familiar one, perhaps a contributing factor to OT's overnight success in phonology. Imagine a person entering a store to buy a pair of shoes.<sup>1</sup> Suppose that precisely three factors matter to this consumer: affordability, comfort and style. If we also make the simplified assumption that these three properties are binary and can be objectively determined, eight possible types of shoes can exist in the world:

Type A	expensive, uncomfortable and ugly
Type B	expensive, uncomfortable and stylish
Type C	expensive, comfortable and ugly
Type D	expensive, comfortable and stylish
Type E	cheap, uncomfortable and ugly
Type F	cheap, uncomfortable and stylish
Type G	cheap, comfortable and ugly
Type H	cheap, comfortable and stylish

If all types of shoe are available at the store, it is unlikely that anyone in their right mind will buy type A, since it truly has nothing going for it. Conversely, type H should fly off the shelves, as it is superior to all other types no matter what one looks for in a shoe. Within the subset of types B to G however, the optimal choice depends on the relative importance of the three factors. A frugal consumer less concerned with style will prefer a type G shoe to a type B or D. Optimality Theoretic writings usually visualize the decision process in a *tableau*. Figure 2.1 shows a tableau for our shoe world under a ranking where cost is the most important factor, followed by comfort and finally style.

Each row in the tableau represents a candidate, and the columns stand for

<sup>1</sup>A very similar example was used in Boersma (1998).

	*EXPENSIVE	*UNCOMFORTABLE	*UGLY
Type A	*!	*	*
Type B	*!	*	
Type C	*!		*
Type D	*!		
Type E		*!	*
Type F		*!	
Type G			*!
☞ Type H			

Figure 2.1: An Optimality-Theoretic tableau for our shoe world, with the ranking  $*EXPENSIVE \gg *UNCOMFORTABLE \gg *UGLY$ .

the constraints, ordered from left to right according to their standing in the hierarchy. Constraint names tend to be set in SMALL CAPS, and are often formulated as prohibitive commandments or negative imperatives: e.g. \*EXPENSIVE is shorthand for ‘Thou shalt not buy expensive shoes’. The content of the cells in a tableau indicate the **violations** that a candidate incurs on a constraint. Additionally, the shading of a tableau may indicate a candidate’s state in an evaluation procedure, which I call EVALUATION BY ELIMINATION. It can informally be stated as follows:

1. Starting with the highest-ranked constraint, count the number of violations it assigns to each active candidate.
2. Eliminate all candidates that incur more violations than the minimum found in the previous step.
3. Count the number of remaining candidates:
  - (a) If only one active candidate remains, this is the optimal candidate and the evaluation terminates.
  - (b) Else, go back to step 1 for the next highest-ranked constraint on the remaining active candidates.
4. If all constraints have been iterated over and more than one candidate remains, this set of candidates is optimal and the evaluation terminates.

All steps of this evaluation procedure are represented in a tableau. Step 1 begins in the leftmost column, at the highest-ranked constraint \*EXPENSIVE. Asterisks mark the number of violations this constraint assigns to each candidate. In step 2, we see that the lowest number of violations incurred by

\*EXPENSIVE is zero; therefore, candidates A to D have incurred a **fatal** violation and are eliminated. Exclamation marks signify that these violations are fatal, and the subsequent cells in these rows are grayed out to indicate that they are eliminated from the evaluation. In step 3, we check how many candidates are left after elimination; as no single optimal candidate has been found yet (3b), we move on to the next column in the tableau, repeating step 1 for constraint \*UNCOMFORTABLE over candidates E to H. Candidates E and F fatally violate this constraint, so that two candidates remain. Finally, \*UGLY eliminates candidate G and we are left with a single optimal candidate (3a): the impeccable affordable, comfortable and stylish shoe H. The pointing finger symbol  $\text{☞}$  indicates its optimality.

The EVALUATION BY ELIMINATION procedure illustrates **strict domination**, an important assumption of OT. Compare candidates D and E: the former violates just one constraint, the latter violates two. Nevertheless, under the ranking of Figure 2.1, D is suboptimal to E. The violation of a high-ranked constraint by D cannot be “outweighed” by its nonviolation of lower-ranked constraints. Strict domination is not a property of all constraint-based frameworks: Harmonic Grammar Legendre et al. (1990) and Maximum Entropy Goldwater and Johnson (2003), among others, use weighted violation marks in order to calculate the harmony of a candidate. In this case, the cumulative weighting of lower-ranked constraints may outweigh a higher-ranked constraint.

To know what the outcome of OT evaluation would be under a different ranking, one may mentally “switch” two columns in a tableau like Figure 2.1, leaving the violation marks (asterisks) intact but erasing and redrawing the fatal violation marks. But even without this operation, a glance at the tableau should make clear that H will be the optimal candidate from this set under any permutation of the constraints – since it never incurs more than the minimum number of violations (zero), it can never be subject to elimination. Likewise, no reordering of the constraints will change the optimality of candidate A, as it will always be suboptimal to any other candidate in the set. It will be more instructive to consider the subset consisting of shoe types B to G in order to see the effect of permuting the constraint hierarchy (Figure 2.2).

Within this more realistic subset of shoe types, the optimal candidate depends on the ordering of the constraints. Under the six possible permutations of the three-constraint set, types D, F and G each emerge twice as the optimal candidate. As long as the constraint that these respective candidates violate is ranked lowest in the hierarchy (e.g. \*UGLY for the inexpensive, comfy and unsightly type G), the order of the first two constraints does not influence the final outcome of evaluation. To frame this in linguistic terms, the same **language** that appoints type G as the optimal candidate may result from two distinct *hierarchies*: \*EXPENSIVE  $\gg$  \*UNCOMFORTABLE  $\gg$  \*UGLY and \*UNCOMFORTABLE  $\gg$  \*EXPENSIVE  $\gg$  \*UGLY. This has important consequences for Optimality Theory both as an analytical tool and as a model of language processing. If multiple constraint rankings yield the same observable (linguis-

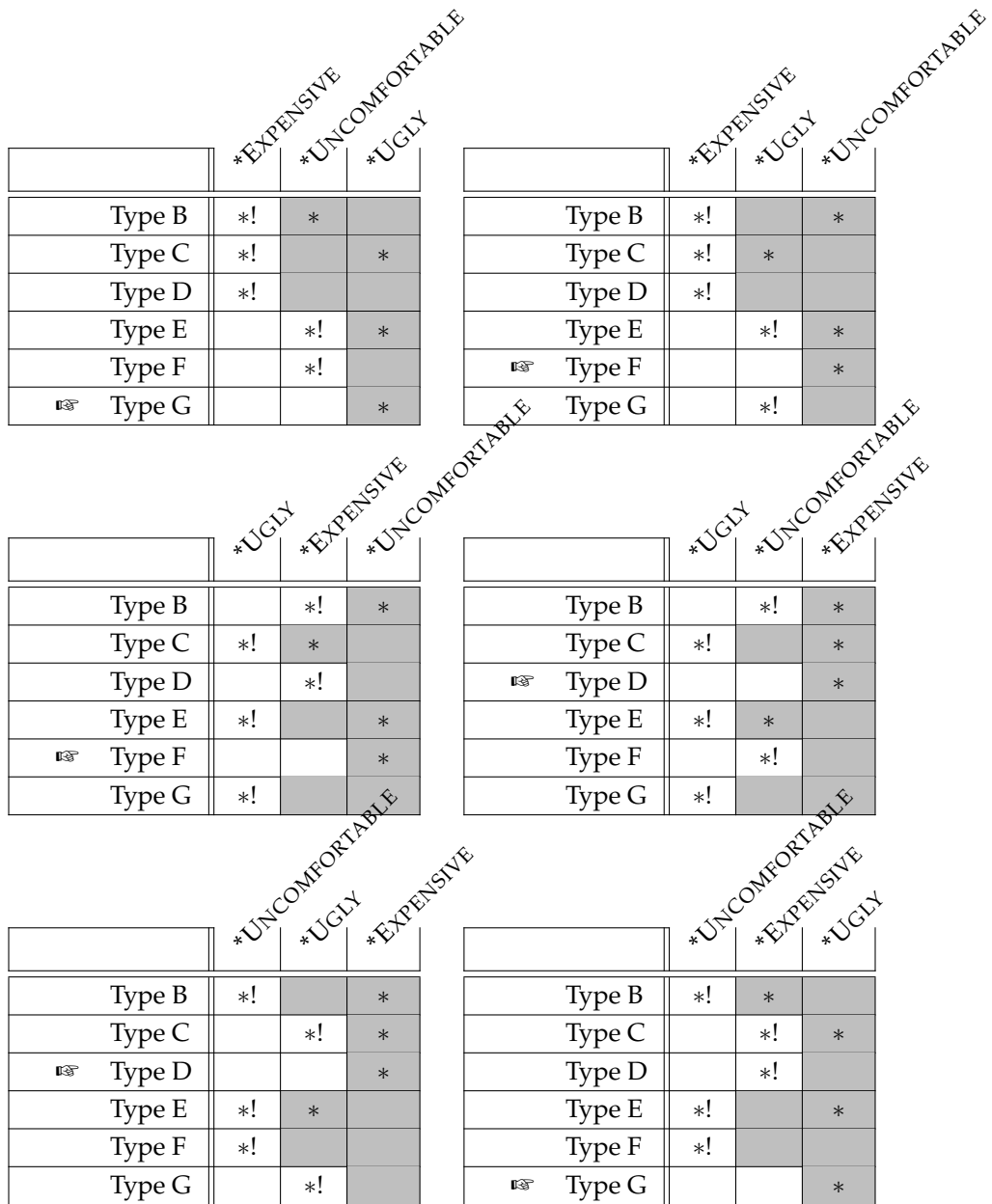


Figure 2.2: Six possible permutations for the tableau in Figure 2.1



tic) behavior, more data are needed in order to arrive at a definitive ranking in a phonological analysis; alternatively, it raises the possibility that actual language users may entertain distinct hypotheses about the underlying rules of the grammar, which then lead to the same surface data. In Chapters 4 and 5, the learnability of distinct rankings resulting in the same surface language is further explored.

Candidates B, C and E illustrate another important concept. Under no ordering of the three constraints will any of these three candidates emerge as optimal. In fact, this follows from comparing these candidates to others in the set. For instance, compare the violations inflicted on candidate B to those inflicted on candidate D. There is no constraint on which B incurs fewer violation marks than D, but there is at least one constraint on which B incurs more violation marks than D. Under step 2 of EVALUATION BY ELIMINATION, this means that a situation where D is eliminated but B is not can never occur, whilst the reverse should eventually occur if all constraints are considered. This guaranteed superiority of one candidate over another is called **harmonic bounding** (Samek-Lodovici, 1992, Samek-Lodovici and Prince, 1999). In this set of candidates, D harmonically bounds B and C, F harmonically bounds B and E, and G harmonically bounds C and E. These same bounding relations hold in the exhaustive candidate set of Figure 2.1. Additionally, A is bounded by all other candidates; and H bounds all other candidates in that set. The invincibility of type H regardless of ranking is a consequence of the latter fact.

Discussing harmonic bounding at length may appear unnecessary –why consider candidates that we know to be perpetually suboptimal?– but in fact the existence of bounded candidates increases the predictive power of the model. If we consider our footwear buying model to be an accurate reflection of a rational shoe consumer’s behavior, we expect that certain types of shoes will forever remain unsold as long as better (unbounded) alternatives are available. Likewise, given a set of constraints, OT predicts that certain linguistic forms will never be observed in human language, because they are universally suboptimal; it bounds the “hypothesis space” for the surface forms that we can expect to encounter in the languages of the world. Encountering such a form means that our hypotheses must be rejected or sharpened. The next section takes a closer look at standard OT’s application to phonological analysis.

### 2.2.2 OT as a linguistic model

In Prince and Smolensky (1993)’s foundational work on OT, an Optimality Theoretic grammar is defined as a composition of two functions. The first is the *generator* function, usually abbreviated GEN, which takes an *input* form that has been retrieved from the lexicon, and generates or derivates outputs as a set of candidate forms. This set is the input for the *evaluation* function, usually shortened to EVAL. On the basis of a constraint hierarchy, EVAL outputs the optimal candidate(s) from the candidate set, for instance through the elim-

/bid/	Max-IO	*VOICEDOBSCODA	Ident-IO[voi]
a. [bid]		*!	
b. [pid]		*!	*
c. [bit]			*
d. [pit]			**!
e. [bi]	*!		

Figure 2.3: Final devoicing in Dutch: underlying /bid/ surfaces as [bit].

ination procedure described above. In the strongest version of OT as originally formulated by Prince and Smolensky (1993), the available set of constraints, or CON, is *universal* in all languages; only the hierarchy over the constraints in CON is language-specific.

The notion that language is shaped by several conflicting constraints, forces or principles is much older than OT. The chief and most clearly opposing two principles are those of *clarity* and *economy* (Passy, 1890; Gabelentz, 1901; Martinet, 1955). We speak in order to be understood, and as such strive to pronounce words clearly and unambiguously. At the same time, we aim to minimize the effort and time spent on speaking. Assuming that the words of a language have some sort of canonical target form which is shared among speakers, maximizing clarity demands that a word be pronounced as closely as possible to this target. On the other hand, the need for economy may distort this canonical form through vowel reduction, consonant lenition, deletion, assimilation, and so forth. Clarity and economy are opposed to the extent that all linguistic forms violate one or the other principle to some degree. Just as the ideally cheap, comfortable and stylish shoe does not exist in the real world, no expression can be maximally distinct to a hearer while exerting minimal effort from a speaker. OT's notion of *violable* constraints allows a formalization of how languages resolve conflicts between these principles. A familiar (e.g. Kager, 1999) and useful example is the process of *final obstruent devoicing*. Figure 2.3 illustrates an Optimality Theoretic analysis of Dutch final devoicing.

Compared to the non-linguistic examples treated up to this point, the tableau introduces some new concepts and notation. The *underlying form* that serves as input to GEN, a string of phonemes traditionally placed between slashes, is written in the top left cell. The output candidates produced by GEN are *surface forms*, usually notated as a broad phonetic transcription between brackets. Candidates are enumerated (with letters instead of numbers) in order to refer to them easily in running text. Finally, note that unlike above, some candidates incur more than one violation mark on a constraint: our current formulation of EVALUATION BY ELIMINATION allows for this.

The three constraints in Figure 2.3 work as follows:

- MAX-IO assigns a violation mark for every segment in the input that is not represented in the output.
- \*VOICEDOBSCODA assigns a violation mark for every voiced obstruent in coda position in the output.
- IDENT-IO[VOI] assigns a violation mark for every segment whose value for the [voice] feature changes between input and output.

These formulations of the relation between constraints and candidates hint at an important dichotomy in the constraint set CON. MAX-IO and IDENT-IO[VOI] concern the relation between *input* and *output*: they respectively militate against deletion and changes in voicing. \*VOICEDOBSCODA on the other hand makes no mention of the input: it only constrains what is allowed on the output. Prince and Smolensky (1993) name the class of constraints concerned with the input–output mapping **faithfulness** constraints, and the class of constraints concerned only with outputs **markedness** constraints. This closely parallels the principles of clarity and economy mentioned above. A completely faithful mapping will be unchanged from the canonical form stored in the lexicon; to violate a faithfulness constraint is to obscure or neutralize some distinguishing feature of the underlying form, reducing its clarity. On the other hand, markedness is concerned with the relative desirability of certain segments or structures appearing in surface forms. For instance, placing a voiced obstruent in coda position appears to be slightly unattractive, given the fact that many languages avoid this altogether. This typological observation may be grounded in phonetic factors: devoicing obstruents in this position requires less articulatory effort or control (Ohala, 1983). Markedness constraints such as \*VOICED-OBSCODA punish forms which violate the principle of economy in this way.

Figure 2.3 shows how ordering a markedness constraint above a faithfulness constraint changes underlying /bid/ “(I) bid” into surface [bit], neutralizing the distinction with underlying /bit/ ‘beet’. Under the hierarchy MAX-IO >> \*VOICEDOBSCODA >> IDENT-IO[VOI], the unfaithful but less marked candidate c. is more optimal than the completely faithful candidate a. Nevertheless, the low-ranked IDENT-IO[VOI] still serves to eliminate candidates with superfluous changes in voicing, such as d. Candidates that would obviate the problem of voicing by deleting segments altogether (such as e.) are eliminated by the high-ranked MAX-IO.

The evaluation is simple enough that we may mentally picture what its tableau should look like for languages that allow voiced obstruents in coda, such as English. Switching the positions of \*VOICEDOBSCODA and IDENT-IO[VOI] makes a. the optimal candidate; this means that an equivalent tableau for underlying /bid/ ‘bead’ will yield faithful [bid]. The distinction with underlying /bit/ ‘beet’ is maintained at the price of a more marked or less economical articulation, although it is unlikely that native speakers of English

are conscious of this increased effort.<sup>2</sup> We can also construct two hierarchies where MAX-IO is dominated by the other two constraints. Under these hierarchies, candidate e. [bi] emerges as optimal. This more radical manifestation of “markedness over faithfulness” is often used to model the speech of young children, where processes like cluster simplification and coda deletion are common (see e.g. Demuth, 1995; Gnanadesikan, 1996; Smolensky, 1996).

Unlike the exhaustive constraint set of the previous section, the tableau in Figure 2.3 lists only a very small subset of what we may presume to be active constraints in Dutch or English phonology. Similarly, the set of candidates is limited to a select number of forms that demonstrate the behavior of the constraints. Indeed, nearly all publications that present Optimality-Theoretic analyses list only a “hand-crafted” set of constraints and candidates that are immediately relevant to the phenomenon at hand. An immediate practical reason is that interesting effects often arise from only a handful of interacting constraints, and enumerating scores of non-crucial constraints and suboptimal candidates would only muddle the presentation. A more problematic theoretical reason also exists. Prince and Smolensky (1993) explicitly formalize EVAL, but are less precise about CON and GEN. Among phonologists working in OT, a broad consensus on the precise contents of CON has yet to emerge.<sup>3</sup> Concerning GEN, under certain formulations it is capable of generating an *infinite* number of candidates, for instance if epenthesis of segments is unbounded. The informal EVALUATION BY ELIMINATION algorithm cannot find the optimal candidate in finite time on an infinite candidate set.

In a *computational* model of OT or other constraint grammars, such as will be pursued in this thesis, these concerns are not trivial. A number of authors (Ellison, 1994; Eisner, 1997; Riggle, 2004), suggest formulating GEN as a regular expression or finite state automaton, which makes evaluation over large or even infinite candidate sets feasible as long as certain conditions on constraints are met. Alternative solutions to the problem of infinite candidate sets have also been proposed by Tesar and Smolensky (2000) and Biró (2006). Regarding CON, an alternative view is that constraints are learned and language-specific instead of innate. This eliminates the problem of establishing a definitive CON, but introduces the question of how constraints may be deduced from linguistic data. While Chapters 3 and 4 employ “hand-crafted” constraint and candidate sets, Chapter 5 explores approaches to formalizing a (finite) GEN and creation of constraints, respectively.

<sup>2</sup>I ignore some phonetic details: the English lenis–fortis distinction is in fact one of aspiration rather than voicing, and vowels are shortened (clipped) whenever a fortis obstruent follows.

<sup>3</sup>A survey by Ashley et al. (2010) found that a total of 1,666 distinct constraints had been proposed in the literature by that time.

## 2.3 Learning and parsing

The previous section showed that different permutations of a constraint set will result in distinct surface preferences, such as absence or presence of final obstruent devoicing in different languages. Different hierarchies may also represent different stages in the acquisition of a language, such as the coda-deleting learner of Section 2.2. In this second case, the constraint *hierarchy* can be equated with a speaker’s internal knowledge of the phonological rules of their language, and the process of learning is then a question of **constraint reranking**. For example, learning to faithfully reproduce coda consonants requires that a learner raise MAX-IO’s standing in the hierarchy. More generally, acquiring the rules of a language can be modeled as permuting the ordering of constraints until a hierarchy is found that correctly reproduces the target language.

The idea of learning as constraint reranking has been pursued since OT’s inception, and various learning algorithms have been proposed in the literature. Perhaps the simplest learning algorithm imaginable is to iterate (randomly) over the possible permutations of a constraint set until the correct language emerges. In a three-constraint grammar like that of Figures 2.3 and 2.4, this algorithm will find a ranking for any of the three possible languages in maximally five iterations. This brute-force search of the **factorial typology** is not expected to scale well to larger constraint sets; even with a modest grammar of ten constraints, more than three million ( $10! = 3628800$ ) total orderings are possible. Nevertheless, this simple algorithm can serve as a baseline to which allegedly “smarter” learning algorithms may be compared (Jarosz, 2013b; see also Chapters 4 and 5).

Tesar (1995) proposed the first learning algorithm in a class that Magri (2012) calls *error-driven reranking algorithms* or EDRA’s. EDRA’s update the hierarchy on the basis of two candidates: one that is optimal for a given input under the current ranking, but considered *incorrect*; and one that is suboptimal under the current ranking, but deemed *correct*, i.e. congruent with the language to be learned. By comparing the constraint violation patterns of both candidates, two sets of constraints can be assembled: a set that is eligible for **promotion** in the hierarchy, and a set that is eligible for **demotion**. Usually, the constraints to be promoted are violated by the currently optimal, but incorrect candidate; and those to be demoted are violated by a correct, but currently suboptimal candidate.

Figure 2.4 gives an example of a *learning tableau*, where a learner of Dutch updates their hierarchy in such a way that it will produce Dutch-like final obstruent devoicing instead of English-like faithful voicing. The constraints selected for demotion and promotion are found by comparing the incorrect optimal candidate (✗) with the target suboptimal candidate (✓). Arrows indicate which constraints are promoted (←) and demoted (→). In a somewhat counterintuitive terminology, the literature often labels the currently optimal but

/bid/	MAX-IO	IDENT-IO[VOI]	*VOICEDOBSCODA
a. $\mathbb{L}$ [bid]			* ←
b. [pid]		*!	*
c. ✓ [bit]		→ *!	
d. [pit]		*!*	
e. [bi]	*!		

Figure 2.4: A learning tableau for Dutch devoicing. Arrows in a column indicate the direction into which the corresponding constraints will shift.

incorrect candidate ( $\mathbb{L}$ ) the *loser* and the non-winning but correct candidate the *winner*. Inspired by the notation of Biró (2013), I will instead abbreviate  $\mathbb{L}$  the *L*-candidate (standing either for *Learner* or *Loser*), and ✓ the *T*-candidate (*Target* or *Teacher* candidate).

### 2.3.1 Parsing hidden structure

Error-driven learners acquire a target language through *positive examples* (Tesar, 1995). For instance, a positive **learning datum** for Dutch may be to hear an adult coupling underlying /bid/ “(I) bid” to the surface pronunciation [bit]. The *T*-candidate /bid/ → [bit] is then fully provided by the linguistic environment. The *L*-candidate is produced by running the same input /bid/ through the current grammar, yielding perhaps /bid/ → [bid]. In this case, a *mismatch* has occurred: the learner notices that their output for /bid/ differs from that of the learning datum, and proceeds to update their grammar accordingly.

In many cases, this may be too simple a view of learning. Acquiring a language often requires interpreting *hidden structure*: the surface data available to learners is not fully specified in terms of its underlying structure. Tesar and Smolensky (2000) use the example of deriving hidden metrical structure from surface stress data. Under the assumption that all metrical feet are binary, a trisyllabic word like Polish [tɛˈlɛfɔn] ‘telephone’ may be parsed as containing either an iamb or a trochee (example from Jarosz, 2013a):

The learning datum /tɛlɛfɔn/ → [tɛˈlɛfɔn] is not congruent with the *L*-candidate /tɛlɛfɔn/ → tɛ(lɛˈfɔn) → \*[tɛlɛˈfɔn], triggering an update of the hierarchy. However, candidates b. and c. are both compatible with the positive example. The learner cannot recover which of the two candidates corresponds to the structure employed by adult speakers. Which should be chosen as the *T*-candidate?

Tesar and Smolensky (1998)’s solution is to select as *T*-candidate the optimal candidate, given the current hierarchy, among those congruent with the learning datum. They name this solution **Robust Interpretive Parsing** (henceforth RIP). We may see RIP as performing a second evaluation on a more re-

	/telefɔn/	ALLFT-R	IAMBIC	TROCHAIC	ALLFT-L
a.	(ˈtɛlɛ)ɔn [ˈtelefɔn]	*!	*		
b.	✓? (tɛˈlɛ)ɔn [tɛˈleɸɔn]	*!		*	
c.	✓? tɛ(ˈlɛɔn) [tɛˈleɸɔn]		*!		*
d.	⊗ tɛ(lɛˈɸɔn) [tɛlɛˈɸɔn]			*	*

Figure 2.5: Parsing problem: surface [tɛˈleɸɔn] may correspond to two candidates.

stricted subset of candidates: the optimal candidate under this second run of EVAL is the *T*-candidate.<sup>4</sup> Tesar and Smolensky (1998) test RIP in combination with the Error-Driven Constraint Demotion algorithm on a large test set of artificial languages with various types of underlying stress. They found that this combination of parsing and updating would at times get stuck in a local optimum, failing to converge on the target grammar. This has led several authors to propose alternative algorithms for both updating and parsing. Chapter 4 will explore the viability of a number of these algorithms.

## 2.4 Stochastic ranking and gradual learning

This section introduces two mechanisms that underpin all the constraint-based simulations used in this thesis: *stochastic ranking*, which introduces an element of randomness in the ranking of CON; and *gradual learning*, which uses stochastic ranking to effect incremental changes in the state of the constraint ranking.

### 2.4.1 Stochastic ranking

Under Prince and Smolensky’s 1993 original formulation, EVAL is fully deterministic: the same hierarchy will always lead to the same optimal candidate(s). Stochastic OT (Boersma, 1997, Boersma, 1998) is an extension of OT that enriches the constraints in CON with a real-valued **ranking value**. A hierarchy is produced by sorting the members of CON by their ranking value, in descending order. At evaluation time, however, each constraint’s ranking value is temporarily distorted by generating a random number from a normal distribution centered around zero. This random number is added to the ranking value. Constraints are then sorted in descending order by the resulting **disharmony** values. Following Jarosz (2013a), we call this pre-evaluation randomization **sampling**.

<sup>4</sup>In Tesar and Smolensky’s formulation, EVAL is *always* performed twice, and an update is triggered iff candidates *T* and *L* are not equal.

/bid/	Max-IO 102.0	Ident-IO[voi] 101.0	*VoicedObsCoda 100.0
☞ [bid]			*
[pid]		*!	*
[bit]		*!	
[pit]		*!*	
[bi]	*!		

Figure 2.6: A tableau with constraints sorted by ranking value.


Stochastic ranking allows constraint grammars to robustly deal with optionality, variation and frequency effects. Figure 2.6 shows a stochastic version of the obstruent devoicing grammar. The three constraints have ranking values of 102.0, 101.0 and 100.0, yielding the ranking MAX-IO  $\gg$  IDENT-IO[voi]  $\gg$  \*VOICEDOBSCODA and giving faithful /bid/  $\rightarrow$  [bid] as the winning candidate. However, after sampling this hierarchy, the resulting disharmonies may well yield a different ranking, as long as the **evaluation noise** parameter (the standard deviation of the normal distribution used to generate random values) is large enough.

Figure 2.7 shows the result of a simple simulation in which 100,000 hierarchies were sampled from the template of Figure 2.6, with evaluation noise set to 2.0. Under these settings, roughly a third of the sampled hierarchies retained the original ranking. The other five possible rankings are all found, albeit not with the same frequency: naturally, hierarchies where MAX-IO is ranked highest are more probable given its higher ranking value. Note that only three winning candidates result from the six rankings: the other two candidates are harmonically bounded. Stochastic ranking of constraints thus allows an OT grammar to exhibit variation for a given input. The relative frequencies (probabilities) of variant outputs are a function of the distance between ranking values of competing constraints, relative to the evaluation noise parameter.


## 2.4.2 Gradual learning

The distribution shown in Figure 2.7 reflects the final obstruent realization patterns of neither Dutch nor English: the former tends towards 100% devoicing, the latter towards 0%. Instead, we might consider this grammar to be a snapshot taken during a young learner's acquisition of English. When attempting to say /bid/ 'bead', this learner produces the correct form about half the time.




/bid/	Max-IO	Ident-IO[voi]	*VoicedObsCoda
 [bid]			*
[pid]		*!	*
[bit]		*!	
[pit]		*!*	
[bi]	*!		


33.82%

/bid/	Ident-IO[voi]	Max-IO	*VoicedObsCoda
 [bid]			*
[pid]	*!		*
[bit]	*!		
[pit]	*!*		
[bi]		*!	


21.13%

/bid/	Max-IO	*VoicedObsCoda	Ident-IO[voi]
[bid]		*!	
[pid]		*!	*
 [bit]			*
[pit]			**!
[bi]	*!		


21.05%

/bid/	*VoicedObsCoda	Max-IO	Ident-IO[voi]
[bid]	*!		
[pid]	*!		*
 [bit]			*
[pit]			**!
[bi]		*!	

8.92%

/bid/	Ident-IO[voi]	*VoicedObsCoda	Max-IO
[bid]		*!	
[pid]	*!	*	
[bit]	*!		
[pit]	*!*		
 [bi]			*

8.92%

/bid/	*VoicedObsCoda	Ident-IO[voi]	Max-IO
[bid]	*!		
[pid]	*!	*	
[bit]		*!	
[pit]		*!*	
 [bi]			*

6.17%

Figure 2.7: Relative frequencies of rankings when sampling the tableau in Figure 2.6 with an evaluation noise of 2.0.

At other times, she incorrectly devoices the final obstruent or even deletes it altogether. The acquisition process is not yet complete.

Besides modeling variation, stochastic ranking allows modeling *gradual learning* (Boersma 1998, Boersma and Hayes 2001). In gradual learning, promoting and demoting constraints is equated with adding to and subtracting from their ranking values. Instead of immediately reordering the hierarchy to accommodate the *T*-candidate, the relative harmony of the *T*-candidate over the *L*-candidate is raised slightly, increasing *T*'s probability of emerging as optimal. For our hypothetical learner of English, increasing the distance between VOICEDOBSCODA and IDENT-IO[voi] will lead to a more adult-like distribution.

The amount that is added/subtracted to the ranking values at each gradual learning step is called the learning **plasticity**, and may be decreased over the course of a learning simulation. Stochastic OT with gradual learning adds an element of randomness to learning which can prevent a learner from being misled by certain forms in the data into a grammar which is suboptimal for the language as a whole. Continuous ranking allows a learner to precisely tune its production to the variation exhibited in the target language.

Boersma's original EDRA for stochastic grammars, the Gradual Learning Algorithm (GLA, Boersma, 1998), divides the plasticity value equally over the promotion and demotion sets. Several alternative methods of distributing the plasticity have since been proposed (see also Magri 2012 for a discussion of the relative merits of different methods). Chapter 4 gives an overview of the **update rules** that are used for the simulations in this thesis.

## 2.5 Multiple levels of representation

The next section will go into more depth about the multi-level aspect of multi-level constraint grammars. It will also go into detail about serial versus parallel evaluation, a theme which is experimentally explored in all subsequent chapters, especially Chapters 3 and 5.

### 2.5.1 Beyond two-level OT

A strong position of Prince and Smolensky (1993)'s original formulation of OT is that evaluation occurs in parallel: CON contains constraints bearing on the input-output relation (faithfulness constraints) as well as constraints bearing on the form of the output (markedness constraints, the alignment constraints seen in 2.3.1). Within this parallel evaluation, any thinkable permutation of these types of constraints is allowed by the theory. Single-step parallel evaluation is a marked departure from the generative framework (Chomsky and Halle, 1968) that spawned OT; in generative phonology, surface structures result from the serial application of rewrite rules, feeding the output of one rule

as input to the next. Strictly speaking, OT works with two levels of representation: the input to GEN is the first of these, the output the second.

Several authors have proposed variants of OT with more than two levels, a class of constraint grammars which I will henceforth refer to as *multi-level constraint grammars* or MLCGs. Within this class, a further distinction can be made between *parallel* MLCGs, which retain original OT's insistence on free interaction for constraints that act on different levels of representation; and *serial* MLCGs, which divide EVAL and CON over the different levels of representation. Figure 2.8 illustrates two-level OT, parallel MLCG and serial MLCG.

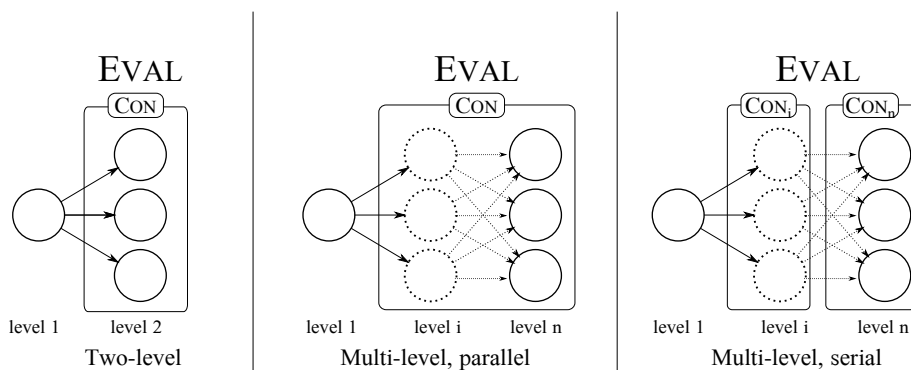


Figure 2.8: Schematic representation of standard two-level constraint evaluation (left), parallel multi-level evaluation (middle) and serial multi-level evaluation (right).

Many parallel OT analyses that have been proposed in the literature implicitly consider more than two levels of representation, and can be recast as taking place in parallel MLCG. For instance, the metrical parsing scenario of Tesar and Smolensky (2000), reproduced in Figure 2.5, introduced an intermediate 'hidden' metrical level of representation in between the underlying form and surface form, with constraints referring to the representations located on this level. However, Tesar and Smolensky (2000) still frame metrical parsing as a two-level input-output relation, where the output level contains both the hidden metrical representation and the overt phonetic representation (Figure 2.9). The mapping from the former to the latter is quite trivial, since it merely involves erasing the parsing brackets.

Figure 2.9 illustrates how some standard, two-level constraint grammar concepts translate to a MLCG view. Functionally speaking, both sides of the figure are equivalent: they visualize the candidate set depicted in the tableau of 2.5, showing the four ways in which underlying /tɛləfɔn/ can be mapped to a phonetic surface form. Conceptually, there are a number of distinctions to be made. Candidates in two-level constraint grammars are *pairs* of input and output like (/tɛləfɔn/, ('tɛlə)fɔn [tɛləfɔn]) whereas candidates in *n*-level

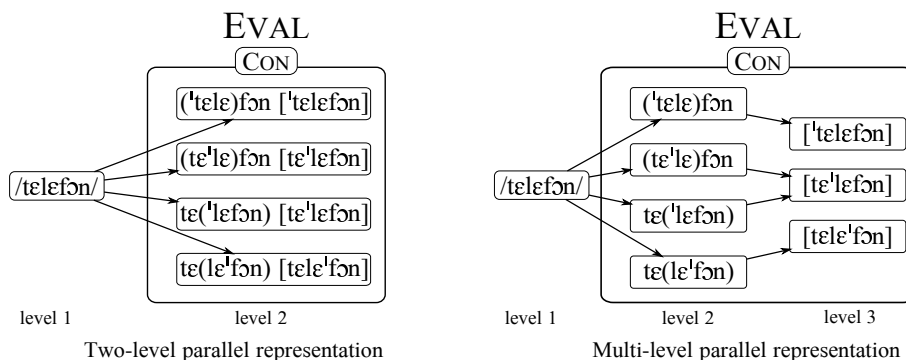


Figure 2.9: Two alternative views of the /tɛləfɔn/ metrical parsing scenario.

grammars are  $n$ -tuples like (/tɛləfɔn/, ('tɛlə)fɔn, [tɛləfɔn]). The latter notation allows us to be more formally precise about properties of candidates and the learning scenario. Expressed as 3-tuples (i.e. triples), the two possible  $T$ -candidates (/tɛləfɔn/, (tɛ'lɛ)fɔn, [tɛ'lɛfɔn]) and (/tɛləfɔn/, tɛ('lɛfɔn), [tɛ'lɛfɔn]) share their third element, which is what makes them suitable parses for the learning datum /tɛləfɔn/ → [tɛ'lɛfɔn].

More generally, the  $n$ -tuple representation allows breaking candidates into *sub-candidates*, together with the constraint violations incurred by these sub-candidates. Together with the graph-like representation hinted at in Figure 2.9, this allows optimization in the evaluation of large multi-level candidate sets. These optimizations are explored further in Chapter 4, and used to advantage in the large-scale case study of Chapter 5.

Serial MLCG models use the constraint evaluation mechanism of OT, but reject Prince and Smolensky (1993)'s assertion that (morpho)phonological evaluation is strictly parallel. An example is Stratal OT (Bermúdez-Otero, 1999; Kiparsky, 2000), which posits that evaluation of phonological forms passes through multiple *cycles* corresponding to the morphological *word*, *stem* and *affix* levels. Such a model deals quite naturally with phonological processes that are sensitive to different types of morphological boundaries. By dividing the complete constraint set CON of the grammar into multiple strata CON<sub>1</sub>...CON <sub>$n$</sub>  for  $n$  levels of representation, an optimal intermediate form is decided at each level up to the output level  $n$ . Stratal OT and other serial constraint evaluation frameworks are thus explicitly multi-level, with both candidates and constraints distinguishing multiple levels of representation, and constraints assigning violation marks only to mappings and forms on specific levels.

One may question the above view of serial OT models where only CON is split into multiple strata: why not consider serial grammars as performing serial *evaluation*, i.e. splitting EVAL into multiple strata together with CON and taking the output of each evaluation as input to the next? In fact the two approaches yield equivalent results: OT's principle of strict domination guar-

antees that a stratified CON will evaluate in a serial manner. As an example, Figure 2.10 depicts a candidate set consisting of four 3-tuples: (1A, 2A, 3A), (1A, 2A, 3B), (1A, 2B, 3B) and (1A, 2B, 3C). Let us suppose that two “markedness” constraints militate against the forms 2A and 2B respectively, and that two “faithfulness” constraints militate against the mappings (2A – 3B) and (2B – 3C). The markedness constraints (operating on forms located on level 2) are located in a higher stratum than the faithfulness constraints (operating on the relation between forms located on level 2 and 3). Figures 2.11 and 2.12 illustrate that a stratified CON with monolithic EVAL will output the same winning candidate (1A, 2A, 3A) as a two-step EVAL/CON.

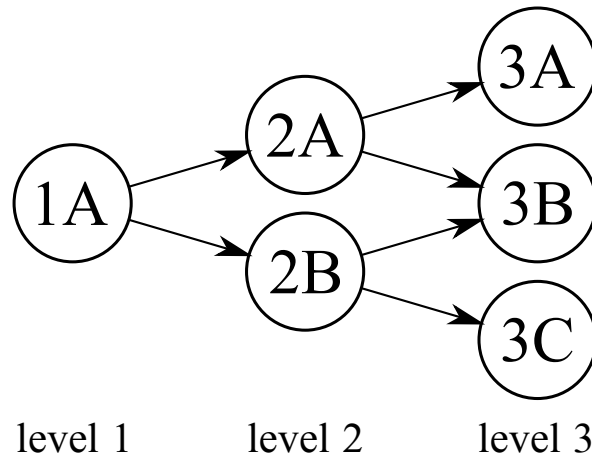


Figure 2.10: Graph representation of a small four-candidate set in a three-level grammar.

Representing serial MLCGs with a single evaluation over a stratified CON offers a number of advantages. First, it stays close to OT’s original conception,

Input: 1A	*2B	*2A	*2A-3B	*2B-3C
<i>stratum:</i>	1	1	2	2
a. (1A, 2A, 3A)		*		
b. (1A, 2A, 3B)		*	*!	
c. (1A, 2B, 3B)	*!			
d. (1A, 2B, 3C)	*!			*

Figure 2.11: Serial evaluation of the candidate set of Figure 2.10, with single pass through EVAL.

	1A	*2B	*2A
a.	(1A, 2A)		*
b.	(1A, 2B)	*!	

	2A	*2A-3B	*2B-3C
a.	(2A, 3A)		
b.	(2A, 3B)	*!	

Figure 2.12: Serial evaluation of the candidate set of Figure 2.10, with two passes through EVAL.

and as such is directly compatible with most superimpositions on OT such as learning with EDRA and stochastic, real-valued ranking. Secondly, for the purposes of this thesis, it allows a very close comparison between serial and parallel evaluation over multi-level grammars with all else being equal: the only difference is that for the latter, all constraints are located within the same stratum. This direct comparison will be explicitly made in Chapter 3.

The constraints of Figures 2.11 and 2.12 were referred to as ‘faithfulness’ and ‘markedness’ constraints; however, these terms refer specifically to the underlying and surface representations of standard two-level phonology in OT. In MLCGs, we can view them as instances of two basic types of constraints: **interlevel** and **intralevel** constraints, respectively. The first type concerns mappings between adjacent levels, the second concerns *forms* within a level. A crucial assumption in the unified MLCG framework presented here is that all constraints must belong to either the interlevel or intralevel type. Constraints that consider forms on more than two levels of representation, or on nonadjacent levels, are not considered in this thesis. This locality restriction allows computational and representational gains, as Chapter 4 will demonstrate.

A second important remark is that constraint strata *need not* correspond to the levels of representation in a serial MLCG, as is the case in Stratal OT. A constraint stratum may cover two or more levels of representation. For instance, in Chapter 3, a four-level MLCG will be introduced which is divided into two (not three) strata. Conversely, different strata may pertain to the same level(s) of representation, as in Ito and Mester (2009).

The next section introduces a specific example of MLCG modeling: the BiPhon framework of speech perception and production, which forms the basis for the simulations described in this thesis.

## 2.6 Bidirectional Phonetics and Phonology

This section introduces Bidirectional Phonetics and Phonology, the theoretical framework central to the simulations conducted in this thesis. It also gives some rationale for going beyond two levels of representation in a model of sound learning.

### 2.6.1 On the BiPhon model

Bidirectional Phonetics and Phonology (henceforth BiPhon) was first conceived as Functional Phonology by Boersma (1998), and has since undergone a name change to reflect its growing scope. In a programmatic paper, Boersma (2011) describes BiPhon as

an Optimality-Theoretic (OT) grammar model that is intended to be capable of handling all of phonology: its representations with their relations, its processes with their relations, its connection to the semantics, its acquisition by the child, its evolution over the generations, and its typology across languages.

Many of the characteristics that define the framework follow from the ambitions stated in this quote. BiPhon is *multi-level*: besides the Underlying and Surface form of two-level OT, it presumes additional levels of representation pertaining to semantics/morphology and phonetics. **Constraints** govern the relations and mappings between representations, and it is explicitly designed as a computational model of **learning** and **variation** through reranking of stochastic constraints.<sup>5</sup> A final important tenet is **bidirectionality**: evaluation over a single constraint ranking governs both perception and production.

Stochastic evaluation and gradual learning have proven to be the most influential properties of the model, inspiring a considerable amount of work on variation and learning in constraint-based grammars. The framework as a whole has seen less adoption in “mainstream” phonology, but BiPhon-based multi-level analyses of various phenomena have been proposed: e.g. loanword adaption in Korean (Boersma and Hamann, 2009), *h-aspiré* forms in French (Boersma, 2007), metrical phonology in Latin and Modern Greek (Apoussidou, 2007), prepositional allomorphy in Czech (Chládková, 2009), as well as a plethora of work exploring the phonetics–phonology interface (e.g. Boersma, 2009a). The Second Language Linguistic Perception model of Escudero (2005), further explored in Chapter 3, is an application of BiPhon to L2 phonology.

Figure 2.13 depicts the six levels of representation and the constraints that govern the relation between them. An extensive walkthrough of the levels

<sup>5</sup>The constraint-based approach is not a necessary ingredient of the framework; since 2013, BiPhon’s bidirectional multi-level approach has also been used in neural network models, replacing forms and constraints with nodes and weighted connections. In this thesis however, BiPhon will be taken to mean its OT-based variant.

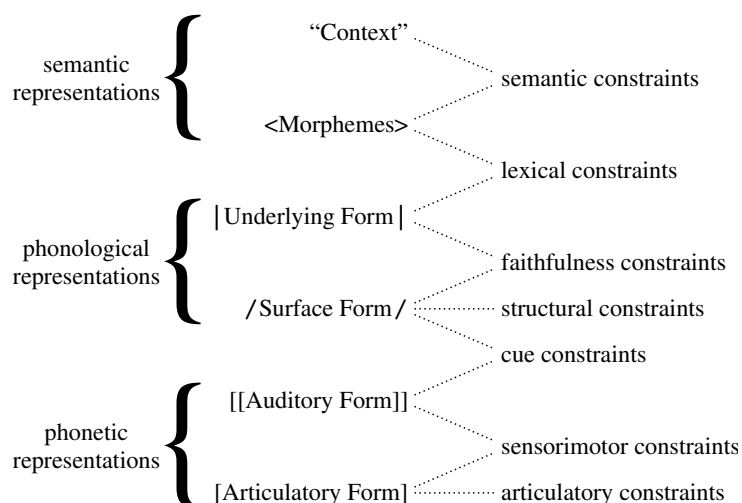


Figure 2.13: Overview of levels and constraints in the BiPhon Model. From Boersma (2009a).

and constraints, including a detailed motivation for their inclusion in a model of speech perception and production, can be found in Boersma (2011). Subsequent chapters will go into more detail about the representations and constraints used in those specific studies. Note however that BiPhon uses a slightly different notation for different levels of representation than standard two-level OT: underlying form is notated with `|square brackets|`, and surface form with `/slashes/`. Unless otherwise stated, this book conforms to the BiPhon notation.

## 2.6.2 Why multiple levels of representation?

Purely-phonological approaches treat phonology as an autonomous component of language, and its “interface” with other modules is accounted for only implicitly. Indeed, some researchers argue that phonological models ought to be completely “substance-free”, devoid of representations and rules that refer to phonetics (Hale and Reiss, 2000; Blaho, 2008). On the other hand, many parallel phonological OT analyses do include constraints referring to other components of the grammar, but do not make a formal distinction between these different components (e.g. Tranel, 1996).

Both approaches are adequate for descriptive analyses of a well-circumscribed phonological process in a given language or set of languages. These processes often extend to novel (loan) or nonce words, indicating an active phonological component and vindicating the explanatory adequacy of these two-level grammars. However, many phonological phenomena display



complex interactions with phonetics, lexicon or morphology, and cannot be fully described without referring to non-phonological representations. This becomes more apparent in cases where we can trace the historical emergence (“phonologization”) of a process.

Including non-phonological representations in our model lets us formalize various forces that shape phonological grammars and inventories of individual speakers: categorical perception, functional load, stylistic variation, second-language acquisition, reanalysis, and so forth. By explicitly modeling the role of the *language learner*, our phonological analyses come a step closer to not only describing but perhaps also *explaining* some sound patterns of spoken languages.

### **Formalizing extra-phonological relations**

A two-level view such as that presented in Section 2.2.2 provides an elegant model of many phonological processes such as final obstruent devoicing, and allows easy comparison between languages that do and do not exhibit this phenomenon. By distinguishing Underlying and Surface representations, we can view the sounds as being underlyingly distinct ( $|d|$  versus  $|t|$ ) yet surface-identical ( $/t/$ ). Neutralization is a many-to-one mapping from segments on UF to a segment on SF. A one-to-many mapping, where a single underlying segment surfaces in distinct ways, describes allophony, a zero-to-one mapping is insertion, its reverse is deletion, and so forth.

By formalizing the relation between elements within and between these two levels through constraints, OT can be used to differentiate and predict possible outcomes of a phonological process. Outside of phonology proper, we may likewise formalize many familiar concepts in terms of inter- and intralevel relations. Taking the same approach to formalizing representations outside phonology allows us to create or generate constraint grammars that can hold precise hypotheses about these non-phonological processes, with constraints expressing preferences for certain interlevel mappings or intralevel configurations.

### **Learners cannot access phonological representations**

The metrical learning model of Tesar and Smolensky (2000) characterizes a difficulty inherent in phonological learning: the phonological representations employed by a speaker cannot be directly accessed by a listener. Often, a phonetic signal is ambiguous, allowing for multiple possible representations on the Surface and Underlying Form levels. To disambiguate and analyze the signal, real listeners rely partly on knowledge of the language, and partly on knowledge about the world and discourse context. A comprehensive model of phonological acquisition must account for this ambiguity.

Phonological representations should not serve directly as input to the learning process, since they are in reality hidden from the learner. This is another ar-

gument for including extra-phonological levels of representation. In the BiPhon model, only the Context and Phonetic Forms are directly accessible to learners. The covert intermediate representations are “parsed” on the basis of the current grammar. As in the shoe world of Section 2.2, a consequence is that distinct grammars (i.e. different hierarchies) may produce the same language, or one which varies from the target language only slightly. In this way, MLCGs can account for reanalysis as a force of language variation and change.

### Data-driven evaluation of models

The possibility of distinct hidden representations leading to the same output forms opens up interesting methodological questions. If two or more instances of an MLCG model can replicate a set of linguistic data, but differ with respect to the hidden representations employed, which is the preferred analysis? One approach is to prune hypotheses based on properties of the models themselves. A well-known criterion is Occam’s Razor: the model that makes the fewest assumptions while accounting for all the facts is superior. Parsimony or explanatory power is also often invoked as an argument in favor of certain analyses: it is desirable that a small set of rules or constraints can account for many phenomena. However, the computational approach that informs the simulations of this book is primarily *data-driven*. Using the constraint-based modeling framework, we consider the learnability of different grammars in the face of large and varied data sets culled from experiment and corpus studies. This gives us an additional criterion to distinguish between models that correctly predict overt forms using different covert representations: we should prefer the model that is learnable on the basis of the data available to the learner.

## 2.7 Description of the simulation framework

In Chapters 3, 4 and 5, various BiPhon-based MLCGs will be trained and tested on different data sets. The EDRA simulation framework, however, is basically the same in all chapters, forming the foundation of the research presented in this thesis. Where possible, the simulations also reuse parameter values of previous work in OT learning. e.g. Boersma and Hayes (2001). Algorithm 1 outlines the basic procedure in pseudocode.

Training data come in the form of *pair distributions*, tables listing meaning-form pairs together with their relative frequency. A single learning trajectory runs for a predefined number of steps. At each step, a **learning datum** ( $D_{In}, D_{Out}$ ) is drawn at random from the data, where  $D_{In}$  is the overt input form from the first level and  $D_{Out}$  is the overt output form from the last level of an MLCG. A pair’s probability of being drawn is derived from its relative frequency in the distribution. The generator function GEN – which is a universal property of the model, rather than an input or parameter of the algorithm

**Algorithm 1** The main training procedure used in the simulations

---

```

1: Input A constraint hierarchy  $H$  and pair distribution  $P$ 
2: Parameters  $nSteps$ ,  $nEpochs$ ,  $evalNoise$ ,  $initPlasticity$ ,  $decay$ 
3:  $stepsPerEpoch \leftarrow nSteps \div nEpochs$ 
4:  $plasticity \leftarrow initPlasticity$ 
5: for  $i \leftarrow 1$  to  $nSteps$  do
6:    $(D_{In}, D_{Out}) \leftarrow P.drawDatum()$ 
7:    $Cans \leftarrow GEN(D_{In})$ 
8:    $H' \leftarrow sample(H, evalNoise)$ 
9:    $L \leftarrow eval(H', Cans)$ 
10:  if  $L_{Out} \neq D_{Out}$  then
11:     $Cans' \leftarrow GEN(D_{In}).filter(contains(D_{Out}))$ 
12:     $T \leftarrow eval(H', Cans')$ 
13:     $H \leftarrow update(H, L, T, plasticity)$ 
14:  if  $i \bmod stepsPerEpoch = 0$  then
15:     $plasticity \leftarrow plasticity \times decay$ 
16: Result A trained hierarchy  $H$ 

```

---

– creates a candidate set  $Cans$  from one half of this pair. Next, the current constraint hierarchy  $H$  is sampled (see Section 2.4.1) yielding  $H'$ . EVAL then computes the “learner candidate”  $L$  on the basis of  $H'$ . If the output overt form  $L_{Out}$  is the same as  $D_{Out}$ , the learner proceeds directly to the next step. If not, the learner has erred and an update will be performed. The optimal teacher candidate  $T$  is parsed by applying EVAL on  $Cans'$ , the subset of  $Cans$  that contains  $D_{Out}$ . One of various update algorithms then shifts the rankings in  $H$  based on the violation profiles of  $L$  and  $T$ , by an amount based on the current plasticity. After the update, the learner proceeds to the next step. The total number of steps is divided into a number of *epochs*. At the end of each epoch, the plasticity value is multiplied by a decay factor known as *plasticity decay*, a positive number less than or equal to 1. As an example, if  $nSteps$  is 10000 and  $nEpochs$  is 4,  $stepsPerEpoch$  will be 2500. Under these settings, with an initial plasticity of 1.0 and a plasticity decay of 0.1, the plasticity will be lowered to 0.1, 0.01 and 0.001 at step 2500, 5000 and 7500.

### Initial state of the grammar

In most of the simulations described in this thesis, the initial ranking of constraints will be equal, set to 100.0.<sup>6</sup> This means that no assumptions are made about an initial “markedness over faithfulness” representation (Gnanadesikan,

<sup>6</sup>An initial ranking of 100.0 is standard in most BiPhon literature. A beneficial effect of choosing a positive number is that such a ranking will also work for a Harmonic Grammar evaluation where constraint weightings are summed; if ranking values are allowed to drop under zero, a violation would actually improve the optimality of a candidate.

1996; Tesar and Smolensky, 2000) . An exception is Chapter 3, where a faithfulness relation between different levels of representation is expressed through a lower constraint ranking for some mappings.

## **2.8 Conclusion**

This chapter presented a brief overview of the theoretical and computational framework that drives the simulations in this thesis. Each of these simulations makes use of gradual learning and stochastic evaluation in a multi-level, constraint-based grammar to investigate aspects of modeling phonological acquisition. It was also explained that multi-level grammars, employing levels of representation outside of the traditional Underlying and Surface Form of two-level OT, may be more adequate to model some cases of sound learning. The subsequent chapters go into more detail about the particular representations and simulation choices made for those specific case studies.



## CHAPTER 3

---

### Learning to perceive and recognize a second language: the L2LP model revised

---

#### *Abstract*

We present a test of a revised version of the Second Language Linguistic Perception (L2LP) model, a computational model of the acquisition of second language (L2) speech perception and recognition. The model draws on phonetic, phonological, and psycholinguistic constructs to explain a number of L2 learning scenarios. However, a recent computational implementation failed to validate a theoretical proposal for a learning scenario where the L2 has less phonemic categories than the native language (L1) along a given acoustic continuum. According to the L2LP, learners faced with this learning scenario must not only shift their old L1 phoneme boundaries but also reduce the number of categories employed in perception. Our proposed revision to L2LP successfully accounts for this updating in the number of perceptual categories as a process driven by the meaning of lexical items, rather than by the learners' awareness of the number and type of phonemes that are relevant in their new language, as the previous version of L2LP assumed. Results of our simulations show that meaning-driven learning correctly predicts the developmental path of L2 phoneme perception seen in empirical studies. Additionally, and to contribute to a long-standing debate in psycholinguistics, we test two versions of the model, with the stages of phonemic perception and lexical recognition being either sequential or interactive. Both versions succeed in learning to recognize minimal pairs in the new L2, but make diverging predictions on learners' resulting phonological representations. In sum, the proposed revision to the L2LP model contributes to our understanding of L2 acquisition, with implications for speech processing in general.

---

### 3.1 Introduction

Adult second language (L2) learners often struggle to understand native speech and to make themselves understood by native speakers. One important reason behind this difficulty seems to be that adult learners rely on the rules and categories of their own native language (L1) when learning to perceive and produce L2 sounds. Numerous experiments have demonstrated the influence of L1 perception and the specific problems it causes for L2 learners: for instance, troublesome English minimal pairs are “rocket” and “locket” for Japanese speakers (Aoyama et al., 2004), “beat” and “bit” for Spanish (Flege et al., 1997) and Portuguese (Rauber et al., 2005) speakers, or “bet” and “bat” for Dutch speakers (Broersma, 2005). The overarching cause of these problems is that these specific sounds do not contrast in these learners’ L1 phoneme repertoires. In other words, novel L2 contrasts are difficult to perceive and produce.

Linguistic experience is therefore at the core of current theories and models of L2 perception and production, which advance proposals and predictions based on how L1 speech sounds compare to those in the new language. Three such theories, the Perceptual Assimilation Model (PAM; Best 1995) and its extension to L2 learning (PAM-L2; Best and Tyler 2007), the Speech Learning model (SLM; Flege 1995; Flege et al. 2003) and the Second Language Linguistic Perception model (L2LP; Escudero 2005, 2009) explain how L1 experience influences L2 sound learning in a number of learning scenarios. We sketch three such scenarios and their predicted result in L2LP, SLM, and PAM-L2 below. Unlike the other two models which account for either naïve non-native and beginning L2 perception (PAM and PAM-L2) or L2 speech learning (SLM), as reviewed in Tyler et al. (2014), L2LP aims at modeling the entire developmental process of L2 speech perception, from naïve, non-native to advanced, native-like performance. L2LP therefore proposes precise learning tasks and developmental trajectories for learners, depending on the learning scenario with which they are confronted, and comes with a computational learning model within the connectionism-inspired learning framework of Stochastic Optimality Theory (Boersma, 1998).

The basis for all predicted L2 learning trajectories in L2LP is the optimal perception hypothesis (Escudero, 2005, 2009). This states that learners will initially perceive L2 sounds in a manner resembling the production of these same sounds in their L1 environment. The L2LP model thus explicitly represents the result of L1 acquisition as the initial state of L2 learning, predicting that acoustical differences and similarities between the phonemes of two languages will shape development. From this starting point, three scenarios can be distinguished. Unlike the SLM which deals with isolated L2 sounds, both the L2LP and PAM make predictions for the perceptual development of sound contrasts. When the majority of productions of an L2 contrast are acoustically closest to typical or average productions of a single L1 sound, learners face

what L2LP calls a NEW scenario and PAM calls single category assimilation (Best, 1995). Learners facing this scenario must either create a new L2 category or split their existing single L1 category. L2LP and PAM predict that this is a difficult scenario for L2 learners, and the experimental studies cited above confirm this. In contrast, when the majority of the tokens of an L2 contrast are acoustically closest to the typical productions of two separate L1 sounds, learners are faced with a SIMILAR scenario (PAM: two-category assimilation). According to the L2LP, in this scenario, the existing L1 categories are simply replicated and then adjusted so that their boundaries will come to match those of the L2 contrast, as there is hardly ever a perfect match between the productions of an L1 and L2 contrast. PAM and L2LP predict that this shifting is less problematic than creating new categories (Escudero et al., 2014), while Flege's SLM predicts that new sounds would be easier to learn than similar or old sounds (Flege, 1995). However, since the SLM focuses on single sounds and not on sound contrasts, as the PAM and L2LP models, a comparison of predictions across models may not be straightforward.

A third possible case only considered by the L2LP and PAM is the SUBSET scenario, which may be comparable to what is called uncategorized or categorized-uncategorized assimilation, depending on how each of the members of the contrast are assimilated to native categories, in PAM. It takes place when a single non-native sound is perceived as more than one L1 category, so-called multiple category assimilation within the L2LP (Escudero and Boersma, 2002; Escudero, 2005). Both the PAM and L2LP models predict that this scenario poses fewer problems than the NEW scenario, since no new contrast has to be created in L2 perception (L2LP) and little discrimination difficulty is predicted (PAM). Given that the PAM and PAML2 use perceptual assimilation data to make predictions for discrimination accuracy, while they do not predict assimilation patterns (Escudero et al., 2014; Tyler et al., 2014; Colantoni et al., 2015), these models would predict little discrimination difficulty for Dutch learners of Spanish from Escudero and Boersma (2002)'s categorization pattern. This is because as reported in Escudero and Boersma (2002), Dutch listeners perceived Spanish /i/ mostly as Dutch /i/ (in average 71% of 25 tokens) and Spanish /e/ mostly as Dutch /ɪ/ (in average 65% of 25 tokens), which according to PAM would lead to a two-category assimilation or a category-goodness scenario <sup>1</sup>, resulting in very good to good discrimination. However, L2LP's architecture allows pinpointing a potential difficulty for learners in this scenario that goes beyond discrimination difficulty: Escudero and Boersma (2002) note that if a learner's L1 contrasts are left intact when acquiring an L2 without this contrast, this may in turn lead to spurious

---

<sup>1</sup>Escudero and Boersma (2002) did not collect goodness of fit ratings together with L1 categorization, which is crucial for establishing the perceptual assimilation types proposed within PAM, which are the start point for the model's discrimination difficulty predictions. However, following Bundgaard-Nielsen et al. (2011), one can conclude that Spanish /i/-/e/ are categorized or assimilated to the Dutch contrast /i/-/ɪ/, given that an L2 vowel is defined as categorized if it was identified as an L1 vowel in more than 50% of presentations.



contrasts at the word level (i.e., lexical contrasts), ultimately hampering the attainment of a fully native-like command of the L2.

If the purpose of speech communication is to understand and to be understood, it seems important to not only concentrate on how perceptual development takes place in an L2 but also to examine how the novel L2 categories are employed to recognize and store new words in the L2 lexicon. Experimental evidence suggests continuity between L2 perceptual and lexical abilities, as difficulties in distinguishing novel L2 sounds are commonly accompanied by difficulties in distinguishing L2 minimal pairs. However, other research has shown dissociation between perceptual and lexical abilities in L2 development. For instance, some studies document that L2 learners fail to encode a novel L2 contrast lexically, despite them being fully able to perceive the L2 contrast (e.g., Curtin et al., 1998), while other studies show that the opposite can also be true: L2 learners may develop distinct lexical representations for words that they cannot reliably discriminate in perception (Weber and Cutler, 2004; Cutler et al., 2006; Escudero et al., 2008) or production (Hayes-Harb and Masuda, 2008). These studies suggest that distinguishing pre-lexical perception from lexical recognition in an L2 model will provide further insight into the processes underlying L2 acquisition. By incorporating separate but linked representations for perceptual and lexical contrast, L2LP can serve as a model to investigate both continuity and discrepancy between perceptual and lexical abilities in L2 acquisition.

The computational architecture of L2LP allows simulating the entire trajectory from naïve to experienced L2 listener in various scenarios. These trajectories can then be compared to empirical data to assess the adequacy of the model. Escudero and Boersma (2004) and Escudero (2009) performed simulations with computer-modeled learners in the L2LP framework, showing that these exhibited developmental paths that are comparable to the performance of Spanish learners of the Southern British (SBE) and Scottish English (SE) /i/-/ɪ/ contrast. These learners face a NEW and SIMILAR scenario respectively, as exemplars of the vowels in SBE are acoustically closest to Spanish /i/, while exemplars of SE are acoustically closest to /i/ and /e/. However, the modeled learners in these studies had direct access to the phonemic or phonological categories of the L2 in the input data. Escudero (2005) argues that ultimately L2 learning should be modeled as meaning-driven or message-driven<sup>2</sup>: learners have no direct access to the phonological categories employed by native speakers of the L2, but rather infer these based on how well they are able to understand the meaning intended by a speaker. This is, in fact, a more ecologically valid proposal. Escudero's theoretical account of this more realistic mechanism for language learning used the SUBSET scenario for Dutch learners of the Spanish /i/-/e/ contrast as a case study. Dutch has three front vowels /i/, /ɪ/, and /ɛ/ in the area of the vowel space where Spanish has only /i/

<sup>2</sup>Escudero (2005)'s original proposal considered learning *message-driven*, but as the learning data for the model described in sections 2 and 3 do not strictly contain messages, we use the term *meaning-driven* here.

and /e/, which according to the L2LP should lead to the multi-category assimilation of Spanish /i/ as Dutch /i/ and /ɪ/, and of Spanish /e/ as Dutch /i/ and /ɛ/, which was confirmed in naïve, beginning, intermediate, and advanced Dutch learners of Spanish (Escudero and Boersma, 2002). The theoretical account predicted that meaning-driven learning would result in a reduction of the middle /ɪ/ category. However, a computational implementation of the model by Weiand (2007) failed to confirm this hypothesis, as the modeled learners mostly did not manage to converge on a more L2-like grammar. A thorough inspection of Weiand's results has led us to believe that by revising some details of learning and representation in the model, meaning-driven category reduction could be borne out.

In the present study, we further investigate the adequacy of the L2LP model, in its theoretical proposal (Escudero, 2005) and earlier computational implementations (Escudero and Boersma, 2004; Weiand, 2007; Boersma and Escudero, 2008), for explaining a case of perception and lexicalization of an L2 contrast. Although PAM-L2 incorporates the role of the lexicon in L2 sound perception, it is limited to hypothesizing that vocabulary size determines L2 sound perception success. Current psycholinguistic models of spoken-word recognition (e.g., McClelland and Elman, 1986; Norris, 1994; Gaskell and Marslen-Wilson, 1997) assume that the process of identifying a word in the lexicon is the result of a process of competition between lexical candidates that are activated at the same time, with each candidate being supported to different degrees by the speech signal. L2LP uses this activation process in a network-like model. Another important feature of L2LP that is compatible with a number of L1 acquisition models (e.g., PRIMIR, Werker and Curtin, 2005) is the assumption of continuity between perceptual and lexical development: perceptual learning is triggered as learners attempt to improve recognition by updating their lexical representations. This trickle-down view of meaning-driven lexical learning and lexicon-driven perceptual learning will be detailed below. In short, L2LP bridges insights from the field of L2 sound acquisition with more general cognitive theories of linguistic processing. These concepts are embedded in a simulation framework that is capable of generating quite specific predictions for various acquisition scenarios.

The present study has two aims. First, we present a revised version of L2LP, changing two crucial details of how learning takes place but retaining the fundamental properties of the model listed above. We assess the explanatory adequacy of this revised L2LP by re-applying it to an instance of lexical and perceptual learning in the Dutch-to-Spanish SUBSET scenario described above. Our hypothesis is that the revisions will improve the L2LP's ability to model the learning process in a multiple-category assimilation case followed by a SUBSET scenario, as observed in real L2 learners by Escudero and Boersma (2002).

Second, we propose two alternative versions of the revised model with regards to information flow from speech signal to lexicon, given that pre-lexical and lexical perception can be implemented as sequential (strictly bottom-up)

or interactive (allowing lexical feedback to lower-level perception). The existence of lexical feedback is a matter of much debate within models of psycholinguistics, as shown by Norris et al. (2000)'s proposal and the many alternatives that emerged in response (McClelland et al., 2006; McQueen et al., 2006). In Escudero (2005)'s account, L2 comprehension is described as sequential, but this is not a necessary property of the model. We will thus contribute to this more general debate by investigating the explanatory adequacy of these alternative views on processing grammars in L2 speech comprehension. Below, we present our revised version of the L2LP model and its specific application to the SUBSET scenario, and demonstrate that it successfully explains L2 learning of perception and lexicalization.

## 3.2 The L2LP model revised

Escudero (2005)'s L2LP model aims at providing a comprehensive platform to explain L2 acquisition, perception, and lexicalization. It grew out of, and co-evolved with, the Bidirectional Phonetics and Phonology framework (Boersma, 1998, 2011; henceforth BiPhon), which itself is an extension of Optimality Theory (OT; Prince and Smolensky, 1993). In this section, we describe how linguistic knowledge, processing, and learning are implemented in a revised version of L2LP, taking care to highlight changes from Escudero (2005)'s description and Weiland (2007)'s implementation.

### 3.2.1 Architecture of the L2LP-revised: levels and connections

Like its predecessors, L2LP is an explicit computational model of the processes driving L2 perception and learning. Modeling the acquisition of pre-lexical phonetic categorization in the L2, as well as the subsequent recognition of L2 categories in stored lexical items, requires units on four levels of representation. Figure 3.1 shows an overview of these four levels and the connections between them.

At the bottom we find the acoustic level, representing incoming speech sounds as they arrive in the peripheral auditory system. The subsequent phonetic level encodes a speaker's language-specific, invariant representations of speech sounds, including context-specific allophonic detail. These intermediate representations are linked to the phonemic level where possible canonical forms of words/morphemes are stored, encoding only contrasts that may change the meaning of a word. Finally, phonemic forms connect to possible meanings at the lexical level<sup>3</sup>. By including an intermediate phonetic level between the acoustic signal and phonemic forms stored in the lexicon, L2LP

<sup>3</sup>Traditionally in BiPhon and L2LP these four levels are known as Auditory, Surface, Underlying, and Lexical Form, respectively. Here we replace these phonological concepts with terms more familiar to psychologists and psycholinguists.

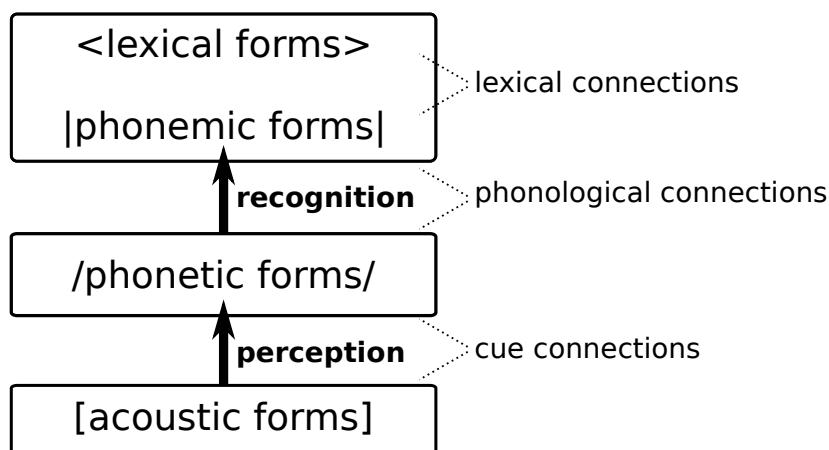


Figure 3.1: The levels of representation and connection types in the L2LP model.

aims to explicitly represent the distinction between the pre-lexical and lexical stages of speech perception, as was described in the Introduction.

Units on adjacent levels are connected, and the process of perceiving and eventually recognizing an incoming word is represented in the model as a four-step path through this network: [acoustic] → /phonetic/ → |phonemic| → <lexical>. The winning or optimal path is decided by relative strength of connections among competing paths. While the units themselves are fixed, the strengths of the connections are altered over the course of learning. This in turn alters the optimal paths from acoustics to lexicon through the network. Knowledge of a language is thus stored in the connection strengths: for instance, a strong |phonemic| → <lexical> connection encodes knowledge of a given lexical item as a meaning-form pair.

A central assumption of L2LP is the *Full Copying* hypothesis (Escudero, 2005): L2 learners initiate their learning process on a duplicate or copy of their L1 perception grammar, so that their L2 grammar is attuned to the sounds and categories of the L1. Over time, exposure to the new language shifts the connections of this copy to a state more suited to perception and recognition of the L2. The next sections elaborate this learning process, showing how perception, recognition, and learning are modeled in the Dutch to Spanish SUBSET scenario that is the focus of this chapter.

### 3.2.2 Evaluating optimal paths

An incoming word is represented as a unit on the [acoustic] level. As this study concerns the L2 acquisition of Spanish *front vowels*, inputs are repre-

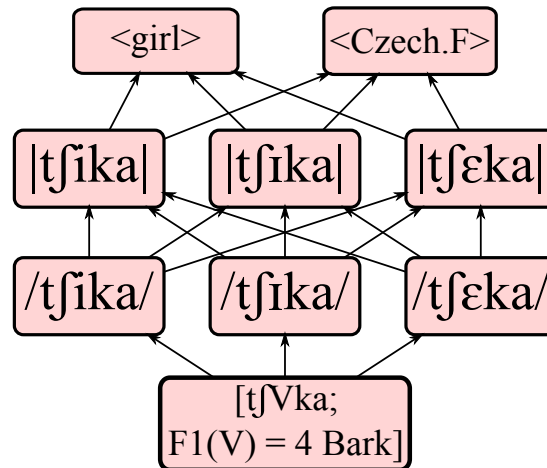


Figure 3.2: Possible mappings for the input  $[tʃVka; F1(V) = 4 \text{ Bark}]$ , via phonetic and phonemic representations, to a lexical form. Each bottom-to-top path through the graph represents a possible pathway of perception and recognition.

sented by two variables, namely a “carrier” word containing a front vowel, and the first formant (F1) of the said vowel, which is the acoustic cue for vowel height. The carrier words (see Appendix A) are always members of a Spanish /i/ - /e/ minimal pair, and are represented as acoustically invariant: they can be seen as narrowing down the available units in the network to those specific to a given minimal pair. The F1 input values do show acoustic variation: they are represented as discretized values on the psychoacoustic Bark scale, ranging from 2 to 8 Bark in steps of 0.1 Bark. For example, the acoustic input  $[tʃVka, F1(V) = 4.0 \text{ Bark}]$  corresponds to a realization of either the Spanish word *chica* ‘girl’ or of *checa* ‘Czech female,’ with an F1 value of 4 Bark for the front vowel (V). Figure 3.2 shows the possible mappings from this particular input form, via phonetic and phonemic representations, to one of two possible lexical meanings. All other combinations of carrier words and front vowel realizations are similarly connected to two possible meanings via the two intermediate levels of representation.

Under the assumption that the L2 grammar is initially a copy of the L1 grammar, the learner may connect the [V] contained in the acoustic input to one of the three different front vowels of Dutch on the /phonetic/ level, embedded in a phonetic representation of the carrier word. Our example input  $[tʃVka, F1(V) = 4.0 \text{ Bark}]$  thus connects to the phonetic representations /tʃika/, /tʃika/ and /tʃɛka/. These connect in turn to three phonemic representations |tʃika|, |tʃika| and |tʃɛka|, which lead to either of two ⟨lexical⟩ items, namely ⟨girl⟩ or ⟨Czech.F⟩. This yields a total of 18 paths ( $3 \times 2 \times 2$ ) from acoustics to

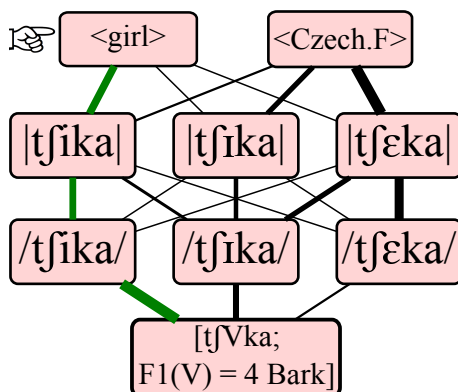


Figure 3.3: Recognizing a lexical form by finding an optimal path. Of the 18 possible routes to from sound to meaning, the optimal path is that whose weakest connection is stronger than the weakest connection of any other path. In this figure, line thickness visualizes connection strength. The input containing a front vowel with an F1 of 4 Bark is perceived as phonetic /tʃika/, phonemic |tʃika|, and ultimately recognized as lexical <girl>.

lexicon for each representable acoustic input. The relative strengths of connections along the paths decide the optimal route. However, the ranking values encoding these connection strengths are distorted slightly at each evaluation step by adding a random value from a normal distribution. This stochastic evaluation (Boersma, 1998) allows the model to deal with probability and variation when mapping from input to output. Stochastic evaluation is also robust to occasional errors in the input data during the learning procedure (detailed in Section Meaning-driven learning below), making it more likely to converge on a target language (Boersma and Hayes, 2001).

Following a central tenet of Optimality Theory, the optimal path from [acoustic] to <lexical> is not defined by the sum of its connection strengths. Rather, a path is as strong as its weakest link, which means that the optimal path is the one containing the least weak connections. Equivalently, one can envision evaluation as iterating through the connections from weakest to strongest, pruning each connection until a single route remains. Figure 3.3 illustrates this evaluation procedure and is further explained below.

The [acoustic] → /phonetic/ connection strengths are initially inherited from the L1 Dutch grammar and thus suitable for the Dutch system with three front vowels. The /phonetic/ → |phonemic| connections are also not arbitrary, as the grammar is biased toward what phonologists refer to as faithful mappings, i.e., the connections between a phonetic representation and its identical phonological counterpart (e.g., /i/ → |i|, /ɪ/ → |ɪ| and /ɛ/ → |ɛ|). The bias is enforced by initializing these connections as stronger than the other six /pho-

netic/  $\rightarrow$  |phonemic| connections. Nevertheless, this is an important conceptual shift from the original architecture proposed by Escudero (2005) and its implementation by Weiland (2007). While they also biased the grammars toward faithful mappings, this bias was qualitative, so that these connections could never be weaker than a non-faithful mapping, and their strength was impervious to learning. In the revised L2LP, this initial bias is quantitative and may diminish or vanish over the course of learning. Thus, our revision retains symbolic representations but has a connectionist perspective on the relation between the /phonetic/ and |phonemic| levels: the two types of representation are of a distinct nature, there is no identity mapping, and the affinities between units on the two levels are gradual.

Finally, the |phonemic|  $\rightarrow$  ⟨lexical⟩ connections are all initialized at equal strength in the L2 grammar, since no knowledge about the lexical meaning of Spanish word forms could be inherited from the L1 grammar. While |phonemic|  $\rightarrow$  ⟨lexical⟩ mappings are specific to the subnetworks selected by the carrier words, the [acoustic]  $\rightarrow$  /phonetic/ and /phonetic/  $\rightarrow$  |phonemic| connection strengths pertain only to the representations of front vowels and are shared between representations regardless of carrier word. An update triggered by our example acoustic input [tʃVka, F1(V) = 4.0 Bark] will therefore also affect the outcome of all other inputs with an F1 of 4 Bark; at the same time the updating of |phonemic|  $\rightarrow$  ⟨lexical⟩ connection strengths affects the outcome for the carrier word [tʃVka] across all F1 input values. This update to both levels of connections triggered by an acoustic input validates the need for both a phonemic and a lexical level within the model.

### 3.2.3 Sequential vs. interactive processing

As discussed in the Introduction, a standing debate in cognitive models of speech processing is whether the outcome of (pre-lexical) perception forms the input to recognition, or whether the two processes are performed in parallel and may interact with one another. Escudero (2005)'s theoretical treatment of L2LP and its implementation by Weiland (2007) is sequential: their learners always evaluate the [acoustic]  $\rightarrow$  /phonetic/ connections of perception before the /phonetic/  $\rightarrow$  |phonemic| and |phonemic|  $\rightarrow$  ⟨lexical⟩ connections of recognition. However, this two-step processing is not a necessary feature of the model, as Boersma (2011) shows that BiPhon (and by extension L2LP) can handle interaction between different levels of representation. By removing the strict ordering of connections in evaluation, recognition may interact with perception.

In our implementation, assigning connections stratum indices besides their ranking value enforces strict sequential ordering. At evaluation time, connections are ordered first by stratum, then by (distorted) ranking value. This means that if we place the [acoustic]  $\rightarrow$  /phonetic/ connections in a higher stratum, perception precedes recognition and we simulate a learner with sequential perception and recognition. Conversely, by placing all connections in

the same stratum, the connections of recognition may influence the outcome of perception. This allows us to compare a purely bottom-up version to an interactive version of the model, all else being equal.

### 3.2.4 Meaning-driven learning

Learning in the L2LP framework equates with updating the connection strengths in the network, and is error-driven: simulated learners attempt to improve perception and recognition of the L2 whenever the current state of the grammar leads to misunderstandings. This is referred to as meaning-driven learning, as described above. After an acoustic input is evaluated and matched to a lexical form (Section *Evaluating Optimal Paths*), the learner is presented with a target ⟨lexical⟩ form encoding the intention of the speaker. If this target form matches the lexical form as understood by the learner, recognition is correct and no action is undertaken. In case of a mismatch, the learner will attempt to decrease the likelihood of a future mismatch by updating their grammar through weakening all connections along the path that led to the incorrect lexical form, and strengthening all connections along the path to the intended target form. If the two paths share subpaths, the net change in the strength of that connection will be zero. The plasticity value that is subtracted and added in order to weaken and strengthen connections, respectively, gradually decreases during learning.

Importantly, the target ⟨lexical⟩ item presented to learners contains no information on the /phonetic/ or |phonemic| categories employed by the speaker. The connection strengths on these intermediate levels must be updated such that future instances of this acoustic input will follow a path to the intended target item. Although the use of minimal pairs restricts possible outputs to two ⟨lexical⟩ items, the learner is confronted with several possibilities for performing this update, and is initially biased toward retaining its three-vowel L1 Dutch system where possible.

Since nine distinct paths lead from any input to each individual lexical form, the learner must first parse a single path to the correct form to decide which connections to strengthen. Finding this parse occurs through interpretive parsing (Tesar and Smolensky, 1998). That is, the learner uses its current grammar to find an alternative path, but this time considers only the subset of nine paths leading to the target form, instead of the full network, as shown in Figure 3.4. Following Jarosz (2013a), and departing from the implementation of Weiand (2007), evaluation noise is re-applied to the connections prior to parsing. Jarosz found that this resampling technique greatly increases the chances of finding a grammar that is compatible with the input data in Optimality-Theoretic, error-driven learning models.

To summarize, the present L2LP-revised model implements Escudero (2005)'s proposal with the following three revisions: (1) the phonologically inspired bias for “faithful” mappings is less restrictive (Section *Evaluating Optimal Paths*), (2) the possibility of interaction between perception and recog-



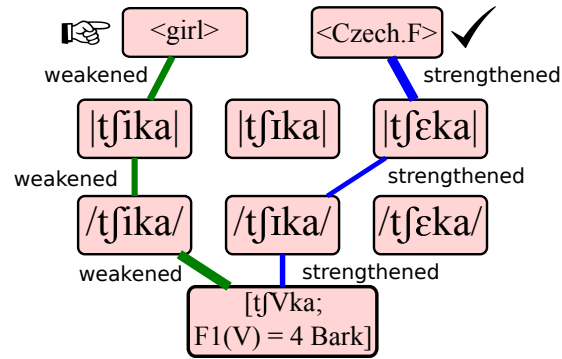


Figure 3.4: Error-driven learning. The learner discovers that it should have recognized lexical  $\langle \text{Czech.F} \rangle$  rather than  $\langle \text{girl} \rangle$ . It performs another evaluation, this time within the subset of paths leading to  $\langle \text{Czech.F} \rangle$ . Learning strengthens connections along that path, and weakens connections along the incorrect path initially found.

nition can be explored (Section *Sequential Vs. Interactive Processing*), and (3) Jarosz (2013a)’s resampling is applied in parsing to enhance the likelihood of convergence. The next section describes the methodology for training and testing our model of the SUBSET scenario using computational simulations.

### 3.3 Computational modeling with the L2LP-revised model

We performed a number of learning simulations to investigate whether the revised model described in Section *The L2LP Model Revised* can successfully implement the meaning-driven SUBSET learning scenario described by Escudero (2005). The simulation program consisted of two phases: L1 training, in order to create the “naïve” L1 starting point from which L2 acquisition proceeds, and L2 training to simulate the acquisition of Spanish categories through error-driven learning on lexical items. This two-stage simulation procedure was applied both for sequential-type learners whose [acoustic]  $\rightarrow$  /phonetic/ connections are always evaluated before all other connections, and for interactive-type learners whose connections pertaining to recognition are allowed to outrank connections pertaining to perception. At various points during both training procedures, learners were given data from a test set in order to investigate to what extent L2 training improves recognition of the Spanish lexical items, as well as how phonemic/phonetic categories were remapped to this end.

Parameter settings were identical to those used in Boersma and Escudero (2008) and Weiand (2007) wherever possible. Ranking values (strength)

for all connections were initialized to an equal value of 100, with the exception of /phonetic/ → |phonemic| connections, which were set to 95 for “faithful” connections that preserved identity across these levels (Section Evaluating Optimal Paths), and to 105 for the other connections. The evaluation noise parameter was set to 2.0, which represented the standard deviation of a random normal distribution (centered around zero) that distorted ranking values before each evaluation. Plasticity was initialized to 0.1 at the start of learning, with a decay rate such that plasticity shrank by a factor 0.7 every 10,000 steps.

### 3.3.1 Acoustic input data for the simulated learners

In both training phases, simulated learners were repeatedly given [acoustic] inputs, each of which represented some word or utterance containing a front vowel. The auditory correlate of the height of these front vowels is its first formant (F1), which the grammar represents on the psychoacoustic Bark scale, from 2.0 to 8.0 Bark in bins of 0.1 Bark. In order to increase the ecological validity of our simulations, we obtained these F1 values from two recent, methodologically similar vowel production studies, as described below.

The F1 values for the L1 Dutch input data were generated by taking all female tokens of the vowels /i/, /ɪ/, and /ε/ from the corpus of van van Leussen et al. (2011), converting the F1 of these tokens to Bark and rounding it to the nearest “bin”. The L2 formant values were likewise generated by taking all female tokens of /i/ and /e/ from Chládková et al. (2011), but these were also paired with a randomly selected carrier word containing either /i/ or /e/ in Spanish. Carrier words were the minimal pairs listed in the Appendix, which were the same as those used in Weiland (2007).

Figure 3.5 shows the distribution of the F1 per category in the L1 and L2 input data.

### 3.3.2 Training and testing procedures

In the L1 Dutch training phase, simulated learners were exposed to [acoustic]–/phonetic/ pairs of binned input F1 values and target vowels, in order to train them directly on the three-way Dutch contrast. In this way, we cast L1 learning as perceptual, as in Boersma and Escudero (2008). This special status for L1 learning is warranted by results in the infant learning literature, which strongly suggest that infants learn language-specific perceptual warping before a lexicon is in place (Werker and Tees, 1984; Polka and Werker, 1994; Maye et al., 2002). An example input-output pair would be [F1 = 3.4 Bark] - /i/. To test whether training resulted in correct Dutch-like perception of /i/, /ɪ/, and /ε/, we used a holdout method where the production tokens described in Section Acoustic Input Data for the Simulated Learners were first split into a training (90%) and testing (10%) subset. A total of 40,000 [acoustic] input tokens was then randomly sampled from these training sets for each learner,

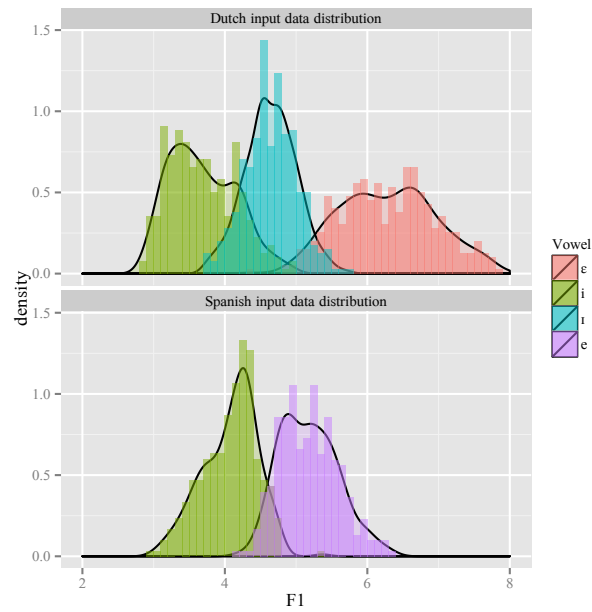


Figure 3.5: Distribution of input data over the F1 continuum for Dutch (above) and Spanish training phases. The histograms represent the “binned input data.

with the grammar updating the ranking in case of an error as described in Section Sequential vs. Interactive Processing. Following Jarosz (2013a), the ranking was resampled (i.e., evaluation noise was applied a second time) after an error, so that the connection strengths used for parsing may differ slightly from those used for the initial evaluation.

To simulate immersion in the L2 environment, the simulated learners were next trained on labeled pairs of binned input F1 values plus invariant carrier words (Section Acoustic Input Data for the Simulated Learners), and output ⟨lexical⟩ forms representing a meaning congruent with the chosen carrier word and vowel token. An example input–output pair would be [tʃVka], F1(V) = 3.8 Bark] – ⟨girl⟩. Learners were given no information about the intermediate /phonetic/ and |phonemic| categories; remapping these representations takes place only on the basis of the target ⟨lexical⟩ form through the parsing strategy described in Section Meaning-driven Learning, and learners began with a system optimally suited to perceiving the L1 training data.

In all other respects, L2 training resembles L1 training: again the input data were split into a training (90%) and testing (10%) subset, and a total of 40,000 training tokens (generated from the training data) was given to learners, who again employed resampling to determine which ranking values to update in case of an error. The (informal) pseudocode below summarizes the learning algorithm performed on the L1 and L2 training datasets.

- 
- 1: **for** each pair ( $input_T \sim output_T$ ) **do**
  - 2:   add evaluation noise to ranking values of all connections
  - 3:   evaluate optimal path ( $input_T \dots output_O$ ) from  $input_T$
  - 4:   **if**  $output_T \neq output_O$  **then**
  - 5:     add evaluation noise to all connection strengths in grammar
  - 6:     evaluate target parse between  $input_T$  and  $output_O$
  - 7:     decrease ranking value for each connection in optimal path by *plasticity*
  - 8:     increase ranking value for each connection in target parse by *plasticity*
  - 9:   decrease plasticity by decay rate
- 

### 3.4 Modeling results

Results were obtained by evaluating tokens from the test sets at various stages of L1 and L2 training. No learning took place on these test tokens. Since there are some elements of randomness in the model and training (specifically in the division of the input data into training and test sets, and the noise employed in evaluation), we ran 50 simulations for both the sequential and interactive versions of the grammar, representing 50 simulated sequential-type and 50

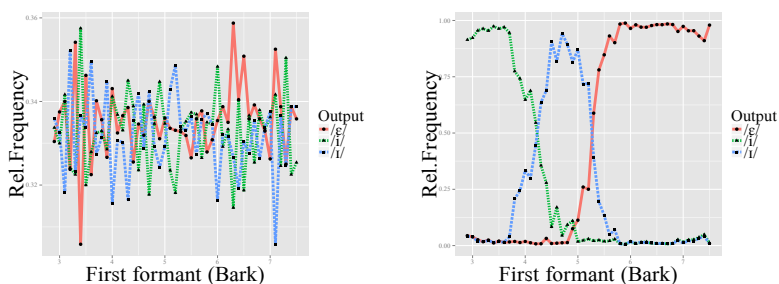


Figure 3.6: Classification of inputs after 0 (left) and 40,000 (right) learning iterations on the L1 input data

simulated interactive-type learners. The results reported here are averaged over these 50 simulated learners per grammar type.

### 3.4.1 L1 learning

As stated above, L2LP assumes the initial state of an L2 grammar to be a copy of the L1 grammar. We simulated this initial state by first training each grammar on the discretized acoustic values coupled to the phonetic categories mentioned above. Since the L1 training concerns only the mapping from [acoustic] inputs to the /phonetic/ level, without involving the lexicon, there is no difference in behavior between the sequential and interactive learners. Figure 3.6 shows how these [acoustic]-/phonetic/ mappings develop over the course of training. At the end of training, the categorization curves matched those of the input distribution of Figure 3.6. Since the distributions of the three vowels on the F1 continuum show some overlap, learners reached a ceiling of about 80% correct recognition of the test set (Figure 3.7, left). This means that without lexical or semantic context, it is not always possible to distinguish these vowels from one another.

### 3.4.2 L2 learning

Both sequential and interactive learners were able to improve their classification of the Spanish minimal ⟨lexical⟩ pairs, arriving at a stable recognition rate of around 85% over time. As in L1 learning, this is probably the peak possible success rate given the fact that the distributions of Spanish /i/ and /e/ overlap, as shown in the original vowel production study (see Chládková et al., 2011 and Figure 3.5). Although sequential learners needed a slightly larger number of input data to attain this peak rate, both types ultimately reach this ceiling (with overlapping confidence intervals) after about 8000 iterations, as shown in Figure 3.7 (right). This slower attainment may be a consequence of the more L1-like representations maintained by sequential learners, as will be

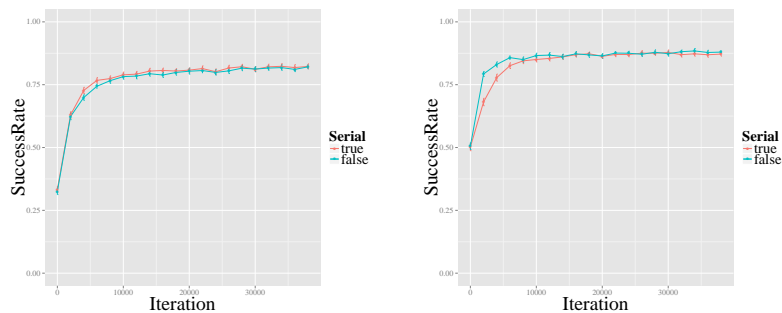


Figure 3.7: Success rates over time for Dutch L1 (left) and Spanish L2 (right) training. Bars indicate 95% confidence intervals.

discussed below.

The success of this new implementation of the model in learning to recognize the L2 confirms our hypothesis that the original L2LP’s predictions as implemented by Weiland (2007) failed because of the phonologically inspired “faithfulness” connections. The current revision, which implements phonetic-phonemic mappings through a more general concept of connection strength, is more successful in modeling empirical L2 learning results. The revised L2LP furthermore shows that the meaning-driven learning of lexical items proposed by Escudero (2005) can account for improved understanding of the L2 through exposure to the language.

Furthermore, the L2LP model makes specific predictions on learners’ phonological categorization of speech sounds over the course of development. All learners shifted the boundaries between /phonetic/ categories during learning: they adapted to the two-vowel L2 system at the cost of the middle /I/ category, as shown by the /phonetic/ categorization of learners over time (Figure 8). This result of the simulations closely resembles the empirical findings of Escudero and Boersma (2002), as well as the modeling results of Boersma and Escudero (2008), which assumed learners access category labels. The revised model however shows that acquiring L2-like representations can also be modeled as meaning-driven, without assuming that a learner has explicit knowledge of the L2 phonological categories, an assumption that was at the core of Boersma and Escudero (2008)’s model.

Without phonetic or phonemic labels in the L2 input data, learners are faced with several options on how to adapt their old perceptual systems to the L2. Interestingly, Figure 3.8 also shows that the sequential and interactive versions of the model do not predict the same extent of perceptual remapping in the L2. For interactive learners, the former Dutch /I/ category eventually falls into complete disuse, so that these L2 Spanish learners are effectively native-like in their perception of front vowels, employing only two categories. The sequential model predicts that perception of /I/ diminishes, but is retained

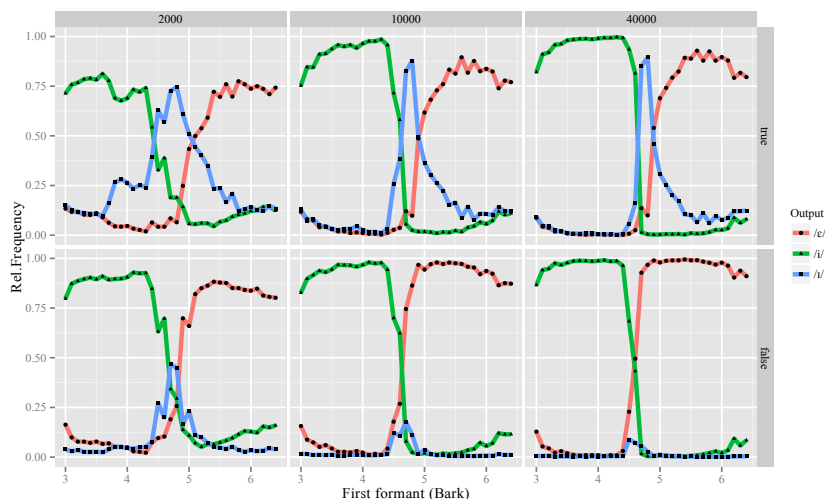


Figure 3.8: Sequential (above) and interactive learners’ categorization of inputs after 2,000 (left), 10,000 (middle) and 40,000 (right) learning iterations.

for certain inputs. This difference in /ɪ/ responses after learning, corresponding to the area under the /ɪ/ curve after 40,000 training tokens, is significant between the two groups of learners<sup>4</sup>.

This difference between the two groups is restricted to a small range of inputs: the phonemic categorizations of the two groups are significantly different for [acoustic] inputs whose F1 lies between 4.5 and 5.1 Bark.<sup>5</sup> This range corresponds to Dutch /ɪ/ and includes the boundary between Spanish /i/ and /e/. The sequential model thus predicts that L2 perception remains filtered by the L1 for these intermediate vowels, with more open vowels usually classified as /e/ but occasionally as /ɪ/. Despite this maintenance of a three-vowel system in their internal L2 representations, sequential learners attain the same recognition rate of Spanish lexical items (Figure 7). These learners appear to consider /ɪ/ an “allophone” of /e/ in Spanish, and store both phones as possible realizations for words containing phonemic |e|. We discuss the implications of these predictions below.

<sup>4</sup>Welch’s t-test, two-sided,  $t = 30.5903$ ,  $df = 69.32$ ,  $p = 5.5 \times 10^{-35}$

<sup>5</sup>Per input F1 value, relative frequency of response for each vowel was summed per learner type (sequential vs. interactive); Pearson’s chi-squared (response type) was significant ( $N1 = N2 = 50$ ,  $df = 2$ ,  $\chi^2 = 7.06$ , Bonferroni-corrected  $p = 0.00047$ ).

### 3.5 Discussion

Experience in one's native language largely shapes the perceptual and lexical acquisition of a second language. We provide a computational, network-like model of L2 perception and lexicalization. The revised L2LP retains (psycho-)linguistic concepts on representations and evaluation of input data, but removes a number of assumptions from theoretical phonology about the way units on these levels of representation are connected. Discarding these assumptions has increased the explanatory power of the model, suggesting that a strictly symbolic view of the phonetics–phonology interface is not consistent with what we know about L2 learning. Another novel aspect was that we trained our simulated learners on data taken directly from vowel production studies, rather than artificial distributions.

Our first aim was to explore the viability of a meaning-driven learning paradigm, in which learners have access to the intended meanings but not to the phonological specifications of the L2 input. Simulated learners showed progress toward native-like perception and recognition of front vowels, progressively adapting to the L2 in a way similar to real-life L2 learners Escudero and Boersma (2002). This mirrors the results of an earlier modeling study (Boersma and Escudero, 2008) but obviates the assumption that overt phonological structure is present in the learning input.

Secondly, the revised model allows us to differentiate between a sequential and an interactive perspective on phonetic (pre-lexical) perception and lexical recognition. While both versions of the model gravitate toward correct recognition of the L2, they make different predictions on the phonetic representations ultimately employed by learners. Specifically, sequential learners are predicted to retain an L1 phonetic category for certain “boundary” stimuli whereas interactive learners ultimately fully adapt their vowel system to the L2. Anecdotal evidence suggests that adult L2 learners only very rarely reach native-like ability, which at first glance seems more in line with the results of our sequential learners (but see Bongaerts, 1999). However, experimental evidence is needed in order to untangle the influence of L1 on the perception of L2 learners. Previous research (e.g., Escudero and Boersma, 2002; Mayr and Escudero, 2010; Escudero et al., 2012) has studied L2 categorization behavior by activating listeners' L2 language mode (Grosjean, 2000). We conjecture that categorical perception effects (discrimination peaks) in the region of the old L1 phonetic categories (e.g., the subsumed Dutch /ɪ/) when perceiving the L2 may provide clues for the accuracy of either the sequential or the interactive model. These effects may be measured with discrimination and identification experiments, presenting the relevant tokens to advanced Dutch learners of Spanish in their Spanish language mode.<sup>6</sup> Experiments can include more sen-

---

<sup>6</sup>A reviewer suggested the possibility that either strategy occurs in real-life L2 learners, and is perhaps a locus of individual differences in L2 acquisition. The potential co-existence of the two types of grammars in the same listener or differences across listeners can also be explored with



sitive measures such as reaction times or event-related potentials to examine whether retaining the extra L1 vowel category negatively affects L2 perception. Indeed, previous studies have shown that the availability of extra phonetic categories affects native and non-native vowel perception (Benders et al., 2012; Elvin et al., 2014). Our results thus offer testable hypotheses that may in turn contribute to the general debate of sequential vs. interactive language processing (Norris et al., 2000; McClelland et al., 2006).

We conclude that L2LP offers a workable and fruitful model of the processes underlying acquisition of non-native sound systems. Compared to alternative models of L2 acquisition, the simulation paradigm illustrated in this study allows L2LP to make very specific predictions on how L1 experience and L2 input shape the outcome of learning. These numerical predictions can be compared to empirical findings and in turn inform new hypotheses. Future work is to investigate whether L2LP's success extends beyond the SUBSET scenario described above for instance, the reverse scenario (which would be an instance of the L2LP NEW scenario) of going from a two-way to a three-way contrast, and would therefore require the creation of a new L2 category rather than the discontinued use of an old L1 category.

## Acknowledgements

We are grateful to Jaydene Elvin and Daniel Williams for comments on this paper. Thanks also go out to Klara Weiand, Paul Boersma and the audience at OCP9 in Berlin for earlier comments on this work. The work of the first author was funded by NWO (The Netherlands Organization for Scientific Research) grant 277-70-008 awarded to Paul Boersma. Collaboration between authors was facilitated by ARC (Australian Research Council) grants DP130102181 awarded to the second author and by the ARC Centre of Excellence for the Dynamics of language (project number CE140100041) where the second author is Chief Investigator.

## CHAPTER 4

---

### Efficient evaluation and learning in multi-level parallel constraint grammars

---

*Abstract*

In multi-level parallel Optimality-Theoretic grammars, the number of candidates (possible paths from the input to the output level) increases exponentially with the number of levels of representation. The problem with this is that with the customary strategy of listing all candidates in a tableau, the computation time for the evaluation (i.e. choosing the winning candidate) and learning (i.e. reranking the constraints on the basis of language data) increases exponentially with the number of levels as well. This paper proposes instead to collect the candidates in a graph in which the number of nodes and the number of connections increase only linearly with the number of levels of representation. As a result, there exist procedures for evaluation and learning that increase only linearly with the number of levels. These efficient procedures help to make multi-level parallel constraint grammars more feasible as models of human language processing. We illustrate visualization, evaluation and learning with a toy grammar for a traditional case that has already previously been analyzed in terms of parallel evaluation, namely French liaison.

---

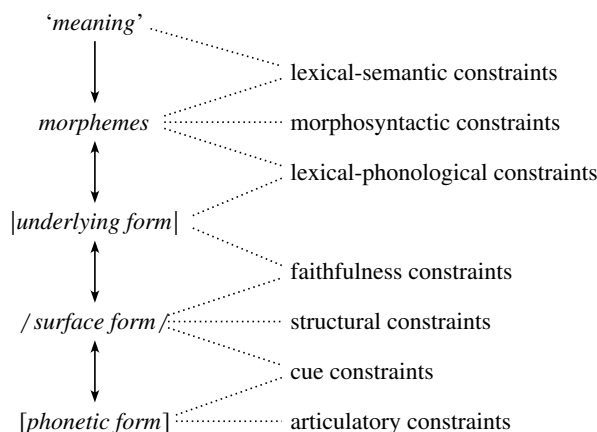


Figure 4.1: The five-level grammar model needed for this paper.

## 4.1 Multi-level parallel constraint grammars

Difficult problems in phonology and its interaction with phonetics, such as the French liaison problem discussed in this paper, often benefit from being expressed in terms of multiple levels of representation. The five levels considered in the present paper are the ones shown in Fig. 4.1: meaning, morphemes, (phonological) underlying form, (phonological) surface form, and phonetic form. In an Optimality-Theoretic (OT) implementation (Boersma, 2007; Apoussidou, 2007), relations between adjacent levels are evaluated by inter-level constraints (here, lexical-semantic, lexical-phonological, faithfulness and cue constraints), and the representations themselves are evaluated by intra-level constraints (here, morphosyntactic, structural, and articulatory constraints). The downward arrows in the figure represent the direction of the production process; the input to this process is an intended meaning and its output is a realized phonetic form.

For the present paper the parallel property of the model is crucial: production is implemented as the evaluation of candidates across all levels of processing in parallel. The speaker starts with an intended meaning and computes an optimal quadruplet of morphemes, underlying, surface, and phonetic form. Cross-level parallelism means that e.g. faithfulness constraints, which evaluate the relation between underlying and surface form, can interact with cue constraints, which evaluate the relation between (phonological) surface form and (auditory-) phonetic form. This cross-level parallelism allows “later” phonetic considerations such as articulatory effort and auditory cue quality to influence “earlier” phonological decisions (Boersma, 2007, 2008). In the present paper, parallelism crucially allows later phonological considerations such as hiatus avoidance to influence earlier choices in the morphology such as gen-

der selection. These types of bottom-up influences in production are depicted with upward arrows in Fig. 4.1.

In learning, cross-level parallelism typically causes the credit (or blame) for overt phenomena to be distributed over multiple levels. Thus, when an underlying  $|an+pa|$  is pronounced as phonetic  $[ampa]$ , the learner will automatically come to interpret this partly as a phonological, partly as a phonetic assimilation (Boersma, 2008).

A potential problem with parallel evaluation is its computational load. If every meaning comes with 10 possible morphemes, 10 possible underlying forms, 10 possible surface forms, and 10 possible phonetic forms, and we assume that all of these can be combined freely, then the number of candidates (quadruplets of morpheme, underlying, surface, and phonetic form) that have to be evaluated for any given meaning is 10,000. In other words, the number of candidates increases exponentially with the number of levels of representation. If we list all of these candidates in a big tableau, and use the usual candidate elimination procedure for tableaux (Prince and Smolensky, 1993), the computational load of choosing the optimal candidate increases exponentially with the number of levels of representation. This problem could become especially prohibitive in a practical comprehensive computational model of linguistic processing, in which there may be more phonological levels than in Figure 4.1 (e.g. an additional word level; Mohanan 1981; Kiparsky 1982; Bermúdez-Otero 2003) and more phonetic levels (e.g. separate articulatory and auditory levels; Boersma 1998), and in which there will certainly be more semantic levels (e.g. separate lexical and sentence levels; Jackendoff (1997) as well as multiple syntactic levels (e.g. deep and surface structure; Chomsky 1957) and some discourse levels (Hengeveld and Mackenzie, 2008); a total of 12 or more levels can easily be imagined.

Fortunately, an exponentially increasing candidate set does not have to come with an exponentially increasing computation time. The computational OT literature has repeatedly shown that a candidate set whose size is exponential in a certain parameter can be evaluated in a time that is linear in that parameter, if certain conditions are met. For instance, in two-level OT, the number of output candidates is typically exponential in the length of the input: if e.g. each input segment either surfaces unchanged, or is deleted, or has something epenthesized before and/or after it, each input segment has four potential surface realizations, and an input string of  $N$  segments has  $4^N$  surface candidates; Ellison (1994), however, showed that if such a candidate set can be represented as a regular expression and appropriate restrictions on the constraints are met ("finite-state OT") the evaluation time is only linear in the length of the input (Riggle 2004 improves on this, by showing that in finite-state OT even an infinite candidate set, e.g. one with unbounded epenthesis, can be evaluated in linear time). Likewise, if we allow subsegmental phonological structure, the number of candidates is exponential in the number of autosegmental tiers, but Eisner (1997) showed that if in finite-state OT the candidate set can be represented on a non-hierarchical set of tiers with a time-

line (“Primitive OT”) the evaluation is linear in the number of tiers.

The same consideration applies to the multi-level OT of Figure 4.1, where the number of candidates is exponential in the number of levels of representation. In this paper we show that if each constraint evaluates either a single level or the relation between two levels, evaluation time becomes linear in the number of levels. As with Ellison (1994), Eisner (1997) and Riggle (2004), this linearity is achieved with graph-theoretic methods, but without requiring the assumptions of finite-state OT. Specifically, we represent the candidate set not in a big tableau but in a candidate graph, which would reduce the 10,000-candidate case above to a graph with only 40 nodes and 310 connections, i.e. with a number of elements that increases only linearly with the number of levels of representation.

The concept of the candidate graph is introduced in section 2. In section 3 we then show how this economical representation naturally leads to efficient procedures for evaluation and learning, which either extend the usual candidate elimination procedure for tableaux (Prince and Smolensky, 1993) or reflect standard procedures for optimization in graphs (Ford 1956; Bellman 1957; Viterbi 1967; for OT: Riggle 2009). As a real-language case to illustrate the procedures of Section 3, Section 2 introduces the case of the interaction of liaison and gender selection in French. In Section 4 we investigate how the learning procedure works out for the French case. In Section 5 we discuss how our findings relate to complexity-reducing proposals for other parts of OT.

## 4.2 Visualization of candidates with multi-level OT tableaux

The usual way to visualize an evaluation process in Optimality Theory is with a tableau, a list of all possible candidate outcomes with their constraint violations (Prince and Smolensky, 1993). For parallel models such as the one in Fig. 4.1, this list can become very long. The present section illustrates this point with the example used throughout this paper, which is the phenomenon of phonological liaison in French and its interaction with morphological gender selection. Liaison has served as a testing ground for many phonological theories (for reviews see Eychenne 2006 and Durand and Lyche 2008); its parallel interaction with gender selection has been noted by Encrevé-Lambert (1971) and Encrevé (1988) and was first addressed within OT by Tranel (1996). We first present a traditional serial analysis, then proceed to the parallel analysis.

### 4.2.1 The serial analysis of gender selection and liaison

The serial analysis of gender selection and liaison in French proceeds as follows, in a stepwise OT evaluation where in each step we heavily restrict the candidate generator (GEN) to providing only the most relevant candidates.

We only model the behaviour of the adjective good in French, which is pronounced [bɔn] when feminine and [bɔ̃] when masculine, except that before vowel-initial nouns it is always [bɔn] (in first approximation). For our serial account we follow the early generative approach by Schane (1968), Dell (1970, 1973) and Selkirk (1972), which posits that in the underlying form there is a single stem |bɔn| for both masculine and feminine gender, and that a gender morpheme is appended to this, which is phonologically empty (|∅|) for the masculine and a schwa (|ə|) for the feminine.

Suppose that one wants to produce the meaning ‘good<sub>i</sub> actor<sub>i</sub>’ in French. In a serial version of Figure 4.1, a French speaker starts by connecting this meaning to the French morphemes ⟨bon-M; acteur<sub>M</sub>⟩, where the subscript M marks the Masculine value of the gender feature of the noun acteur, and the appended “-M” indicates the masculine ending of French adjectives. In OT, this could look like the following tableau:

(1) Mapping meaning to morphemes

“good <sub>i</sub> actor <sub>i</sub> ”	*⟨FM⟩
☞ ⟨bon-M; acteur <sub>M</sub> ⟩	
⟨bon-F; acteur <sub>M</sub> ⟩	*!

where the constraint \*⟨FM⟩ militates against a gender mismatch within the morphemic level. In the serial account, the working of this constraint is quite trivial: since it interacts with nothing else, it will always force a gender identity between adjective and noun, as it does here.

Next, the serial speaker connects the morpheme sequence to an underlying form by using a ranking of lexical-phonological constraints (Boersma and Hayes 2001; Apoussidou 2007) that makes sure that the underlying forms proposed by Schane, Dell and Selkirk are selected. For the present case, the result is |bɔn+∅#aktœʁ|, where “+” denotes a word-internal morpheme boundary and “#” a word boundary:

(2) Mapping morphemes to an underlying form

⟨bon-M; acteur <sub>M</sub> ⟩	*⟨bon-M⟩   bɔ̃	*⟨bon-F⟩   bɔ̃	*⟨F⟩   ∅	*⟨M⟩   ə	*⟨bon-M⟩   bɔn	*⟨bon-F⟩   bɔn	*⟨F⟩   ə	*⟨M⟩   ∅
☞  bɔn+∅#aktœʁ					*			*
bɔn+ə#aktœʁ				*!	*			
bɔ̃+∅#aktœʁ	*!							*
bɔ̃+ə#aktœʁ	*!			*				

Here, the lexical-phonological constraint \*⟨*bon*-M⟩ |b̃| militates against connecting the morpheme ⟨*bon*-M⟩ to the underlying form |b̃|, the lexical-phonological constraint \*(F)|∅| militates against connecting morphemic femininity to an underlying null form, and so on. In the serial account, the workings of the lexical constraints are quite simple: since they interact with nothing else, the winning underlying form of a morpheme is always the one that violates the lowest-ranked lexical constraint.

Now that the underlying form is known, the speaker connects it to the surface structure /b̃.nak.tœʁ./, where “.” denotes a syllable boundary and where the final underlying |n| of |b̃n| has been moved to the onset of the same syllable that contains the first two segments of *acteur* (a case of *liaison*):

(3) Mapping the underlying form to a surface form

b̃n+ə#aktœʁ	*    /ə/	* ə  / /	*/V.V/	*/n./	* ̃  /ɔn/	* ̃n  /̃/
☞ /b̃.nak.tœʁ./						
/b̃.āk.tœʁ./				*!		*
/b̃.nə.āk.tœʁ./	*!			*		
/b̃.ə.āk.tœʁ./	*!			**		*

Here we see four faithfulness constraints, in a generalized notation suitable for multi-level approaches: \*| | /ə/ militates against schwa insertion, \*|ə| / / against schwa deletion, \*|̃| /ɔn/ against n-insertion, and \*|̃n| /̃/ against n-deletion. There are also two structural constraints: \*/n./ against /n/ in coda, and \*/V.V/ against hiatus. Some of these constraints become relevant only in later tableaux.<sup>1</sup>

Finally, the speaker pronounces the surface form as the overt phonetic form [b̃naktœʁ], with the help of cue constraints (faithfulness-like constraints for phonetic implementation; Boersma, 2007) and constraints against articulatory effort.

(4) Mapping the surface form to a phonetic form

/b̃.nak.tœʁ./	*/̃/ [ɔn]	*/ɔn/ [̃]	*[ə]	*/ə/ [ ]	*/ / [ə]
☞ [b̃naktœʁ]					
[b̃āktœʁ]		*!			
[b̃nəktœʁ]			*!		*
[b̃āktœʁ]		*!	*		*

Here, the cue constraint \*/̃/ [ɔn] militates against pronouncing a phonological nasal vowel as a phonetic nasal consonant, and \*/ / [ə] militates against

<sup>1</sup>We ignore in this paper the potential multiplicity of syllabification candidates. Specifically, we assume that there are high-ranking constraints that dictate that *bonne maison* good house is better represented as /b̃n.m̃.z̃./ than as /b̃.n.m̃.z̃./, and, conversely, that *bon oiseau* good bird is /b̃.n.wa.zo/ rather than /b̃n.wa.zo/

pronouncing a phonetic schwa without a phonological correspondent (the articulatory constraint \*[ə] is needed below).

To summarize tableaux (1) through (4), one can now express the full route from meaning to sound as the winning candidate path ‘good<sub>i</sub> actor<sub>i</sub>’ – ⟨bon-M; acteur<sub>M</sub>⟩ – |bɔn+ɔ#aktœʁ| – /.bɔ.nak.tœʁ./ – [bɔnaktœʁ].

While the word *acteur* was masculine and vowel-initial, we now proceed to the word *mari* ‘husband’, which is also masculine, but consonant-initial. The meaning *good<sub>i</sub> husband<sub>i</sub>* shows up as [bɔ̃maʁi], without any [n], the idea being that /n/ can phonologically show up only before vowel-initial forms such as ak.tœʁ./, and not before consonant-initial forms such as /ma.ʁi./. In a serial account, the first two mappings have no knowledge of this phonological conditioning, so that their workings are identical to what they were in the ‘good<sub>i</sub> actor<sub>i</sub>’ case:

(5) *Serial account: Mapping meaning to morphemes is insensitive to phonology*

“good <sub>i</sub> husband <sub>i</sub> ”	*⟨FM⟩
☞ ⟨bon-M; mari <sub>M</sub> ⟩	
⟨bon-F; mari <sub>M</sub> ⟩	*

(6) *Serial account: Mapping morphemes to an underlying form is insensitive to phonology*

⟨bon-M; mari <sub>M</sub> ⟩	*⟨bon-M⟩   bɔ̃	*⟨bon-F⟩   bɔ̃	*⟨F⟩	*⟨M⟩	*⟨bon-M⟩    bɔn	*⟨bon-F⟩    bɔn	*⟨F⟩    ə	*⟨M⟩
☞  bɔn+ɔ#maʁi					*			*
bɔn+ə#maʁi				*!	*			
bɔ̃+ɔ#maʁi	*!							*
bɔ̃+ə#maʁi	*!			*				

So the underlying form is |bɔn+ɔ#maʁi|, with the same initial two morphemes as ‘good<sub>i</sub> actor<sub>i</sub>’, with the inclusion of an underlying bɔn. Now that phonological material is available, the coalescence of the vowel and the nasal (ɔn → ɔ̃) can enter the derivation. All authors mentioned above (Schane, Dell, Selkirk) agree that this is an early phonological rule. In OT, the phonological production phase can indeed enforce this change. The constraint against coda nasals (\*/n./) forces underlying |bɔn| to surface as /bɔ̃/, thus violating a faithfulness constraint:

(7) *Serial account: Preconsonantal vowel-nasal coalescence in masculine forms must take place no earlier than in the phonology*



bɔn+ə#mavɪ	*  /ə/	* ə  / /	*/n./	*/V.V/	* ɔ  /ɔn/	* ɔn  /ɔ/
/bɔn.ma.vi./			*!			
<sup>ESP</sup> /bɔ̃.ma.vi./						*
/bɔ̃.nə.ma.vi./	*!					
/bɔ̃.ə.ma.vi./	*!			*		*

Finally, the phonetic implementation phase offers no surprises, because there is a candidate that violates none of the constraints:

(8) *Serial account: Mapping the surface form to a phonetic form*

/bɔ̃.ma.vi./	*/ɔ̃/ [ɔn]	*/ɔn/ [ɔ̃]	*[ə]	*/ə/ [ ]	*/ / [ə]
[bɔ̃mavɪ]	*!				
<sup>ESP</sup> [bɔ̃mavɪ]					
[bɔ̃nəvavɪ]	*!		*		*
[bɔ̃əvavɪ]			*!		*

The whole path from meaning to sound can be summarized as ‘good<sub>i</sub> husband<sub>i</sub>’ – ⟨bon-M; mari<sub>M</sub>⟩ – |bɔn+ə#mavɪ| – /bɔ̃.ma.vi./ – [bɔ̃mavɪ].

We are now left with the feminine case, where [n] shows up despite a subsequent consonant, as in [bɔ̃nvvatyɐ] ‘good<sub>i</sub> car<sub>i</sub>’. As in the masculine case, the \*⟨FM⟩ constraint enforces gender agreement at the morpheme level (at least in the serial account):

(9) *Serial account: Mapping meaning to feminine morphemes*

“good <sub>i</sub> car <sub>i</sub> ”	*⟨FM⟩
⟨bon-M; voiture <sub>F</sub> ⟩	*!
<sup>ESP</sup> ⟨bon-F; voiture <sub>F</sub> ⟩	

The early generative accounts mentioned above posit a schwa in the underlying form:

(10) *Serial account: Underlying feminine forms have schwa*

⟨bon-F; voiture <sub>F</sub> ⟩	*⟨bon-M⟩   bɔ̃	*⟨bon-F⟩   bɔ̃	*⟨F⟩   ə	*⟨M⟩   ə	*⟨bon-M⟩   bɔ̃	*⟨bon-F⟩   bɔ̃	*⟨F⟩   ə	*⟨M⟩   ə
bɔn+ə#vvatyɐ			*!			*		
<sup>ESP</sup>  bɔn+ə#vvatyɐ						*	*	
bɔ̃+ə#vvatyɐ	*!	*						
bɔ̃+ə#vvatyɐ		*!					*	

The existence of schwa prevents the deletion of /n/, because /n/ is not in coda:

(11) Serial account: Schwa shows up in feminine forms

bɔn+ə#vwa.tyɤ	*  /ə/	* ə  / /	*/n./	*/V.V/	* ɔn  /ɔ̃/	* ɔn  /ɔ̃/
/ .bɔn.vwa.tyɤ./		*!	*			
/ .bɔ̃.vwa.tyɤ./		*!				*
☞ / .bɔ.nə.vwa.tyɤ./						
/ .bɔ̃.ə.vwa.tyɤ./				*!		*

In these serial generative accounts, schwa is dropped later in the derivation, i.e. after coda-n deletion. In an OT account with only two phonological levels, as here, the drop of schwa has to be relegated to the phonetic implementation:

(12) Serial account: Schwa drop in phonetic implementation

/ .bɔ.nə.vwa.tyɤ./	* ɔ̃/ [ɔn]	*/ɔn/ [ɔ̃]	*[ə]	*/ə/ [ ]	*/ / [ə]
☞ [bɔnvwa.tyɤ]					*
[bɔ̃vwa.tyɤ]		*!			*
[bɔnəvwa.tyɤ]			*!		
[bɔ̃əvwa.tyɤ]		*!	*		

Note the crucial ranking of \*[ə] over \*/ə/ [ ], i.e. it is worse to have a schwa in the phonetics than to drop a phonological schwa from the phonetics.

A source of complexity in this serial analysis is its crucial reliance on the existence of two phonological and/or phonetic mappings. In the original rule-ordering approach, the rule of schwa deletion had to be ordered after the rule of vowel-nasal coalescence. In OT, such a situation of counterfeeding interaction cannot be modelled with a single mapping, if the two processes (here, schwa deletion and vowel-nasal coalescence) are governed by separate faithfulness constraints (here, \*|ə| / / and \*|ɔn| /ɔ̃/; Smolensky 1995; Orgun 1995; Kirchner 1995; Gnanadesikan 1997; Moreton and Smolensky 2002. In our French example there is no ranking of the constraints in (11) that yields schwa deletion in the phonology proper, i.e. it is impossible to derive a phonological / .bɔn.vwa.tyɤ./ with the given constraints and levels. With a high-ranked \*/ə/ in (11), / .bɔ̃.vwa.tyɤ./ would win, because \*/n./ outranks \*|ɔn| /ɔ̃/, a ranking that is crucial to make (7) work. In other words, no ranking of \*/ə/, \*/n./, and \*|ɔn| /ɔ̃/ will produce both vowel-nasal coalescence in / .bɔ̃.ma.ʁi./ and the surfacing of the /n/ in / .bɔn.vwa.tyɤ./, and schwa drop can only take place in the “later” phonetic implementation phase. If you insist on having schwa deletion in the phonology instead, perhaps because other phonological rules interact with it (e.g. Dell 1973: 188), you will need an intermediate phonological level of representation, such as the word level proposed by theories of lexical phonology (Kiparsky, 1982; Bermúdez-Otero, 2003); schwa drop would then take place in the postlexical phonology).<sup>2</sup> The conclusion is that given

<sup>2</sup>Or one could introduce another source of complexity and add a high-ranked conjoined faithfulness constraint along the lines of Smolensky (1995), i.e. \*|ɔn| /ɔ̃/ & \*|ə| / /.

our straightforward constraint set the opacity of the interaction between vowel-nasal coalescence and schwa drop cannot be handled with only two levels of phonological and/or phonetic representation, and that it can be handled with three levels, as it is here.

#### 4.2.2 The parallel analysis of gender selection and liaison

A parallel account may look entirely differently at the matter than the serial account. According to Encrevé (1988), Encrevé-Lambert (1971) proposed that French speakers opt for gender disagreement if this benefits the phonology: the feminine phrase *l'idée* ('the idea') has (a reduced form of) the masculine article *le* instead of the feminine article *la*, because the schwa of *le* ( $|\text{lə}|$ ), but not the full vowel of *la* ( $|\text{la}|$ ), can be deleted before vowels ( $/.li.de./$ ); *un vieil acteur* (an old actor) has the feminine adjective *vieille* rather than the masculine adjective *vieux*, because *vieille* ( $|\text{vjɛj}+\text{ə}|$ ) can provide an onset to *acteur* ( $|\text{.vjɛ.jak.tœʁ.}|$ ) whereas *vieux* ( $|\text{vjø}+\text{ø}|$ ) cannot; *mon idée* (my idea) has the masculine possessive pronoun *mon* rather than the feminine *ma*, because *mon* ( $|\text{mɔ̃}(n)|$ ) can provide an onset to *idée* ( $/.mɔ̃.ni.de./$ ) whereas *ma* ( $|\text{ma}|$ ) cannot.

Translating this idea to OT, Tranel (1996) proposed the gender agreement constraint that we write here as  $*\langle\text{FM}\rangle$ , and had it interact with phonological constraints equivalent to  $*/n./$  and  $*/V.V/$ . While Tranel made no attempt to formalize this situation with more than two levels of representation, it can straightforwardly be formulated in a principled manner within the model of Fig. 4.1, where the morphosyntactic constraint of gender disagreement and the structural constraints of syllable onsets and codas play their roles at different levels of representation. Crucially, a constraint at a "later" level of representation (surface form) dominates a constraint at an "earlier" level of representation (morphemes), a situation that can only occur if the levels are handled in parallel (or **interactively**) rather than serially.

In the present example, we have the option of generalizing the *ma* ~ *mon* alternation to *bon* ~ *bonne*, regarding the latter pair as suppletive and ignoring any phonological relationship. Under that view, French speakers wanting to produce the meaning good actor have the option of choosing the morpheme  $\langle\text{bon-F}\rangle$  instead of  $\langle\text{bon-M}\rangle$ . The resulting morpheme sequence  $\langle\text{bon-F}; \text{acteur}_M\rangle$  does violate the morphosyntactic constraint against gender disagreement between adjective and noun, but does have the advantage that  $\langle\text{bon-M}\rangle$  can take the underlying form  $|\text{bɔ̃}|$ , violating no faithfulness constraints in  $/.bɔ̃.ma.bi./$ , and  $\langle\text{bon-F}\rangle$  can take the underlying form  $|\text{bɔn}|$ , violating no faithfulness constraints in  $/.bɔn.vwa.tyʁ./$  or  $/.bɔ.nak.tœʁ./$ . The following tableau shows how the constraint against hiatus in the phonological surface form can force the selection of the feminine morpheme  $\langle\text{bon-F}\rangle$  for the masculine noun  $\langle\text{acteur}_M\rangle$ :

(13) Parallel account: Phonology influences morphology

"good <sub>i</sub> actor"	*⟨bon-M⟩ bɔn	* ɔ̃  /ɔn/	* /V.V /	*⟨FM⟩	* /ɔn/ [ɔ̃]	*⟨bon-F⟩ bɔ̃ ,
⟨bon-M; acteur <sub>M</sub> ⟩  bɔn+ɔ̃#aktœʁ  /.bɔ.nak.tœʁ./ [bɔnaktœʁ]	*!					
⟨bon-M; acteur <sub>M</sub> ⟩  bɔn+ɔ̃#aktœʁ  /.bɔ̃.ak.tœʁ./ [bɔ̃aktœʁ]	*!		*		*	
⟨bon-M; acteur <sub>M</sub> ⟩  bɔ̃+ɔ̃#aktœʁ  /.bɔ.nak.tœʁ./ [bɔnaktœʁ]		*!				
⟨bon-M; acteur <sub>M</sub> ⟩  bɔ̃+ɔ̃#aktœʁ  /.bɔ̃.ak.tœʁ./ [bɔ̃aktœʁ]			*!			
☞ ⟨bon-F; acteur <sub>M</sub> ⟩  bɔn+ɔ̃#aktœʁ  /.bɔ.nak.tœʁ./ [bɔnaktœʁ]				*!		*

The working of the morphosyntactic constraint \*⟨FM⟩ is no longer trivial, as it was in §4.2.1, and the workings of the lexical-phonological constraints are no longer simple, as they were in §4.2.1. Instead, the morphosyntactic constraint and the lexical-phonological constraints now interact in interesting ways with constraints at a “later” level, namely faithfulness constraints and a structural constraint.

When we compare the serial account of §4.2.1 with the parallel account of §4.2.2, we see several differences. In the parallel account, surface “ghost” (i.e. unpronounced) schwas as in (11) are no longer needed, even for feminine forms. Instead, there is suppletive allomorphy at the underlying level: |bɔ̃| is the masculine form, |bɔn| the feminine form. In ‘good<sub>i</sub> actor<sub>i</sub>’, an additional allomorphy on the basis of gender takes place: the feminine form is selected because a gender change (violating \*⟨FM⟩) is less bad than hiatus (vi-

olating \*/V.V/). To sum up, the parallel analysis gets rid of the ghost segment /ə/, and no longer do any nontrivial processes have to take place in phonetic implementation; these two advantages (and the advantage of being able to simultaneously account for the *ma* ~ *mon* suppletion) come at the cost of losing any generalization about a phonological relationship between /b̃/ and /b̃n/.

Another apparent disadvantage of the parallel model is the gigantic size of a full list of candidates with their violations. For the sake of brevity, tableau (13) has been reduced to the bare minimum number of constraints and candidates: it includes only 6 of the 20 constraints of §4.2.1, assumes a trivial relation between surface and phonetic form, and otherwise excludes many potentially relevant candidate paths. A full version of tableau (13) would contain hundreds of candidates and thousands of violation marks. This disadvantage, however, is only apparent: in §4.2.3 we introduce the “candidate graph”, a method that captures all the possible candidate paths in a much more concise and visually informative manner.

### 4.2.3 Candidate graphs

All of the three winning candidate paths for the serial analysis in §4.2.1, as well as the three winning candidate paths for the parallel analysis in §4.2.2, can be visualized as paths through the candidate graph in Figure 4.2. In a candidate graph, the candidates are paths that share connections and forms with each other. The winning candidate paths of the parallel analysis of §4.2.2, for instance, are drawn in Figure 4.2 as thick lines.

## 4.3 Efficient evaluation and learning

This section discusses how Optimality-Theoretic evaluation and learning can be done with candidate graphs.

### 4.3.1 Constraints in the graph

We have seen in Figure 4.2 how candidates in a multi-level OT grammar can be described as paths along a directed (left-to-right) graph, where each *node* represents a form on some level of representation, and each *connection* (the graph-theoretical term is “edge”) represents a possible mapping between two forms on adjacent levels of representation. Each path in Figure 4.2 visits five nodes and follows four connections. The graphs can visualize not only candidates, but also constraint violations: the intra-level constraints of Figure 4.1 and section 2 militate against visiting some nodes, and the inter-level constraints of Figure 4.1 and section 2 militate against following some connections.

For illustrating how evaluation and learning work, we take a simplified version of our French grammar, containing a smaller number of constraints

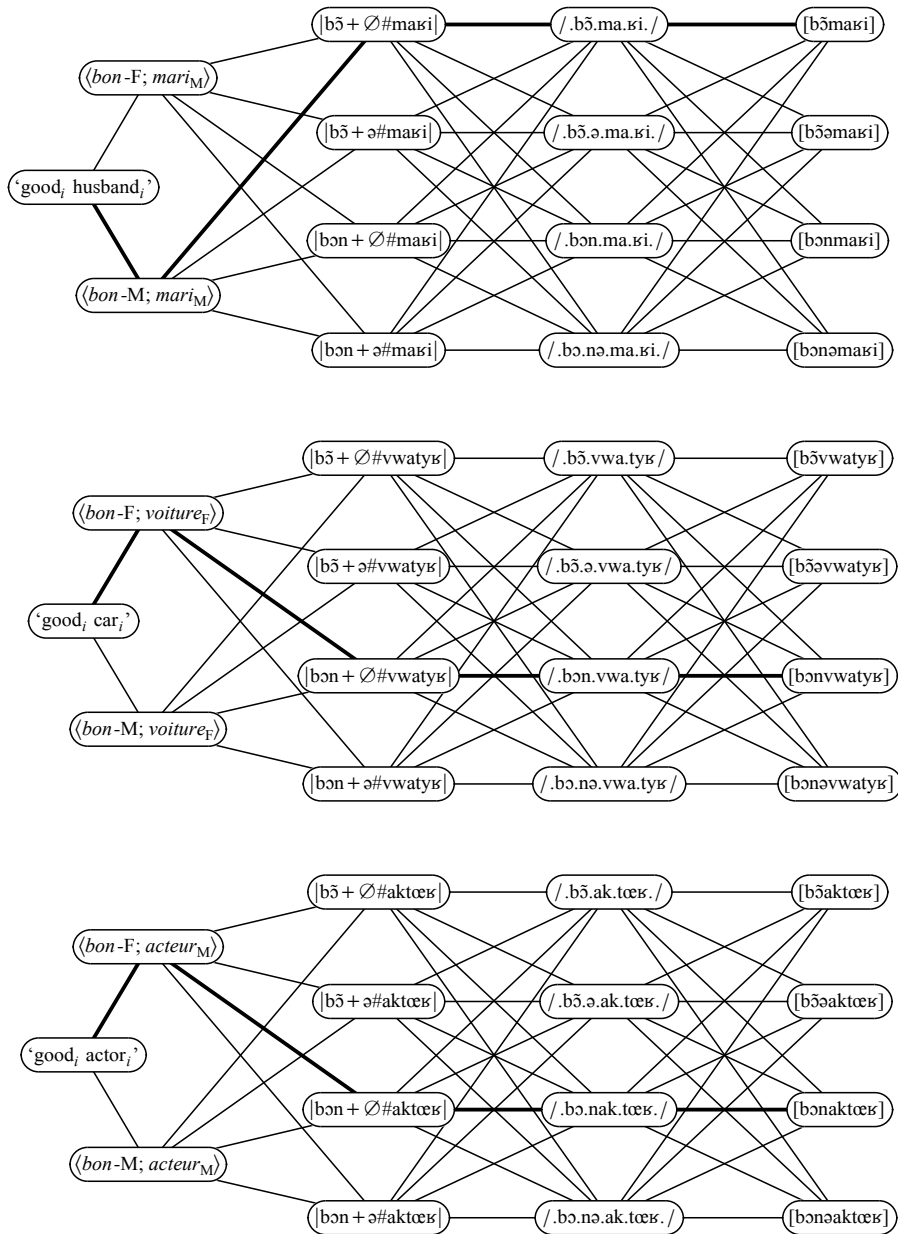


Figure 4.2: Some French graphs of representations. Each path from left to right is a candidate production. The thick paths illustrate the parallel gender allomorphy analysis of §4.2.2

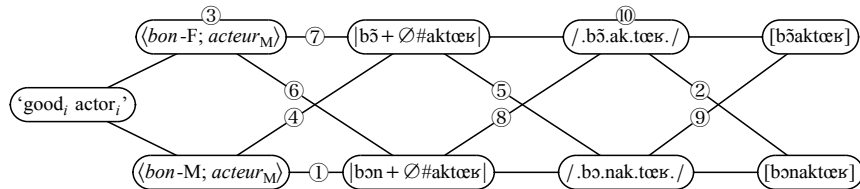


Figure 4.3: The candidate graph for the meaning  $good_i actor_i$ . Also the starting point for the evaluation procedure of §4.3.2.

and forms (e.g. only two phonetic forms instead of four). The constraints are indicated by numerical labels ① through ⑩):

- |   |                                       |                                 |
|---|---------------------------------------|---------------------------------|
| ① | * $\langle bon-M \rangle  b\bar{o}n $ | lexical-phonological constraint |
| ② | * $ \bar{o}  / \bar{o}n/$             | cue constraint                  |
| ③ | * $\langle FM \rangle$                | morphosyntactic constraint      |
| ④ | * $\langle bon-M \rangle  b\bar{o} $  | lexical-phonological constraint |
| ⑤ | * $ \bar{o}  / \bar{o}n/$             | faithfulness constraint         |
| ⑥ | * $\langle bon-F \rangle  b\bar{o}n $ | lexical-phonological constraint |
| ⑦ | * $\langle bon-F \rangle  b\bar{o} $  | lexical-phonological constraint |
| ⑧ | * $ \bar{o}n  / \bar{o}/$             | faithfulness constraint         |
| ⑨ | * $/\bar{o}n/ [\bar{o}]$              | faithfulness constraint         |
| ⑩ | * $/V.V/$                             | structural constraint           |

Figure 4.3 shows the graph for the meaning  $good_i actor_i$ . The two intra-level constraints \* $\langle FM \rangle$  and \* $/V.V/$  are indicated by labels ③ and ⑩ placed on the relevant nodes, while the remaining (inter-level) constraints are indicated by labels placed on the relevant edges.

The graph contains sixteen paths that run from the meaning at the left of the graph toward the phonetic forms at the right. One of these sixteen paths is the optimal candidate under the given constraint ranking. The next section explains how the search for the optimal candidate (the “evaluation”) proceeds.

### 4.3.2 Efficient evaluation: the elimination version

The optimal candidate in a graph is the path whose nodes and edges incur the least serious violations. Analogously to the familiar procedure for evaluating candidates in tableaux (Prince and Smolensky, 1993), there exists a procedure for graphs that eliminates candidates by iterating through the constraint hierarchy, starting with the highest ranked constraint. This slightly informal procedure is presented in §4.3.2, while §4.3.2 presents a more formal account in terms of dynamic programming.

### Evaluation by elimination of edges and nodes

The first evaluation method for candidate graphs that we discuss is analogous to the usual tableau evaluation: at the start, all nodes and edges in the graph are “active”, i.e. they can potentially contribute to the optimal path; next, nodes and edges are iteratively eliminated in an order dictated by the constraint hierarchy, until one or more optimal paths are left.

The procedure starts by considering the highest ranked constraint. The edges or nodes associated with the constraint are tentatively deactivated. Next, the algorithm checks how many paths are left, i.e. via how many routes the right side of the graph can be reached from the left side (this can be done efficiently, with an algorithm that involves each edge only once).<sup>3</sup> If the number of remaining paths is zero, we must conclude that all the candidates are apparently “tied” on the constraint; the tentative deactivation is then undone, and the algorithm proceeds to the next constraint. In the other case, i.e. if the number of remaining paths is greater than zero, the deactivation of the nodes or edges is made permanent (for the duration of this evaluation); under the stipulation that paths cannot follow deactivated edges and cannot visit deactivated nodes, this step eliminates all candidate paths that fatally violate the constraint. If the number of remaining paths is one, we must conclude that this sole remaining path represents the optimal candidate, and the algorithm terminates successfully. Finally, if multiple paths remain, the above steps are repeated with the next-highest ranked constraint. These iterations are repeated down the constraint hierarchy, until either a single path remains or the lowest-ranked constraint has been handled. Should there still be multiple paths through the graph after all constraints have been handled, we must conclude that all these remaining paths are optimal candidates.

Figures 4.4 - 4.9 take us stepwise through the elimination procedure for the case of Figure 4.3. The set of possible candidates at the start of evaluation is defined by Figure 4.3. The constraints of 4.3.1 are ranked from high to low by their label, i.e. with  $*(\text{bon-M}) \text{ |b}\text{ɔ}\text{n}|$  ranked highest and  $*/\text{V.V}/$  lowest. The first step, now, is that the highest ranked constraint, i.e. constraint ①, deactivates the connection between the morpheme sequence  $\langle \text{bon-M}; \text{acteur}_M \rangle$  and the underlying form  $\text{|b}\text{ɔ}\text{n} + \text{ɔ}\#\text{akt}\text{œ}\text{v}|$ , eliminating all 4 paths that travel this connection. Figure 4.4 depicts this elimination by using dotted lines.

---

<sup>3</sup>The number of paths can be computed iteratively from left to right: for a node  $X$  on level  $n$ , the number of paths that lead to it from the single node on level 1 is zero if the node is deactivated, and otherwise it is the sum of the numbers of paths to those nodes on level  $n - 1$  for which the edges to node  $X$  are not deactivated.



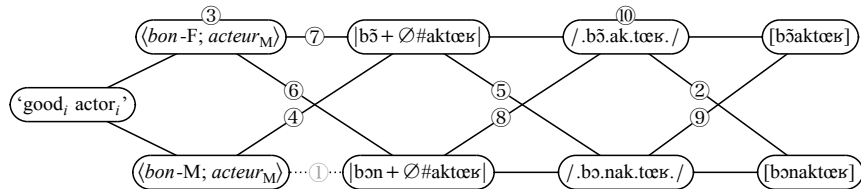


Figure 4.4: The graph after handling constraint ①; 12 out of 16 paths are left.

In the next iteration (Figure 4.5), constraint ② severs the connection between the surface form /bɔ̃.ak.tœɛ./ and the phonetic form [bɔ̃aktœɛ|].

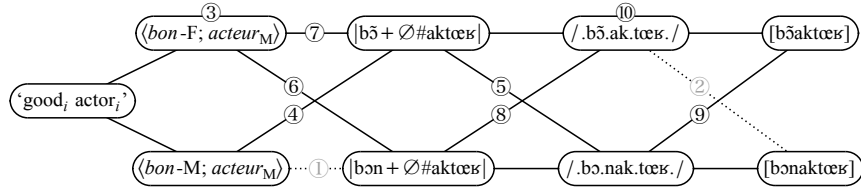


Figure 4.5: The graph after handling constraint ②; 9 paths are left.

Next, the intra-level constraint ③ deactivates the morpheme sequence  $\langle bon-F; acteur_M \rangle$ , eliminating all paths that visit this node (Figure 4.6). After this, only three paths remain out of the original 16.

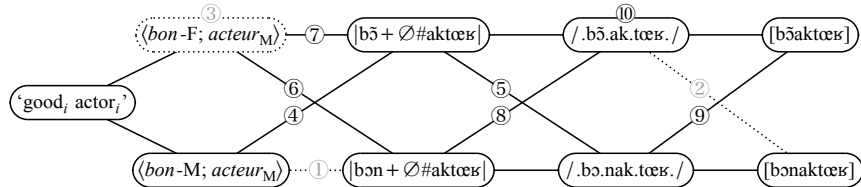


Figure 4.6: The graph after handling constraint ③; 3 paths are left.

In Figure 4.7, constraint ④ tentatively deactivates the connection between  $\langle bon-M; acteur_M \rangle$  and  $|bɔ̃+Ø\#aktœɛ|$ .

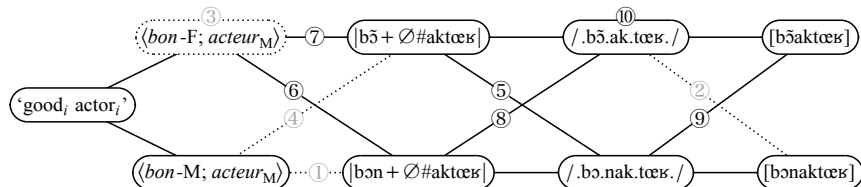


Figure 4.7: The graph after trying to handle constraint ④; no paths are left.

As this step reduces the number of paths to zero (one can no longer get from left to right through the graph in Figure 4.7), this connection is reactivated, as seen in Figure 4.8 (the lines become solid again, but constraint ④ stays grayed out).

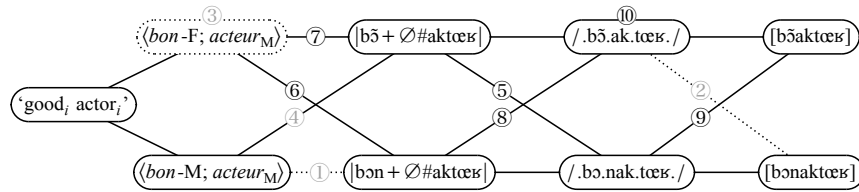


Figure 4.8: The graph after fully handling constraint (4); again, 3 paths are left.

In Figure 9, constraint (5) eliminates the connection between  $|b5 + \emptyset\#aktœɛ|$  and  $/.bɔ.nak.tœɛ./$ , and a single path remains. This optimal path corresponds to the candidate ‘good<sub>i</sub> actor<sub>i</sub>’ –  $\langle bon-M; acteur_M \rangle$  –  $|b5 + \emptyset\#aktœɛ|$  –  $/.b5.ak.tœɛ./$  –  $[b5aktœɛ]$  and is marked by thick lines in the figure.

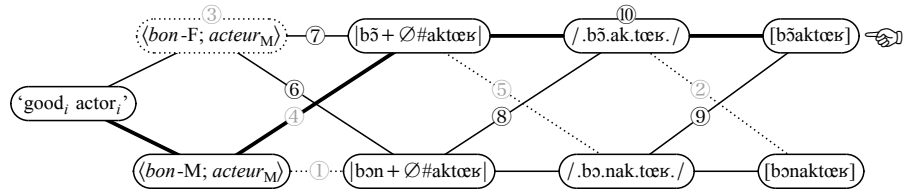


Figure 4.9: The graph after handling constraint (5); a single path (thick lines; pointing finger) is left.

The algorithm is much faster than a full search of all candidates would be, especially for more realistic numbers of constraints or amounts of data than contained in the toy example of this section. The complexity of the algorithm therefore no longer depends exponentially on the number of levels of representation, as it did in Boersma’s and Apoussidou’s simulations.

**Evaluation by dynamic programming**

Computationally, the overall structure of the graph (a *trellis*) invites a practical implementation in terms of a “dynamic programming” algorithm, whose complexity is linear in the number of edges and nodes. Dynamic programming computes the “cheapest” path from the input to (in our case) the right side of the graph. While originally designed for additive cost functions (Ford, 1956; Bellman, 1957; Dijkstra, 1959) or multiplicative probabilities (Viterbi, 1967), dynamic programming applies to OTs lexicographic constraint order as well (Riggle 2004: 159-161; Riggle 2009).

To start off, we associate each edge and each node with a cost, which is a vector of constraint violations, sorted by ranking. For instance, in Figure 4.4 the cost of the node  $\langle bon-F; acteur_M \rangle$  is  $(0, 0, 1, 0, 0, 0, 0, 0, 0)$ , because this node violates only the third-ranked constraint  $\langle FM \rangle$ , and the cost of the edge from  $|bɔn + \emptyset\#aktœɛ|$  to  $/.b5.ak.tœɛ./$  is  $(0, 0, 0, 0, 0, 0, 0, 1, 0, 0)$ , because this edge violates only the eighth-ranked constraint  $\langle \text{ɔn} \mid / \text{ɔ} / \rangle$ . Costs can be added, which goes element by element, e.g., if the cost  $A$  equals  $(1, 0, 2)$  and the cost  $B$  equals  $(1, 1, 0)$ , then  $A + B$  equals  $(2, 1, 2)$ . Finally, costs can be com-

pared (OTs lexicographic ordering: Prince and Smolensky 1991) by checking the first element (counted from the left) in which they differ: in the example of the previous sentence,  $A$  is less than  $B$ , because the first element in which  $A$  and  $B$  differ is their second element, which is smaller in  $A$  (namely 0) than in  $B$  (where it is 1); as a result, we can also talk about the *minimum* of a set of costs, a notion we need in the algorithm. As shown by Riggle (2009), the existence of well-defined addition and “minimum” operations is precisely what makes OT suitable for dynamic programming.

In our formalization, we number the levels, and within the levels we number the nodes. The input level is level 0; its number of nodes is  $N_0 = 1$ . Then there follow  $L$  (in our example: 4) levels of representation, each with  $N_l$  nodes ( $l = 1 \dots L$ ). Each node  $n$  at level  $l$  comes with a cost  $nodeCost[l, n]$ , where  $n$  runs from 1 to  $N_l$ , and each connection from node  $m$  at level  $l - 1$  to node  $n$  at level  $l$  comes with a cost  $edgeCost[l, m, n]$ . At the right edge there is a single invisible node at level  $L + 1$ , which is connected to each node at the phonetic level without violating any constraints ( $N_{L+1} = 1$ ). Here is the pseudocode for the initialization of the relevant variables in the algorithm (the scope of **for** blocks is determined by indentation):

---

**Algorithm 2**


---

```

1: nodeCost[0,1] := the violation vector of the input node, sorted by constraint ranking
2: for  $l = 1 \dots L$  do
3:   for  $n = 1 \dots N_l$  do
4:     for  $m = 1 \dots N_{l-1}$  do
5:        $edgeCost[l, m, n]$  := the violation vector of edge $_{l,m,n}$ ,
6:       sorted by constraint ranking
7:        $nodeCost[l, n]$  := the violation vector of node $_{l,n}$ ,
8:       sorted by constraint ranking
9:   for  $m = 1 \dots N_L$  do
10:     $edgeCost[L + 1, m, 1]$  := (0, 0, ...)           ▷ a vector of C zeroes
11:  $nodeCost[L + 1, 1]$  := (0, 0, ...)

```

---

After this initialization, we employ a dynamic programming algorithm that takes into account the possibility of tied candidates (i.e. equally good paths). In left-to-right graphs like the ones in the present paper, dynamic programming works iteratively from left to right, in the following way. Suppose that for each node  $Y$  on level  $n - 1$  the best paths (and their cost) for going from the input node (the single node on level 0) to  $Y$  is known. The best paths for going from the input node to a node  $X$  on level  $n$  are then the paths through that node  $Y$  for which the cost to  $Y$  plus the cost of going from  $Y$  to  $X$  is less than (or equal to) the cost for all other nodes on level  $n - 1$ . Since the best path is trivial for all nodes on level 1, and the best path can be computed for level

$n$  if it is known for level  $n - 1$ , the best path can be computed for all nodes at levels, including for the single invisible node at the last level. The algorithm uses the minimum operation (“min”) and the addition operation (“+”) for the cost vectors:

---

**Algorithm 3**


---

```

1:  $nodeRoutes [0, 1] := 1$     ▷ the violation vector of the input node, sorted by
   constraint ranking
2:  $nodePathCost [0, 1] := nodeCost [0,1]$ 
3: for  $l = 1 \dots L + 1$  do
4:   for  $n = 1 \dots N_l$  do
5:     for  $m = 1 \dots N_{l-1}$  do
6:        $edgePathCost[m] := nodePathCost[l - 1, m] + edgeCost[l, m, n]$ 
7:        $minimumEdgePathCost := \min_{m=1}^{N_{l-1}} edgePathCost[m]$ 
8:       for  $m = 1 \dots N_{l-1}$  do
9:         if  $edgePathCost[m] = minimumEdgePathCost$  then
10:           $edgeRoutes [l, m, n] := nodeRoutes [l - 1, m]$ 
11:        else
12:           $edgeRoutes [l, m, n] := 0$ 
13:         $nodeRoutes [l, n] := \sigma_{m=1}^{N_{l-1}} edgeRoutes [l, m, n]$ 
14:         $nodePathCost[l, n] := minimumEdgePathCost + nodeCost [l, n]$ 

```

---

After this,  $edgeRoutes[l, m, n]$  contains the number of best paths from the input node to node  $n$  at level  $l$  that go through node  $m$  at level  $l-1$ , and  $nodeRoutes[l, n]$  contains the number of best paths from the input node to node  $n$  at level  $l$ .

Now that we know the number of best routes to any node, including the rightmost (invisible) node, we can randomly choose an optimal path back from that rightmost node to the input node, by iteratively choosing edges from right to left, with probabilities proportional to the number of best routes along the edges at each level:

---

**Algorithm 4**


---

```

1:  $optimalNode [L + 1] := 1$     ▷ the invisible node at the end
2: for  $l = L \dots 1$  do    ▷ backtracking, i.e., counting down
3:    $chosenRoute := randomInteger(1, nodeRoutes [l + 1, optimalNode [l + 1]])$ 
4:    $node := 0$ 
5:    $routes := 0$ 
6:   while  $routes < chosenRoute$  do
7:      $node := node + 1$ 
8:      $routes := routes + edgeRoutes [l + 1, node, optimalNode [l + 1]]$ 
9:    $optimalNode[l] := node$ 

```

---

Here,  $randomInteger(a, b)$  is a function that chooses a whole number between  $a$  and  $b$  (inclusively), all with equal probability. When the algorithm finishes, the optimal path is given by  $optimalNode[1 \dots L]$ ; for instance, in Figure 4.9 the optimal path is given by  $optimalNode[1] = 2$ ,  $optimalNode[2] = 1$ ,  $optimalNode[3] = 1$ , and  $optimalNode[4] = 1$ , at least if at each level of representation the nodes are numbered from top to bottom.

### Range of application

The informal account of §4.3.2 and the more formal account of §4.3.2 are equivalent: basically, the former travels the constraint hierarchy in the outer loop, whereas the latter does that in the inner loop (in the addition and in the comparison function).

For the algorithm to work, the constraints have to honor a strong restriction, namely that each constraint evaluates either nodes on a single level or connections between adjacent levels. For instance, a constraint that is violated only for paths that follow a specific UF node *and* a specific SF node is not allowed, and a constraint that is violated only for a specific connection from e.g. UF through SF to PF is not allowed either. All constraints proposed within the early version of OT that just map UF to SF automatically satisfy this restriction, and all the constraints mentioned here in section 2 do so as well.<sup>4</sup>

There is no limitation on the number of violations per constraint. In the informal account of §4.3.2 we tacitly assumed that every constraint could be violated only once (although the account can easily be extended by considering a doubly violated constraint as a constraint that is ranked slightly higher than the same constraint if singly violated), but the formal account of §4.3.2 makes clear that no such assumption is needed.

The algorithm is compatible not only with the notion of tied *paths* (see §4.3.2), but also with the notion of crucially tied *constraints* (Prince and Smolensky 1993: fn. 31; Anttila 1997; Tesar and Smolensky 1998: 241). Both in the informal account of §4.3.2 and in the formal account of §4.3.2 two constraints that are ranked at the exact same height can be collapsed, before evaluation starts, into a single constraint. However, our simulations exclusively use Stochastic OT Boersma (1998), in which each constraints ranking is numeric and contains a bit of noise, so that constraint ties have probability zero of occurring.

In the French example, given an input, all nodes are exhaustively connected to all nodes at the next level. This is not a real restriction: to allow unconnected nodes between consecutive levels, one can set the relevant *edgeCost* to a vector of positive infinity in the algorithm of §4.3.2, or alternatively, the algorithm can be easily adapted to visit only the connected nodes. The efficiency of the algorithm, as compared to the efficiency of enumerating all possible paths, then depends on the degree to which nodes tend to be connected

<sup>4</sup>The restriction of adjacent levels can be lifted. If the graph has e.g. direct connections between UF and PF, Dijkstra (1959)'s algorithm, which is more general than the Viterbi algorithm employed here, can do the trick; the informal method of §4.3.2 would also still work.

to multiple nodes at the previous level: in the worst case, i.e. if each non-input node is connected to only one node at the previous level, big-list evaluation and graph evaluation are equally efficient; in the best case, i.e. if each node is connected to all nodes at the previous level (as in our French case), the efficiency gain of graph evaluation over big-list evaluation is maximal.

### 4.3.3 Efficient learning from meaning–sound pairs

The reader may have noticed that the winning phonetic form [bɔ̃aktœʋ] in Fig. 4.9 is not correct French. The inclusion of this failure in the present paper is deliberate: it highlights the fact that correct French is not something that speakers start doing automatically when they are born in France; instead, French children have to *learn* to produce French, presumably from the language input they receive from their environment. In this section we describe how a learner can process French data in such a way that she does come to produce correct French.

What a learner of French hears are overt phonetic utterances. If we assume that the learner is capable of inferring the meaning of these utterances, we can say that the learner obtains knowledge of *pairs* of meaning and phonetic form. For our toy French example, the relevant learning data then consists of the following form–meaning pairs:

- [bɔ̃nvwatyʋ] paired with ‘*good<sub>i</sub> car<sub>i</sub>*’
- [bɔ̃naktœʋ] paired with ‘*good<sub>i</sub> actor<sub>i</sub>*’
- [bɔ̃mavɪ] paired with ‘*good<sub>i</sub> husband<sub>i</sub>*’

We have seen in Figures 4.3–4.9 how the learner computes an optimal phonetic form by eliminating connections and forms from the graph until a single candidate path from meaning to phonetic form remains. However, we may assume that a learner does not start out with a constraint hierarchy that will pair every meaning to the correct phonetic form; instead, the OT learning literature has proposed that either all constraints start out being ranked at the same height (Tesar and Smolensky 1993) or with all constraints on nodes outranking all constraints on connections (e.g. “markedness  $\gg$  faithfulness”: Levelt 1995). Following the OT literature on learning from overt forms (Tesar and Smolensky 1998, 2000; Apoussidou and Boersma 2004; Biró 2013; Jarosz 2013b), we regard it as the learners job in our French example to rearrange the constraint hierarchy until the phonetic form produced for each of the three meanings corresponds to the form presented in the learning data.

Following the idea of error detection in OT learning Tesar and Smolensky (1998), the learner can learn by comparing two paths: the path that she would produce herself, which for our example is simply the path in Figure 9, and the path that she considers correct. We will now explain in detail first how the learner determines the “correct” path, and then how she learns from it.

To determine the “correct” path the learner follows a procedure that generalizes the idea of *robust interpretive parsing* (Tesar and Smolensky, 2000). When given a meaning–form pair, the learner determines which of the many possible paths from the meaning to the given (correct) phonetic form satisfies the constraints best, i.e. the path between given meaning and form that is optimal according to the learner’s current constraint ranking. This is done by a procedure analogous to the one for evaluation in §4.3.2, with one important additional step at the beginning, namely the deactivation of all phonetic forms that are not given in the data. Figure 4.10 illustrates the procedure with the pair ‘good<sub>i</sub> actor<sub>i</sub>’ ~ [bɔ̃naktœʁ]. The first step is that all phonetic forms except [bɔ̃naktœʁ] are deactivated; since our example has only one such form, namely [bɔ̃aktœʁ], this step deactivates only the node [bɔ̃aktœʁ], as can be seen in Figure 4.10a.<sup>5</sup> Typically, after this initial deactivation, there remain many possible paths through the graph (in Figure 10a, there are eight). To find the optimal path from among these, the learner follows the exact same efficient procedure as in §4.3.2, namely to deactivate nodes and connections in an order determined by the constraint ranking: constraints ① and ② sever two connections, and constraint ③ deactivates a node, as illustrated in Figure 4.10b. As with the evaluation procedure in §4.3.2, only a single path is left (or, in the general case, a few equally optimal paths). Because of the initial deactivation of non-French phonetic forms, this path must run from the meaning given in the data to the phonetic form given in the data. In the figure, this path is drawn with thick lines and provided with a check mark.

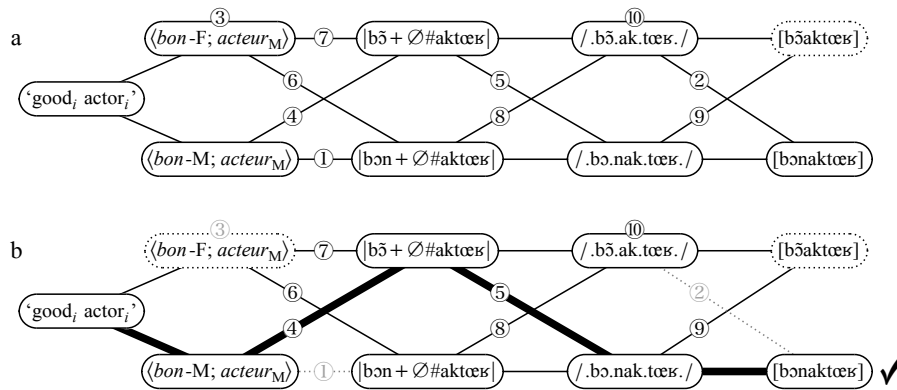


Figure 4.10: Robust interpretive parsing: initial deactivation of non-French phonetic forms, followed by the elimination of edges and nodes by the constraint hierarchy. One path remains (very thick lines).

The learner now regards the path of Figure 4.10 as the *correct path*, at least for the purpose of handling the current form–meaning pair.

<sup>5</sup>In later sections we apply robust interpretive parsing to the larger graphs of Figure 4.2. In those cases, this first step deactivates three phonetic nodes at once.

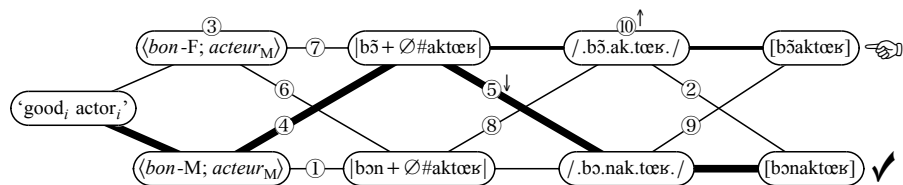


Figure 4.11: Learning by comparing the produced path of Figure 4.9 to the “correct” path of Figure 4.10b.

Now that the learner knows the “correct” path, she proceeds to learn from it, i.e. to potentially change her constraint ranking. To achieve this, she starts by comparing the “correct” path of Figure 4.10b to the produced path of Figure 4.9. If the two paths are the same, the learner sees no reason to change her grammar (and indeed, her produced phonetic form is correct French). If the two paths are different, the learner decides that she made an *error* and that her grammar has to change.<sup>6</sup> Analogously to learning from the violation profiles of one “winner” and one “loser” candidate in an OT tableau (Tesar and Smolensky, 1998), the learner initiates the change by comparing the violation profile of her produced path to that of the “correct” path, as in Figure 4.11. She will then update her grammar by applying one of several OT update rules that have been proposed in the literature: following the procedure used by Boersma (2007) and Apoussidou (2007), the rankings of the constraints violated by the production path are raised, and the rankings of all the constraints violated by the “correct” path are lowered. In Figure 4.11, the two paths differ in the right half of the figure, so that constraint ⑤ has to be lowered and constraint ⑩ has to be raised, as indicated by the arrows.

We have found efficient procedures for evaluation (§4.3.2) and for learning (section 4.3). The next sections investigate how these procedures handle the French sample problem.

## 4.4 Learning French gender allomorphy

In section 4.2, we showed through an example with 3 meanings and 20 constraints how an aspect of French liaison may be described in a multi-level parallel OT grammar. The present section illustrates how the evaluation and learning procedures of section 4.3, which were illustrated there with only one meaning and 10 constraints, work out for the complete problem of section 4.2. We investigate what parallel and serial analyses are possible, and which of these are found by computer-simulated learners.

<sup>6</sup>This is appropriate, because if the two paths are different, their phonetic forms must also be different. Intuitively, this must be true because if the phonetic forms of the two paths are identical, the paths must be identical as well (Bellman’s principle of optimality: if the overall optimal path contains phonetic form X, then this optimal path must be in the set of all possible paths to phonetic form X, and it must be the best of those paths).



### 4.4.1 The constraints

Our first grammar contains the constraints of section 4.2, except the constraint against phonetic schwa ( $*[\text{ə}]$ ). We summarize the whole set of 19 constraints here:

**Lexical-semantic constraints (0 in CON).** All such constraints are in the candidate generator GEN, i.e. they are so strong that not even a connection is allowed that would violate them. For instance, the meaning ‘husband’ cannot connect to the morpheme  $\langle \text{acteur}_M \rangle$  or  $\langle \text{voiture}_F \rangle$ , in this simplified world. The effect of this is equivalent to the effect of inviolable lexical-semantic constraints like  $*\langle \text{husband} \langle \text{acteur}_M \rangle$ . The connections between actor and  $\langle \text{acteur}_M \rangle$  and between good and  $\langle \text{bon} \rangle$  therefore do not violate any constraints in CON.

**Morphosyntactic constraints (1).** Here we only have the gender disagreement constraint  $*\langle \text{FM} \rangle$ , which is violated in  $\langle \text{bon-F}; \text{acteur}_M \rangle$ ,  $\langle \text{bon-F}; \text{husband}_M \rangle$ , and  $\langle \text{bon-M}; \text{car}_F \rangle$ .

**Lexical-phonological constraints (8).** Many of these are in GEN, but the ones relevant to the allowed connections in Figure 2 are  $*\langle \text{bon-F} \rangle | \text{b}\text{ɔ}\text{n} |$ ,  $*\langle \text{bon-M} \rangle | \text{b}\text{ɔ}\text{n} |$ ,  $*\langle \text{bon-F} \rangle | \text{b}\text{ɔ} |$ ,  $*\langle \text{bon-M} \rangle | \text{b}\text{ɔ} |$ ,  $*\langle \text{M} \rangle | \text{ə} |$ ,  $*\langle \text{M} \rangle | \text{ɔ} |$ ,  $*\langle \text{F} \rangle | \text{ə} |$ , and  $*\langle \text{F} \rangle | \text{ɔ} |$ . The last of these is violated e.g. in the connection from  $\langle \text{bon-F}; \text{mari}_M \rangle$  to  $| \text{b}\text{ɔ}\text{n} + \text{ɔ} \# \text{ma} \text{ʁ} \text{i} |$ . This constraint subset is formulated in an exhaustive way, and handles allomorphy at the morpheme level, i.e. whether feminine and masculine stems are allowed to be different in underlying form.

**Faithfulness constraints (4).** We have  $*|\text{ɔ}| / \text{ɔ}\text{n} /$ ,  $*|\text{ɔ}\text{n}| / \text{ɔ} /$ ,  $*|\text{ə}| / /$  (against schwa deletion) and  $*| / \text{ə} /$  (against schwa insertion). Exhaustive connectivity between underlying and surface form would also require anti-faithfulness constraints such as  $*|\text{ɔ}| / \text{ɔ} /$  and  $*|\text{ɔ}\text{n}| / \text{ɔ}\text{n} /$ , but these are not included in our example.

**Structural constraints (2).** We have the hiatus-avoiding constraint  $*\langle \text{V.V} \rangle$ , which is violated by the second syllable in  $/ \text{b}\text{ɔ} \text{.} \text{a} \text{k} \text{.} \text{t}\text{œ} \text{ʁ} \text{.} /$ , the third syllable in  $/ \text{b}\text{ɔ} \text{.} \text{n}\text{ə} \text{.} \text{a} \text{k} \text{.} \text{t}\text{œ} \text{ʁ} \text{.} /$ , and by any form that starts with  $/ \text{b}\text{ɔ} \text{.} \text{ə} /$  (i.e.  $/ \text{b}\text{ɔ} \text{.} \text{ə} \text{.} \text{a} \text{k} \text{.} \text{t}\text{œ} \text{ʁ} \text{.} /$  violates it twice). We also have the constraint  $*\langle \text{n} \rangle$ , which militates against  $/ \text{n} /$  in coda.

**Cue constraints (4).** We have  $*\langle \text{ɔ} \rangle / [ \text{ɔ}\text{n} ]$ ,  $*\langle \text{ɔ}\text{n} \rangle / [ \text{ɔ} ]$ ,  $*\langle \text{ə} \rangle / [ ]$  and  $*\langle / \rangle / [ \text{ə} ]$ . Exhaustive connectivity between surface and phonetic form would also require “perverse” constraints such as  $*\langle \text{ɔ} \rangle / [ \text{ɔ} ]$  and  $*\langle \text{ɔ}\text{n} \rangle / [ \text{ɔ}\text{n} ]$ , but these are not included in our example so as not to introduce too much arbitrariness at the phonology–phonetics interface.

### 4.4.2 Grammars that work

For each of the three meanings there are 32 routes through the graph to arrive at the correct French phonetic form. When only looking at the graphs and ignoring constraints, we would therefore predict that there are  $32^3 = 32768$  possible analyses of the French data. The 19 constraints of §4.4.1 severely re-

strict this number. A brute-force search<sup>7</sup> shows that with this constraint set, there are only six different analyses that produce the correct phonetic forms for each of the three meanings. Three of these analyses are shown in Figures 4.12 through 4.14 each of these figures shows a triplet of graphs that represents a grammar-cum-analysis of the French data (just as a set of tableaux does in two-level OT).

The analyses in Figures 4.12–4.14 constitute two crucially parallel analyses and one “serial” analysis.

The first crucially parallel analysis (“PU”) is seen in Figure 4.12. The morphemes ⟨bon-M⟩ map to the underlying form |bɔ̃n| if followed by a vowel in the surface form, but to the underlying form |bɔ̃| if followed by a consonant in the surface form. Hence, a “later” consideration in the phonology, namely the constraint \*/V.V/ at the surface level, influences an “earlier” choice at the underlying level. Such bottom-up influences in production can only occur in a parallel evaluation, not in a serial evaluation.

The second crucially parallel analysis (“PG”) is seen in Figure 4.13. The meaning good, applied to a masculine noun, maps to the morphemes ⟨bon-M⟩ if followed by a consonant in the surface form, but to the morphemes ⟨bon-F⟩ if followed by a vowel in the surface form. This is Encrevé-Lamberts and Tranel’s gender allomorphy of section 4.2.2. Again, a “later” consideration influences an “earlier” choice, this time at the morphemic level. This type of analysis can again only occur in a parallel evaluation, not in a serial evaluation.

The “serial” analysis, i.e. an analysis in our parallel model that could also occur in a serial model, is seen in Figure 4.14. From the meaning level to the morpheme level, the contrast between masculine and feminine is maintained in that good maps to ⟨bon-M⟩ before masculine nouns and to ⟨bon-F⟩ before feminine nouns. From the morphemic to the underlying level, the contrast between masculine and feminine is maintained in that the morphemes ⟨bon-M⟩ map to underlying |bɔ̃| and the morphemes ⟨bon-F⟩ map to underlying |bɔ̃n|. The phonological constraint \*/V.V/ enforces its influence only at the surface level, where it converts underlying |bɔ̃| to surface /bɔ̃n/ if a vowel follows. This serial analysis (labeled “SN” as a reference to n-insertion) cannot handle data beyond the current toy example, because it cannot explain why |bɔ̃| should become /bɔ̃n/ but |mɔ̃|— may become /mɔ̃n/ (Dell, 1970; Selkirk, 1972).

---

<sup>7</sup>Technically, this was done by applying Batch Constraint Demotion (Tesar and Smolensky, 1995) on each of the 32768 analyses and seeing whether the algorithm converged. Cycling through all possible analyses like this is feasible for the current toy example, but stops being feasible if the number of forms per level grows much larger.

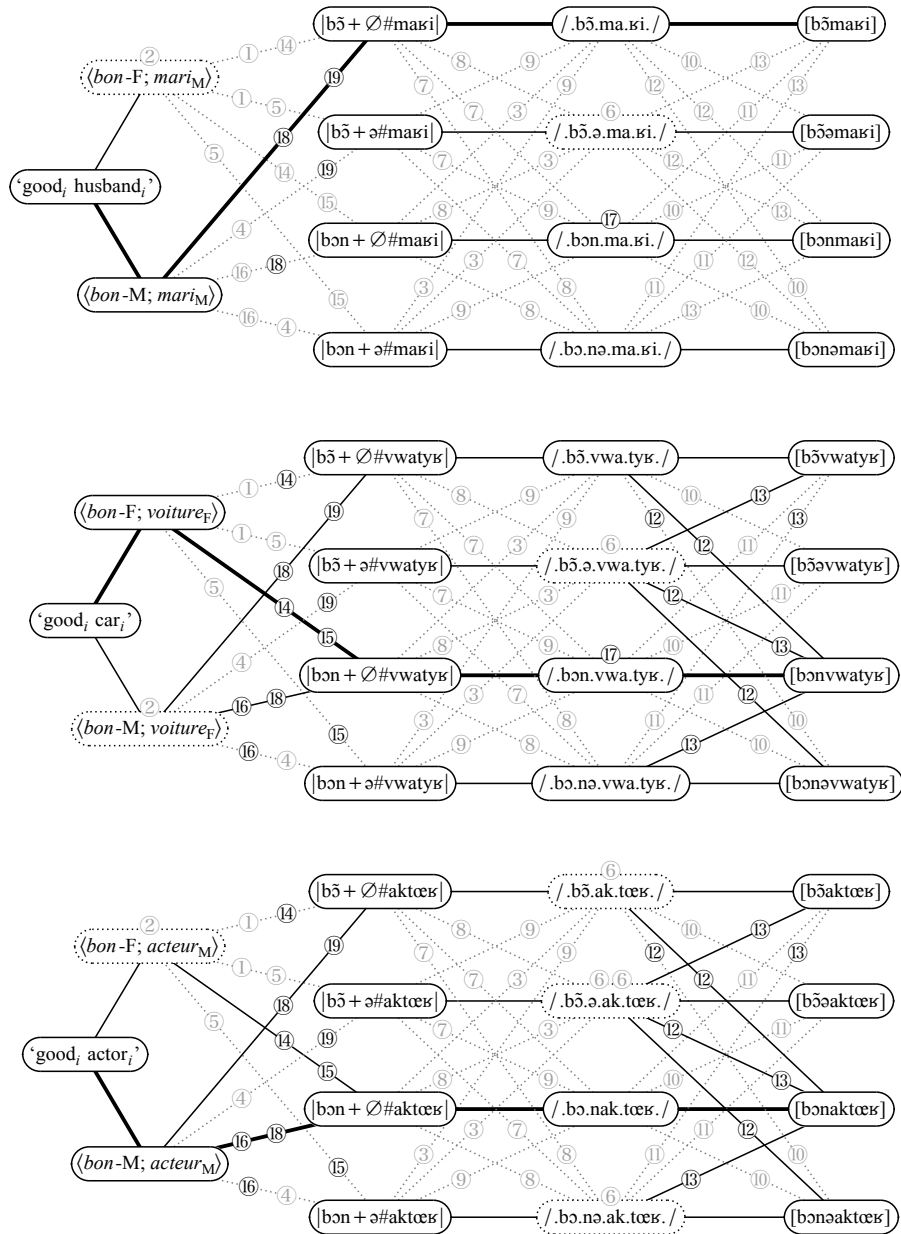


Figure 4.12: Parallel analysis with allomorphy in the underlying form (“PU”).  
 Ranking: ① \*(bon-F) |bɔ̃| >> ② \*(FM) >> ③ \*/ɔn| /ɔ̃/ >> ④ \*(M)|ə| >> ⑤ \*(F)|ə| >> ⑥ \*/V.V/ >> ⑦ \*/ɔ̃/ɔn/ >> ⑧ \*/|/ə/ >> ⑨ \*/ə|// >> ⑩ \*/|/ [ə] >> ⑪ \*/ɔn/ [ɔ̃] >> ⑫ \*/ɔ̃/ [ə] >> ⑬ \*/ə/ [ ] >> ⑭ \*(F) |Ø| >> ⑮ \*(bon-F)|bɔn| >> ⑯ \*(bon-M)|bɔn| >> ⑰ \*/n./ >> ⑱ \*(M) |Ø| >> ⑲ \*(bon-M)|bɔ̃|.

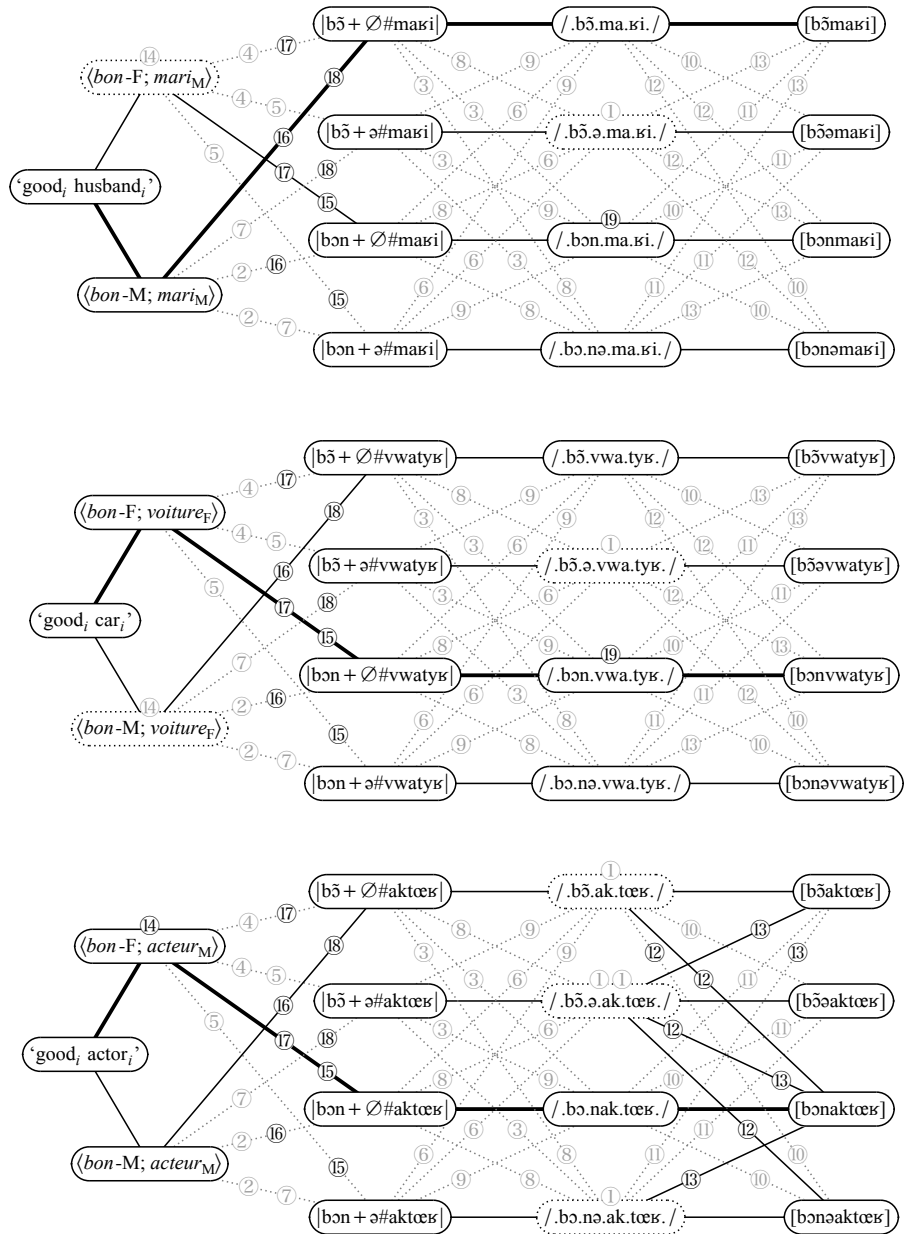


Figure 4.13: Parallel analysis with true gender allomorphy (‘PG’). Ranking:  
 ① \*/V.V/ >> ② \*(bon-M)bɔn >> ③ \*|ɔ̃|/ɔn/ >> ④ \*(bon-F) |bɔ̃| >> ⑤ \*(F)|ə| >> ⑥ \*|ɔn|/ɔ̃/ >> ⑦ \*(M)|ə| >> ⑧ \*| | /ə/ >> ⑨ \*|ə|// >> ⑩ \*/ / [ə] >> ⑪ \*/ɔn|/ɔ̃| >> ⑫ \*/ɔ̃|/ɔn| >> ⑬ \*/ə/ [ ] >> ⑭ \*(FM) >> ⑮ \*(bon-F)bɔn >> ⑯ \*(M) |ə| >> ⑰ \*(F)|ə| >> ⑱ \*(bon-M) |bɔ̃| >> ⑲ \*/n./.

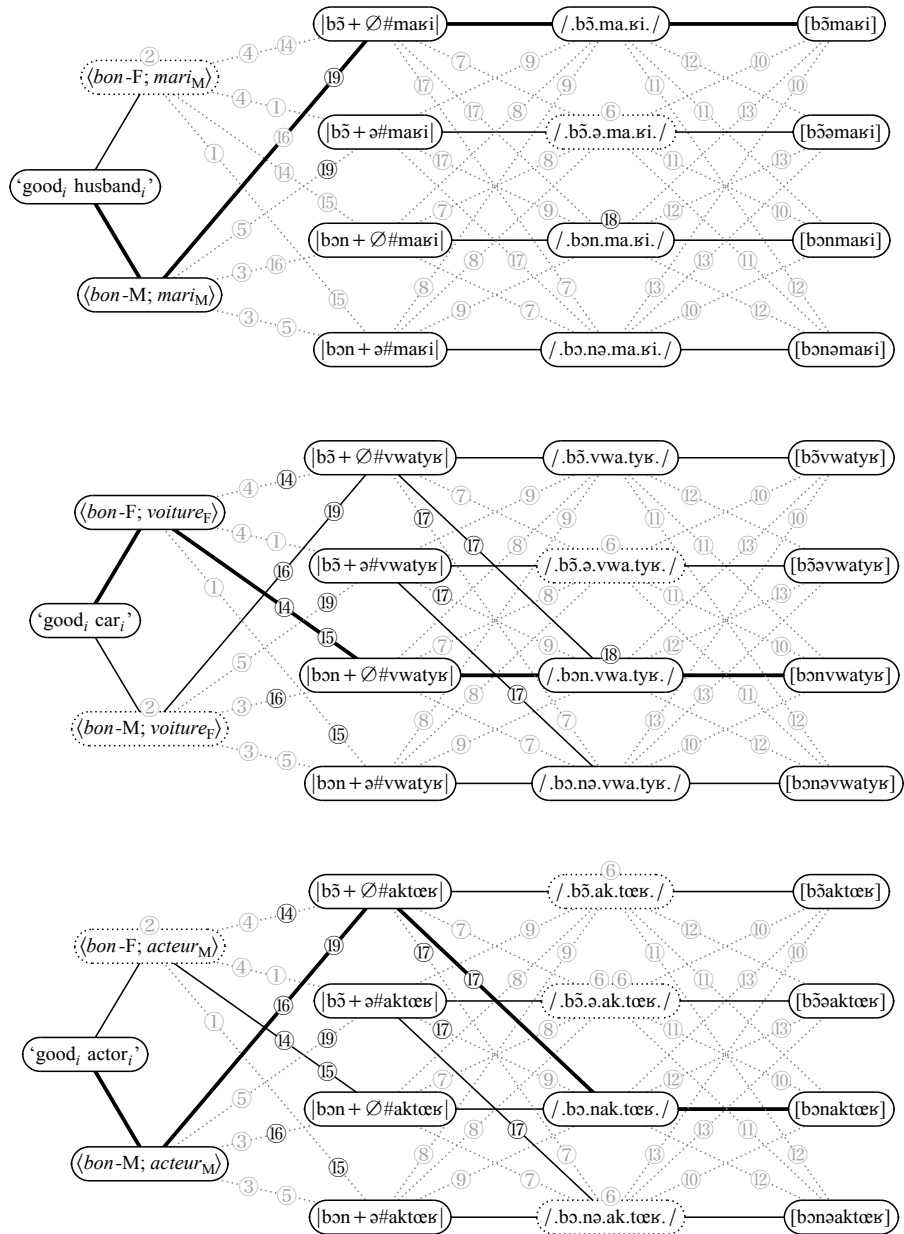


Figure 4.14: “Serial” analysis: phonological n-insertion (“SN”). Ranking: ①  $*(F)|ə|$  >> ②  $*(FM)$  >> ③  $*(\text{bon-M})bɔn$  >> ④  $*(\text{bon-F})|bɔ̃|$  >> ⑤  $*(M)|ə|$  >> ⑥  $*/V.V/$  >> ⑦  $*/|ə/$  >> ⑧  $*/ɔn/ɔ̃/$  >> ⑨  $*/ə//$  >> ⑩  $*/ə/[$  >> ⑪  $*/ɔ̃/|ɔn|$  >> ⑫  $*/ / [ə]$  >> ⑬  $*/ɔn/ɔ̃/$  >> ⑭  $*(F)|ə|$  >> ⑮  $*(\text{bon-F})bɔn$  >> ⑯  $*(M)|ə|$  >> ⑰  $*/ɔ̃/ɔn/$  >> ⑱  $*/.n./$  >> ⑲  $*(\text{bon-M})|bɔ̃|$ .

The remaining three analyses are similar to those of the graphs of Figs. 4.12 through 4.14, but have a schwa in some masculine intermediate forms. PU (Fig. 4.12) has a variant (“PU<sub>ə</sub>”) with schwa in both masculine underlying forms:  $|b\tilde{o}+\partial\#ma\beta i|$  and  $|b\tilde{o}n+\partial\#akt\alpha\beta|$ .<sup>8</sup> PG (Fig. 4.13) has a variant (“PG<sub>ə</sub>”) with schwa in the consonantal masculine underlying form  $|b\tilde{o}+\partial\#ma\beta i|$  only.<sup>9</sup> Finally, SN has a variant (“SN<sub>ə</sub>”) with schwa in both masculine underlying forms, which are now  $|b\tilde{o}+\partial\#ma\beta i|$  and  $|b\tilde{o}+\partial\#akt\alpha\beta|$ .<sup>10</sup> One can say that for all analyses it is immaterial whether the underlying form has  $|\emptyset|$  or  $|\partial|$ ; the ranking of  $\ast\langle M \rangle |\emptyset|$  with respect to  $\ast\langle M \rangle |\partial|$  and  $\ast|\partial| / /$  determines the masculine underlying ending, but this ranking does not influence anything else in the analysis. The reason why we see no schwas in the feminine underlying forms is that  $\ast/V.V/$  is capable of “deleting” underlying schwas from  $|b\tilde{o}+\varepsilon\#ma\beta i|$ ,  $|b\tilde{o}+\varepsilon\#akt\alpha\beta|$ , and  $|b\tilde{o}n+\varepsilon\#akt\alpha\beta|$ , but not from  $|b\tilde{o}n+\varepsilon\#vwa\tau y\beta|$  (but see 4.6).

The serial analysis of §4.2.1 does not appear here, because  $\ast[\partial]$  has not been included in the constraint set yet. See 4.6 (and Figure 15) for analyses that become possible when the constraint set is larger.

### 4.4.3 Error-driven learning

We have seen that the set of 19 constraints of 4.4.1 is compatible with six distinct analyses that produce the correct phonetic form for each meaning. Some of those analyses had been proposed before, but some were novel. It remains to be investigated which (if any) of these six analyses are learnable. After all, it is possible that there exist analyses that are allowed by factorial typology (i.e. representable by a constraint ranking) but for which no learning path exists. It then becomes interesting to see which of the six analyses virtual, i.e. computer-generated, learners will find: starting with a certain initial state and following a certain reranking algorithm, will they come up with a serial analysis known from the literature, such as the one in 4.2.1, or with a parallel analysis known from the literature, such as the one in 4.2.2, or with one of the novel analyses?

To assess learnability, we supply a virtual learner with 10,000 pairs of meaning and phonetic form randomly drawn from the French data in §4.3.3, i.e. each of the three meaning–form pairs will occur approximately 3300 times in the learners input. During this process, the learner is equipped only with a constraint set (the 19 constraints), a candidate generator (the graphs of Figure 4.2), a single current grammar hypothesis (i.e. a current constraint ranking),

<sup>8</sup>A possible ranking is:  $\ast\langle F \rangle |\partial| \gg \ast\langle FM \rangle \gg \ast/V.V/ \gg \ast|\tilde{o}/\alpha n/ \gg \ast\langle \text{bon } F \rangle |b\tilde{o}| \gg \ast\langle M \rangle |\emptyset| \gg \ast| / \partial / \gg \ast|\alpha n| / \tilde{o} / \gg \ast/\tilde{o} / [\alpha n] \gg \ast/\alpha n / [\tilde{o}] \gg \ast/\partial / [ ] \gg \ast//[\partial] \gg \ast\langle \text{bon } M \rangle |b\tilde{o}n| \gg \ast\langle \text{bon } F \rangle |b\tilde{o}n| \gg \ast\langle F \rangle |\emptyset| \gg \ast\langle M \rangle |\partial| \gg \ast/n./ \gg \ast|\partial| / / \gg \ast\langle \text{bon } M \rangle |b\tilde{o}|$ .

<sup>9</sup>A ranking is:  $\ast\langle F \rangle |\partial| \gg \ast\langle M \rangle |\emptyset| \gg \ast\langle \text{bon } M \rangle |b\tilde{o}n| \gg \ast/V.V/ \gg \ast| / \partial / \gg \ast\langle \text{bon } F \rangle |b\tilde{o}| \gg \ast|\tilde{o} / \alpha n / \gg \ast|\alpha n| / \tilde{o} / \gg \ast/\tilde{o} / [\alpha n] \gg \ast/\partial / [ ] \gg \ast//[\partial] \gg \ast/\tilde{o} n / [\tilde{o}] \gg \ast\langle FM \rangle \gg \ast/n./ \gg \ast\langle \text{bon } M \rangle |b\tilde{o}| \gg \ast\langle M \rangle |\partial| \gg \ast|\partial| / / \gg \ast\langle \text{bon } F \rangle |b\tilde{o}n| \gg \ast\langle F \rangle |\emptyset|$ .

<sup>10</sup>A ranking is:  $\ast\langle M \rangle |\emptyset| \gg \ast\langle \text{bon-M} \rangle |b\tilde{o}n| \gg \ast\langle \text{bon-F} \rangle |b\tilde{o}| \gg \ast\langle FM \rangle \gg \ast/V.V/ \gg \ast\langle F \rangle |\partial| \gg \ast| / \partial / \gg \ast|\alpha n| / \tilde{o} / \gg \ast/\alpha n / [\tilde{o}] \gg \ast/\partial / [ ] \gg \ast/\tilde{o} / [\alpha n] \gg \ast / / [\partial] \gg \ast\langle \text{bon-M} \rangle |b\tilde{o}| \gg \ast\langle F \rangle |\emptyset| \gg \ast|\tilde{o} / \alpha n / \gg \ast\langle \text{bon-F} \rangle |b\tilde{o}n| \gg \ast\langle M \rangle |\partial| \gg \ast/n./ \gg \ast|\partial| / /$ .

Analysis	PU	PG	SN	PU <sub>ə</sub>	PG <sub>ə</sub>	SN <sub>ə</sub>	SN~PU	SN~SN <sub>ə</sub>	Total
Proportion of random baseline learners	25%	7%	46%	8%	2%	12%	0	0	100%
Proportion of “weighted uncanceled” learners	1.7%	0?	21%	0.01%	0?	0.9%	0.16%	0.04%	24%

Table 4.1: Proportions in which the 6 possible analyses are found by two kinds of learners. For random baseline learners, see §4.4; for “weighted uncanceled” learners, see §4.5.

and at most a single datum (meaning–form pair) that she is currently handling. On each datum, the learner performs a virtual production, i.e. she computes the optimal path from the given meaning according to the evaluation procedure of 4.3.2, and compares this with the path computed by robust interpretive parsing (4.3.3); if the paths are different, the learner takes action by changing her constraint ranking (this is therefore error-driven learning). After 10000 data, we stop the learner and check whether the learners final constraint ranking is correct, i.e. whether the ranking maps the three French meanings to the correct French phonetic forms. Note that the learning procedure works *online*: the learner maintains a single ranking at each time and considers a single data pair at a time, without any memory of previous hypotheses or previous data.

#### 4.4.4 The random baseline learner

It has been argued that learning algorithms should be checked against a random baseline (for parameter setting: Berwick and Niyogi 1996; for OT: Jarosz 2013b). In our case, the random baseline learner starts by randomly choosing a constraint ranking out of the 19-factorial possible rankings, and when her ranking fails on a datum, she randomly chooses a new constraint ranking. When we simulated 1,000 learners in this way, they all turned out to have a correct French grammar after 10,000 data. Counts of their resulting analyses are in the middle row of Table 4.1.

We conclude that the learners have a preference for the “serial” analyses (58%) over the parallel analyses (42%), and for the schwa-less analyses (78%) over the analyses that posit a schwa somewhere along the route (22%). Since in this algorithm a grammar no longer changes after it is correct, these preferences are easy to explain: they correspond to the number of rankings (out of the 19! possible ones) that yield each analysis.

The fact that all 1,000 learners succeeded can be explained by the probability of guessing a correct ranking by chance. When we drew 1,000,000 random rankings of the 19 constraints, about 0.62 percent of these rankings rep-

resented a correct grammar for the French data. This means that a random baseline learner will run into a correct grammar after making on average 160 errors, a number that is reached after at most approximately 480 pieces of data (the worst case, which occurs if all errors are made for only one of the three data pairs). The probability that a learner has not encountered a correct grammar after 10,000 data, if the incidence of such grammars is 0.48%, is approximately  $10^{-9}$ .

The problem with the random baseline, however, is that it does not scale. Berwick and Niyogi (1996) took Gibson and Wexler (1994) grammar of 3 parameters and showed that a random error-driven selection out of the 23 (= 8) possible grammars results in more (namely, 100%) successes and faster convergence than Gibson and Wexler's "local and greedy" algorithm, which takes on the order of  $3^2$  (= 9) error steps; the problem here is that Berwick and Niyogi's baseline would probably fail if the number of parameters were 30 instead of 3, because  $2^{30}$  (= 1,073,741,824) is much more than  $30^2$  (= 900). Likewise, the random baseline in the OT case tends to scale exponentially with the number of constraints: there is no hope that e.g. the 322 cue constraints simulated by Boersma and Hamann (2008), or the large number of lexical constraints suggested by Boersma (2001), are ranked correctly by random selection within any learners lifetime.

In the following section we therefore try out the "local and greedy" reranking procedure of §4.3.3, and several variants on it.

#### **4.4.5 Incremental learning procedures**

The opposite of a random reselection procedure is an incremental procedure: in case the current grammar hypothesis fails on the incoming datum, only a few constraints are reranked, and the remainder retains their current ranking. One example of this is the reranking procedure in 4.3.3, which is the standard version of the Gradual Learning Algorithm (Boersma and Hayes, 2001). This assumes Stochastic OT, in which all constraints have numerical ranking values, and indeed the initial state of the learner in our simulations is that all constraints are ranked at the same height of 100.0. Evaluation proceeds by temporarily adding to each constraints ranking a random value drawn from a Gaussian distribution with standard deviation 2.0 (the evaluation noise). When a learning pair comes in, the learner uses the evaluation noise to determine a constraint ranking, and then uses this ranking to compute both her virtual production and her robust interpretive parsing ("correct" path). If these two paths differ, all constraints that prefer (i.e. have fewer violations in) the produced path (✗) are demoted by a value of 1.0 (the plasticity), and all constraints that prefer the "correct" path (✓) are promoted by 1.0. The effect is that the grammar moves closer to a ranking where the supposedly correct candidate may win.

When we simulate 100 learners with this "symmetric all" reranking procedure, however, none of them finds a correct grammar of French after 10,000



data. The learners typically get stuck in various types of limit cycles, alternating between a grammar that handles ‘good actor’ and ‘good car’ correctly and a grammar that handles ‘good actor’ and ‘good husband’ correctly. Figures 15 and 16 illustrate a type of limit cycle that is known from the literature (Tesar and Smolensky 2000, p. 67, Boersma and Pater 2016). The constraint ranking in Figure 15 correctly produces good husband as [bɔ̃mavɛ] (not shown), but incorrectly produces good car as [bɔ̃vwatɥ], as shown by the semi-thick path and the pointing finger. When the correct phonetic form [bɔ̃nvwatɥ] comes in, leading to the “correct” thick path, a comparison between the two paths will lead to a demotion of constraints (16) \*|ɔ̃|/ɔ̃n/ and (19) \*/ə/ [ ] and a promotion of constraint (17) \*/ə/ / /.

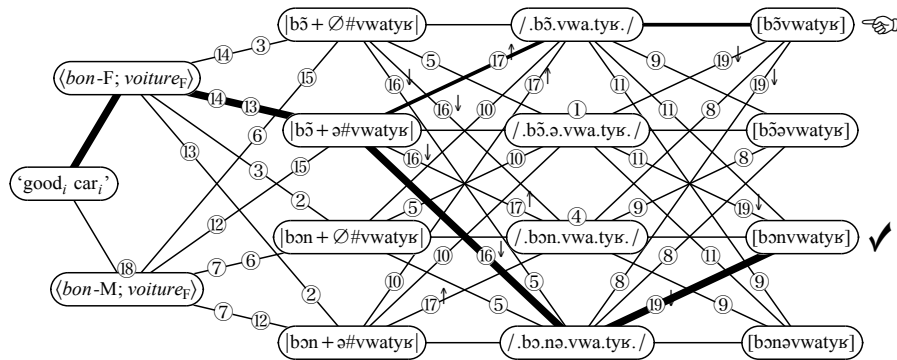


Figure 4.15: A ranking appropriate for [bɔ̃mavɛ] but not for [bɔ̃nvwatɥ]: upon encountering [bɔ̃nvwatɥ], the learner will promote constraint (17) and demote constraints (16) and (19).

The combination of movements of (16) \*|ɔ̃|/ɔ̃n/ and (17) \*/ə/ / / will typically result in these constraints becoming ranked in the opposite order. This change is shown in Figure 16, where \*/ə/ / / is now labeled (16) and \*|ɔ̃|/ɔ̃n/ is labelled (17). This new ranking, however, now causes the grammar to fail on the input good husband, producing [bɔ̃nəmavɛ] as shown in the figure. When the correct phonetic form [bɔ̃mavɛ] now comes in, the learning algorithm will demote (16) \*/ə/ / / and promote (17) \*|ɔ̃|/ɔ̃n/.

This movement will restore the original ranking of \*|ɔ̃|/ɔ̃n/ over \*/ə/ / /, leading again to a grammar that handles good husband correctly but fails on good car.

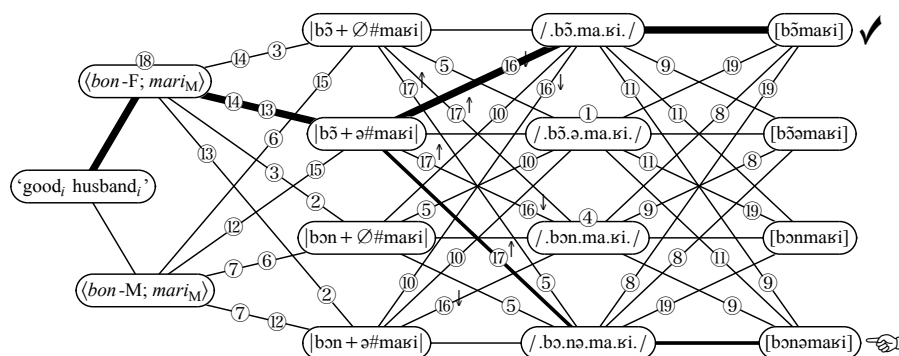


Figure 4.16: The ranking that results from the constraint movements in figure 4.15, which will now incorrectly produce [bɔ̃əmaksɪ]. When confronted with [bɔ̃maksɪ] the learner promotes one of the constraints demoted in figure 4.15 and demotes the constraint that was promoted in Figure 4.15.

We can identify two problems with the events in Figures 4.15 and 4.16. The first problem is that  $*|\bar{\sigma}| / \bar{\sigma}n/$  and  $*|\bar{\sigma}| / /$  will swap places eternally, never allowing the grammar to leave this cycle. The second problem is that the net result of the Figure 4.15 – Figure 4.16 cycle is that constraint (19)  $*|\bar{\sigma}| / [ ]$ , which was already bottom-ranked, has moved further down the hierarchy; in a situation of stochastic ranking this fruitless eternal downward “pumping” reduces the possibility that  $*|\bar{\sigma}| / [ ]$  will ever play a role again. We will now discuss various attempts to improve on this problem.

The pumping effect can be reduced by a variant reranking procedure, namely “weighted uncanceled” (Apoussidou, 2007: 174). This update rule is similar to “symmetric all”, except that the value subtracted from the ranking of  $\bar{\sigma}$ -preferring constraints is now equal to the plasticity divided by the number of these constraints, and the value added to the ranking of  $\checkmark$ -preferring constraints equals the plasticity divided by the number of these constraints (if either of these numbers of constraints is zero, no constraint moves). In Figure 15, for instance, constraints (16) and (19) are demoted by only 1/2 each and constraint (17) is promoted by 1, and the combination of Figures 15 and 16 leads to (17) staying fixed, (16) rising by 1/2, and (19) falling by 1/2. In general, this scheme leads to less downward pumping than “symmetric all”, because the amount by which constraints can fall is kept within bounds by the amount that other constraints can rise, and this is in turn bounded by the other constraints in the grammar: in our example, for instance, constraint (16)  $*|\bar{\sigma}| / \bar{\sigma}n/$  can never reach beyond the top of the hierarchy, because if this constraint becomes high-ranked the learner will choose as the “correct” path a path that does not violate it. The result is that both (16) and (19) get more chances to interact with other constraints under the “weighted uncanceled” scheme than under the “symmetric all” scheme, and this is a general difference between

the schemes that is not specific to the toy case at hand.

For our French example, it is indeed the case that the “weighted uncanceled” reranking procedure works better than “symmetric all”: about 24% of 10,000 virtual learners succeeded. The analyses that the succeeding learners came up with are summarized in the bottom row of Table 1. Most found the “serial” analysis SN, 90 found SN<sub>∅</sub>, and 170 found PU (same preference as the random baseline learners). A few (16) learners came up with a variable grammar in which  $\langle \text{bon-M} \rangle | \text{b} \circ \text{n} |$  and  $\langle \text{∫} \rangle / \circ \text{n} /$  were ranked at the same height, leading to a 5050 variation between the SN and PU analyses; in this example we see that the ranking difference between a “serial” and a crucially parallel grammar can be minimal (the learner cannot notice this; to a learner, all grammars are parallel and the potential existence of a serial equivalent is inaccessible). Finally, 4 of the 10,000 learners came up with a variable grammar in which  $\langle \text{M} \rangle | \circ |$  and  $\langle \text{M} \rangle | \circ |$  were ranked at the same height, leading to a 5050 variation between the SN and SN<sub>∅</sub> analyses. We see that only three of the six analyses seen in §4.2 emerge with any frequency in the Optimality-Theoretic grammars, and that variable grammars that never occur for random baseline learners (for whom two rankings have probability zero of being equal) are also possible.

A point of concern in figure 15 is the demotion of (19): demoting constraints that are ranked below the highest-ranked  $\checkmark$ -preferring constraint ((17)) cannot really help to improve the decision. Tesar and Smolenskys (1998) Error-Driven Constraint Demotion (EDCD) therefore prevents the demotion of (19): in this scheme, all  $\checkmark$ -preferring constraints ranked above the highest-ranked checkmark-preferring constraint are demoted to a value just below this highest-ranked  $\checkmark$ -preferring constraint, and no constraint is promoted. In figure 15, (16)  $\langle \text{∫} \rangle / \circ \text{n} /$  will be demoted below (17)  $\langle \text{∫} \rangle / /$ , and in Figure 4.16,  $\langle \text{∫} \rangle / /$  will be demoted below  $\langle \text{∫} \rangle / \circ \text{n} /$ . These two constraints will continue to tumble down in this way, and once they pass constraint (19)  $\langle \text{∫} \rangle / \circ / [ ]$ , they will drag it along down the hierarchy. The chances for these three constraints to interact with the rest of the hierarchy will diminish even faster than with symmetric all, and indeed in our simulations with EDCD (with stochastic ranking; Boersma 2009b), none of the 10,000 simulated learners succeeded in correctly learning a ranking for the three French pieces of data.

Magri (2012)’s update rule limits some of the demotions of EDCD, while retaining the advantage of not demoting (19) in Figure 4.15: in this scheme, all  $\checkmark$ -preferring constraints ranked above the highest-ranked  $\checkmark$ -preferring constraint are demoted by 1.0, whereas all constraints that prefer the correct path are promoted by 1.0 multiplied by (the number of constraints being demoted) and divided by (the number of constraints being promoted plus one). In Figure 4.15, this means that (16) falls by 1 and (17) rises by 1/2, and in the combination of Figures 4.15 and 4.16, both constraints end up being demoted by 1/2, while (19) is pumped down once the other two constraints have reached it. Re-

garding this behavior, this scheme can be expected to work better than EDCD but not better than weighted uncanceled, and indeed none of 10,000 simulated learners (with stochastic ranking) turned out to be able to learn from the data.

The relative success of the various update rules corroborates earlier comparisons in the literature, where symmetric all had more success than EDCD (Boersma, 2003) and “weighted uncanceled had more success than symmetric all (Apoussidou, 2007).

A general strategy for improving solutions to difficult optimization problems is to add randomness (Kirkpatrick et al., 1983), and this indeed turns out to help in the current case. We performed an additional series of simulations for the 19-constraint grammar of section 4.4.1 under the Generalized Robust Interpretive Parsing (GRIP) strategy introduced by Biró (2013). Recall that under regular RIP, the sets of constraints eligible for promotion and demotion are decided by comparing an (incorrectly) optimal candidate with a parsed target candidate containing a target form. Under GRIP, the optimal candidate is instead compared with a Boltzmann-weighted mean of the entire set of candidates containing the target form. Biró (2013) argues that by maintaining multiple hypotheses over the correct parse in this manner, the learner is more likely to converge on a grammar consistent with all data. The settings for simulations performed with GRIP were identical to those reported above, except with regard to evaluation noise: as Tamás Biró (pers. comm.) suggests, the repeated shuffling of the hierarchy through evaluation noise may not be compatible with GRIP’s notion of a decreasing temperature vector. Indeed, we found that no GRIP learners converged on a correct French grammar within 40,000 data under the standard evaluation noise of 2.0. On the other hand, when the evaluation noise was set to a very small value of 10<sup>-9</sup>, GRIP turned out to outperform regular RIP for our virtual learners of French: 100 out of 100 weighted uncanceled learners, 84 out of 100 symmetric all learners, 33 out of 100 learners using Magri (2012)’s update rule, and 0 out of 100 EDCD learners converged to a correct grammar under GRIP. Again, we see the usual success relations between the algorithms.

As stated in the beginning of this section, the virtual learners in our simulations find optimal candidates through a hierarchy that is stochastically influenced by evaluation noise, and the same draw of the evaluation noise is used for virtual production and for interpretive parsing. A study by Jarosz (2013a) suggests, however, that virtual production and interpretive parsing are performed with different draws of the evaluation noise. Jarosz reports a case in Stochastic OT where this resampling improves learning. When repeating our simulations described in the beginning of this section with resampling (and reranking the constraints only if the phonetic form differed between the two paths), we found that for all update rules the chances of a learner settling on a correct grammar improved: 83 out of 100 “weighted uncanceled” learners, 39 out of 100 symmetric all learners, and 45 out of 100 learners using Magri (2012) update rule succeeded in finding a correct French grammar. The increase in learning success reported by Jarosz thus seems to extend to

multilevel grammars.

#### 4.4.6 A grammar with a constraint against phonetic schwa

In all, the success of the learners with the 19 constraints of section 4.4.1 is modest. A possible cause is that the number of constraints is too low. We therefore tried the constraint set of section 4.2.1, which contains the same 19 constraints as the grammar of section 4.4.1, plus an articulatory constraint \*[ə] that is violated by every occurrence of a schwa in the phonetic form (the constraint seen in the serial account of tableau (12)). Again, only “weighted uncanceled” learners succeeded, in this case 80 out of 100. The analysis found most frequently by the learners (32 times) was one not possible in the original 19-constraint grammar: a variant of SN (figure 4.14) in which the female form contains a schwa on the underlying and surface levels, which is deleted in the phonetic form.

The traditional serial of analysis of section 4.2.1 was found by 1 of the 80 learners, and it is shown in Figure 17. The ranking of Figure 17 is of course not the only ranking that yields these paths. A shift of the four relevant constraint classes such that the rankings between these classes become {lexical-semantic, morphosyntactic lexical-phonological  $\gg$  faithfulness, structural  $\gg$  cue constraints, \*[schwa] } leads to the same winning paths as long as the rankings *within* the classes are preserved; precisely this property is what allows us to call the analysis serial. A fairly large group of learners found a crucially parallel analysis in which a feminine underlying schwa was deleted from the surface form by the ranking of the lower-level constraints \*[ə] and \*/ə/ [ ] over the higher-level constraints \*|ə| / / and \*/n./.

#### 4.4.7 Harmonic Grammar

The algorithm in section 3.2 is specific to the constraint-ranking decision mechanism of OT. It has been suggested that learners that instead use the decision mechanisms of Harmonic Grammar (HG) or one of its variants might perform better on multilevel problems than OT learners (Boersma and Pater, 2016). HG grammars employ weighted constraints instead of the ranked constraints of OT. In our graph-based representation, the evaluation procedure reduces in HG to finding the shortest path, which is efficiently done with dynamic programming, as in the case of OT (section 3.3).

We carried out a number of simulations in which candidates were evaluated using HG (Legendre et al., 1990) and its variants Maximum Entropy (Goldwater and Johnson, 2003), Exponential HG (Boersma and Pater, 2016), Positive HG (Boersma and Pater, 2016), and Linear OT (Keller, 2000). All of these were tested with two of the update rules discussed before, namely, “symmetric all” and “weighted uncanceled” (the other update rules we used above are specific to OT). The result was that for all of these decision mechanisms (except in some cases Exponential HG), the “weighted uncanceled” update

rule was successful for all 100 learners (both with the 19-constraint set and with the 20-constraint set), and the “symmetric all” update rule was successful for 99 learners (with 20 constraints), 78 learners (with 19 constraints in HG and Maximum Entropy), or 50 learners (with 19 constraints in Positive HG and Linear OT). Indeed, HG-style learners performed better in the simulations than OT learners (replicating a tendency also reported in Boersma and Pater 2016), and, again, the weighted uncanceled update rule performed better than symmetric all and learners with an additional constraint against phonetic schwa again outperformed learners who had only the original 19 constraints. Typologically speaking, HG tends to come with a larger set of possible languages than OT. While for the 19-constraint case only six analyses were possible using an OT constraint hierarchy with strict domination, the HG learners found several more weighted constraint rankings: besides five of the possible OT analyses of section 4.2 (the three analyses that the OT learners found in section 4.2, plus the one of figure 9, plus the true gender allomorphy analysis of figure 7), the learners found three analyses that rely on the possibility of having negative constraint weights. An example of an additional triplet of optimal candidates found by a small percentage of HG learners is the following:

‘good<sub>i</sub> car<sub>i</sub>’ ⟨bon-F; voiture<sub>F</sub>⟩ |b̃+ə#vwa.tyɐ| / .bɔ.nə.vwa.tyɐ./ [bɔnvwa.tyɐ]  
 ‘good<sub>i</sub> actor<sub>i</sub>’ ⟨bon-M; acteur<sub>M</sub>⟩ |b̃+ə#aktœɐ| / .bɔ.nak.tœɐ./ [bɔnaktœɐ]  
 ‘good<sub>i</sub> husband<sub>i</sub>’ ⟨bon-M; mari<sub>M</sub>⟩ |b̃+ə#ma.ʁi| / .bɔ.ma.ʁi./ [bɔma.ʁi]

This analysis combines phonological n-insertion with schwa deletion at the phonetic level, and is not similar to any analysis previously proposed in the literature; the cause is, for example, that the path between / .bɔ.nə.vwa.tyɐ./ and [bɔnvwa.tyɐ] requires a negative weight for the constraint \*/ə/ [ ], something that is impossible in OT. As the typological consequences of negative constraint weights are probably undesirable in general (Pater, 2009), it is worthwhile to look at the analyses found by HG learners whose constraint weights are restricted to being positive. The analyses found by learners with such decision mechanisms (Exponential HG, Positive HG, Linear OT) all fell within the set of six OT-compatible analyses of section 4.2, which supports Pater (2009, 2016)’s claim that the typology of positive versions of HG is not so different from the typology of OT. In the end, while learnability criteria seem to favor HG over OT, as suggested both by Boersma and Pater (2016) and by the simulations in this article (as measured by the success rates of the learners), the choice between HG and OT has to be determined also by which of the two produces the better typologies.

#### 4.4.8 Conclusion

In the OT framework, the “symmetric all”, “Magri”, and EDCD learners succeeded with neither constraint set, whereas the “weighted uncanceled” update rule was moderately successful. Whether these differences between the update rules for the present case reflect genuine differences in quality between

the update rules or whether they are peculiar to the present case cannot yet be determined; a whole-language simulation performed on a large corpus of French data may shed light on this question.

Such a larger simulation is also needed if we want to find out whether a serial analysis or a parallel analysis is more appropriate for French. For instance, the serial analysis regards the *bon* ~ *bonne* alternation as phonological, whereas the parallel analysis regards this alternation as suppletive. Including in the simulation some data that are uncontroversially suppletive, such as the *ma* ~ *mon* or the *vieux* ~ *vieil(le)* alternation, may shift the preference of the virtual learners in the direction of a parallel analysis, whereas including more data that are possibly regarded as phonological, such as the alternation between the nouns *chien* [ʃjɛ̃] ‘dog’ and *chienne* [ʃjɛ̃n] ‘bitch’, may shift the preference in the direction of a serial analysis. Such data, and more realistic and extensive data on schwa drop and schwa insertion, pose an interesting subject of future research. A specific expectation is that the analyses with underlying masculine schwas found in the present limited example will vanish.

## 4.5 The relation to complexity reductions for other parameters than number of levels

In this article, we constructed an algorithm for parallel evaluation across multiple levels of representations whose complexity is linear in the number of levels, while the size of the candidate set is exponential in the number of levels. As mentioned in section 1, this stands in a tradition of reducing exponential candidate sets to linear by graph-theoretic means: linearity has been achieved for the number of segments in the input and the number of autosegmental tiers (Ellison, 1994; Eisner, 1997; Riggle, 2004). There is a difference in the kind of grammar and the kind of candidate generator between our work and this earlier work: when describing the relation between adjacent levels, we worked with small lists of candidates (enumerated in tableaux), just as in Prince and Smolensky (1993)’s formalization and in most practical work in OT, whereas Ellison, Eisner, and Riggle achieved their linearity results under the restriction that candidate sets can be represented as regular expressions, the finite-state assumption. A super-efficient comprehensive model of evaluation in parallel multilevel OT would preferably be subexponential in the number of segments in the input and in the number of autosegmental tiers and in the number of levels of representation. Can this be achieved?

The finite-state models by Riggle (2004) achieve the evaluation of an infinite candidate set. One would like to apply that method to multilevel evaluation. However, finite-state transducers have a single input alphabet and a single output alphabet, whereas our French example works with at least four different alphabets (the underlying form and the surface form may both be written in the same phonological alphabet, but the other three levels are in-

commensurable with the phonological levels and with each other), and once one includes syntactic and semantic representations, the number of alphabets will increase again. One could represent the French case as a concatenation of four finite-state transducers, but such an apparatus would perform only serial multilevel evaluation. For parallel evaluation, one would need a single giant transducer, with meaning as the input and phonetic form as the output; the alphabets of the intermediate levels would then remain unexpressed. Unfortunately, the size of the transducer seems to have to become impractically large. According to (Riggle, 2004, p.100), the number of states in an OT transducer is exponential in the number of constraints if the constraints work on different kinds of structures. In a multilevel case, constraints that work at different levels of representations do not share structures, so that the number of states seems to have to be exponential in the number of levels, which is exactly what the present account wants to avoid; for our 19-constraint case, the states probably number many thousands. Future work by finite-state and/or multilevel theorists may find a solution to this problem.

## **4.6 Conclusion**

In this article, we illustrated that a parallel multilevel constraint grammar can be represented as a graph with a number of connections that is linear in the number of levels, although the number of candidate paths is exponential in the number of levels. We illustrated how this leads to efficient evaluation procedures and learning mechanisms whose computation times are also linear in the number of levels. Although for the time being we have to stay agnostic about whether our French example is best described with a serial or with a parallel grammar, the linear computation time helps to make parallel multilevel evaluation and learning feasible as a method of modeling phonological processing and acquisition. This kind of linearity may well become essential when we scale up to more realistic problems – for example, when we apply parallel multilevel evaluation in whole-language simulations.





## CHAPTER 5

---

# Learning from corpus data in multi-level constraint grammars

---

### *Abstract*

This chapter builds on the efficient evaluation approach of the previous chapter. The size of the learning data set, constraint sets and candidate spaces is increased greatly. A large corpus of spoken French serves as the empirical basis for this scaled-up analysis of liaison in the multi-level BiPhon framework. Several variant models are tested for their ability to correctly model the patterns found in the data. The results of these simulations shed new light on the factors that are at play in French liaison. First, a random baseline learner was unable to find a correct ranking for this enlarged constraint set. Error-driven learners equipped with a serial production grammar are unable to learn the patterns in the data, whereas fully interactive grammars do succeed. Successful learners also show an overwhelming preference for a lexical analysis, using a consonant-initial allomorph of certain nouns to avoid vowel hiatus. More generally, the results demonstrate the viability of the multi-level modeling approach for analysing complex phonological phenomena.

---

### 5.1 Introduction

In this chapter, the multi-level modeling approach to French liaison is expanded. The previous chapter introduced an efficient evaluation algorithm and tested it on a toy language containing three phrases. This chapter takes the logical next step: utilizing these computational gains by enlarging the scale and scope of our multi-level model. This will be done in two ways. First, the learning data set is expanded so that learners are confronted with a va-

riety of forms both with and without liaison. Second, the candidate graphs for each individual learning datum are generated procedurally rather than crafted by hand, in a slightly augmented version of the model used in the previous chapter. These two adaptations result in much larger candidate sets and, consequently, give virtual learners more possible analyses to consider. With this enlarged data set and hypothesis space, we may test and validate our “data-driven” approach on a learning problem that does more justice to the complexity and variation faced by real learners.

### 5.1.1 A multi-level model of liaison

As discussed in Chapter 4, liaison in modern French has been the subject of extensive analysis in various theoretical frameworks. The main focus of Chapter 4 was the comparison between the serial phonological analyses of e.g. Dell (1973) and parallel morphophonological analyses of Encrevé-Lambert (1971) and Tranel (1996). However, it has also long been noted that non-phonological factors play a role: e.g. syntax (Selkirk, 1974; Bonami et al., 2004), morphology proper (Morin, 2003), lexical frequency (Bybee, 2001), orthography (Chevrot and Malderez, 1999; Laks, 2005), and others (e.g. Eychenne, 2011). Moreover, corpus studies have pointed out that there is substantial variation in the realization of liaison (Durand and Lyche, 2003); many contexts considered obligatory in earlier formal analyses turn out to be optional, sometimes outright rare. Extragrammatical (sociolinguistic, regional, stylistic and idiosyncratic) factors also influence the probability of liaison appearing (Ågren, 1973). As Durand and Lyche (2003) put it, “[d]ealing with liaison requires stepping into all the components of the grammar, while tackling at the same time the quicksands of variation.”

From a modeling perspective, the BiPhon framework is a good fit to that description. Indeed, Boersma (2011) states that BiPhon should ultimately do *whole-language simulations* in order to achieve explanatory adequacy. We saw in the previous chapter that the presence of lexical and morphosyntactic representations and constraints enables analyses that explicitly acknowledge non-phonological factors. Additionally, variation and gradient phenomena can be represented through stochastic ranking. The simulations of Chapter 4 showed how a BiPhon MLCG implementation with these properties can efficiently model liaison. However, the specific candidate and constraint sets used in those simulations were created in a somewhat ad-hoc fashion, to represent a small set of serial and parallel analyses. Such an approach is unsustainable on the scale needed to model liaison in its full complexity. Instead, this chapter presents a method for dynamically creating constraint sets and candidate spaces on the basis of a sizable input data set, edging closer to the ultimate goal of whole-language simulation.

### 5.1.2 Using data from the PFC corpus

The *Phonologie du français contemporain* project (PFC; Durand and Lyche, 2003; Durand et al., 2009; Detey et al., 2016) was created to establish a large, reliable and accessible corpus for studying the sound patterns of modern spoken French. The PFC corpus contains annotated recordings of speakers from a variety of locations in France and other Francophone countries. While a large part of the corpus consists of free-flowing unguided dialogue between researcher and informant, the data collection and annotation particularly focus on three phenomena of special phonological interest: the phonemic inventory, usage of schwa, and (non-)realization of liaison.

This focus on liaison is reflected in the data structure of the corpus. Through a web interface<sup>1</sup>, or alternatively by downloading the data together with a software tool, one may specifically search for utterances that contain potential contexts for liaison, and extract transcripts and recordings of these utterances. This study leverages these transcripts and accompanying orthographic transcriptions to generate the *meaning–sound pairs* that serve as input to our error-driven learning model. In this way, we can use the PFC data to scale up Chapter 4’s investigation of liaison and hopefully gain more insight in the factors that influence its manifestation.

### 5.1.3 Aims and limitations of this study

The approach and aim of this work resemble those of Chapters 3 and 4. We start by describing and computationally implementing a meta-model, which subsumes several models that generate distinct hypotheses about linguistic representations and processes. These models are then trained on data reflecting the overt evidence available to language learners. The results of training will indicate which (if any) of the models are capable of learning, representing and reproducing the patterns found in the empirical data. Moreover, after training we can test whether any solution found by the virtual learners generalizes to unseen data or a holdout test set. As with the serial and parallel L2 learners of Chapter 3, this can help to make a distinction between models that at first glance seem to exhibit the same surface behaviour on the training data, in spite of architectural differences. Finally, the state of the constraint grammar can be inspected during and after learning. This provides insight in the possible pathways to learning surface-correct grammars, and how different approaches to error-driven constraint learning may govern the success or failure to learn certain patterns in the data.

Our computational approach thus provides us with a tool to compare competing hypotheses of speech perception and production. By basing the input on lab-collected or spontaneously recorded corpus data, we ground the outcome of this comparison in empirical linguistic behaviour, including variation and frequency effects. This study is therefore partly *methodological* in nature: it

---

<sup>1</sup><https://www.projet-pfc.net/>

aims to showcase the viability of this data-driven model selection approach, on a large data set representing a well-studied and intricate phenomenon. This is not to say that the meta-model covers the gamut of analyses that have been proposed in the prodigious literature on liaison. Although we are relatively agnostic on the locus of liaison in the grammar, representational choices still unquestionably shape and restrict the hypothesis space available to our virtual learners. The learning data will likewise be restricted to a subclass of noun phrases manifesting liaison, limiting the explanatory power of our models for liaison as a whole. We return to these questions in the discussion (Section 5.5).

#### 5.1.4 Outline

The rest of this chapter is structured as follows. The next section motivates expanding the model from that of the previous chapter, and informally describes the procedures for candidate and constraint generation. Architectural choices and restrictions mentioned above are also addressed. Section 5.3 follows, describing a series of simulations with the expanded model, starting with a handful of training items resembling the toy liaison dataset of the previous chapter. This dataset is then augmented with more input data and surface variation, allowing us to address some speculative comments made in the previous chapter. Finally, Section 5.4 describes a series of simulations on a much larger data set culled from the PFC. The implications of the simulation results are discussed in Section 5.5.

## 5.2 Model and data

### 5.2.1 Reducing teleological bias

A majority of OT analyses in the literature mention only a fraction of the total constraint and candidate set presumed to be active in an evaluation. This is for good reason: the authors analyse a particular phenomenon in some particular language(s), and take care to construct concise examples to illustrate the merits of their approach. However, the ambition of handling a larger and more varied set of liaison data in our model requires an alternative approach to candidate and constraint creation. Not only would it be laborious and error-prone to manually design candidate graphs for thousands of inputs, but as Bane and Riggle (2012) point out, there is considerable risk of accidentally omitting a constraint or candidate which breaks a given analysis. Perhaps more critically, knowing the desired outcome(s) is likely to introduce a teleological bias in our view of the learning problem. The toy liaison grammars of the previous chapter are no exception. The “candidate graph” representation allowed compact visualization of multi-level evaluation over a reasonably large candidate set (128 candidates per input). Nonetheless, the possible interlevel mappings and constraints were designed by hand, and many thinkable sub-candidates

<i>ours</i>	/ .uʁs. /	<i>loup</i>	/ .lu. /
<i>un ours</i>	/ .œ̃.nuʁs. /	<i>un loup</i>	/ .œ̃.lu. /
<i>les ours</i>	/ .le.zuʁs. /	<i>les loups</i>	/ .le.lu. /
<i>des ours</i>	/ .de.zuʁs. /	<i>des loups</i>	/ .de.lu. /

Table 5.1: Syllabification of the nouns *ours* and *loups* when combined with various determiners

were precluded from the analysis or presented as forbidden by inviolable constraints. The candidate graphs encompassed several distinct analyses that have been proposed in the liaison literature, and contained just enough forms and constraints to represent these. Relative success of EDRA's to find these covert structures served as a metric for *learnability* of one analysis over another, given the same overt and ambiguous input data.

Presumably, real learners acquiring the phonology of French face even greater ambiguity. For one, they are not born into the world with knowledge of boundary representations for specific French words and morphemes. Rather, they infer these from the linguistic input; at the age of 8.5 months infants already show sensitivity to word boundaries (Jusczyk and Aslin, 1995; Pelucchi et al., 2009), and in the second year of life are able to distinguish between similar-sounding words (Werker et al., 2002). Despite this remarkable feat of unsupervised learning, children (and indeed adults) do sometimes continue to entertain spurious word or morpheme boundaries. In French, cues for word boundaries can be obscured by two related sandhi phenomena: liaison and *enchainement*. The latter refers to the syllabification of historically word-final liaison consonants to the onset of the following word. The forms in Table 5.1 illustrate.

The underlying forms of *ours* and *loup* can be presumed to be |ʁs| and |lu|, respectively. This can also be seen when they occur in isolation. The articles *les* and *des* have liaising forms ending in /z/, which roughly speaking only surface when the following noun begins with a vowel. When preceding a consonant-initial noun such as *loup*, the /z/ does not surface and |lez+lu| becomes / .le.lu. / On the basis of these and similar data, a learner assuming that word boundaries coincide with syllable boundaries might well be led to believe that the plural definitive article takes the form |le|, and that /zuʁs/ is an allomorph or prefixed plural form of |ʁs|. Surface / .le.lu. / and / .le.zuʁs. / would be restructured as underlying |le+lu| and |le+zuʁs|. Such restructuring becomes even more plausible with words that are chiefly encountered in the plural, accompanied by indefinite *des* or definite *les*, and rarely in the singular or in isolation. Indeed, the speech of young learners of French sometimes betrays such restructuring. Chevrot et al. (2009) cite the example of a child re-analyzing *les arbres* as *les zarbres* and *un ours* as *un nous*. Several French-based creoles contain vocabulary items originating from reanalyzed word boundaries, e.g. *zanimu* "animal" and *zistwar* "story" in Mauritian Creole (Grant

and Guillemin, 2012; Bonami and Henri, 2012).<sup>2</sup> Beginning with Gougenheim (1935), some authors consider this joining of liaison consonants to a following word to be active in L1 adult grammars. In particular, a plural prefix |z| has been proposed to appear before vowel-initial nouns (Morin and Kaye, 1982).

If our learning models are to provide insight in the acquisition patterns of liaison by real French learners, we should not bias the virtual learners toward the historically or orthographically correct form of the words in the input data – these forms are after all not available to learners at the time they are acquiring the language. Instead, the approach we take is to generate various (sub)candidates and constraints over these hidden structures on the basis of *semantic* and *phonemic* features overtly present in the input, through a computational implementation of GEN and CON.

Algorithmically generating a finite number of candidates is the same approach that was taken in the more limited simulation world of Chapter 3, and many other studies of learning in constraint grammars, e.g. Tesar and Smolensky (1998) and Jarosz (2013a). The resulting candidate space still derives from features present in the overt learning data, whose representations are in turn inspired by the specific phenomenon we want to analyze. However, by generating rather than designing the hypothesis space that our MLCG learners will explore, the grammars should contain less bias toward known solutions to the problem.

### 5.2.2 Formalizing levels of representation and forms

To get a good understanding of the representations and structures involved in our multi-level liaison model, we may re-examine a candidate quintuplet from Chapter 4's toy grammar:

“good<sub>i</sub>, actor<sub>i</sub>” –  $\langle \text{bon-M}; \text{acteur}_M \rangle$  – |bɔn+∅#aktœʁ| – /.bɔ.nak.tœʁ/ – [bɔnak.tœʁ]

This notation concisely expresses properties from various components of the grammar which supposedly play a role in liaison: syntactic agreement, inflection, lexical access, syllabic structure, phonemic and phonetic content, and so on. Many of these properties were mentioned in passing or only implicitly in Chapter 4. The linear notation also obscures the essentially hierarchical structure within some forms. Moreover, it is clear that there is a correspondence between adjacent forms within the candidate quintet, but the nature of this correspondence was not made explicit. To implement procedural generation of such multi-level candidates, as well as many others, it is necessary to be more precise about the features and structures that make up the forms on each level of representation, and the way features correspond on adjacent levels. This holds true especially for semantic, syntactic and morphological features, which traditionally have not been the focus of the BiPhon framework.

<sup>2</sup>Another interesting example comes from the *Verlan* argot, which creates words by inverting their syllables: *rabza* “Arab”, apparently inversed from *\*zarabe*.

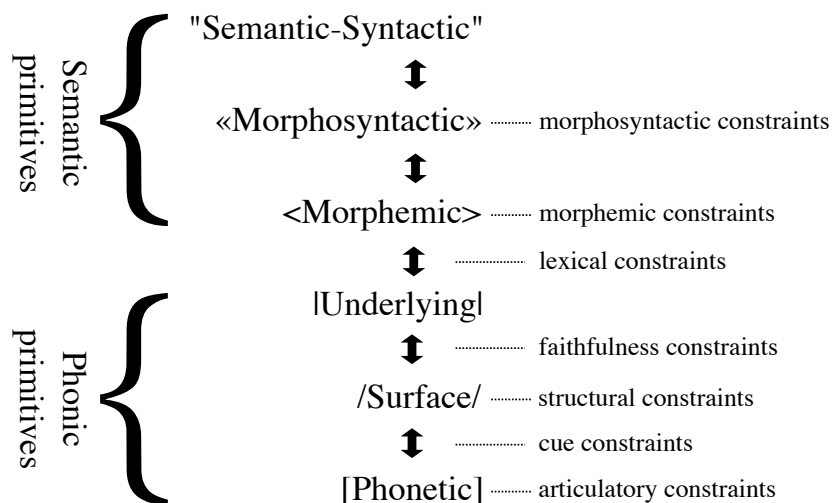


Figure 5.1: The extended BiPhon model used in this chapter

Figure 5.1 shows an overview of the extended BiPhon model that will be employed in this chapter. The most salient and important addition is an extra level of representation. To make the nature of the mappings, features and constraint interactions of a morphological nature more explicit, the Morpheme level of the previous chapter has been split into two levels, *Morphosyntactic* and *Morphemic*. While this split is not strictly necessary to represent the liaison grammars used here, it has an explanatory benefit in distinguishing between different types of processes. The next sections will walk through the extended model, using forms familiar from the previous chapter to illustrate the procedure by which GEN composes candidates and CON differentiates between them.

### 5.2.3 A dynamic and local GEN

Chapters 2 and 4 stated a general restriction to the MLCG framework utilized in this thesis: constraints must evaluate either *forms* on a single level, or *form mappings* between adjacent levels. With this restriction, candidate evaluation over multiple levels can be defined as the 'sum' of multiple independent and local sub-evaluations. This dynamic programming approach can also be applied to candidate *generation*: for a grammar with  $n$  levels of representation, we decompose GEN into a series of functions  $GEN_1 \dots GEN_{n-1}$ . The output of one such sub-GEN functions as input on its successor. This approach to candidate generation is also used in serial constraint grammar frameworks such as Stratal OT (Bermúdez-Otero, 1999; Kiparsky, 2000), which is however restricted to phonology proper – that is, purely concerned with levels that are



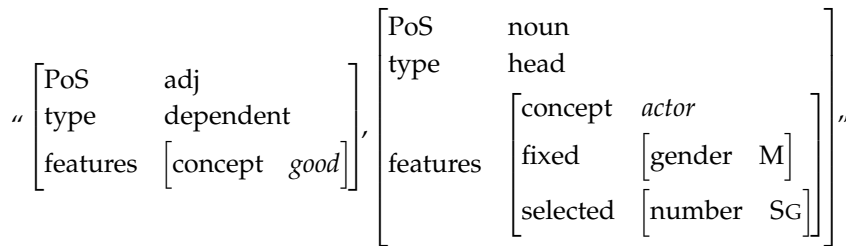


Figure 5.2: An attribute-value matrix representation of a Semantic-Syntactic Form: an ordered pair of S-words. The set of features is typed for concept, fixed, selected or contextual.

considered “hidden” in BiPhon. Other approaches such as Harmonic Serialism (McCarthy, 2000) and Simulated Annealing for Optimality Theory (Biró, 2006) also take a dynamic approach to candidate generation, but apply the same GEN repeatedly rather than iterating over different sub-GENs.

In this chapter, rather than using numeric level indices in our notation, we use the abbreviated level of the input forms. For example,  $\text{GEN}_{\text{UF}}$  stands for the sub-GEN that takes Underlying Forms as input.

## 5.2.4 Level-by-level walkthrough of the extended model

### Syntactic-semantic form

The topmost level in Figure 5.1 is Semantic-Syntactic Form (SSF). Like the Semantic Form of the previous chapter, it formalizes the “semantic” or “contextual” information presented to learners. One difference with “traditional” BiPhon SemF is that SSF contains not only semantic but also rudimentary syntactic information, about word order and syntactic category. On the basis of SSF, learners go on to infer possible morphosyntactic representations. SSF forms are structured hierarchically: at the highest level they are an ordered collection of *syntactic words* (**S-words**), complex structures that encode semantic and syntactic information about dependency and optional morphosyntactic features. In linguistic formalisms, complex feature structures of this type are often represented using *attribute-value matrices* (AVMs): see (Sag and Polard (1987) ) As an example, the Semantic Form “ $\textit{good}_i \textit{actor}_i$ ” of the previous chapter might be represented as in Figure 5.2, a pair of S-words in SSF in an AVM:

An S-word, then, is composed of a syntactic category (PoS), a binary headedness value also derived from syntax (either head or dependent) and a set of typed features, where a feature is understood to mean a combination of an *attribute* and a *value* for that attribute. Of these features, the concept type is obligatory: it encodes the *meaning* of the S-word, and may theoretically take

on any value that language users are able to represent conceptually.<sup>3</sup>

The other features might be called **morphosyntactic** features. Structurally, they are attribute-value pairs, where both attribute and value come from a small closed set. Morphosemantic features come in three types<sup>4</sup>: **Fixed** or lexical features are inherently determined by the expressed meaning. In French, we may consider the grammatical gender of nouns to be fixed: any noun belongs to either the class of masculine or feminine nouns, irrespective of the semantic or grammatical context in which it is used.<sup>5</sup> **Selected** features are those that are dictated by semantic context, i.e. the meaning conveyed by a phrase. The grammatical number of *acteur* in the noun phrase *bon acteur* is singular, since the phrase refers to a single actor. **Contextual** features are those that are assigned by grammatical context: through agreement or government. Contextual features are actually not yet present on the SSF level: they are imposed onto non-head S-words on the next level, Morphosyntactic Form.

The treatment of headedness as a binary feature is deliberately simplistic and theory-neutral. Nichols (1986) calls headedness “a theory-independent notion which in fact figures as a primitive in all theories”. Since this chapter is confined to small nominal phrases in French, a simple marking of S-words as head or dependent allows us to generate candidates and constraints related to agreement, as the next section will explain.

### Mapping from Semantic-Syntactic to Morphosyntactic Form

In mapping from Semantic-Syntactic Form to Morphosyntactic Form (MSF), the morphosyntactic features of SSF may spread out from the head S-word to the dependent S-word(s), a process of *agreement*. While feature *attributes* are always copied from head to dependent in MSF, the dependents’ features may take on other *values* than those of the head S-word, including a *null value*. In this way, agreement as well as disagreement and non-expression of gender and number on the determiner may all be generated as hypotheses – recall from the previous chapter that a possible analysis of liaison requires that a learner select an incongruent feminine form of the adjective with some masculine nouns, in order to prevent vowel hiatus.

The figure illustrates one possible mapping from “Actor.M, SG” to an MSF form. Copied features are printed bold. In this particular mapping, the number feature is faithfully copied from the head S-word, but the value of the gender feature is changed to F, resulting in non-agreement.

<sup>3</sup>In practice, because the concept feature is given in the input and cannot change in subsequent levels of representation, the number of values for this feature is finite, limited by the total set of inputs available to a learner in the simulation.

<sup>4</sup>This taxonomy of features derives from Kibort (2007).

<sup>5</sup>The presence of a fixed gender feature on SSF violates the premise that it only contains features that can be determined by context. This choice was made deliberately in order to be able to simulate the role of gender in liaison, without dealing with the extra complication of having to learn a gender feature.

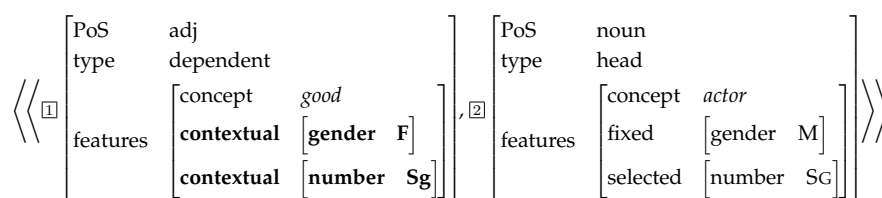


Figure 5.3: An attribute-value matrix representation of a Morphosyntactic Form. An example of typed features is seen in the second S-word: the feature of type *fixed* has the attribute *gender* and the value *M*.

The  $\text{GEN}_{\text{SSF}}$  function thus generates a set of possible output SSF–MSF mappings for a given SSF. Each feature whose attribute is copied from a head to a dependent S-word may assume one from a closed set of possible feature values. The Cartesian product of these sets yields the set of possible feature combinations that may be added to the dependent S-words of the MSF; each combination corresponds to a possible SSF–MSF mapping.

The tableau in Figure 5.4 lists all possible mappings from SSF to MSF for the example form of the previous section. It also illustrates the two types of constraints active in this mapping: EXPRESS constraints and AGREE constraints. Both are *intralevel* constraints militating against certain features or feature combinations on the Morphosyntactic Forms.

AGREE constraints operate on the level of the entire phrase. They punish disagreement between a morphosyntactic feature on a dependent S-word and the corresponding feature on its head S-word. They are specified for part of speech and morphosyntactic attribute. For example, AGREE ADJ-G inflicts a violation for an adjective whose value for the gender feature is not in agreement with that of the head S-word.

EXPRESS constraints are similarly specified for part of speech and morphosyntactic attribute, but inflict a violation for a dependent S-word not *expressing* that attribute, i.e. when the value for that feature is null.

In the context of the problem researched in the simulations, these two constraint types serve to shape learners’ hypotheses on the role of “gender allomorphy” in liaison. For instance, AGREE-ADJ-G must be ranked low in order to allow for the hypothesis that *bon acteur* employs a syntactically feminine allomorph of ⟨Good⟩. MSF forms with null feature values, forbidden by EXPRESS constraints, allow for the hypothesis that some parts of speech are not marked at all for a given morphosyntactic feature.

### Mapping from Morphosyntactic Form to Morphemic Form

In this mapping, the list of S-words from MSF is transformed to a list of **M-words** on Morphemic Form (MF), retaining the word order. An M-word consists of a PoS feature inherited from the S-word, and an *ordered* collection of

	<i>Express-Adj-G</i>	<i>Express-Adj-Num</i>	<i>Agree-Adj-G</i>	<i>Agree-Adj-Num</i>
“{Good} <sub>Adj</sub> – {Actor.M, SG} <sub>N</sub> ”				
☞ <<{Good, M, SG} <sub>Adj</sub> – {Actor.M, SG} <sub>N</sub> >>				
<<{Good, M, PL} <sub>Adj</sub> – {Actor.M, SG} <sub>N</sub> >>				*!
<<{Good, M, ∅} <sub>Adj</sub> – {Actor.M, SG} <sub>N</sub> >>	*!			
<<{Good, F, SG} <sub>Adj</sub> – {Actor.M, SG} <sub>N</sub> >>			*!	
<<{Good, F, PL} <sub>Adj</sub> – {Actor.M, SG} <sub>N</sub> >>			*!	*
<<{Good, F, ∅} <sub>Adj</sub> – {Actor.M, SG} <sub>N</sub> >>	*!		*	
<<{Good, ∅, SG} <sub>Adj</sub> – {Actor.M, SG} <sub>N</sub> >>	*!			
<<{Good, ∅, PL} <sub>Adj</sub> – {Actor.M, SG} <sub>N</sub> >>	*!			*
<<{Good, ∅, ∅} <sub>Adj</sub> – {Actor.M, SG} <sub>N</sub> >>	*!	*		

Figure 5.4: Tableau illustrating a mapping from SSF to MSF. Note how features may spread from the head noun in SSF to the dependent adjective in MSF, violating constraints when the feature attribute is null (EXPRESS) or non-congruent (AGREE).

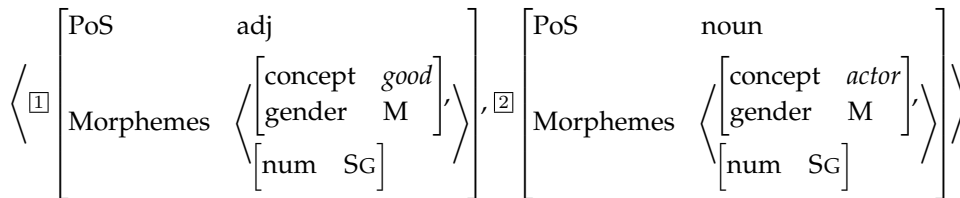


Figure 5.5: An attribute-value matrix representation of a Morphemic Form.

unordered feature sets. These feature sets are named Morphemes.

The output of GEN<sub>MSF</sub> is defined by a procedure that maps an input MSF to a set of MFs that retain the ordering of S-words from the input, lose the headedness feature, and introduce a partial order on the set of morphosyntactic features contained in each S-word. This procedure takes three steps:

1. Each possible *partition* (set of non-empty subsets) for the feature set of each MSF S-word is generated. A partition represents a possible subdivision of the S-word’s non-null features into one or more Morphemes, where each feature must be a member of exactly one Morpheme; an additional restriction is that fixed features must remain attached to their *concept* feature.
2. For each partition into Morphemes, all possible *orderings* of these Mor-

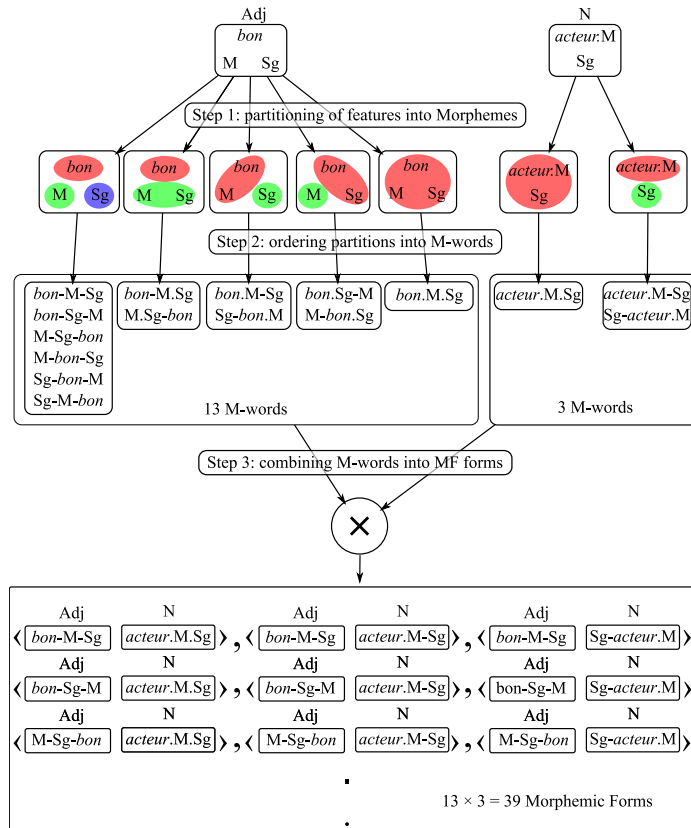


Figure 5.6: A schematic example of a mapping from Morphosyntactic Form to Morphemic Form.

phemes are generated. This ordering, combined with the part of speech of the parent S-word, is called an M-word. (The total number of M-words created from a given partition is the factorial of the partition's size.)

3. The list of S-words from the MSF has now been turned into a list of sets of M-words. We take the Cartesian product of this list of sets – that is, all possible combinations that can be created by combining a member from each set in this list of sets. This operation yields the output set of Morphemic Forms.

Figure 5.6 illustrates this algorithm for a specific MSF. It will be clear from this description that the number of MFs generated grows more than exponentially in the number of features present in the MSF.

Two types of intralevel constraints operate on the mappings generated by  $\text{GEN}_{\text{MSF}}$ . The first is the family of ANALYZE constraints. These are specified

for part of speech and morphological feature, and militate against synthesizing that feature with another. Thus, ANALYZE-ADJ-NUM inflicts a violation on the Morpheme GOOD.SG because its number feature is coalesced with the concept feature *good*. The second type is the family of ORDER constraints, concerned with the ordering of Morphemes within the M-word. It too is specified for part of speech and morphological feature. It inflicts a violation when the relative ordering of two Morphemes in an M-Word differs from that specified in the constraint. For instance, \*A(NUM <G) inflicts a violation on an Adjective M-Word when it has a Morph with a Number attribute ordered before a Morph with a Gender attribute. The tableau below illustrates how these different constraint types may interact in the MSF → MF mapping. It is worth noting that no two subcandidates have the same violation profile. As in the previous mapping, this is a result of the close relation between GEN and CON in our dynamically generated grammar.

In the context of our liaison model, an MF representation contained in the candidate for a given meaning–form pair represents a hypothesis on the coalescence of morphosyntactic features into single morphemic elements, as well as the placement of inflectional elements relative to a stem. By treating inflectional elements as possibly separate morphemes, we are able to represent hypotheses about their interaction with liaison, as also done by e.g. Bonami et al. (2004). At the same time, the model allows suppletive analyses through the coalescence of morphosyntactic features with stems, yielding e.g. *good.M* as a single morphemic unit which has two allomorphs |bõ| and |bɔn|.

### Morphemic Form to Underlying Form

Through this mapping, the abstract Morphemes forming the minimal parts of words are mapped to a string of *Morphs* making up an Underlying Form (UF). A Morph is in turn a string of segments representing the phonemic content of a Morpheme. In other words, the MF–UF relation concerns the *lexicon*: each mapping from MF to UF contains hypotheses about the segmental representations associated with the Morphemes contained in an MF.

Unlike the previous two sub-GENs, GEN<sub>MF</sub> cannot be defined strictly in terms of elements present on the input level; the phonemic segments on UF are written in a different “alphabet”. Instead, the mapping from MF to UF is determined by a learner’s **lexicon**, which contains the set of allomorphic hypotheses for a Morpheme and so bounds the set of subcandidates. We return to this lexicon in Section 5.2.5, after this level-by-level walkthrough of the grammar.

### Constraints

The only constraint type active in this mapping is that of interlevel LEX constraints (Apoussidou 2007; see also Chapter 3). An individual LEX constraint forbids a particular mapping from a single Morpheme to a single Morph.

«A{bon M SG} N{acteur.M SG} »	Analyze A-CONCEPT	Analyze A-num	Analyze A-g	Analyze N-num	Analyze N-g	Analyze N-CONCEPT	*A(g < num)	*A(g < CONCEPT)	*A(num < g)	*A(num < CONCEPT)	*A(CONCEPT < g)	*A(CONCEPT < num)	*N(g < num)	*N(num < g)	*N(num < CONCEPT)	*N(CONCEPT < num)
⟨bon <sub>A</sub> + SG <sub>A</sub> + M <sub>A</sub> ; acteur.M <sub>N</sub> + SG <sub>N</sub> ⟩					*	*			*		*	*	*!			*
⟨bon <sub>A</sub> + SG <sub>A</sub> + M <sub>A</sub> ; acteur.M.SG <sub>N</sub> ⟩			*!	*	*			*		*	*					
⟨bon <sub>A</sub> + SG <sub>A</sub> + M <sub>A</sub> ; SG <sub>N</sub> + acteur.M <sub>N</sub> ⟩				*	*			*		*	*		*	*		
⟨bon <sub>A</sub> + M.SG <sub>A</sub> ; acteur.M <sub>N</sub> + SG <sub>N</sub> ⟩	*!	*		*	*					*	*	*				*
⟨bon <sub>A</sub> + M.SG <sub>A</sub> ; acteur.M.SG <sub>N</sub> ⟩	*!	*	*	*	*					*	*					
⟨bon <sub>A</sub> + M.SG <sub>A</sub> ; SG <sub>N</sub> + acteur.M <sub>N</sub> ⟩	*!	*		*	*					*	*		*	*		
⟨M <sub>A</sub> + bon.SG <sub>A</sub> ; acteur.M <sub>N</sub> + SG <sub>N</sub> ⟩	*!	*		*	*	*	*						*			*
⟨M <sub>A</sub> + bon.SG <sub>A</sub> ; acteur.M.SG <sub>N</sub> ⟩	*!	*		*	*	*	*									
⟨M <sub>A</sub> + bon.SG <sub>A</sub> ; SG <sub>N</sub> + acteur.M <sub>N</sub> ⟩	*!	*		*	*	*	*						*	*		
⟨M.SG <sub>A</sub> + bon <sub>A</sub> ; acteur.M <sub>N</sub> + SG <sub>N</sub> ⟩	*!	*		*	*	*	*		*				*			*
⟨M.SG <sub>A</sub> + bon <sub>A</sub> ; acteur.M.SG <sub>N</sub> ⟩	*!	*	*	*	*	*	*		*							
⟨M.SG <sub>A</sub> + bon <sub>A</sub> ; SG <sub>N</sub> + acteur.M <sub>N</sub> ⟩	*!	*		*	*	*	*		*				*	*		
⟨SG <sub>A</sub> + M <sub>A</sub> + bon <sub>A</sub> ; acteur.M <sub>N</sub> + SG <sub>N</sub> ⟩				*	*		*!	*	*				*			*
⟨SG <sub>A</sub> + M <sub>A</sub> + bon <sub>A</sub> ; acteur.M.SG <sub>N</sub> ⟩			*!	*	*	*	*	*								
⟨SG <sub>A</sub> + M <sub>A</sub> + bon <sub>A</sub> ; SG <sub>N</sub> + acteur.M <sub>N</sub> ⟩				*	*		*!	*	*				*	*		
⟨SG <sub>A</sub> + bon <sub>A</sub> + M <sub>A</sub> ; acteur.M <sub>N</sub> + SG <sub>N</sub> ⟩				*	*			*!	*	*			*			*
⟨SG <sub>A</sub> + bon <sub>A</sub> + M <sub>A</sub> ; acteur.M.SG <sub>N</sub> ⟩			*!	*	*	*	*	*	*							
⟨SG <sub>A</sub> + bon <sub>A</sub> + M <sub>A</sub> ; SG <sub>N</sub> + acteur.M <sub>N</sub> ⟩				*	*			*!	*	*			*	*		
⟨M <sub>A</sub> + bon <sub>A</sub> + SG <sub>A</sub> ; acteur.M <sub>N</sub> + SG <sub>N</sub> ⟩				*	*	*!	*				*	*				*
⟨M <sub>A</sub> + bon <sub>A</sub> + SG <sub>A</sub> ; acteur.M.SG <sub>N</sub> ⟩			*!	*	*	*	*				*					
⟨M <sub>A</sub> + bon <sub>A</sub> + SG <sub>A</sub> ; SG <sub>N</sub> + acteur.M <sub>N</sub> ⟩				*	*	*!	*				*	*	*	*		
⟨bon.M.SG <sub>A</sub> ; acteur.M <sub>N</sub> + SG <sub>N</sub> ⟩	*!	*	*	*	*								*			*
⟨bon.M.SG <sub>A</sub> ; acteur.M.SG <sub>N</sub> ⟩	*!	*	*	*	*	*										
⟨bon.M.SG <sub>A</sub> ; SG <sub>N</sub> + acteur.M <sub>N</sub> ⟩	*!	*	*	*	*								*	*		
⟨bon <sub>A</sub> + M <sub>A</sub> + SG <sub>A</sub> ; acteur.M <sub>N</sub> + SG <sub>N</sub> ⟩				*	*	*!				*	*	*				*
⟨bon <sub>A</sub> + M <sub>A</sub> + SG <sub>A</sub> ; acteur.M.SG <sub>N</sub> ⟩			*!	*	*	*				*	*					
⟨bon <sub>A</sub> + M <sub>A</sub> + SG <sub>A</sub> ; SG <sub>N</sub> + acteur.M <sub>N</sub> ⟩				*	*	*!				*	*		*	*		
⟨SG <sub>A</sub> + bon.M <sub>A</sub> ; acteur.M <sub>N</sub> + SG <sub>N</sub> ⟩	*!	*		*	*			*	*				*			*
⟨SG <sub>A</sub> + bon.M <sub>A</sub> ; acteur.M.SG <sub>N</sub> ⟩	*!	*	*	*	*	*		*	*							
⟨SG <sub>A</sub> + bon.M <sub>A</sub> ; SG <sub>N</sub> + acteur.M <sub>N</sub> ⟩	*!	*		*	*	*		*	*				*	*		
⟨bon.SG <sub>A</sub> + M <sub>A</sub> ; acteur.M <sub>N</sub> + SG <sub>N</sub> ⟩	*!	*		*	*	*		*	*	*			*			*
⟨bon.SG <sub>A</sub> + M <sub>A</sub> ; acteur.M.SG <sub>N</sub> ⟩	*!	*		*	*	*		*	*							
⟨bon.SG <sub>A</sub> + M <sub>A</sub> ; SG <sub>N</sub> + acteur.M <sub>N</sub> ⟩	*!	*		*	*	*		*	*		*	*	*	*		
⟨bon.M <sub>A</sub> + SG <sub>A</sub> ; acteur.M <sub>N</sub> + SG <sub>N</sub> ⟩	*!	*		*	*	*		*	*		*	*	*	*		*
⟨bon.M <sub>A</sub> + SG <sub>A</sub> ; acteur.M.SG <sub>N</sub> ⟩	*!	*	*	*	*	*		*	*		*	*	*	*		
⟨bon.M <sub>A</sub> + SG <sub>A</sub> ; SG <sub>N</sub> + acteur.M <sub>N</sub> ⟩	*!	*		*	*	*		*	*		*	*	*	*		*
⟨M <sub>A</sub> + SG <sub>A</sub> + bon <sub>A</sub> ; acteur.M <sub>N</sub> + SG <sub>N</sub> ⟩				*	*	*!	*		*	*		*	*	*		*
⟨M <sub>A</sub> + SG <sub>A</sub> + bon <sub>A</sub> ; acteur.M.SG <sub>N</sub> ⟩			*!	*	*	*	*		*	*		*	*	*		
⟨M <sub>A</sub> + SG <sub>A</sub> + bon <sub>A</sub> ; SG <sub>N</sub> + acteur.M <sub>N</sub> ⟩				*	*	*!	*		*	*		*	*	*	*	*

Figure 5.7: A tableau illustrating a mapping from Morphosyntactic Form to Morphemic Form.


Good.M SG, Actor.M SG	*GOOD.M b̄on	*GOOD.M b̄ō	*ACTOR.M aktœr	*Nom.SG ∅	*Nom.SG z	*Adj.SG ∅	*Adj.SG z
b̄on+∅#aktœr+∅	*!	*	*	*	*	*	*
b̄on+∅#aktœr+z	*!	*	*	*	*	*	*
b̄on+z#aktœr+∅	*!	*	*	*	*	*	*
b̄on+z#aktœr+z	*!	*	*	*	*	*	*
b̄ō+∅#aktœr+∅	*	*	*!	*	*	*	*
b̄ō+∅#aktœr+z	*	*	*!	*	*	*	*
b̄ō+z#aktœr+∅	*	*	*	*	*!	*	*
 b̄ō+z#aktœr+z	*	*	*	*	*	*	*

Figure 5.8: A tableau illustrating a mapping from Morphemic Form to Underlying Form.

Morphs may simply be the phonologically empty *null affix*  $|\emptyset|$ . Thus, the constraint \*ACTOR.M - |aktœr| forbids mapping the Morpheme ACTOR.M to the Morph |aktœr| and inflicts a single violation mark on every MF–UF mapping realizing this lexical relation. Ranking this constraint lower will increase the likelihood that this Morph is retrieved for this Morpheme. The learner’s lexicon determines the size of the LEX constraint family, since each allomorphic hypothesis is represented by a lexical constraint forbidding it. Recall from the previous chapter that, in a parallel analysis, phonological considerations (e.g. hiatus-forbidding structural constraints) may interact with LEX constraints, giving rise to phonologically conditioned allomorphy.

The sub-GEN active on the MF–UF mapping works quite simply. For each Morpheme that is present on the MF level, the lexicon retrieves a set of Morphs (allomorphs). A list of Morphemes in a particular MF thus yields an ordered list of such sets, and the set of subcandidates generated for an MF is the Cartesian product of this list of sets. CON contains a unique LEX constraint for each Morpheme–Morph submapping, so that again each MF–UF mapping has a unique violation profile. Figure 5.8 gives an example, which is exhaustive under the assumption that the lexicon contains one possible allomorph for *actor.M* and two possible allomorphs for each of other three Morphemes in the input.



### Underlying Form to Surface Form

This mapping is probably the most familiar to readers: it works analogously to that of the previous chapter, and is generally closest to the two-level, purely phonological processes often analysed with “vanilla” two-level OT (see Chapter 2). On the Surface Form (SF) level, the Morphs present on the Underlying Form level are spelled out to strings of individual phonemes. These concatenated phoneme strings are then syllabified. Interlevel *faithfulness* and intralevel *structural* constraints govern the outcome of this mapping. As in the previous chapter, we model sound changes on the *segmental* rather than the featural level. While this may fail to capture some generalisations (e.g. that the submappings  $|\text{ɔn}| \rightarrow /ɔ̃/$  and  $|\text{ɛn}| \rightarrow /ɛ̃/$  both involve the same transformation), it seems justifiable in this study which is primarily concerned with the appearance or disappearance of complete segments. One novel addition compared to the previous chapter is that a single UF may map to multiple segmentally identical *syllabifications*, rather than using the “Max Onset Principle” (Kahn, 1976).

The faithfulness constraints take the form  $*|\alpha| \rightarrow /\beta/$  where a violation mark is inflicted for every instance where the UF segment(s)  $|\alpha|$  are realized as the SF segments(s)  $/\beta/$ . Insertion and deletion are represented as  $|\emptyset| \rightarrow / \beta/$  and  $|\alpha| \rightarrow / \emptyset/$ , respectively. Structural constraints forbid entire codas or onsets, inflicting a violation for each syllable on SF that contains such a coda or onset, including empty ones:  $* / .V/$  is equivalent to the constraint traditionally named ONSET. The constraint set in itself thus does not contain a preference for the absence or presence of consonants in onset or coda.

As in Chapter 4, and unlike Chapter 3, there are no constraints militating against a fully faithful mapping between UF and SF. There is an implicit correspondence relation between the two levels, and only when a segment is not mapped to an identical segment on the other level will a faithfulness violation be inflicted. Otherwise, the segments on UF are mapped left-to-right to identical correspondents on SF.

$\text{GEN}_{\text{UF}}$  creates Surface Forms through the following (informally defined) procedure:

1. The Morphs of UF are “spelled out” to an intermediate, linearized segmental representation that retains markers for morpheme and word boundaries.
2. The boundary markers in these intermediate representations serve as anchor points for context-sensitive transformational rules. The set of all possible combinations of rule applications is generated, with the stipulation that each boundary marker may be the locus of at most one transformation.
3. For each member of this set, a set of syllabifications is generated, with the rule that each syllable must contain precisely one vowel.

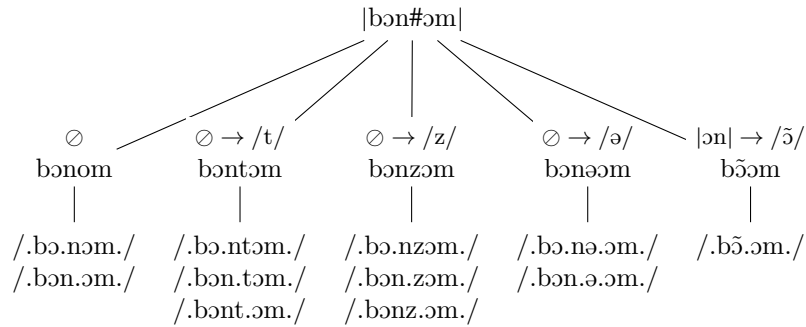


Figure 5.9: An illustration of the mapping from UF to SF. First a phonological rule is applied at a word edge; for the resulting string of segments, all possible syllabifications are created.

The rewrite rules that are applied in the second step come from a finite set of rules, including a “zero rule” that does not rewrite any segments. As with the lexical repository of  $GEN_{MF}$ , the provenance of this rule repository is discussed in the next section. Figure 5.9 illustrates how the procedure may create 11 distinct SFs from the UF  $|bɔn#ɔm|$ .

### Surface Form to Phonetic Form

$GEN_{SF}$  constitutes the final mapping in the production grammar, from a syllabified Surface Form to a flat string of segments on the Phonetic Form (PF) level. As in the previous chapter, the SF–PF mapping is rather similar to the UF–SF mapping: a PF is here represented as a string of phones, rather than the more fine-grained phonetic-articulatory representations of Chapter 3. As on the previous level, mappings from SF to PF are generated on the basis of rewrite rules, operating on edges between segments. In this case, they operate at syllable edges. In the simulations described below, the rules mostly concern insertion or deletion of schwa segments. Interlevel CUE constraints take on the same form as the faithfulness constraints of the previous mapping; intralevel ART constraints punish certain segments or combinations thereof. The role of cue and articulatory constraints involving schwa are crucial in our multi-level version of Dell’s “abstract serial” analysis of liaison (see also Chapter 4).

The constraints and candidate sets shown in Figures 5.10 and 5.11 are again not complete, but serve to illustrate the workings of the constraints and their interaction. The number of mappings created on  $GEN_{UF}$  and  $GEN_{SF}$  depends for a large part on the particular transformations that are allowed. The next section explains in more detail where the transformation rules come from.

$[b\text{on}+\emptyset\#\text{akt}\text{er}+\emptyset]$	$n \rightarrow / \emptyset /$	$\emptyset \rightarrow /z/$	*.V	*.C	*CC	*V.
/bɔn.ak.tɛr/			*!	***		
/bɔn.akt.ɛr/			*!	**	*	
/bɔ.nak.tɛr/				**!		*
$\rightarrow$ /bɔ.nakt.ɛr/				*	*	*
/bɔ.ak.tɛr/	*!		*	**		*
/bɔ.akt.ɛr/	*!		*	*	*	*
/bɔz.ak.tɛr/		*!		***		
/bɔz.akt.ɛr/		*!		**	*	
/bɔ.zak.tɛr/		*!		**		
/bɔ.zakt.ɛr/		*!		*	*	

Figure 5.10: A tableau illustrating a mapping from Underlying Form to Surface Form.

/bɔ.nak.tɛr/	$\emptyset \rightarrow \text{ɔ}$	$\text{ɔ} \rightarrow \text{z}$	$\emptyset \rightarrow \text{n}$	$\text{n} \rightarrow \text{ɹ}$
$\rightarrow$ [bɔnaktɛr]				
[bɔnaktɛrə]	*!			
[bɔaktɛr]			*!	
[bɔ̄aktɛr]		*!	*	
[bɔaktɛrə]	*!		*	
[bɔ̄aktɛrə]	*!	*	*	
[bɔnakɛr]				*!
[bɔakɛr]			*!	*

Figure 5.11: A tableau illustrating a mapping from Surface Form to Phonetic Form.

### 5.2.5 Constraining GEN and generating CON

One of the stated aims of this chapter was to decrease the teleological bias of our simulations by allowing for more hypotheses. At the same time, we had to place clear bounds on GEN and CON by introducing a lexical repository that enumerates possible allomorphic submappings, rather than allowing any Morpheme to map to an arbitrary string of segments.

We also posited rule repositories in order to limit the number of segmental substitutions, deletions, and insertions that could be in effect between the segmental levels of representation. These bounds are necessary to prevent the generation of infinite candidate sets. The problem of infinite candidate generation has been resolved for two-level OT grammars within a Finite State OT context by Riggle (2004). However, those results do not seem applicable to our multi-level parallel OT model: a subcandidate which is harmonically bounded may be part of a full candidate that is possibly optimal.

A pragmatic balance must be struck between minimizing bias and limiting generative power. We introduce a *deductive* bias on the lexical forms that can be hypothesized by learners. Only lexical hypotheses for which there is some evidence in the overt input data may enter the lexical repository. We first detail this deductive procedure below, then consider whether this approach might also be suitable for creating a phonological “rule repository”.

#### Deducing lexical hypotheses

Section 5.2.2 showed the dual nature of input to the learners. The features of the Semantic-Syntactic representations, representing the context to an utterance, are largely unordered (except the top-level S-words), and organised hierarchically. The Phonetic Forms that represent the acoustic signal are a string, a flat ordered list of primitive symbols. The lexicon as represented by the MF–UF mapping forms the link between those two types of representations, but the primitives on the two “overt” levels of representation are of a quite different nature. What is a sensible way to constrain the set of possible lexical mappings?

In the context of all spoken languages, the relation between sound and meaning is almost completely arbitrary and unlimited (Saussure, 1916). However, this relation is *not* arbitrary for the members of a linguistic community: among speakers of the same language, things have a name, and similar events and entities will provoke similar sequences of speech sounds. The details of this process are the subject of much debate and ongoing research, but we may at least assume that infants are sensitive to these patterns and use them to learn word segmentation and acquire a lexicon of the target language.

When generating lexical hypotheses in the grammar of our virtual learners, we therefore limit the possibilities to those for which there is some evidence in the input data. Rather than simulate the formation of these hypotheses ‘on-line’ as part of the learning procedure, we generate them by iterating

over the complete input data set and building possible representations “from the outside in”.<sup>6</sup> The procedure can be informally described as follows:

1. For each SSF–PF learning datum, generate all possible SSF → MSF → MF mappings.
2. For each unique Morpheme in the resulting MFs, add an entry in a one-to-many map with the Morpheme as key and the co-occurring PFs as values.
3. For each Morpheme in the resulting map, take the set of PFs associated with it and calculate the *longest common substring* (LCS) of phones over this set. (A Morpheme’s LCS may have length zero).
4. Create all possible *alignments* of MFs with co-occurring PFs such that:
  - (a) every Morpheme is at least aligned with the longest common substring calculated in step 3
  - (b) every phone is aligned to exactly one Morpheme
  - (c) alignments are non-crossing
5. For each possible alignment, add the resulting Morpheme-Morph mappings to the lexical repository.

As an example, take the following simplified data set:

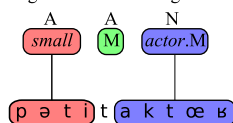
SemF	MF	PF
“good <sub>i</sub> , husband <sub>i</sub> ”	$\langle (GOOD-M)_A (HUSBAND.M)_N \rangle$ $\langle (GOOD.M)_A (HUSBAND.M)_N \rangle$	[bõmavĩ]
“small <sub>i</sub> , husband <sub>i</sub> ”	$\langle (SMALL-M)_A (HUSBAND.M)_N \rangle$ $\langle (SMALL.M)_A (HUSBAND.M)_N \rangle$	[pætĩmavĩ]
“good <sub>i</sub> , car <sub>i</sub> ”	$\langle (GOOD-F)_A (CAR.F)_N \rangle$ $\langle (GOOD.F)_A (CAR.F)_N \rangle$	[bõnvvatyʋ]
“small <sub>i</sub> , car <sub>i</sub> ”	$\langle (SMALL-F)_A (CAR.F)_N \rangle$ $\langle (SMALL.F)_A (CAR.F)_N \rangle$	[pætĩvvatyʋ]
“good <sub>i</sub> , actor <sub>i</sub> ”	$\langle (GOOD-M)_A (ACTOR.M)_N \rangle$ $\langle (GOOD.M)_A (ACTOR.M)_N \rangle$	[bõnaktœʋ]
“small <sub>i</sub> , actor <sub>i</sub> ”	$\langle (SMALL-M)_A (ACTOR.M)_N \rangle$ $\langle (SMALL.M)_A (ACTOR.M)_N \rangle$	[pætĩtaktœʋ]

In step 3 of the algorithm, the following longest common substrings would be found for each Morpheme:

<sup>6</sup>In principle, the procedure could also be applied on-line, with new lexical hypotheses appearing as more input data are offered to learners. However, this would somewhat complicate our general learning framework, where GEN and CON are represented as fully formed from the beginning.

Morpheme	PFs	LCS
(Husband.M) <sub>N</sub>	[bõmaʝi] [pətimavɨ]	maʝi
(Car.F) <sub>N</sub>	[bõnvwatyʝ] [pətitvwatyʝ]	vwatyʝ
(Actor.M) <sub>N</sub>	[bõnaktœʝ] [pətitaktœʝ]	aktœʝ
(Good) <sub>A</sub>	[bõmaʝi] [bõnvwatyʝ] [bõnaktœʝ]	b
(Good.F) <sub>A</sub>	[bõnvwatyʝ]	bõn
(Good.M) <sub>A</sub>	[bõmaʝi] [bõnaktœʝ]	b
(Small) <sub>A</sub>	[pətimavɨ] [pətitvwatyʝ] [pətitaktœʝ]	pəti
(Small.F) <sub>A</sub>	[pətitvwatyʝ]	pətit
(Small.M) <sub>A</sub>	[pətimavɨ] [pətitaktœʝ]	pəti
(M) <sub>A</sub>	[bõmaʝi] [pətimavɨ] [bõnaktœʝ] [pətitaktœʝ]	∅
(F) <sub>A</sub>	[bõnvwatyʝ] [pətitvwatyʝ]	∅

Morphemes are first aligned with their longest common phone substring



Then, all possible non-crossing alignments are generated for unaligned phones

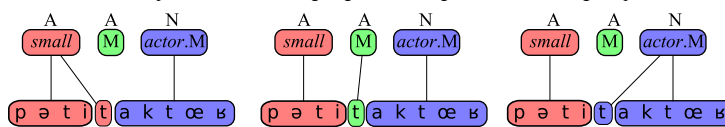


Figure 5.12: Creating alignments on the basis of longest common substrings found for Morphemes.

Then the possible alignments for the partition  $\langle (SMALL), (M), (ACTOR.M) \rangle \dots [bõnaktœʝ]$  are shown in Figure 5.12.

On the basis of these alignments, the following lexical hypotheses will be added to the repository:

<i>small</i>	→	pəti
<i>small</i>	→	pətit
<i>Adj.M</i>	→	∅
<i>Adj.M</i>	→	t
<i>actor.M</i>	→	taktœʁ
<i>actor.M</i>	→	aktœʁ

We see that the LCS constraint on GEN still allows for various hypotheses on how the liaison consonant should be analyzed: in one alignment the latent *t* of *petit* is reanalysed as belonging to the morph for *actor*. This allows for the type of reanalysis discussed in section 5.2, where liaison consonants are considered a prefix or part of an allomorph of the noun that follows the modifier.

### Deducing phonological rules?

We described above that the mappings generated by GEN<sub>UF</sub> and GEN<sub>SF</sub> are mostly determined by SPE-style phonological rewrite rules. The provenance of these rules was not yet made clear. It would be an interesting endeavour to also extract these on the basis of the input data, rather than craft them by hand. In fact, by assuming a deductive bias for our learners similar to that used for lexical hypotheses, auto-generation of rewrite rules is possible to some extent.

In the BiPhon framework, the UF–SF mapping is usually considered the locus of *allophony*: alternations involving distinct realizations of a single underlying phoneme, conditioned by the phonological environment. In fact, we have already lists of alternations at our disposal in the lexical repository. These alternations might be hypothesized to be the result of a rule on the phoneme level, instead of lexically based. For instance, from an alternation resulting in two stored allomorphs |pətit| and |pəti|, we can infer two possible rules: one inserting a /t/, perhaps before a vowel, and one deleting a /t/, perhaps before a consonant. Likewise, the presence of allomorphs |bɔ̃n| and |bɔ̃| in the lexicon would inspire two rules |ɔ̃n| → /ɔ̃/ and |ɔ̃| → /ɔ̃n/.

Rules for the SF–PF mapping could be generated in the same manner, with the distinction that they be limited by alternations found within the same phrase: i.e. if the phrase “good daughter” is sometimes realised as [bɔ̃nfij], sometimes as [bɔ̃nəfij], this would constitute evidence for two rules /ə/ → ∅ and ∅ → [ə].

Unfortunately, the sheer number of phonological rules generated by this deduction process grows quite large for bigger data sets. Suppose that for the morpheme ACTOR.M, the lexicon contains the set of Morphs |aktœʁ|, |laktœʁ|, |zaktœʁ|, |taktœʁ| and |naktœʁ|. If we consider each of these Morphs the possible result of a phonological transformation from another Morph, we end up with  $5 \times 4 = 20$  rules. Other lexical hypothesis sets may yield yet more rules, many of which seem outlandish from a phonological perspective (e.g. |ez| → ð). The additional SF and PF forms generated in this way exploded the candidate space, and as a result put too much processing and memory demand on

the simulation framework. For this reason, a more restrictive approach was taken, which will be described in the following subsection. However, we will return to a slightly similar rule generation method in Section 5.4.4. We discuss the implications of these restrictions in 5.5

### Predefined phonological rules for liaison and schwa

A rule repository was crafted by hand for  $\text{GEN}_{\text{UF}}$  and  $\text{GEN}_{\text{SF}}$  to allow for phonologically and phonetically based factors in liaison. We consider the following consonants to play an active role in optional and obligatory liaison: /t/, /z/ and /n/. The other consonants sometimes marked as liaising, /r/, /p/ and /g/, are not taken into consideration, since they are quite rare and mostly restricted to a handful of constructions (see Durand et al., 2011).

The full rule set used is shown below.

Sub-GEN	Formulation	Name
$\text{GEN}_{\text{UF}}$	$ z  \rightarrow \emptyset$	z-deletion
$\text{GEN}_{\text{UF}}$	$ t  \rightarrow \emptyset$	t-deletion
$\text{GEN}_{\text{UF}}$	$ \emptyset  \rightarrow \emptyset$	Schwa deletion
$\text{GEN}_{\text{UF}}$	$ \text{ɔ}n  \rightarrow /̃/$	ɔ-nasalisation
$\text{GEN}_{\text{UF}}$	$\emptyset \rightarrow /z/$	z-insertion
$\text{GEN}_{\text{UF}}$	$\emptyset \rightarrow /t/$	t-insertion
$\text{GEN}_{\text{UF}}$	$ \tilde{ɔ}  \rightarrow /ɔn/$	denasalisation/n-insertion
$\text{GEN}_{\text{SF}}$	$/\emptyset/ \rightarrow \emptyset$	Schwa deletion

## 5.3 Simulation 1: toy French revisited

The previous sections described a model of liaison in the multi-level BiPhon framework, expanded from that of Chapter 4 in order to handle data sets of any size and account for more facets of the phenomenon. As a proof of concept, this expanded model will first be tested on a small data set, similar to the one used in Chapter 4. This exercise also serves to finetune the learning parameters to values suitable for a larger candidate and constraint set. Next, this toy set will be slightly altered to test the capability of the expanded model to handle variation in the input.

The previous chapter tested various update algorithms, parsing strategies, and cost functions (i.e. OT, HG, MaxEnt and so on). For the simulations described in this and the following section, we will stick to a narrower range of parameter settings: generally, those that gave the best results in the previous chapter. However, the interested reader is welcome to test different parameter settings by downloading the code and data from the author's website.<sup>7</sup>

<sup>7</sup><http://jwvl.eu/ssen>



### 5.3.1 Data set 1A: three-noun toy grammar

#### Input data and resulting grammars

The input data for this first set of simulations is a slightly augmented version of the three-pair grammar of Chapter 4. The same nouns are used, but varied with a second liaising adjective as well as the definite article *le/la/l'*. These extra data serve mainly to differentiate a phonological analysis from an allomorphic one in the results; if the data contain only one noun and one adjective (albeit in varying forms), there is little sense in calling an alternation ‘phonological’. Table 5.2 shows the input data used, together with their orthographic form (not used in learning) and frequency within the dataset. Here, the frequency is the same for all pairs in the distribution: that is, each has a probability of  $100/900 \approx 11.11\%$  of being drawn for an evaluation.

Orthographic	SSF	PF	Freq
bonne voiture	(good) <sub>Adj</sub> ; (car.F, SG) <sub>N</sub>	[bɔ̃nvwatyʁ]	100
bon mari	(good) <sub>Adj</sub> ; (husband.M, SG) <sub>N</sub>	[bɔ̃mari]	100
bon homme	(good) <sub>Adj</sub> ; (man.M, SG) <sub>N</sub>	[bɔ̃nɔm]	100
petite voiture	(small) <sub>Adj</sub> ; (car.F, SG) <sub>N</sub>	[petitvwatyʁ]	100
petit mari	(small) <sub>Adj</sub> ; (husband.M, SG) <sub>N</sub>	[petimari]	100
petit homme	(good) <sub>Adj</sub> ; (man.M, SG) <sub>N</sub>	[petitɔm]	100
la voiture	(the) <sub>Det</sub> ; (car.F, SG) <sub>N</sub>	[lavwatyʁ]	100
le mari	(the) <sub>Det</sub> ; (husband.M, SG) <sub>N</sub>	[ləmari]	100
l’homme	(the) <sub>Det</sub> ; (man.M, SG) <sub>N</sub>	[ləm]	100

Table 5.2: The input distribution for simulation 1A.

#### Learning procedure

The learning procedure and hyperparameters used conform to those described in Chapter 2; for a number of parameters, several settings were tried. Evaluation noise was set to 1.0, and the number of learning steps was 20,000 in every simulation, divided into four epochs. Three values were tried for the initial plasticity: (0.5, 1.0, 2.0). For plasticity decay, three values were tried: (0.25, 0.5 and 0.75). The final parameter that was varied was the update algorithm: either “weighted uncanceled” or “all up, high down” (see Chapter 4). The “re-sampling” parsing strategy of Jarosz (2013a) was applied in all simulations, since it was shown to notably increase success rates of BiPhon learners in both Chapters 3 and 4.

300 learners were trained for each combination of these ( $3 \times 3 \times 2 =$ ) 12 possible configurations. To assess the extent to which the resulting grammars matched the input data, the error rate of the resulting grammar was tested by evaluating 1,000 forms drawn randomly from the training distribution under

the same evaluation noise of 1.0, and counting the proportion of winning candidates that correctly reproduced the target form pair. A learning success was defined as an error of less than five percent on these test data.

### Quantitative analysis of results

Results show that a majority of learners were labeled “successful” on this data set, regardless of update algorithm, initial plasticity or plasticity decay. Table 5.3 shows the percentage of successful learners under the various parameter combinations discussed above. Figure 5.14 shows the averaged error rates (regardless of success/failure). The results appear to indicate a slight advantage for the AllUpHighDown algorithm, although the difference is small.

Update algorithm	W.Uncancelled			AllUpHighDown		
Initial plasticity	0.5	1.0	2.0	0.5	1.0	2.0
Plasticity decay: 0.75	87.7	88.3	83.3	88.7	87.3	86.0
Plasticity decay: 0.5	87.3	86.3	87.7	89.0	88.0	89.0
Plasticity decay: 0.25	89.7	87.3	88.0	83.3	<b>90.7</b>	86.7

Table 5.3: Successes as % under varying parameter settings for data set A

The data-based GEN used in these simulations leads to a much larger candidate space per input compared to that of Chapter 4, with many more non-canonical candidates (as discussed in Section 5.2.1). Contrary to what might be expected, this enlarged candidate space seems to improve rather than diminish an EDRA’s chance of success; recall that the success rates in Chapter 4 were 83 out of 100 for “weighted uncanceled” learners. Nevertheless, about one learner in ten still appears to become stuck in a local optimum from which it cannot recover. Figure 5.14 illustrates this: it tracks the success rates *over the course of training*, separating learners which eventually emerge successful from those which fail to acquire the target language. As the figure illustrates, the initial stages of learning turn out to be crucial: after just 1,000 evaluations, a large gulf in error rates can be seen between ultimately successful and ultimately unsuccessful learners. The sources of randomness in the training framework (drawing input data, evaluation noise) may irretrievably force the grammar into a locally optimal, globally suboptimal state. We return to this failure of convergence in Section 5.5.1.

### Qualitative analysis of results

The previous section described how the procedurally generated candidate and constraint sets should be compatible with various analyses for the liaison forms: selection of gender-mismatching forms, allomorphy, and phonologically-based consonant deletion or insertion. Only the strictly serial analysis of Dell (1970, 1973) is not available to the learners in this simulation, since it requires

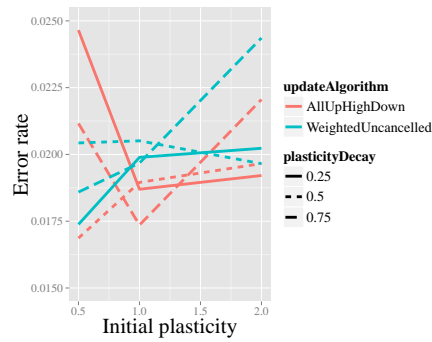


Figure 5.13: Average error rates over various parameter settings for data set A.

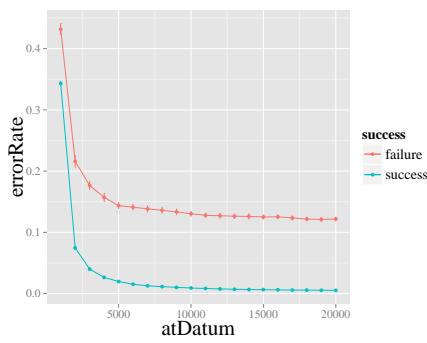


Figure 5.14: Average error rates versus number of data processed, split by eventual success.

postulating an underlying schwa which cannot be reconstructed from the input data. An inspection of the output after training reveals that the “less biased” candidate and constraint generator of this chapter does not lead to a pronounced preference for one solution to the exclusion of others. Morphosyntax, lexicon and phonology are all implicated as the locus of liaison alternations. Figure 5.15 visualizes the various solutions found for the forms *bon homme*, *petit homme* and *petite voiture*. For the liaising forms with *homme*, learners unanimously select  $[b\text{ɔ}n]$  as an underlying form, but disagree as to its morphological makeup (masculine, feminine or not specified for gender). Interestingly, a considerable proportion of learners consider *voiture* to have an underlying allomorph  $[tvwaty\text{v}]$ , whose  $/t/$  is then deleted in phonology.

### Random baseline learner

In Chapter 4, the results of OT/EDRA learners were compared to a baseline “random learner”, which simply shuffles its ranking around until it stumbles upon a hierarchy that is compatible with all learning data. It turned out that for the 19-constraint, 3-datum grammar of Chapter 4, the random baseline learner performed no worse than the error-driven reranking algorithms. However, the prediction was made that such an approach would no longer be feasible if the number of forms and constraints were to be increased. This prediction was tested by applying the random baseline learning algorithm to the expanded model used in this chapter, again using the input data shown in Table 5.2.

After 10 million shufflings of the grammar, the random learner had still not found a hierarchy compatible with all 9 pairs in the input data. Grammars that correctly reproduced 5, 6 and 7 pairs were found after respectively 616, 27444 and 290457 tries. This justifies the need for an efficient learning/evaluation algorithm such as that of Chapter 4: random learning of hierarchies does not appear to scale to larger grammars and data sets. In the rest of this chapter, this assertion will be taken for granted, and random learners will no longer be used as a baseline.

## 5.3.2 Data set 1B: variation in schwa

### Data and method

Having established that our model is capable of learning a variant of the toy grammar of the previous chapter, a next step is to test its performance in the face of variation. So far, we have been simulating speakers of “Standard” French. In this dialect, the orthographic *e* in feminine adjectives such as *petite* and *bonne* is almost invariably mute in normal speech. This is not necessarily the case in all dialects: a prominent feature of Midi (Southern) French is a much higher rate of schwa realization. Armstrong and Unsworth (1999) report that for younger speakers of a variety of Southern French, schwa deletion

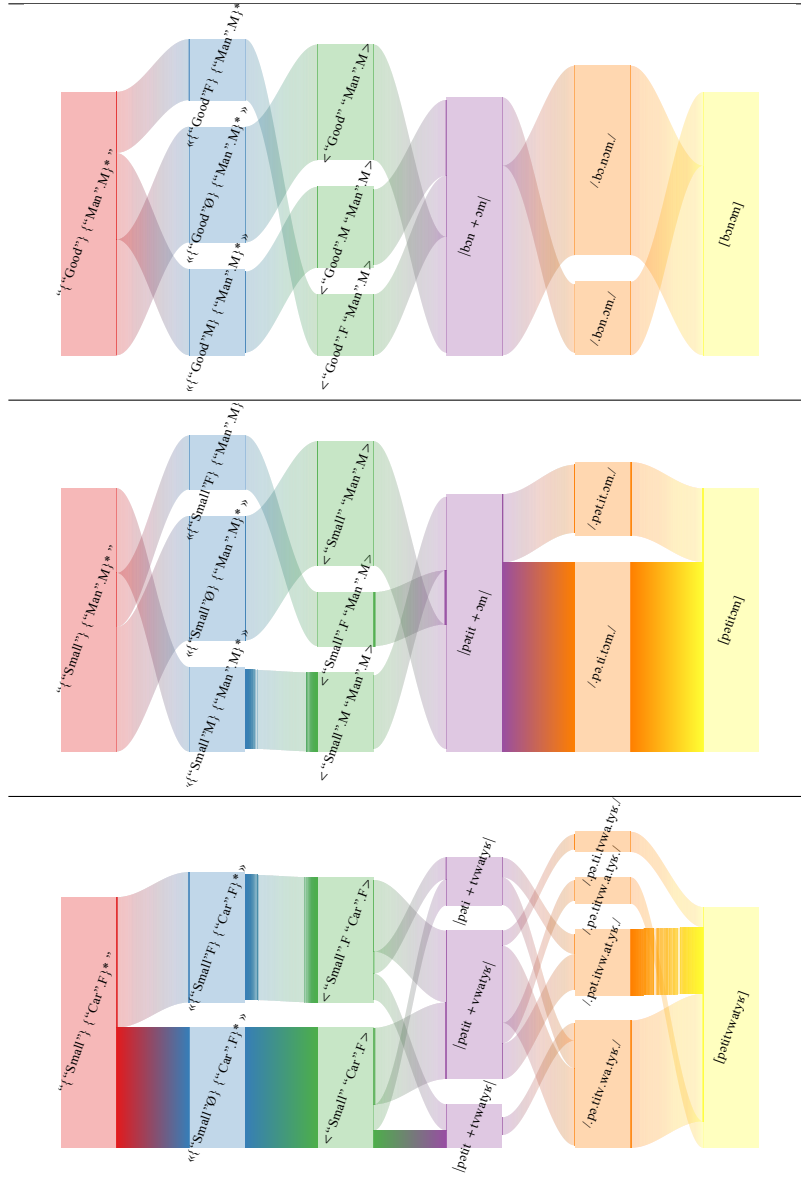


Figure 5.15: Sankey diagrams for some output candidates of successful learners of data set A. A wider band indicates a more popular mapping or form.

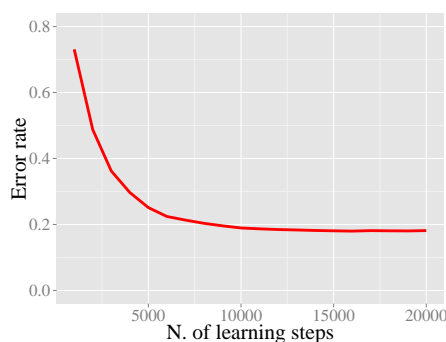


Figure 5.16: Average error rates for all learners on data set B

occurs in approximately 37% of cases in a C.C context in conversational style.

We create a (grossly simplified) “Midi version” of the toy language of data set 1A by applying this same schwa retention rate to the feminine adjectives of our data set. Table 5.4 shows the distribution.

## Results

Figure 5.16 shows the average error rate over time for all learners. Given the variation displayed for some SSF forms, learners’ L-candidates will inevitably keep deviating from the form pairs drawn from the distribution, even when a learner’s language closely mirrors the target language. An error rate of less than five percent therefore cannot be expected on this data set: a hypothetical minimum error rate of about eight percent could be reached by learners who always output the more likely AudF for those SSFs with variable realizations. However, the error-driven learning framework is not conducive to converging on such a best-guess learner. Table 5.4 instead compares the original input distribution with the average output distributions of learners after learning.

These ‘averaged’ results appear to indicate that learners achieve some success in mirroring the variation found in the data, although a perfect match seems hard to achieve, unlike the probability-matching learners of e.g. Boersma and Hayes (2001). Given the variety of factors that decide schwa realization in real French, and the multitude of candidates that lead to schwa-realizing forms in our model, perhaps this outcome is not entirely undesirable: the results indicate that increased complexity in the input leads to more variation among learners. This at least appears to be true for schwa realization in French, which has inspired numerous phonological analyses and shows substantial inter- and intraspeaker variation (see e.g. Dell, 1973; Selkirk, 1978; Tranel, 1987; Côté, 2001; Bayles et al., 2016).

The inclusion of schwa-final Phonetic Forms in the input distribution increases the number of hypothesized lexical entries available to the learners:

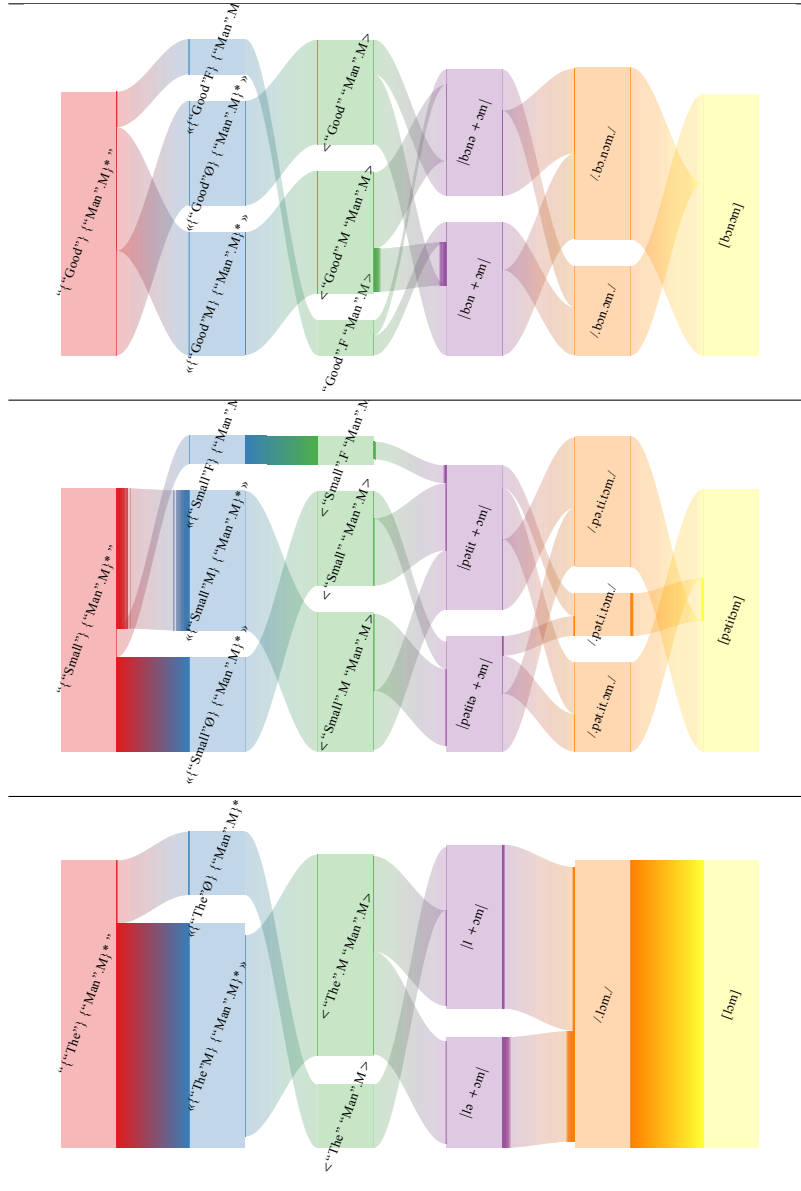


Figure 5.17: Sankey diagrams for some output candidates of successful learners on data set B.

Orthographic	SSF	AudF	Target	Avg (sd)
bonne voiture	(good) <sub>Adj</sub> ; (car.F, SG) <sub>N</sub>	[bɔnɔvwatɥ]	63	52.55 ± 17.65
	(good) <sub>Adj</sub> ; (car.F, SG) <sub>N</sub>	[bɔnvwatɥ]	37	40.41 ± 15.98
bon mari	(good) <sub>Adj</sub> ; (husband.M, SG) <sub>N</sub>	[bɔmari]	100	93.11 ± 12.43
bon homme	(good) <sub>Adj</sub> ; (man.M, SG) <sub>N</sub>	[bɔnɔm]	100	99.16 ± 3.01
petite voiture	(small) <sub>Adj</sub> ; (car.F, SG) <sub>N</sub>	[petitɔvwatɥ]	63	55.40 ± 15.67
	(small) <sub>Adj</sub> ; (car.F, SG) <sub>N</sub>	[petitvwatɥ]	37	39.49 ± 13.59
petit mari	(small) <sub>Adj</sub> ; (husband.M, SG) <sub>N</sub>	[petimari]	100	92.91 ± 13.92
petit homme	(good) <sub>Adj</sub> ; (man.M, SG) <sub>N</sub>	[petitɔm]	100	99.6 ± 0.74
la voiture	(the) <sub>Det</sub> ; (car.F, SG) <sub>N</sub>	[lavwatɥ]	100	96.49 ± 10.23
le mari	(the) <sub>Det</sub> ; (husband.M, SG) <sub>N</sub>	[ləmari]	100	99.45 ± 1.89
l'homme	(the) <sub>Det</sub> ; (man.M, SG) <sub>N</sub>	[lɔm]	100	100 ± 0.00

Table 5.4: Original input distribution compared with learners' outputs for data set B.

there is now evidence for underlying forms [bɔnɔ] or [pɛtitɔ]. Figure 5.17 shows that some learners indeed entertain these forms as allomorphs of these adjectives, resulting in analyses that somewhat resemble the serial analysis of Dell (1973) as formalized in Chapter 4. The difference is that these virtual learners delete in the mapping from UF to SF, rather than from SF to PF. Also, some learners have schwa-final underlying forms for the *masculine* form of *bon* and *petit*. Nevertheless, the fact that candidates with underlying schwa can emerge as optimal, despite their incurring an additional constraint violation, indicates an abstract role similar to that of Dell's analysis.

To summarize, the simulation model introduced in this chapter is capable of reproducing variability in the input data. Furthermore, the adapted dataset demonstrates the influence of the input data on the hypothesis space available to learners: a previously unavailable analysis, involving an "abstract" schwa in some forms, is chosen by a considerable amount of learners. Given that the focus is on liaison rather than schwa deletion/retention, variation in schwa realization will not be further examined in this chapter.

### 5.3.3 Data set 1C: Gender-allomorphic forms

The extended BiPhon model of this chapter is able to generate candidates displaying "true" gender allomorphy, where gender agreement is sacrificed for phonological well-formedness in liaising forms. As Chapter 4 mentioned in passing, there is a small set of forms that lends more weight to this analysis. A well-known triplet of adjectives, namely [nuvo]-[nuvɛl] 'new', [vjø]-[vjɛj] 'old' and [bo]-[bɛl] 'beautiful', have alternations that cannot be derived from more general synchronic phonological rules. Interestingly, vowel-initial male nouns are preceded by forms that are surface-identical to the feminine form of the adjective. The same happens for the possessive pronouns [ma]-[mɔ] 'my', [ta]-[tɔ] 'your (SG)' and [sa]-[sɔ] 'his/her'. A single process of phonologically-driven gender selection, violating a constraint on agreement, can account for all these



alternations (as well as previously seen *petit(e)* and *bon(ne)*). A ‘lexical allomorphy’ analysis requires learning several separate masculine allomorphs for the above irregular forms. From a constraint learning perspective, the gender selection analysis requires fewer crucial constraint orderings and thus seems more attractive. The procedural grammar generation procedure of this chapter allows us to test these intuitions, by including these adjectives to the input data.

### Data and method

Table 5.5 shows the input distribution. It expands on data set A in two ways. First, the adjectives *nouveau/nouvelle*, *vieux/vieille* and *beau/belle* were added. Additionally, a vowel-initial feminine noun *onde* ‘wave’ was added, so that the data clearly reflects that both phonological and morphosyntactic factors influence the phonetic realization of the adjective. Each noun was crossed with each adjective as well as the definite article.

Orthographic	SSF	AudF	Freq
bonne voiture	(good) <sub>Adj</sub> ; (car.F, SG) <sub>N</sub>	[bɔ̃nvwatyʁ]	100
bonne onde	(good) <sub>Adj</sub> ; (wave.F, SG) <sub>N</sub>	[bɔ̃nɔ̃d]	100
bon mari	(good) <sub>Adj</sub> ; (husband.M, SG) <sub>N</sub>	[bɔ̃maʁi]	100
bon homme	(good) <sub>Adj</sub> ; (man.M, SG) <sub>N</sub>	[bɔ̃nɔ̃m]	100
petite voiture	(small) <sub>Adj</sub> ; (car.F, SG) <sub>N</sub>	[pətɪtvwatyʁ]	100
petite onde	(small) <sub>Adj</sub> ; (wave.F, SG) <sub>N</sub>	[pətɪtɔ̃d]	100
petit mari	(small) <sub>Adj</sub> ; (husband.M, SG) <sub>N</sub>	[pətɪmaʁi]	100
petit homme	(good) <sub>Adj</sub> ; (man.M, SG) <sub>N</sub>	[pətɪtɔ̃m]	100
nouvelle voiture	(new) <sub>Adj</sub> ; (car.F, SG) <sub>N</sub>	[nuvɛlvwatyʁ]	100
nouvelle onde	(new) <sub>Adj</sub> ; (wave.F, SG) <sub>N</sub>	[nuvɛlɔ̃d]	100
nouveau mari	(new) <sub>Adj</sub> ; (husband.M, SG) <sub>N</sub>	[nuvomɑʁi]	100
nouvel acteur	(new) <sub>Adj</sub> ; (man.M, SG) <sub>N</sub>	[nuvɛlaktœʁ]	100
vieille voiture	(old) <sub>Adj</sub> ; (car.F, SG) <sub>N</sub>	[vjɛjvwatyʁ]	100
vieille onde	(old) <sub>Adj</sub> ; (wave.F, SG) <sub>N</sub>	[vjɛjɔ̃d]	100
vieux mari	(old) <sub>Adj</sub> ; (husband.M, SG) <sub>N</sub>	[vjømaʁi]	100
vieil homme	(old) <sub>Adj</sub> ; (man.M, SG) <sub>N</sub>	[vjɛjɔ̃m]	100
belle voiture	(beautiful) <sub>Adj</sub> ; (car.F, SG) <sub>N</sub>	[belvwatyʁ]	100
belle onde	(beautiful) <sub>Adj</sub> ; (wave.F, SG) <sub>N</sub>	[belɔ̃d]	100
beau mari	(beautiful) <sub>Adj</sub> ; (husband.M, SG) <sub>N</sub>	[bomaʁi]	100
bel homme	(beautiful) <sub>Adj</sub> ; (man.M, SG) <sub>N</sub>	[belɔ̃m]	100
la voiture	(the) <sub>Det</sub> ; (car.F, SG) <sub>N</sub>	[lavwatyʁ]	100
l’onde	(the) <sub>Det</sub> ; (wave.F, SG) <sub>N</sub>	[lɔ̃d]	100
le mari	(the) <sub>Det</sub> ; (husband.M, SG) <sub>N</sub>	[ləmaʁi]	100
l’acteur	(the) <sub>Det</sub> ; (man.M, SG) <sub>N</sub>	[ləm]	100

Table 5.5: The input distribution for simulation 1C.

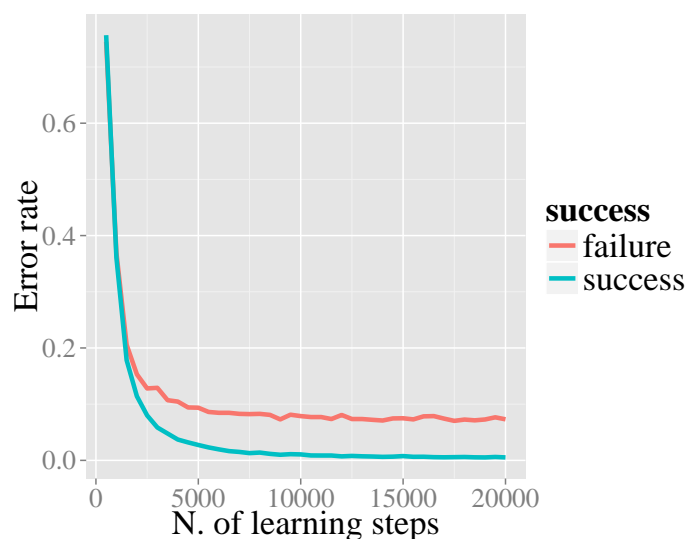


Figure 5.18: Average error rates over time for successful and unsuccessful learners on data set C.

The learning procedure and parameter settings were otherwise identical to those used on data set B: 100 learners were trained for 20,000 learning steps divided over four epochs, using the AllUpHighDown update algorithm and RRIP parsing, with evaluation noise at 1.0, initial plasticity at 1.0 and plasticity decay at 0.5.

### Results

As in Simulation 1A, a majority of learners converged on a “correct” target-replicating grammar. Learning success (less than 5 percent error when testing on the training set) was reached for 88 out of 100 learners. This is an encouraging result, given that the larger training set increases not only the number of constraints but also the size of the auto-generated lexicon. For instance,  $|\varepsilon_j|$  and  $|\varepsilon_l|$  will be considered as morphs for the Morpheme SG, and thereby increase the number of candidates generated from any SSF. Again, separating the eventually successful from the failing learners makes clear that the initial stage of learning is decisive for the outcome (Figure 5.18).

#### 5.3.4 Data set 1D: Plural forms

The final data set in this section enriches the SSF forms with a selected morphosyntactic *number* feature on nouns, and adding both singular and plural forms to the input data. Grammatical number is an important factor in liai-

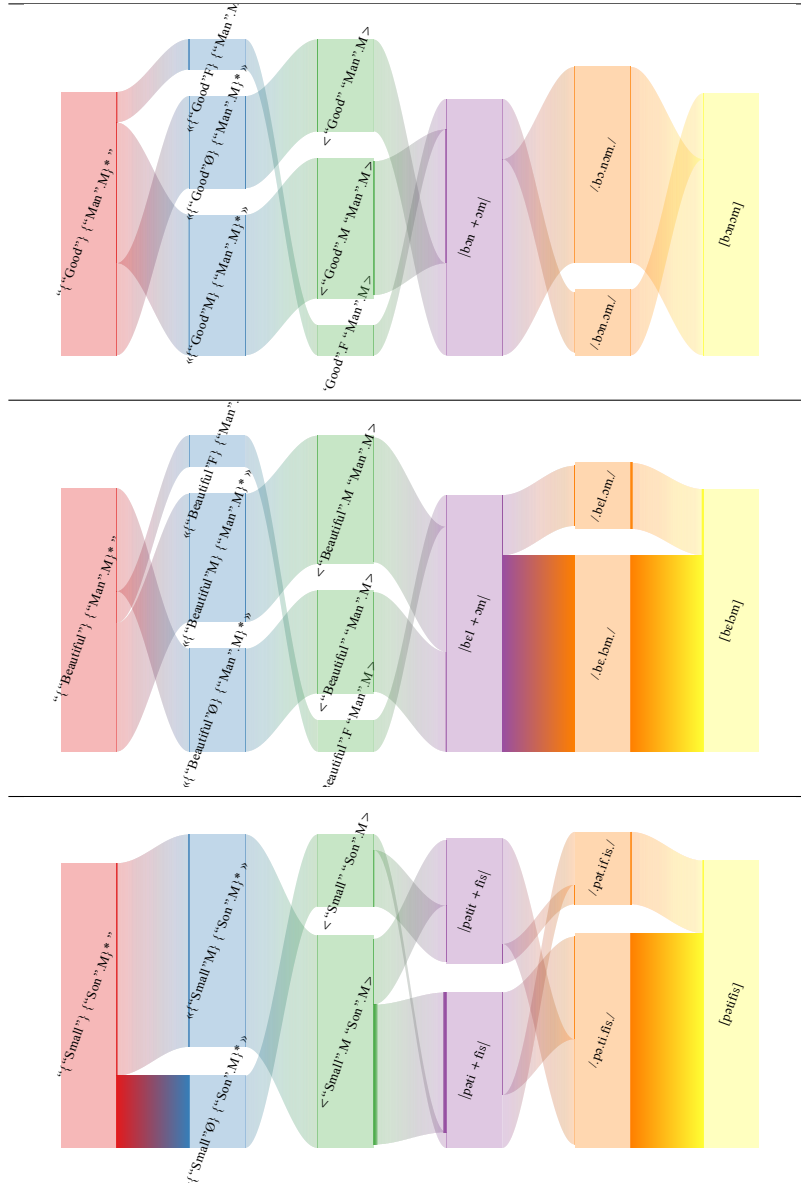


Figure 5.19: Sankey diagrams for some output candidates of successful learners on data set C. A minority of learners entertains a “gender allomorphy” analysis.

son realization, as the plural morpheme  $|z|$  (orthographically  $s$  or  $x$ ) added to nouns, adjectives and determiners is subject to deletion in many phonological contexts. As such, this data set serves as a test for the model's ability to represent additional morphological factors in liaison. Unfortunately, this added morphological complexity adds substantial computational cost, which meant that the generative power of the model had to be toned down, as the next section explains.

### Data and method

The distribution for data set 1D is shown in table 5.6. It resembles data set 1A, except that a plural variant has been added to each form, marking grammatical number as a *selected* morphosemantic feature on nouns in the SSF forms. Its value is either singular (SG) or plural (PL).

This single added feature on SSF leads to a great increase of possible Morphemes on MF compared to those generated in data sets 1A, 1B and 1C. In turn, the size of the lexicon that defines possible forms generated by  $GEN_{MF}$  explodes, and this size increase is propagated to the segmental forms of representation further down the "candidate graph".

The resulting candidate and constraint sets were thus several orders of magnitude larger than those generated without a number feature on SSF. Unfortunately, the time and space complexity of simulations with these large candidate and constraint sets became such that, despite the efficient evaluation procedure detailed in the last chapter, the resulting candidate space could not be traversed in a reasonable amount of time on any machine available to the author. Therefore, to allow the simulations of this and subsequent sections to run their course, the automatic lexicon generation procedure was reduced in generative power. This was done by restricting the number of possible Morpheme to Morph mappings as follows: after automatically generating a lexicon according to the procedure detailed in Section 5.2.5, all non-concept features (or combinations thereof) were restricted to generating either null, schwa, or one of the consonants implicated in liaison:  $|t|$ ,  $|z|$  and  $|n|$ . Other hypotheses, such as the aforementioned  $|ɛj|$  and  $|ɛl|$  allomorphs for the Morpheme SG or M.SG, were deleted from  $GEN_{MF}$ . This restriction only applied to non-concept features; lexical mappings containing concept features were left intact. Restricting  $GEN_{MF}$  in this manner sufficed to make generation and evaluation computationally tractable. The learning parameters were the same as those used for data sets 1B and 1C. The next section describes the results obtained by training virtual learners on the candidate and constraint sets obtained by this procedure.

### Results

A convincing majority of learners found a grammar capable of producing the correct forms: 97 out of 100 learners arrived at an error rate of less than 5

Orthographic	SSF	AudF	Freq
bonne voiture	(good) <sub>Adj</sub> ; (car.F, SG) <sub>N</sub>	[bɔ̃nvwatyʁ]	100
bonne onde	(good) <sub>Adj</sub> ; (wave.F, SG) <sub>N</sub>	[bɔ̃nzɔ̃d]	100
bon mari	(good) <sub>Adj</sub> ; (husband.M, SG) <sub>N</sub>	[bɔ̃maʁi]	100
bon homme	(good) <sub>Adj</sub> ; (man.M, SG) <sub>N</sub>	[bɔ̃nɔ̃m]	100
petite voiture	(small) <sub>Adj</sub> ; (car.F, SG) <sub>N</sub>	[pətivwatyʁ]	100
petite onde	(small) <sub>Adj</sub> ; (wave.F, SG) <sub>N</sub>	[pətizɔ̃d]	100
petit mari	(small) <sub>Adj</sub> ; (husband.M, SG) <sub>N</sub>	[pətimɑʁi]	100
petit homme	(good) <sub>Adj</sub> ; (man.M, SG) <sub>N</sub>	[pətizɔ̃m]	100
la voiture	(the) <sub>Det</sub> ; (car.F, SG) <sub>N</sub>	[lavwatyʁ]	100
l'onde	(the) <sub>Det</sub> ; (wave.F, SG) <sub>N</sub>	[lɔ̃d]	100
le mari	(the) <sub>Det</sub> ; (husband.M, SG) <sub>N</sub>	[ləmaʁi]	100
l'homme	(the) <sub>Det</sub> ; (man.M, SG) <sub>N</sub>	[lɔ̃m]	100
bonnes voitures	(good) <sub>Adj</sub> ; (car.F, PL) <sub>N</sub>	[bɔ̃nvwatyʁ]	100
bonnes ondes	(good) <sub>Adj</sub> ; (wave.F, PL) <sub>N</sub>	[bɔ̃nzɔ̃d]	100
bons maris	(good) <sub>Adj</sub> ; (husband.M, PL) <sub>N</sub>	[bɔ̃maʁi]	100
bons hommes	(good) <sub>Adj</sub> ; (man.M, PL) <sub>N</sub>	[bɔ̃nɔ̃m]	100
petites voitures	(small) <sub>Adj</sub> ; (car.F, PL) <sub>N</sub>	[pətivwatyʁ]	100
petites ondes	(small) <sub>Adj</sub> ; (wave.F, PL) <sub>N</sub>	[pətizɔ̃d]	100
petits maris	(small) <sub>Adj</sub> ; (husband.M, PL) <sub>N</sub>	[pətimɑʁi]	100
petits hommes	(good) <sub>Adj</sub> ; (actor.M, PL) <sub>N</sub>	[pətizɔ̃m]	100
les voitures	(the) <sub>Det</sub> ; (car.F, PL) <sub>N</sub>	[levwatyʁ]	100
les ondes	(the) <sub>Det</sub> ; (wave.F, PL) <sub>N</sub>	[lezɔ̃d]	100
les maris	(the) <sub>Det</sub> ; (husband.M, PL) <sub>N</sub>	[ləmaʁi]	100
les hommes	(the) <sub>Det</sub> ; (man.M, PL) <sub>N</sub>	[lezɔ̃m]	100

Table 5.6: The input distribution for Simulation 1D

percent. As in data sets 1B and 1C, increasing the number of pairs in the input appears to increase, not decrease, the probability of convergence on all pairs. Qualitative inspection of the results reveals a peculiar bias in those learners that converged on a “working” grammar: they are heavily biased toward synthetic forms, which fuse the morphosyntactic number and concept features with the concept feature, yielding a single Morpheme on MF and consequently a single Morph on UF. Figure 5.20 shows a few example analyses. For instance, all learners analyze *petits hommes* as having an underlying form of [pəti+zɔ̃m], where [zɔ̃m] is the “suppletive” Morph for the Morpheme (Man.M.PL). Interestingly, phonological rules are nevertheless active in some learners’ grammars. A minority of learners analyses (Son.M.SG) as [əfis], with the underlying schwa retained in *le fils* [ləfis] but deleted in *bon fis* [bɔ̃fis], apparently for reasons of syllabic wellformedness.

While consistent with the data set and constraint set available to learners, these analyses are quite unorthodox. Intuitively, we can think of several rea-

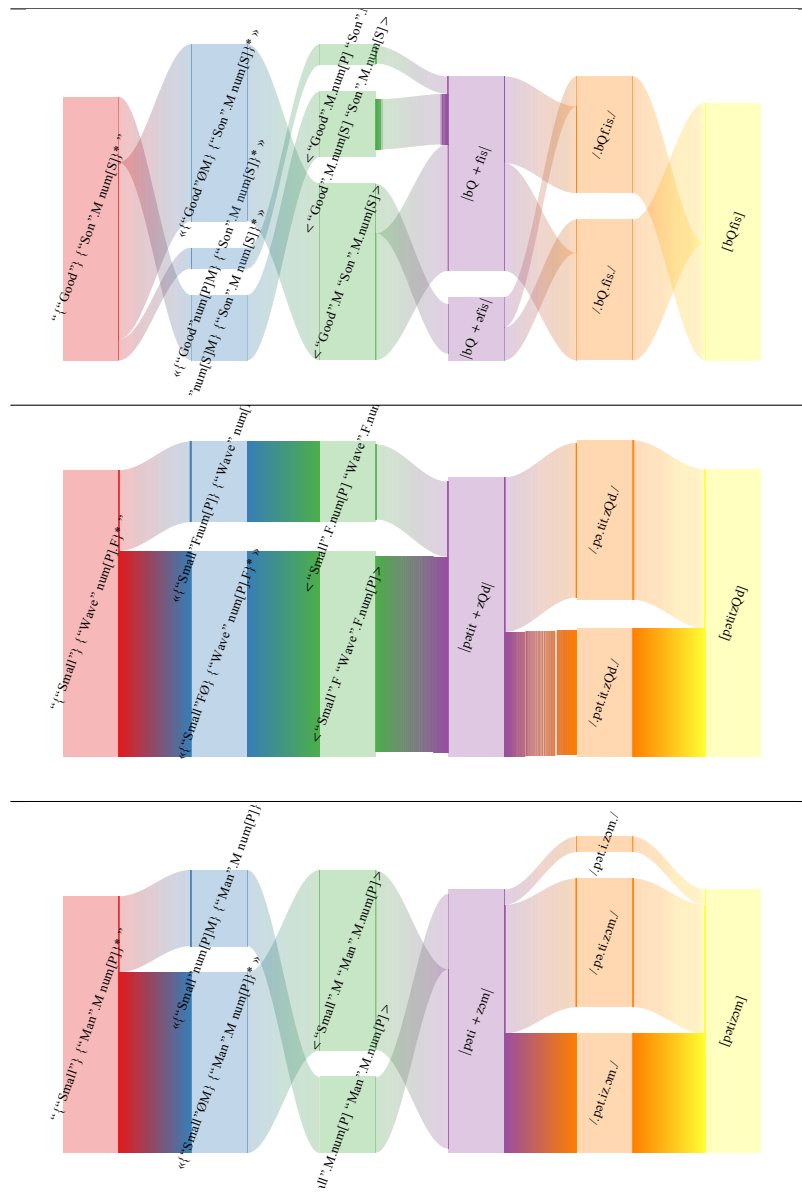


Figure 5.20: Sankey diagrams for some output candidates of succesful learners on data set 1D. A majority of learners uses synthetic plural forms like `|zom|` and `|zɔd|`.

sons why these learners favor a largely lexical analysis of liaison:

- The small size of the data set, and the resulting small lexical constraint set, makes it relatively unexpensive to find a ranking of LEX constraints that conforms to the data. With a larger number of concepts a “general” explanation rooted in phonological rules becomes more attractive (would require less crucial constraint orderings).
- The model used in this chapter is not biased against entertaining multiple allomorphs for any morpheme. On the contrary, the initial equal ranking of LEX constraints favors conditioned allomorphy.
- The definition of the constraints on MSF and MF creates a bias towards synthetic forms.

Since the main objective of Simulations 1A–1D was to investigate the feasibility of the six-level extended BiPhon model for analysing liaison, we will not pursue these intuitions further here. Rather, having demonstrated that the efficient evaluation and learning procedures can find various liaison analyses even in a much larger candidate space, we move on to the next section, where the extended BiPhon model will be tested on larger-scale, more realistic and more varied data sets.

## 5.4 Simulation 2: PFC dataset

### 5.4.1 Preprocessing of input data

The source data used for Simulation 2 were downloaded by FTP from the Projet PFC server in September 2016. To ensure uniformity of dialect in the input data, only speakers from the Île-de-France region were selected. In this manner, transcriptions of interviews with 35 informants were collected. Utterances by the interviewers were left out of consideration.

The analysis in this chapter is restricted to two-word modifier-noun phrases, where a modifier means either a determiner (DET), an adjective (ADJ), or a numeral (NUM). Unfortunately, not all data in the PFC were tagged for part of speech at the time the source data were collected. To gather relevant word pairs, possible modifier-noun phrases were extracted from the data with the help of the LEXIQUE lemma database (New et al., 2004). All combinations of two adjacent words that were potential modifier-noun phrases were extracted from the transcriptions. Using the morphological information and associated transcription for each word in Lexique, together with the liaison annotation for this word in PFC, an SSF–PF pair was automatically generated for this word pair. As in 5.3.4, nouns in the SSF forms were marked for (inherent/lexical) gender and (selected) number, whereas these features were not marked on the determiner. All nouns were marked as head, and the modifiers as dependent. The resulting pairs were then manually checked by the

	All items	Liaising items	Pct liaising
Types	2260	147	6.50 %
Tokens	5335	382	7.16 %

Table 5.7: Type and token frequencies of all and liaising items in the PFC-derived dataset.

SSF	PF	Frequency
"(ArtInd) <sub>Det</sub> ; (zoom.M, PL) <sub>N</sub> "	[dezuɪm]	2
"(ArtInd) <sub>Det</sub> ; (objet.M, PL) <sub>N</sub> "	[dezɔbʒɛ]	1
"(ArtInd) <sub>Det</sub> ; (objet.M, PL) <sub>N</sub> "	[deɔbʒɛ]	2
"(ArtDef) <sub>Det</sub> ; (an.M, PL) <sub>N</sub> "	[leɑ̃]	3
"(ArtDef) <sub>Det</sub> ; (an.M, PL) <sub>N</sub> "	[lezɑ̃]	4
"(ArtDef) <sub>Det</sub> ; (animal.M, PL) <sub>N</sub> "	[leanimo]	1
"(ArtDef) <sub>Det</sub> ; (animal.M, PL) <sub>N</sub> "	[lezanimɔ]	2
"(ArtDef) <sub>Det</sub> ; (concours.M, PL) <sub>N</sub> "	[lekɔkuʁ]	1
"(ArtDef) <sub>Det</sub> ; (concours.M, SG) <sub>N</sub> "	[ləkɔkuʁ]	1
"(PossM) <sub>Det</sub> ; (fille.F, SG) <sub>N</sub> "	[mafij]	1
"(PossM) <sub>Det</sub> ; (frère.M, SG) <sub>N</sub> "	[mɔ̃fʁɛʁ]	1
"(petit) <sub>Adj</sub> ; (ami.F, SG) <sub>N</sub> "	[pɔ̃titami]	1
"(petit) <sub>Adj</sub> ; (banc.M, SG) <sub>N</sub> "	[pɔ̃tibɑ̃]	1

Table 5.8: A small fragment of the input distribution offered to learners in Simulation 2. Items exhibiting liaison are highlighted.

author for spurious items and erroneous gender or number assignments. The resulting data set contains 2260 distinct form pair types. Of these only 147 manifest liaison, about 6.5%. However, since the liaising constructions are often quite frequent, their token proportion in the distribution is actually a bit larger, about 7.16% (Table 5.7). Table 5.8 shows a fragment of the pair distribution, also demonstrating the phonetic variation exhibited by some phrases within the dataset.

#### 5.4.2 Method

The parameter settings used for the PFC dataset simulations are largely the same as those used in Section 5.3.4. Again, the presence of both number and gender features meant that some measures had to be taken to prune the size of the resulting candidate graphs, by manually enumerating the possible mappings of non-concept features. The value of the *plasticity* parameter was raised from 1.0 to 4.0 to account for the larger number of constraints in CON, especially in the number of LEX constraints. *Plasticity decay* was set to 0.75, over four epochs. The reranking algorithm used was *Weighted Uncancelled*, as it is more compatible with learning in a stratified grammar.



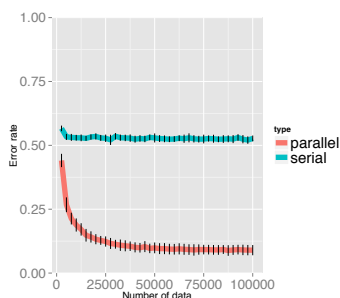


Figure 5.21: Average error rates over time for parallel and serial learners.

### 5.4.3 Simulation 2A: Learnability of serial and parallel grammars

In this first experiment on the PFC dataset, we will compare a stratified serial grammar with a completely parallel one. A similar comparison was made in Chapter 3. In the stratified grammar, the MSF and MF levels are located in the same stratum: as explained in Section 5.2.4, the separation of these two levels in the BiPhonX model is chiefly for explanatory reasons. Putting the evaluations for  $GEN_{SSF}$  and  $GEN_{MSF}$  in distinct strata would render a majority of candidates eternally suboptimal, since completely faithful, non-disagreeing sub-candidates would always win in the sub-evaluation from SSF to MSF. Other than this, each following sub-GEN was placed in a lower stratum than the previous for the serial grammar.

The parallel and serial learners were represented by two groups of 25 virtual learners each. They were fed 100,000 learning data from the data set described above, with the other hyperparameters as described in simulations 1A-1D.

#### Results

The average error rates of parallel and serial learners over the course of learning is shown in Figure 5.21. A very clear difference can be seen between the two types of learner. Serial learners were universally unable to deal with the data set: all of them showed only a small amount of improvement over the initial state of the grammar, and remained stuck at an error rate of over 50%.

Parallel learners, on the other hand, reached an average error rate of 8.9% after 100,000 learning data. This is a marked difference with the results of Chapter 3, where both serial and parallel grammars were capable of learning the input distribution, albeit with different hidden representations.

### Parallel grammar: results on liaison forms

Since the focus of this modeling study is on liaison, the error rate on the totality of forms is in itself not entirely informative; after all, only a small fraction of the data consist of potentially liaising phrases. Furthermore, given the variation present in the learning data, we can expect learners to produce forms that are not attested in our sample, but not strictly incorrect under a view of optional liaison. We are therefore not just interested in how well the grammars obtained by learners mirrored the full input distribution, but also their performance on the subset of the input data allowing potential liaison forms.

In order to assess the acquisition of liaison by our learners after training, a subset of the input distribution was taken whose phrases were marked in the PFC corpus as potential liaison sites. 10,000 items were drawn from this distribution and evaluated with the post-training grammars of each learner. The resulting pairs were classified according to these features of the Phonetic Form:

- attested in the training data
- “acceptable”, i.e. containing no spurious insertions or deletions except a canonical liaison consonant for that phrase
- exhibiting liaison

For example, a Phonetic Form [leopozã] would be marked as acceptable, not exhibiting liaison, and not attested, as only the liaising form [lezopozã] is found in the data; a form like [deztvatyr] would be marked as unacceptable, exhibiting liaison, and unattested. Forms marked acceptable but unattested were manually selected by the author. Table 5.9 shows the results of drawing 10,000 SSF forms from the ‘possibly liaising’ subset of the training distribution, for this simulation and that of the subsequent simulations/

	Attested	Acceptable	Liaising
2a	90.2% ±2.6	95.2% ±2.1	71.6% ±4.8
2b	89.2% ±8.3	94.1% ±8.1	70.7% ±8.0
2c	88.6% ±3.9	92.6% ±3.0	71.3% ±8.9
2d	78.9% ±5.4	94.5% ±2.1	60.6 % ±11.1

Table 5.9: Results on liaison items for simulations 2A-2D. Percentages represent averages over 25 virtual learners, with standard deviations.

These results indicate that learners perform reasonably well at replicating the liaison pattern found in the input data. Moreover, learners seem to regularize their knowledge to phrases unattested in the data, producing liaisons not attested in the training data.

### Analyses found by learners

As in Section 5.3, the post-training grammars can be tested on the input distribution not only to see whether they produce a correct AudF for a given SSF, but also to inspect what analysis learners prefer when training is completed. Of course, the analyses found for liaising forms are of particular interest here: do learners select a “gender-allomorphic”, lexical, phonological or phonetic solution, or some combination thereof?

Given the large number of possible liaison contexts in the input data, and the multitude of analyses possible for any meaning-form pair in our parallel MLCG (see also Chapter 4), analyzing all outcomes in detail is not feasible. We restrict ourselves to pointing out some trends in the analyses found by the virtual learners. Figure 5.22 shows a few sample results. These and other outcome patterns show that learners have an overwhelming preference for a *lexical* approach to liaison, where the liaison consonants are considered part of an allomorph of a noun. In such an analysis, for the phrase *leurs économies* [lœʁzɛkɔnɔmi], the Morpheme ⟨ECONOMY.F.PL⟩ maps to the Morph |zɛkɔnɔmi|. The LEX constraints deciding between these two forms are close in ranking value, and structural constraints may force the choice of one allomorph over the other in a given phonemic context. The phonological and phonetic rewrite rules that operate in GEN<sub>UF</sub> and GEN<sub>SF</sub> do not play any significant role in this analysis.

This lexical approach occasionally results in erroneous forms such as \*[lenã] *les (n)ans* or \*[œzɔʁdinatœʁ] *un (z)ordinateur*, where the liaising allomorphs surface in a phonologically inappropriate context. Such errors are actually consistent with L1 acquisition data of French – see for instance the examples cited in Wauquier-Gravelines and Braud (2005); Wauquier (2009). However, as these errors tend to vanish from the language of older speakers in real L1 acquisition, the lexical analyses that underlie them are usually considered part of a developmental stage. This stage eventually gives way one of more rule-based, generalized analyses (Chevrot et al., 2009; Wauquier, 2009). Our virtual learners do not seem to progress from a lexical analysis to a more general (phonological) one. Nevertheless, the results indicate that a lexical analysis of liaison can go a long way towards replicating the patterns found in the PFC corpus.

#### 5.4.4 Simulation 2B: liaison consonants as abstract segments

Section 5.2.5 mentioned a restriction on the lexical hypotheses generated by the learners in this study: only segments found in the phonetic surface data could be linked to a given morpheme. This restriction rules out “abstract” analyses of liaison, where underlying forms contain segments that may surface either as zero or as a liaising consonant /z/, /t/ or /n/.

To test the effect of allowing such analyses, a simulation was performed in which GEN is enriched allowed for abstract liaison segments. This was done

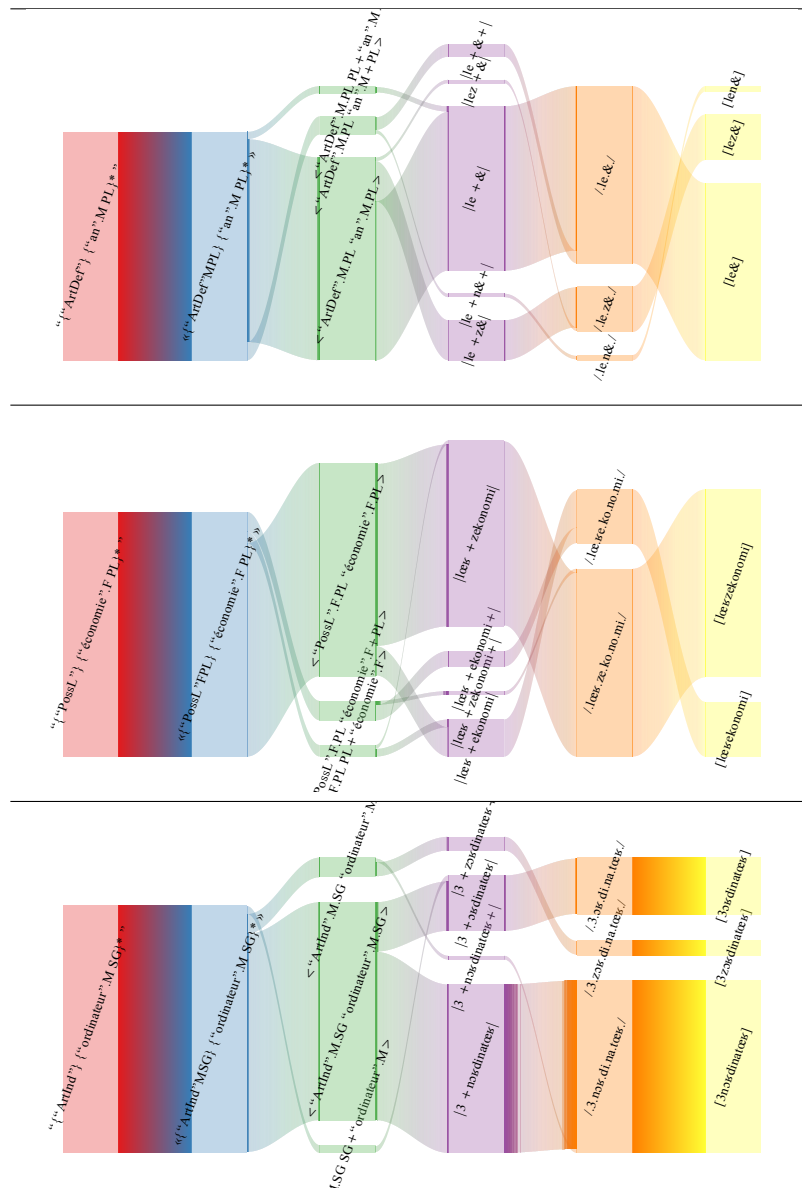


Figure 5.22: Sankey diagrams for some output candidates of parallel learners in Simulation 2A.

as follows. After creating the lexical repository on the basis of the input distribution, the sets of Morphs (partial Underling Forms) generated in the lexicon for each Morpheme are inspected. If any pair within this set differs only in the presence of a  $|z|$ ,  $|t|$  or  $|n|$  at the edge, an additional Morph is added to the set with an “abstract” variant  $|Z|$ ,  $|T|$  or  $|N|$  at that edge. Taking the example from Figure 5.12, if the Morpheme  $\langle Small_A \rangle$  contains mappings both to  $|p\text{ø}ti|$  and  $|p\text{ø}tit|$ , these two UFs differ only in the presence of a segment  $|t|$  at the right edge. The abstract segment generator will then add a new mapping  $\langle Small_A \rangle \rightarrow |p\text{ø}tiT|$  to the lexicon.

On  $GEN_{UF}$ , additional rewrite rules were added that force these abstract segments to be either deleted or realized as their non-abstract counterpart. These rewrite rules come with corresponding constraints, i.e.  $FAITH *|T| \rightarrow /t/$  and  $FAITH *|T| \rightarrow /\emptyset/$ . As  $GEN_{UF}$  forces the abstract segments to be deleted or realized, any Morph containing an abstract segment will incur a violation on one of these constraints in the mapping to SF. The effect of these additions to  $GEN$  is that learners may hypothesize certain morphemes to contain a “latent” consonant whose eventual manifestation is governed by phonological context. In such an analysis, *petit* would have a single Underlying Form  $|p\text{ø}tiT|$ . On the other hand a word like *net*, whose final consonant is always pronounced, should only have a Morph  $|n\text{et}|$  in the lexicon.

## Results

Inspecting the results (Table 5.9), it appears that the availability of ‘abstract’ allomorphs in the lexicon does not diminish learners’ preference for a lexical-allomorphic analysis. The error rates and candidates found for potentially liaising forms are similar to those of Simulation 2A, and Underlying Forms containing abstract segments are rarely utilized in the final grammars of learners.

### 5.4.5 Simulation 2C: merely-lexical grammars

The grammar framework of this chapter allows for liaison consonants to either appear or disappear at the lexical level, when mapping to Underlying Form, or as a result of phonological or phonetic processes, when mapping to Surface and Phonetic Forms. To tease apart these distinct loci for liaison, we ran a simulation where the phonological rewrite rules are “switched off”: the rule sets described in Section 5.2.5 were deactivated, so that only Structural constraints played an active role in the mapping from Underlying Form via Surface Form to Phonetic Form. This effectively means that virtual learners can only realize variation in liaison realization through allomorphy.

## Results

Simulations 2A and 2B demonstrated that parallel learners in this framework are predisposed toward a lexical analysis of liaison. It is therefore not surpris-

ing that learners with a merely-lexical grammar are still capable of reflecting the input data fairly well. Compared to the learners of Simulation 2A, quantitative and qualitative inspection of the results show a slight propensity toward producing “unacceptable” forms with spurious consonants inserted. Presumably this is because there are no phonological rules to delete these spurious consonants for the sake of well-formedness, giving less opportunity for repair. However, the small number of simulations performed means that these differences cannot be called significant.

#### 5.4.6 Simulation 2D: lexically limited grammars

As a final variant on the experiments with the PFC data set, we run a simulation where the lexical hypotheses about the nouns in the phrases are restricted to the “canonical” forms of these nouns – that is, without prefixed liaison consonants. This rules out the type of analyses often found by learners in experiments 2A–2C, where learners analyse plural forms of vowel-initial nouns as beginning with a  $|z|$ .

#### Results

Simulations 2A, 2B and 2C resulted in an overwhelming preference for lexically-based analyses, where liaison consonants were part of a nominal allomorph on Underlying Form. When this possibility is ruled out from the grammar, a slightly different picture emerges. In terms of the number of “acceptable” liaison forms produced, these learners perform similar to the previous variants; but they appear to be less inclined to reproducing forms attested in the input data, and also produce fewer liaising forms. However, due to the small number of simulations performed, these quantitative differences between types of forms produced cannot be called significant.

A closer, qualitative inspection of the candidates produced after learning reveals that these learners combine two approaches to liaison: allomorphy on the determiner (e.g.  $|lez|$  for  $\langle \text{ArtDef.PI} \rangle$ ) and, to a lesser extent, insertion of  $/z/$  in the mapping from Underlying Form to Surface Form. The latter is especially interesting: no other insertion rules are applied, making  $/z/$  a kind of default sandhi consonant for these learners. The fact that this strategy still results in the production of mostly “acceptable” forms indicates that liaison in our determiner-noun data mostly concerns  $/z/$ . Nevertheless,  $/z/$ -insertion sometimes leads to erroneous forms. These lexically limited learners are forced to adopt a general phonological rule for hiatus avoidance, but its limited applicability leads to a diminished production of liaison forms in favor of  $/V.V/$  constructions.

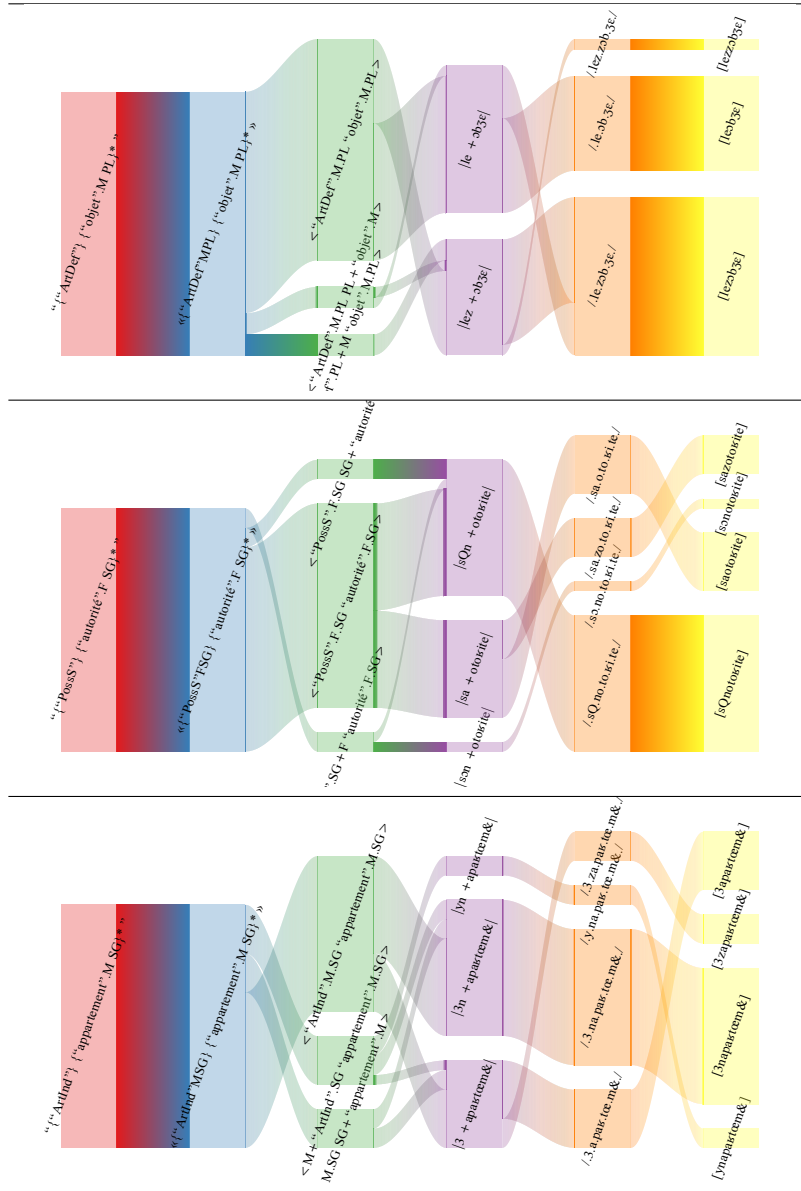


Figure 5.23: Sankey diagrams for some output candidates of learners in Simulation 2D.

## 5.5 Conclusion and discussion

### 5.5.1 Interpretation of simulation results

#### Simulation 1

In Section 5.3, learning simulations were performed on a number of meaning-form pair distributions representing a kind of minimal liaison data set, similar in size to that of Chapter 4. However, the procedural approach to multi-level GEN and CON resulted in much larger candidate and constraint sets. Results show that the efficient evaluation algorithm introduced in the previous chapter keeps learning feasible in these large grammars, although some restrictions to the generation procedure were necessary when more morphosyntactic features were employed. This is an encouraging result: the MLCG/EDRA simulation framework scales to larger and less teleologically biased candidate sets. A data set exhibiting phonetic variation was also learnable.

While serving mainly as a prelude to the corpus data of Simulation 2, some interesting observations can also be made on the basis of the results of Simulation 1. First, the enlarged hypothesis space does not harm the probability of convergence on a surface-correct grammar. On the contrary, generating more forms and constraints resulted in higher success rates than were obtained in Chapter 4. Those learners that fail to converge get stuck in a local optimum in the earliest stages of learning: they acquire a grammar that is able to reproduce some, but not all overt forms in the data.

#### Simulation 2

For the second series of simulations, a much larger input data set was created on the basis of the PFC corpus. Various grammars were generated on the basis of the meaning-form pairs in the input data, and compared for learnability. The results only showed a categorical difference in learnability for one comparison: that between a serial and a parallel grammar. Virtual learners equipped with a serial grammar invariably failed to acquire a ranking consistent with the input data. On the other hand, learners with a completely unstratified grammar did manage to reproduce the patterns in the input training data, both for items with and without liaison. Naturally, this does not constitute definitive proof for an interactive view of speech processing; it remains to be seen whether peculiarities of the representations, constraint set, learning algorithm, or learning data used made a serial analysis unreachable. Nevertheless, given the fact that the number of possible rankings is much larger when all constraints are placed in a single stratum, this result vindicates the feasibility of a parallel analysis in the BiPhon framework and affirms the power of error-driven constraint learning even over a very large, procedurally generated CON.

A closer inspection of the grammars obtained by successful parallel learn-



ers indicates an overwhelming preference for a “lexical”/allomorphic approach to liaison. This preference still holds when a more general abstract analysis with explicit “latent” consonants is available. Only in a variant that removes the possibility of nominal allomorphy do learners utilize a more phonological approach, at the cost of a slightly less faithful reproduction of the input data. As before, architectural choices may have influenced this outcome; in particular, the relatively straightforward phonological/phonetic levels of representation excluded many analyses that have been proposed (e.g. autosegmental analyses etcetera). In particular, the assumption that there exists an *identity* relation between segments on UF, SF and PF seems to favor analyses that faithfully map between these levels. We saw in Chapter 3 that learnability of certain configurations increased when we abandoned this assumption of identity.

Finally, the choice to restrict the analysis to determiner-noun phrases in the learning data may have obscured the generality of liaison, favoring a lexical analysis. The interaction with other phonological processes, in particular variability in the insertion and deletion of schwa, was not explored either.

For these and other reasons, the simulation results from this chapter fall short as a complete analysis of liaison in modern Standard French. Rather, their contribution is as follows. First, they show the feasibility of analysing complex phonological phenomena in multi-level constraint grammars, profiting from the efficient algorithm introduced in Chapter 4. On the basis of a large data set culled from the PFC corpus, a largely lexical analysis of liaison emerges as the most viable approach for representing its patterns and variation. This analysis aligns with some theoretical proposals from the literature, as described in Section 5.2.1. Second, this chapter demonstrated a methodology for exploring and selecting hypotheses about sound patterns. By procedurally generating hypotheses about hidden structures on the basis of corpus data, we reduce the risk of teleological bias, cherry-picking, or oversights. By measuring learnability as we add or remove components in the resulting grammars, we may evaluate which assumptions about hidden structures are necessary and sufficient to learn and reproduce the empirical data. This allows us to refine our models of speech perception and production. The next section sketches some approaches that could be taken on the basis of these results.

### 5.5.2 Suggestions for future research

Even within the confines of the six-level BiPhon framework used in this chapter, the possible approaches to analyzing liaison are far from exhausted. A marked advantage of the procedural candidate generation described above is that including more data should be a relatively small effort. Scaling up the data set, by including more speakers or widening the scope of the analysis to other types of phrases, should increase confidence in the generality of the results.

One particular piece of low-hanging fruit that suggests itself on the basis

of the results of this chapter is to embed the learners in an agent-based iterative learning simulation. With a small adaptation, namely using the output distribution of one learner as the input for another, the stability and learnability of certain liaison patterns over time could be tested. The iterative learning paradigm has been used both in computer simulations (de Boer, 2001; Lopopolo and Biró, 2011) and lab experiments (Verhoef et al., 2014), including several studies in the BiPhon framework (Boersma and Hamann, 2008; van Leussen, 2008).

The data-driven approach of this chapter is relatively “agnostic” about the locus of liaison in the grammar, but the candidate and constraint sets generated still fall short of comprising all representations and processes that have been proposed to play a role in liaison. As long as these analyses can be represented in a MLCG, the framework of this chapter can be adapted to investigate how these alternative grammars fare in learning and reproducing the patterns found in the data.

Alternatively, we may take the opposite approach: removing assumptions about the nature and structure of the hidden representations between meaning and (phonetic) form. The mapping from Semantic Form to Phonetic Form can be viewed as a kind of translation from one code to another. Recent advances in machine learning, in particular *deep learning* models, have achieved remarkable results on this kind of task. By training a sequence-to-sequence model (Sutskever et al., 2014) on the data of this chapter, we could learn to what extent our assumptions about intermediate (morphological and phonemic) representations are necessary, and which features in the input data are essential to learning the patterns. The downside of a deep learning approach is that it is much harder to interpret the intermediate representations acquired by the trained model.

Each of the above approaches suggests taking the training data and results of this chapter as a kind of benchmark. The methodology advocated by this chapter, and in this thesis as a whole, is to compare the relative merits of phonological models not by their intrinsic properties or ability to handle ad-hoc phenomena, but by measuring them against empirical data. A fair comparison between competing models of similar phonological phenomena should ideally use the same test data. For metrical phonology, there is a *de facto* benchmark in the form of Tesar and Smolensky (2000); but to my knowledge, no comparable test set exists for segmental phonology. This study has shown that data from a corpus like the PFC can serve as such a benchmark. With liaison as a testing ground for theories about sound patterns and phonological representations, this chapter evaluated various assumptions about these patterns and representations within the BiPhon framework. By using the same data set – or a variation or enrichment thereof – to test other computational models of phonological acquisition, this evaluation method could be expanded to make a comparison *between* frameworks.



## CHAPTER 6

---

### Conclusion

---

#### 6.1 Summary

Chapter 1 presented the aim that unites the case studies in this thesis: moving towards *whole-language simulation* in a model of phonological acquisition. Chapter 2 laid down a theoretical and computational basis for simulating learning in multi-level constraint grammars. This basis was then used to describe BiPhon, the Optimality-Theoretic framework in which the simulations in subsequent chapters are based.

#### Chapter 3

Chapter 3 applied this framework to second-language learning, in a revised version of Escudero 2005, 2009's L2LP framework (itself an application of BiPhon to second-language learning). In particular, the acquisition and perception of Spanish front vowels by native speakers of Dutch was simulated. Data from empirical studies were used to model the input of first-language acquisition and second-language perception. A previous computational implementation of this scenario in the L2LP model (Weiand, 2007) failed to computationally confirm the predictions of Escudero (2005). The revised model deviated from Weiand (2007)'s implementation in two ways. First, the notion of categorical "faithfulness" between BiPhon's UF and SF levels was replaced by a more gradient measure of similarity between features on these levels. Second, Jarosz (2013a)' resampling method was used to parse optimal learning candidates. With these two revisions, two versions of the model were tested:

a fully parallel one in which constraints over all levels of representation could interact, and a serial one where constraints on pre-lexical perception were evaluated prior to constraints that govern lexical recognition. Both the serial and the parallel versions of the revised L2LP borne out the predictions of Escudero (2005), with the L2 learners successfully transitioning from three to two front vowels. The two versions differed in the underlying representations used to arrive at this surface behaviour. Chapter 3 suggested some ways to tease apart these variant predictions in an empirical study, in order to assess the relative ability of either variant to account for real listeners' behaviour.

### Chapter 4

Chapter 4 introduced an efficient procedure for learning and evaluation in multi-level constraint grammars, that grows linearly (rather than exponentially) in the number of levels of representation. The procedure was illustrated graphically, as well as explained more formally. To test the algorithm, a small (toy-scale) simulation was carried out over a multi-level grammar of French liaison. Various variants of error-driven constraint learning were used to test the performance of the new method. In the context of this thesis, the chapter also served as a prelude to the scaled-up study presented in Chapter 5.

### Chapter 5

In Chapter 5, the efficient procedure laid out in the preceding chapter was put to use in a larger scale study of French liaison. The chapter used a BiPhon-based multi-level model based on that of Chapter 4, but introduced a method for generating candidates in this space on the basis of data from the PFC corpus (Durand and Lyche, 2003; Durand et al., 2009; Detey et al., 2016), with some (but not all) of these candidates manifesting liaison on the overt phonetic level. Virtual learners were then trained on the corpus-based data in order to see how these overt liaisons would be represented on the covert (lexical, morphological and/or phonological) levels of representation. As in Chapter 3, a comparison was made between a serial and parallel version of the model. For these corpus data, and unlike Chapter 3, the serial model was unable to replicate the patterns of the input data. The parallel version of the model, on the other hand, did successfully learn to reproduce the data. It turns out that these learners overwhelmingly preferred a *lexical* analysis of liaison, where liaising words come with two or more lexical variants, whose manifestation is conditioned by the phonological context.

## 6.2 Implications and limitations

In this section, the results of the simulation studies are given some further interpretation in light of the general aim of this thesis: investigating how to

bring multi-level models like BiPhon closer to “whole-language simulation”.

## 6.2.1 Implications

### Serial versus interactive processing

A recurring question in each of the simulation chapters is whether phonological processes should be seen as serial, with distinct components of the grammar serving as links in a chain of inputs and outputs, or as interactive, with constraints on both components influencing one another. The results do not give any definitive answer on this question, but offer some interesting insights on how this question can be pursued in multi-level grammars.

The L2 Spanish model of Chapter 3 allowed for both a serial and an interactive explanation, whereas Escudero (2005)'s original model was strictly serial. However, the two versions of the model gave slightly different interpretation of the role played by covert representations. Further empirical study, or simulations involving other scenarios that the L2LP model is designed to handle, should shed more light on the adequacy of either variant.

The French liaison studies of Chapters 4 and 5 differ somewhat in their prognostications on serial versus parallel processing. The second part of Chapter 4 tested the efficient model of evaluation and learning on a toy grammar of French liaison, and expressly included a number of analyses previously made in the literature. The model used was itself fully interactive, in that no restrictions were placed on the ordering of constraints acting over different levels of representation. Within this interactive model, the simulated learners that emerged successfully generally preferred a “serial” analysis similar to that made in the generative model of Dell (1973).

In Chapter 5, a more direct comparison was made: a fully interactive model of liaison production was compared with one where evaluation was serial over levels of representation. As stated, here only the parallel model was capable of learning to replicate the input data. This appears to contradict the finding of Chapter 4. A number of explanations can be proposed. First, the amount of data covered in Chapter 5 is much larger than that of the toy problem in Chapter 4. It is possible that a serial grammar can no longer account for the variety of forms manifesting liaison in this larger data set. If so, this would constitute a vindication of the “data-based” approach over modeling toy problems. Another possible cause may be that the level-by-level seriality attempted in Chapter 5 is too strict. There are more ways to “serialize” a grammar – compare the serial model of Escudero (2005) as implemented in Chapter 3, which divides four levels of representation into two serial stages of evaluation.

### **Resampling rankings and increasing candidate space**

One outcome supported by both the L2LP and the liaison studies is that the resampling method of Jarosz (2013a) had a positive impact on learning success, as defined by the probability that virtual learners converge on a “correct” constraint ranking that replicates the input data. Another, more subtle phenomenon also seems to have an effect: it was noted in Chapter 5 that the three-form toy grammars used in the first series of simulations therein were similar to those used in Chapter 4, but with a larger candidate space because grammars were automatically generated rather than hand-crafted. Perhaps unexpectedly, learners confronted with this enlarged candidate space were more likely to find a “successful” constraint ranking than with the similar toy grammar of Chapter 4. Taken together, these results seem to point to a cure for constraint grammars getting stuck in local minima: introducing more sources of randomness.

## **6.2.2 Limitations of the presented studies**

There are a number of ways in which the simulations performed in this thesis were limited. This section briefly highlights some of these limitations and how they might be overcome in follow-up research.

### **Limited range of parameter settings**

Any study based in simulation is necessarily limited – there is usually no end to the number of different parameter combinations that can be tested. Chapter 2 gave an overview of different constraint-based models; some, but not all, of these were used in the simulations of Chapters 3 to 5. The most exhaustive number of combinations was made in Chapter 4, but that study was still limited to Optimality Theoretic and Harmonic Grammar learners, ignoring other evaluation mechanisms (e.g. MaxEnt, Goldwater and Johnson 2003). Another avenue which remains mostly unexplored is the role of initial constraint rankings, such as those reflecting “markedness over faithfulness” (Gnanadesikan, 1996). Mostly these choices were made for the sake of brevity, since comparing different variants of multi-level constraint grammars was not the main focus of the study. It should be relatively easy to re-implement simulations over the same datasets and grammars using these different “flavors” of evaluation.

### **Limited view of acquisition process**

Chapters 3 and 5 made some efforts to increase realism in learners’ inputs: the former by basing the Auditory level inputs on empirical production data, the latter by algorithmically generating features on the Lexical and Underlying levels of representation based on the information available to real learners. However, the simulations in either study still gloss over many difficulties faced by real learners – in particular, by framing phonological categories

as already being in place, rather than built on the basis of the phonetic and contextual evidence available to real learners. I propose that the simulations presented in this thesis go some way towards increasing realism, but a theory of category formation in the BiPhon model remains elusive and should be the subject of further (modeling) studies. Some efforts in this direction have been undertaken by Boersma et al. (2003) and Boersma et al. (2018).

### Limited scope of phenomena studied

In each of the simulation chapters, linguistic phenomena are studied largely in isolation from that language as a whole. This is most obvious and deliberate in the toy liaison grammar of Chapter 4, but Chapter 5 likewise limits its scope to liaison, largely ignoring (for example) insertion/deletion of schwa and other salient phonological features of French. Chapter 3 studies the L2 acquisition of Spanish front vowels, without considering their place in the full five-vowel system that is to be acquired. Such limitations in scope are common in studies of phonological phenomena, but somewhat limit the claim that we are doing *whole-language simulation*. The efficient framework introduced in Chapter 4, combined with a grammar and constraint generation method like that proposed in Chapter 5, could conceivably be used to perform simulations on a larger number of data, to further enlarge the scope of these investigations.

## 6.3 Suggestions for future research

Taken as a whole, the simulation studies of Chapters 3 to 5 suggest a methodology of *data-driven* testing of phonological models, in particular complex ones such as can be represented in a grammar over multiple levels of representation. By computationally implementing formal models of morphology and phonology, their often complex predictions can be put to the test. Next, these implementations can be tested not only over a small set of data tailored to the phenomenon at hand, but a larger data set, collected from empirical production studies or corpora. This data-driven approach allows researchers to test the validity of their constraint-based formal models in a systematic way, in the well-established framework of learning by reranking constraint grammars. The methodology also allows testing several variants of a model (e.g. regarding initial constraint ranking or serial versus parallel interaction) to see which better aligns with the empirical data. The implementations should also allow predicting what will be done with novel data (unseen by virtual learners), yielding predictions that can be tested in empirical experiments with real speakers.

It was suggested at the end of Chapter 5 that formal linguistics might in this respect take a cue from the disciplines of machine learning and natural language processing. For various tasks of interest in these fields (parsing, sentiment analysis, co-reference resolution and many others), gold data sets



and competitions exist on which researchers from different institutions may test newly developed methods and algorithms. Relative performance on these data sets gives an objective benchmark for comparing various approaches to these problems. Of course, (formal) linguistics is not an engineering discipline, and improving performance on a given learning problem should not be in itself the end goal of phonological research. Nevertheless, the establishment of freely available standard problem sets in various areas of phonology may allow for more direct and fruitful comparisons of competing views on representation, computation and learning – including more general learning models not rooted in linguistics, such as deep neural networks. The same goes for sharing and reusing code between research groups. I hope that, as computational implementation of formal models grows more common in linguistics, such sharing of resources, code and goals will become commonplace.

## 6.4 Conclusion

This thesis has presented a general framework for multi-level constraint grammars, as well as a number of simulation studies within a particular MLCG, the BiPhon framework. These simulations have shed more light on some previously undecided properties of the framework, and presented a novel methodology for learning and evaluation in multi-level grammars on an unprecedented scale. More broadly, the studies in this thesis demonstrate the feasibility of using formal multi-level grammars on realistic data sets based on corpora, and show how this allows exploring the viability of variant theories within a constraint-based framework. This methodology allows for a tighter coupling of theory and data, bringing together the findings of empirical studies and formal linguistic theory.

## APPENDIX A

---

### List of minimal pairs used as target lexical items in Chapter 3

---

This list is identical to that used in Weiand (2007)

checa	'Czech.F'	chica	'girl'
checo	'Czech.M'	chico	'boy'
fecha	'date'	ficha	'token'
gres	'stoneware'	gris	'gray'
lega	'layman'	liga	'league'
lema	'motto'	lima	'file'
meca	'Mecca'	mica	'mica'
mesa	'table'	misa	'Mass'
memo	'fool'	mimo	'mime'
reto	'dare'	rito	'rite'
rezo	'prayer'	rizo	'curl'
veda	'prohibition'	vida	'life'
peso	'weight'	pisó	'floor'



---

## Bibliography

---

- Ågren, J. (1973). *Etude sur quelques liaisons facultatives dans le français de conversation radiophonique: fréquences et facteurs*. PhD thesis, Uppsala University.
- Anttila, A. (1997). Deriving variation from grammar. In Frans Hinskens, Roeland van Hout, L. W., editor, *Variation, change and phonological theory*, pages 35–68. John Benjamins Publishing, Amsterdam.
- Aoyama, K., Flege, J. E., Guion, S. G., Akahane-Yamada, R., and Yamada, T. (2004). Perceived phonetic dissimilarity and L2 speech learning: the case of Japanese /r/ and English /l/ and /r/. *Journal of Phonetics*, 32(2):233–250.
- Apoussidou, D. (2007). *The learnability of metrical phonology*. PhD thesis, University of Amsterdam.
- Apoussidou, D. and Boersma, P. (2004). Comparing two Optimality-Theoretic learning algorithms for Latin stress. In *Proceedings of the 23rd Western Coast Conference on Formal Linguistics*, pages 29–42.
- Ashley, K. C., Disch, L., Ford, D. C., MacSaveny, E., Parker, S., Unseth, C., Williams, A. M., Wong, R., and Yoder, B. (2010). How many constraints are there? A preliminary inventory of OT phonological constraints. *Graduate Institute of Applied Linguistics Occasional Papers in Applied Linguistics*, 9.
- Bane, M. and Riggle, J. (2012). Consequences of candidate omission. *Linguistic Inquiry*, 43(4):695–706.
- Bayles, A., Kaplan, A., and Kaplan, A. (2016). Inter-and intra-speaker variation in French schwa. *Glossa: a Journal of General Linguistics*, 1(1).
- Bellman, R. (1957). Dynamic programming. Technical report, Princeton, NJ.
- Benders, T., Escudero, P., and Sjerps, M. J. (2012). The interrelation between acoustic context effects and available response categories in speech sound

- categorization). *The Journal of the Acoustical Society of America*, 131(4):3079–3087.
- Bermúdez-Otero, R. (1999). *Constraint interaction in language change: quantity in English and Germanic*. PhD thesis, University of Manchester.
- Bermúdez-Otero, R. (2003). The acquisition of phonological opacity. In *Variation within Optimality Theory: proceedings of the Stockholm workshop on Variation within Optimality Theory*, pages 25–36.
- Berwick, R. C. and Niyogi, P. (1996). Learning from triggers. *Linguistic Inquiry*, 27(4):605–622.
- Best, C. T. (1995). A direct realist view of cross-language speech perception. In Strange, W., editor, *Speech perception and linguistic experience: Issues in cross-language research*, chapter 6, pages 171–204. Timonium, MD: York Press.
- Best, C. T. and Tyler, M. D. (2007). Nonnative and second-language speech perception: Commonalities and complementarities. In Bohn, O. S. and Munro, M. J., editors, *Language experience in second language speech learning: In honor of James Emil Flege*, pages 13–34. Amsterdam: John Benjamins.
- Biró, T. (2006). *Finding the right words: implementing Optimality Theory with simulated annealing*. PhD thesis, University of Groningen.
- Biró, T. (2013). Towards a robust interpretive parsing. *Journal of Logic, Language and Information*, 22(2):139–172.
- Biró, T. and Gervain, J. (2011). Optimality Theory as a general cognitive architecture. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 33.
- Biró, Tamás, T. (2011). Optimal religion: Optimality Theory accounts for ritual dynamics. In *Changing minds. Religion and cognition through the ages*, Groningen Studies in Cultural Change, 42, pages 155–191. Peeters: Leuven.
- Blaho, S. (2008). *The syntax of phonology: a radically substance-free approach*. PhD thesis, Center for Advanced Study in Theoretical Linguistics, Universitetet i Tromsø.
- Blutner, R., Hendriks, P., and de Hoop, H. (2003). A new hypothesis on compositionality. In Slezak, P. P., editor, *Proceedings of the Joint International Conference on Cognitive Science, Sydney, Australia, 13-17 July 2003*, pages 53–57. Sydney: University of New South Wales.
- Boersma, P. (1997). How we learn variation, optionality, and probability. In *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam*, volume 21, pages 43–58.

- Boersma, P. (1998). *Functional phonology*. Holland Academic Graphics, Den Haag.
- Boersma, P. (2001). Phonology-semantics interaction in OT, and its acquisition. In Robert Kirchner, W. W. and Pater, J., editors, *Papers in Experimental and Theoretical Linguistics*, volume 6, pages 24–35. Edmonton: University of Alberta.
- Boersma, P. (2003). Review of Bruce Tesar and Paul Smolensky (2000). Learnability in Optimality Theory. *Phonology*, 20(3):436–446.
- Boersma, P. (2007). Some listener-oriented accounts of h-aspiré in French. *Lingua*, 117(12):1989–2054.
- Boersma, P. (2008). Emergent ranking of faithfulness explains markedness and licensing by cue. *Rutgers Optimality Archive* 954.
- Boersma, P. (2009a). Cue constraints and their interactions in phonological perception and production. In Boersma, P. and Hamann, S., editors, *Phonology in Perception*, pages 55–110. Mouton De Gruyter.
- Boersma, P. (2009b). Some correct error-driven versions of the constraint demotion algorithm. *Linguistic Inquiry*, 40(4):667–686.
- Boersma, P. (2011). A programme for bidirectional phonology and phonetics and their acquisition and evolution. In Benz, A. and Mattausch, J., editors, *Bidirectional Optimality Theory*, pages 33–72. John Benjamins.
- Boersma, P., Benders, T., and Seinhorst, K. (2018). Neural network models for phonology and phonetics. (in revision).
- Boersma, P. and Escudero, P. (2008). Learning to perceive a smaller L2 vowel inventory: an Optimality Theory account. In Avery, P., Dresher, E., and Rice, K., editors, *Contrast in phonology: theory, perception, acquisition*, pages 271–301. Berlin: Mouton de Gruyter.
- Boersma, P., Escudero, P., and Hayes, R. (2003). Learning abstract phonological from auditory phonetic categories: An integrated model for the acquisition of language-specific sound categories. In *Proceedings of the 15th International Congress of Phonetic Sciences*, pages 1013–1016.
- Boersma, P. and Hamann, S. (2008). The evolution of auditory dispersion in bidirectional constraint grammars. *Phonology*, 25(02):217–270.
- Boersma, P. and Hamann, S. (2009). Loanword adaptation as first-language phonological perception. In Calabrese, A. and Wetzels, W., editors, *Loan phonology*, pages 11–58. Benjamins.
- Boersma, P. and Hayes, B. (2001). Empirical tests of the Gradual Learning Algorithm. *Linguistic Inquiry*, 32(1):45–86.

- Boersma, P. and Pater, J. (2016). Convergence properties of a gradual learning algorithm for Harmonic Grammar. In McCarthy, J. J. and Pater, J., editors, *Harmonic Grammar and Harmonic Serialism*, pages 389–434. Sheffield: Equinox.
- Bonami, O., Boyé, G., and Tseng, J. (2004). An integrated approach to French liaison. In *Proceedings of Formal Grammar*, pages 29–45.
- Bonami, O. and Henri, F. (2012). Predicting article agglutination in Mauritian. *Formal Approaches to Creole Studies*, 3.
- Bongaerts, T. (1999). Ultimate attainment in L2 pronunciation: the case of very advanced late L2 learners. In Birdsong, D., editor, *Second language acquisition and the critical period hypothesis*, pages 133–159. Mahwah, NJ: Lawrence Erlbaum.
- Bresnan, J. (2000). Optimal syntax. In *Optimality Theory: phonology, syntax and acquisition.*, pages 334–385. Oxford University Press.
- Broersma, M. (2005). Perception of familiar contrasts in unfamiliar positions. *The Journal of the Acoustical Society of America*, 117(6):3890–3901.
- Bundgaard-Nielsen, R. L., Best, C. T., and Tyler, M. D. (2011). Vocabulary size matters: the assimilation of second-language Australian English vowels to first-language Japanese vowel categories. *Applied Psycholinguistics*, 32(01):51–67.
- Bybee, J. (2001). Frequency effects on French liaison. *Typological Studies in Language*, 45:337–360.
- Chevrot, J.-P., Dugua, C., and Fayol, M. (2009). Liaison acquisition, word segmentation and construction in French: a usage-based account. *Journal of Child Language*, 36(03):557–596.
- Chevrot, J.-P. and Malderez, I. (1999). L'effet Buben: de la linguistique diachronique à l'approche cognitive (et retour). *Langue française*, pages 104–125.
- Chládková, K. (2009). Auditory cues determine allomorphy: vocalized and non-vocalized prepositions in Czech. Master's thesis, University of Amsterdam.
- Chládková, K., Escudero, P., and Boersma, P. (2011). Context-specific acoustic differences between Peruvian and Iberian Spanish vowels. *The Journal of the Acoustical Society of America*, 130:416.
- Chomsky, N. (1957). *Syntactic structures*. Mouton.
- Chomsky, N. and Halle, M. (1968). *The sound pattern of English*. Studies in Language. Harper & Row, New York.

- Colantoni, L., Steele, J., and Escudero, P. (2015). *Second language speech: theory and practice*. Cambridge: Cambridge University Press.
- Côté, M.-H. (2001). *Consonant cluster phonotactics: a perceptual approach*. PhD thesis, Massachusetts Institute of Technology.
- Curtin, S., Goad, H., and Pater, J. V. (1998). Phonological transfer and levels of representation: the perceptual acquisition of Thai voice and aspiration by English and French speakers. *Second Language Research*, 14(4):389–405.
- Cutler, A., Weber, A., and Otake, T. (2006). Asymmetric mapping from phonetic to lexical representations in second-language listening. *Journal of Phonetics*, 34(2):269–284.
- de Boer, B. (2001). *The origins of vowel systems*. Oxford University Press.
- De Swart, H. (2009). *Expression and interpretation of negation: an OT typology*, volume 77. Springer Science & Business Media.
- Dell, F. (1973). *Les règles et les sons*. Paris: Hermann.
- Dell, F. C. (1970). *Les règles phonologiques tardives et la morphologie dérivationnelle du français: topics in French phonology and derivational morphology*. PhD thesis, Massachusetts Institute of Technology, Cambridge.
- Demuth, K. (1995). Markedness and the development of prosodic structure. *NELS*, 25:13–25.
- Detey, S., Durand, J., Laks, B., and Lyche, C. (2016). *Varieties of spoken French*. Oxford University Press.
- Dijkstra, E. W. (1959). A note on two problems in connexion with graphs. *Numerische Mathematik*, 1(1):269–271.
- Durand, J., Laks, B., Calderone, B., and Tchobanov, A. (2011). Que savons-nous de la liaison aujourd’hui? *Langue française*, (1):103–135.
- Durand, J., Laks, B., and Lyche, C. (2009). Le projet PFC (phonologie du français contemporain): une source de données primaires structurées. In Durand, J., Laks, B., and Lyche, C., editors, *Phonologie, variation et accents du français*, pages 19–61. Hermès.
- Durand, J. and Lyche, C. (2003). Le projet ‘Phonologie du Français contemporain’ (PFC) et sa méthodologie. *Corpus et variation en phonologie du français*, pages 213–276.
- Durand, J. and Lyche, C. (2008). French liaison in the light of corpus data. *Journal of French Language Studies*, 18(1):33–66.



- Eisner, J. (1997). Efficient generation in Primitive Optimality Theory. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, pages 313–320. Association for Computational Linguistics.
- Ellison, T. (1994). Phonological derivation in Optimality Theory. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING)*.
- Elvin, J., Escudero, P., and Vasiliev, P. (2014). Spanish is better than English for discriminating Portuguese vowels: acoustic similarity versus vowel inventory size. *Frontiers in Psychology*, 5:1188.
- Encrevé, P. (1988). *La liaison avec et sans enchaînement: Phonologie tridimensionnelle et usages du français.*, volume 15. Seuil, Paris.
- Encrevé-Lambert, M.-H. (1971). A propos de l'élision en français. *Rééducation orthophonique*, 60:245–251.
- Escudero, P. (2005). *Linguistic Perception and Second Language Acquisition. Explaining the attainment of optimal phonological categorization.* PhD thesis, Utrecht University.
- Escudero, P. (2009). Linguistic perception of SIMILAR L2 sounds. In Boersma, P. and Hamann, S., editors, *Phonology in Perception*, pages 151–190. Mouton de Gruyter.
- Escudero, P. and Boersma, P. (2002). The Subset problem in L2 perceptual development: multiple-category assimilation by Dutch learners of Spanish. In *Proceedings of the 26th Annual Boston University Conference on Language Development*, pages 208–219.
- Escudero, P. and Boersma, P. (2004). Bridging the gap between L2 speech perception research and phonological theory. *Studies in Second Language Acquisition*, 26(04):551–585.
- Escudero, P., Hayes-Harb, R., and Mitterer, H. (2008). Novel second-language words and asymmetric lexical access. *Journal of Phonetics*, 36(2):345–360.
- Escudero, P., Simon, E., and Mitterer, H. (2012). Dialectal differences in vowel production lead to differences in cross-language and L2 perception: Dutch and Flemish learners of English vowels. *Journal of Phonetics*, 40:280–288.
- Escudero, P., Sisinni, B., and Grimaldi, M. (2014). The effect of vowel inventory and acoustic properties in Salento Italian learners of Southern British English vowels. *The Journal of the Acoustical Society of America*, 135(3):1577–1584.
- Eychenne, J. (2006). *Aspects de la phonologie du schwa dans le français contemporain.* PhD thesis, Toulouse: Université de Toulouse-Le Mirail.

- Eychenne, J. (2011). La liaison en français et la théorie de l'optimalité. *Langue française*, (1):79–101.
- Flege, J. E. (1995). Second language speech learning: theory, findings, and problems. In Strange, W., editor, *Speech Perception and Linguistic Experience: Issues in Cross-language Research*, pages 233–277. Timonium, MD: York Press.
- Flege, J. E., Bohn, O.-S., and Jang, S. (1997). Effects of experience on non-native speakers' production and perception of English vowels. *Journal of Phonetics*, 25(4):437–470.
- Flege, J. E., Schirru, C., and MacKay, I. R. (2003). Interaction between the native and second language phonetic subsystems. *Speech Communication*, 40(4):467–491.
- Ford, L. R. (1956). Network flow theory. Technical report, Rand Corporation, Santa Monica CA.
- Gabelentz, G. v. d. (1901). *Die Sprachwissenschaft; ihre Aufgaben, Methoden und bisherigen Ergebnisse*. Leipzig: Tauchnitz.
- Gaskell, M. G. and Marslen-Wilson, W. D. (1997). Integrating form and meaning: a distributed model of speech perception. *Language and Cognitive Processes*, 12(5-6):613–656.
- Gibson, E. and Wexler, K. (1994). Triggers. *Linguistic Inquiry*, 25(3):407–454.
- Gnanadesikan, A. E. (1996). Child phonology in Optimality Theory: Ranking markedness and faithfulness constraints. In *Proceedings of the 20th Annual Boston University Conference on Language Development*, volume 1, pages 237–248. Cascadilla Press Somerville, MA.
- Gnanadesikan, A. E. (1997). *Phonology with ternary scales*. PhD thesis, University of Massachusetts, Amherst.
- Goldwater, S. and Johnson, M. (2003). Learning OT constraint rankings using a maximum entropy model. In Jennifer Spenader, A. E. and Dahl, Ö., editors, *Proceedings of the Stockholm workshop on variation within Optimality Theory*, pages 111–120. Stockholm: Stockholm University, Department of Linguistics.
- Gougenheim, G. (1935). *Éléments de phonologie française: étude descriptive des sons du français au point de vue fonctionnel*. Publications de la Faculté des Lettres de l'Université de Strasbourg.
- Grant, A. P. and Guillemin, D. (2012). The complex of creole typological features: the case of Mauritian Creole. *Journal of Pidgin and Creole Languages*, 27(1):48–104.

- Grosjean, F. (2000). The bilingual's language modes. In Nicol, J., editor, *One mind, two languages: bilingual language processing*, pages 1–22. Oxford: Blackwell.
- Hale, M. and Reiss, C. (2000). "Substance abuse" and "dysfunctionalism": current trends in phonology. *Linguistic Inquiry*, 31(1):157–169.
- Hayes-Harb, R. and Masuda, K. (2008). Development of the ability to lexically encode novel second language phonemic contrasts. *Second Language Research*, 24(1):5–33.
- Hengeveld, K. and Mackenzie, J. L. (2008). *Functional Discourse Grammar: A typologically-based theory of language structure*. Oxford University Press.
- Ito, J. and Mester, A. (2009). Lexical classes in phonology. In *The Oxford Handbook of Japanese Linguistics*. Oxford University Press.
- Jackendoff, R. (1997). *The architecture of the language faculty*. Number 28. MIT Press.
- Jarosz, G. (2013a). Learning with hidden structure in Optimality Theory and Harmonic Grammar: beyond Robust Interpretive Parsing. *Phonology*, 30(01):27–71.
- Jarosz, G. (2013b). Naive parameter learning for Optimality Theory – The hidden structure problem. In *Proceedings of the 40th Annual Meeting of the North East Linguistic Society*.
- Jones, D. (2003). The generative psychology of kinship: part 2. Generating variation from universal building blocks with Optimality Theory. *Evolution and Human Behavior*, 24(5):320–350.
- Jusczyk, P. W. and Aslin, R. N. (1995). Infants' detection of the sound patterns of words in fluent speech. *Cognitive Psychology*, 29(1):1–23.
- Kager, R. (1999). *Optimality Theory*. Cambridge University Press.
- Kahn, D. (1976). *Syllable based generalizations in English phonology*. PhD thesis, Massachusetts Institute of Technology.
- Keller, F. (2000). *Gradience in grammar: experimental and computational aspects of degrees of grammaticality*. PhD thesis, University of Edinburgh.
- Kibort, A. (2007). What are morphosyntactic features. <http://www.surrey.ac.uk/lis/smg/morphosyntacticfeatures.html>.
- Kiparsky, P. (1982). From cyclic phonology to lexical phonology. In van der Hulst, H. and Smith, N., editors, *The structure of phonological representations*, pages 131–175. Dordrecht: Foris.

- Kiparsky, P. (2000). Opacity and cyclicity. *The Linguistic Review*, 17(2-4):351–366.
- Kirchner, R. (1995). Going the distance: synchronic chain shifts in Optimality Theory. *Ms., Rutgers Optimality Archive*.
- Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, 220(4598):671–680.
- Laks, B. (2005). La liaison et l’illusion. *Langages*, (2):101–125.
- Legendre, G., Grimshaw, J., and Vikner, S. (2001). *Optimality-theoretic syntax*. MIT Press.
- Legendre, G., Miyata, Y., and Smolensky, P. (1990). 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*. Cambridge, MA: Lawrence Erlbaum.
- Lopopolo, A. and Biró, T. (2011). Language change and SA-OT. The case of sentential negation. *Computational Linguistics in the Netherlands Journal*, 1:21–40.
- Magri, G. (2012). Convergence of error-driven ranking algorithms. *Phonology*, 29(02):213–269.
- Martinet, A. (1955). *Economie des changements phonétiques. Traité de phonologie diachronique*. Bern: Francke.
- Maye, J., Werker, J., and Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, 82(3):101–111.
- Mayr, R. and Escudero, P. (2010). Explaining individual variation in L2 perception: rounded vowels in English learners of German. *Bilingualism: Language and Cognition*, 13(03):279–297.
- McCarthy, J. J. (2000). Harmonic serialism and parallelism. In *Proceedings of the North East Linguistics Society*, volume 40.
- McClelland, J. L. and Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, 18(1):1–86.
- McClelland, J. L., Mirman, D., and Holt, L. L. (2006). Are there interactive processes in speech perception? *Trends in Cognitive Sciences*, 10(8):363–369.
- McQueen, J. M., Norris, D., and Cutler, A. (2006). Are there really interactive processes in speech perception? *Trends in Cognitive Sciences*, 10(12):533–533.
- Mohanan, K. P. (1981). *Lexical phonology*. PhD thesis, Massachusetts Institute of Technology.

- Moreton, E. and Smolensky, P. (2002). Typological consequences of local constraint conjunction. In Mikkelsen, L. and Potts, C., editors, *Proceedings of the Twenty-First West Coast Conference on Formal Linguistics.*, pages 306–319. Cascadilla Press.
- Morin, Y.-C. (2003). Remarks on prenominal liaison consonants in French. *Living on the Edge*, 28:385–400.
- Morin, Y.-C. and Kaye, J. D. (1982). The syntactic bases for French liaison. *Journal of Linguistics*, 18(2):291–330.
- New, B., Pallier, C., Brysbaert, M., and Ferrand, L. (2004). Lexique 2: a new French lexical database. *Behavior Research Methods, Instruments, & Computers*, 36(3):516–524.
- Nichols, J. (1986). Head-marking and dependent-marking grammar. *Language*, pages 56–119.
- Norris, D. (1994). Shortlist: a connectionist model of continuous speech recognition. *Cognition*, 52(3):189–234.
- Norris, D., McQueen, J., and Cutler, A. (2000). Merging information in speech recognition: feedback is never necessary. *Behavioral and Brain Sciences*, 23(03):299–325.
- Ohala, J. J. (1983). The origin of sound patterns in vocal tract constraints. In Macneilage, P., editor, *The production of speech*, pages 189–216. Springer.
- Orgun, C. O. (1995). Correspondence and identity constraints in two-level Optimality Theory. Technical Report [ROA-62-0000], University of California, Berkeley.
- Passy, P. (1890). *Étude sur les changements phonétiques et leurs caractères généraux*. Paris: Firmin-Didot.
- Pater, J. (2009). Weighted constraints in generative linguistics. *Cognitive Science*, 33(6):999–1035.
- Pater, J. (2016). Universal grammar with weighted constraints. In McCarthy, J. J. and Pater, J., editors, *Harmonic Grammar and Harmonic Serialism*, pages 1–46. London: Equinox Press.
- Pelucchi, B., Hay, J. F., and Saffran, J. R. (2009). Statistical learning in a natural language by 8-month-old infants. *Child Development*, pages 674–685.
- Polka, L. and Werker, J. F. (1994). Developmental changes in perception of nonnative vowel contrasts. *Journal of Experimental Psychology: Human Perception and Performance*, 20(2):421.

- Prince, A. and Smolensky, P. (1993). Optimality Theory: constraint interaction in generative grammar. Manuscript, Rutgers University and University of Colorado at Boulder. Available at Rutgers Optimality Archive.
- Rauber, A. S., Escudero, P., Bion, R. A. H., Baptista, B. O., et al. (2005). The interrelation between the perception and production of English vowels by native speakers of Brazilian Portuguese. In *Proceedings of Interspeech*, pages 2913–2916.
- Riggle, J. (2009). Violation semirings in Optimality Theory. *Research on Language and Computation*, 7:1–12.
- Riggle, J. A. (2004). *Generation, recognition, and learning in finite state Optimality Theory*. PhD thesis, University of California, Los Angeles.
- Sag, I. A. and Pollard, C. J. (1987). *Head-driven phrase structure grammar*. Morgan Kaufmann.
- Samek-Lodovici, V. (1992). Universal constraints and morphological gemination: a crosslinguistic study. Ms., Brandeis University.
- Samek-Lodovici, V. and Prince, A. (1999). Optima. *Rutgers Optimality Archive*, 363.
- Saussure, F. d. (1916). *Cours de linguistique générale*. Paris: Payot.
- Schane, S. (1968). *French phonology and morphology*. Cambridge, MA: MIT Press.
- Selkirk, E. (1974). French liaison and the X notation. *Linguistic Inquiry*, pages 573–590.
- Selkirk, E. (1978). The French foot: on the status of mute e. *Studies in French Linguistics*, 1(2):141–150.
- Selkirk, E. O. (1972). *The phrase phonology of English and French*. PhD thesis, Massachusetts Institute of Technology.
- Smolensky, P. (1995). On the internal structure of the constraint component Con of UG. Ms., University of California, Los Angeles.
- Smolensky, P. (1996). On the comprehension/production dilemma in child language. *Linguistic Inquiry*, 27(4):720–731.
- Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, volume 27, pages 3104–3112.
- Tesar, B. (1995). *Computational Optimality Theory*. PhD thesis, University of Colorado.

- Tesar, B. and Smolensky, P. (1995). The learnability of Optimality Theory: an algorithm and some basic complexity results. Manuscript.
- Tesar, B. and Smolensky, P. (1998). Learnability in Optimality Theory. *Linguistic Inquiry*, 29(2):229–268.
- Tesar, B. and Smolensky, P. (2000). *Learnability in Optimality Theory*. Cambridge, MA: MIT Press.
- Tranel, B. (1987). *The sounds of French: an introduction*. Cambridge University Press.
- Tranel, B. (1996). French liaison and elision revisited: A unified account within Optimality Theory. *Aspects of Romance Linguistics*, pages 433–455.
- Tyler, M. D., Best, C. T., Faber, A., and Levitt, A. G. (2014). Perceptual assimilation and discrimination of non-native vowel contrasts. *Phonetica*, 71(1):4–21.
- van Leussen, J., Williams, D., and Escudero, P. (2011). Acoustic properties of Dutch steady-state vowels: Contextual effects and a comparison with previous studies. In *Proceedings of the 17th International Congress of Phonetic Sciences*.
- van Leussen, J.-W. (2008). Emergent optimal vowel systems. Master's thesis, University of Amsterdam. Available as ROA-1006 at <http://roa.rutgers.edu>.
- Verhoef, T., Kirby, S., and de Boer, B. (2014). Emergence of combinatorial structure and economy through iterated learning with continuous acoustic signals. *Journal of Phonetics*, 43:57–68.
- Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2):260–269.
- Wauquier, S. (2009). Acquisition de la liaison en L1 et L2: stratégies phonologiques ou lexicales? *Aile... Lia 2*, pages 93–130.
- Wauquier-Gravelines, S. and Braud, V. (2005). Proto-déterminant et acquisition de la liaison obligatoire en français. *Langages*, (2):53–65.
- Weber, A. and Cutler, A. (2004). Lexical competition in non-native spoken-word recognition. *Journal of Memory and Language*, 50(1):1–25.
- Weiand, K. (2007). Implementing Escudero's model for the SUBSET problem. *Rutgers Optimality Archive 913*.
- Werker, J. F. and Curtin, S. (2005). PRIMIR: a developmental framework of infant speech processing. *Language Learning and Development*, 1(2):197–234.

- Werker, J. F., Fennell, C. T., Corcoran, K. M., and Stager, C. L. (2002). Infants' ability to learn phonetically similar words: effects of age and vocabulary size. *Infancy*, 3(1):1-30.
- Werker, J. F. and Tees, R. C. (1984). Cross-language speech perception: evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development*, 7(1):49-63.





### **The emergence of French phonology**

This thesis presents a computational framework for modeling the acquisition of sound patterns. It argues that in order to model this acquisition, not just phonological but also extra-phonological levels are necessary. To build on previous literature of phonological acquisition, especially in Optimality Theory, the framework is *constraint-based*. The thesis then presents a number of studies that utilize one such multi-level constraint-based framework, Bidirectional Phonetics and Phonology, to simulate a number of scenarios in both first and second language acquisition.

The first simulation presented in the thesis concerns the acquisition of Spanish as a second language by native speakers of Dutch, focusing on how these Dutch learners acquire the Spanish front vowels /i/ and /e/. This process is modelled with a revised version of the Second Language Linguistic Perception model, which improves on an earlier iteration of that model. Two versions of the revised model were tested: a parallel one in all levels of representation could interact, and a serial one where pre-lexical perception precedes lexical recognition. Both the parallel and serial variant were able to model the gradual acquisition of Spanish-like perception of the two front vowels. However, the two versions of the model make different predictions of how these vowels are represented in a learner's grammar. Further perception studies may elucidate whether the parallel or serial grammar more accurately describes the acquisition process.

The second simulation study describes an efficient method to evaluate and learn in a multi-level constraint-based framework, and illustrates this method by modeling the first-language acquisition of a small set of forms that concisely illustrate the phenomenon of *liaison* in French. This efficient method is then used to advantage in the final study described in this work, where the acquisition of *liaison* is tested on a large dataset whose forms are taken from a large corpus of spoken contemporary French. A lexical solution to li-

aison emerges from these simulations, where learners entertain different allomorphs for a given lexical item and choose the one which best fits the phonological context. Such an analysis has been proposed before in the theoretical literature, and contrasts with other analyses which explain liaison in mostly phonological terms.

This thesis, then, demonstrates the viability of a multi-level and data-driven approach to phonology. By training models on large datasets, we can compare competing explanations for a given phonological phenomenon, on the basis of learnability and compatibility with empirical data.

---

## Samenvatting

---

### The emergence of French phonology

Dit proefschrift presenteert een computationeel kader om de verwerving van spraakklanken te modelleren. Het betoogt dat hiervoor niet alleen fonologische, maar ook niet-fonologische niveaus van representatie noodzakelijk zijn. Om aan te sluiten op de bestaande literatuur, met name waar die de Optimaliteitstheorie gebruikt, is gekozen voor een analyse in termen van *constraints* (beperkingen). Vervolgens wordt een specifiek constraintgebaseerd *multilevel* model, te weten Bidirectional Phonetics and Phonology, toegepast om een aantal scenario's in de eerste- en tweedetaalverwerving te bestuderen door middel van computersimulaties.

De eerste studie in dit proefschrift betreft het leren van Spaans als tweede taal door moedertaalsprekers van het Nederlands, en spitst zich toe op de verwerving van de Spaanse voorklinkers /i/ en /e/ door deze leerders. Om het verwervingsproces te modelleren wordt een herziene versie van het Second Language Linguistic Perception-model gebruikt. Van dit herziene model worden twee varianten getest: een parallel model waarin alle niveaus van representatie met elkaar in interactie zijn, en een serieel model waarin prelexicale perceptie voorafgaat aan lexicale herkenning. Zowel de parallelle als de seriële variant kunnen de geleidelijke verwerving van een "Spaans-achtige" perceptie van de twee klinkers modelleren. De beide varianten doen echter verschillende voorspellingen over de uiteindelijke representaties in de fonologische grammatica van de taalleerder. Nader empirisch onderzoek zou kunnen verhelderen welke van de twee varianten het verwervingsproces getrouwer beschrijft.

Voor de tweede studie wordt een methode beschreven om efficiënt te leren en evalueren in een constraintgebaseerd model met meer dan twee niveaus. Ter illustratie wordt een verschijnsel in de Franse fonologie, *liaison*, gemodelleerd aan de hand van een klein aantal voorbeelden. De efficiënte methode wordt vervolgens ten volle benut in de derde en laatste simulatie van het

proefschrift, door de verwerving van *liaison* te modelleren aan de hand van een omvangrijke dataset die onttrokken is aan een corpus van het hedendaagse gesproken Frans. Uit deze simulaties komt een *lexicale* analyse van *liaison* naar voren. In zo'n analyse hanteren leeders verscheidene allomorfen voor een lexicaal item, en kiezen zij hieruit de meest geschikte op basis van de fonologische context. Dergelijke analyses zijn bekend uit de theoretische literatuur, en contrasteren met andere theorieën die *liaison* primair vanuit de fonologie verklaren.

De studies in dit proefschrift pleiten derhalve voor een multi-level, datagestuurde benadering in de fonologie. Het trainen van computermodellen op grote datasets levert een instrument waarmee concurrerende analyses vergeleken kunnen worden, op basis van hun leerbaarheid en overeenstemming met empirische gegevens.