Review of: Bruce Tesar & Paul Smolensky, *Learnability in Optimality Theory*. The MIT Press, Cambridge, Massachusetts, 2000.

1. In recent years, the linguistic framework of optimality theory (OT) became of lively interest not only among phonologists. Students of morphology, syntax and natural language interpretation likewise became sensitive to the opportunities and challenges of the new framework. The reasons for this growing interest in OT are empirical and conceptual. First, it proved that a series of empirical generalizations and observed phenomena can be expressed very naturally within this framework; this especially holds for phonology where in-depth analyses of many languages have given a much better insight into cross-linguistic tendencies than we had before the invention of OT. Second, and perhaps much more important in linking scientists into a new research paradigm, there are the conceptual reasons, which are manifold in the present case: (i) the aim to decrease the gap between competence and performance, (ii) interest in an architecture that is closer to neural networks than to the standard symbolist architecture, (iii) the aim to overcome the gap between probabilistic models of language / speech and standard symbolist models, (iv) the problem of learning hidden structure and the logical problem of language acquisition.

The present book addresses the latter point, effectively concentrating on the issue of learnability. In this way, it tries to develop a rather strong argument favoring OT over alternative linguistic frameworks such as the Principle and Parameter model (PP). This is not a book arguing with an impressive bundle of empirical data. Most examples are used only demonstratively and aim to illustrate the main ideas. For the sake of completeness, they correspond to the basic CV syllable theory (Sections 2, 3, 7 & 8), the distribution of clausal subjects (Section 2 & 3), a basic metrical stress grammar (Section 4), German syllable-final devoicing (Section 5). Although there may be good reasons against writing a linguistic book of this kind, in the present case this strategy proves compelling and powerful: it allows them to present the conceptual points and theoretical ideas in a fairly condensed, rigorous, and authoritative way (in only 140 pages, where 12 pages are allotted to endnotes, references and index.)

I must admit that I'm a little confused by the organization of the book. The authors seem to be aware of that problem: they give a large preamble (Section 1) which explains what the book is about, what the larger context of the work is, and how to read the book. However, consistently having to refer back to this excellent outline is a suboptimal solution for this organizational problem. *Learnability in Optimality Theory* is based on some general merits

that immediately arise from the basic principles of OT. These merits concern the idea of robust parsing, the idea of constraint demotion and the inclusion of these two ideas in a bootstrap mechanism for learning hidden structure. This review will be organized around these central ideas, deviating a bit from the organization of the book.

2. OT respects the generative legacy in two important methodological aspects: the strong emphasis on formal precision in grammatical analysis and a striving towards restricting the descriptive power of linguistic theory. In the present book, both aspects are developed and realized in a way that highlights the main advantages of OT over PP models.

Seeing themselves within the Generative tradition, most representatives of OT adopt the fundamental distinction between Universal Grammar (UG) and a language-specific part of Grammar. UG describes the innate knowledge of language that is shared by normal humans, and aims both to describe the universal properties of language and the range of variation possible among languages. The language-specific part of Grammar typically consists of the lexicon and a system reflecting the structural specifics of the particular language. Within the generative tradition, the concrete theoretical realization of this distinction has changed over the years. In the PP model, for example, UG is conceptualized as a system of (inviolable) principles which are parameterized to demarcate the space of possible forms. The fixing of these parameters (triggered by specific language data) determines the grammar of the particular language. OT realizes an essentially different view of this distinction.

However, before we can consider the overall aims of this book, we must first give some background on the nature of OT. In this section, I follow the exposition given in the book's second section that provides a concise overview of OT.

Not unlike other models of grammars, OT sees a grammar as specifying a function that assigns to each input (underlying representation of some kind) a structural description or output. For example, in the basic CV syllable theory, an input is a string of **C**onsonants and **V**ocals, such as

(1) /VCVC/

An output is a parse of the string into syllables. Examples are

(2) a. .V.CVC. an onsetless open syllable followed by a closed syllable

b. ⁺V‚.CV.⁺C‚ one open syllable; the initial V and final C are not parsed into syllable structure; this is indicated by ⁺ ‚

c. .□V.CV.⁺C‚ a sequence of two open syllables. The onset of the first syllable is unfilled (notated □). Phonetically, this is realized as an epenthetic consonant.

The other case in point is Grimshaw and Samek-Lodovici's theory of the distribution of clausal subjects (e.g. Grimshaw & Samek-Lodovici 1995). Here an input is a lexical head with a mapping of its argument structure into other lexical heads, plus a tense specification. The input also specifies which arguments are foci and which arguments are coreferent with the topic. An example is

(3) <sing(x), x=topic, x=he; T=pres perf>

It represents the predicate *sing*, with a pronominal argument that is the current discourse topic. A possible output is an X-bar structure realizing an extended projection of the lexical head. Examples are

(4) a. [$_{IP}$ has [ sung]] a clause with no subject

b. [$_{IP}$ he$_i$ has [t$_i$ sung]] a clause with subject *he*, co-indexed with a trace in SpecVP

c. [$_{IP}$ has [t$_i$ sung] he$_i$ ]] *he* right-adjoined to VP, co-indexed with a trace in SpecVP

The general idea of standard versions of generative phonology / syntax is to define the acceptable (grammatical) input-output pairs via a system of rules and transformations. In order to restrict the descriptive power of linguistic theory, the role of constraints is added. All of these constraints have been viewed as inviolable within the relevant domain. The inviolable constraints have themselves proved to be problematic and this has led to the "parameterization" of certain constraints, with one parametric setting for one language and another parametric setting for another language.

In OT the "generative part" of the Grammar is reduced to a universal function *Gen* that, given any input *I*, generates the set *Gen*(*I*) of candidate structural descriptions for *I*. The central idea of OT is to give up the inviolability of constraints and to consider a set *Con* of

violable constraints. Furthermore, a strict ranking relation o is defined on *Con*. This relation makes it possible to evaluate the candidate structural descriptions in terms of the total severity of the violations they commit, as determined by the ranking of the constraints. If one constraint $C_1$ outranks certain constraints $C_2$, ..., $C_i$ (written $C_1$ o {$C_2$, ..., $C_i$}), then *one* violation of $C_1$ counts more than as many violations of $C_2$, ..., $C_i$ as you like. The evaluation component selects the optimal (least offending, most harmonic) candidate(s) from the set *Gen*(*I*). The grammar favors the competitor that best satisfies the constraints. Only an optimal output is taken as an appropriate (grammatical) output; all suboptimal outputs are taken as ungrammatical. This idea makes the grammaticality of a linguistic object dependent on the existence of a competitor that better satisfies the constraints.

Constraints are of two different kinds: *markedness constraints* that affect outputs only and *faithfulness constraints* that relate to the similarity between input and output. The main representatives of the faithfulness family are (i) PARSE prohibiting *underparsing* ("underlying input material is parsed into output structure") and (ii) FILL exhibiting *overparsing* ("the elements of the output must be linked with correspondents in the underlying input"). In the case of OT-syntax the corresponding constraint is called FULL-INT(ERPRETATION): "the elements of the output must be interpreted." Markedness constraints are inherently connected with the domain under discussion. By way of example, we consider two constraints in the case of the basic CV syllable theory:

(5) a. ONSET "syllables have onsets"
b. NOCODA "syllables do not have codas"

In the case of OT syntax (distribution of clausal subjects) we have to consider two other constraints:

(6) a. SUBJ "the highest A-specifier in an extended projection must be filled"
b. DROP-TOPIC "arguments coreferent with the topic are structurally unrealized"

To complete this short introduction into OT, let's consider two typical OT tableaux relating to the input-output pairings (1) & (2) and (3) & (4), respectively. In the first example, the following constraint hierarchy is assumed:

(7) ONSET o NOCODA o $\text{FILL}^{\text{NUC}}$ o PARSE o $\text{FILL}^{\text{ONS}}$

The corresponding constraint tableau (8) shows the competing candidates in the left column. The other columns are for the constraints, each indicated by the label at the top of the column (with decreasing rank from left to right.) Constraint violations are indicated with asterisk, one for each violation.

(8)

| /VCVC/ | ONSET | NOCODA | FILL$^{NUC}$ | PARSE | FILL$^{ONS}$ |
|---|---|---|---|---|---|
| (a) .V.CVC. | * | * | | | |
| (b) +V$_{\prime}$.CV.+C$_{\prime}$ | | | | ** | |
| ☞ (c) .□V.CV.+C$_{\prime}$ | | | | * | * |

It is not difficult to see that the assumed constraint hierarchy determines a language in which all syllables have the *overt* form CV (onsets required, codas forbidden.) This is exemplified by the parse (8c), where the famous little hand marks the optimal candidate. If another constraint hierarchy had been chosen, say

(9) FILL$^{NUC}$ o PARSE o FILL$^{ONS}$ o ONSET o NOCODA

a completely different language would be effected, one that admits all types of open and closed syllables, (8a) would exemplify an optimal candidate in this case.

For the selection of an optimal output within *OT syntax*, the following constraint hierarchy is assumed:

(10) FULL-INT o DROP-TOPIC o PARSE o SUBJ

As it can be seen from tableau (11), this ranking yields an Italian-like behavior in which topicalized subjects are suppressed; this is exemplified by the optimal parse (11a).

(11)

| <sing(x), x=topic, x=he; T=pres perf> | FULL-INT | DROP-TOPIC | PARSE | SUBJ |
|---|---|---|---|---|
| ☞ (a) [$_{IP}$ has [ sung]] | | | * | * |
| (b) [$_{IP}$ he$_i$ has [t$_i$ sung]] | | * | | |
| (c) [$_{IP}$ has [t$_i$ sung] he$_i$ ]] | | * | | * |

This behavior would change to an English-like performance if we chose the following hierarchy:

(12) PARSE o SUBJ o FULL-INT o DROP-TOPIC

where PARSE and SUBJ outrank DROP-TOPIC. In this case, (11b) would arise as the optimal candidate.

The architecture of OT suggests a nice realization of the fundamental distinction between UG on the one hand and the language specific part of Grammar on the other hand: UG consists of *Gen* (the generator) and *Con* (the set of constraints); the language-particular aspect of Grammar is determined by a particular ranking of the constraints. This proposal bolsters the way for defining a factorial typology:

> *Typology by reranking*: Systematic crosslinguistic variation is due entirely to variation in language-specific total rankings of the universal constraints in *Con*. Analysis of the optimal forms arising from all possible total rankings of *Con* gives the typology of possible human languages. UG may impose restrictions on the possible rankings of *Con.* (p. 27)

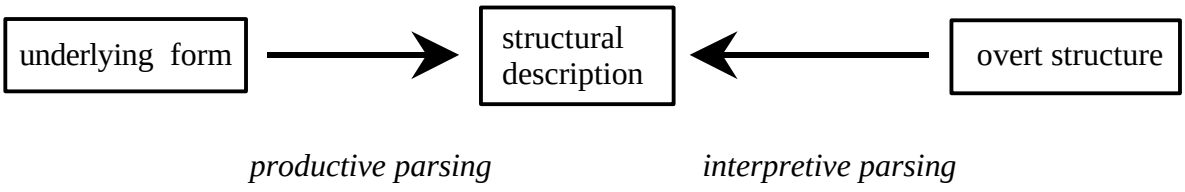
As already shown in Prince & Smolensky (1993), analysis of all rankings of the constraints considered in the basic CV syllable theory reveals a typology that explains Jakobson's (1962) typological generalizations. In the case of OT-syntax, Grimshaw & Samek-Lodovici (1995) were the first who performed an analysis involving all rankings of the above constraints[1] and derived a typology of subject distribution on this way.

3. Typology by reranking is the most famous but not the only pleasant consequence from the general architecture of OT. Another consequence is the idea of robust interpretive parsing, which is substantial for understanding the central claim of the book.

Although the term *parsing* is used more commonly in the context of language comprehension, Tesar & Smolensky treat it as the general issue of assigning structure to input, an issue relevant to both comprehension and production. To be sure, the canonical perspective of an OT grammar is related to production – taking the input as an underlying form, and the output structural description as including the surface form. This type of parsing is called "productive parsing", and it is schematically represented in diagram (13) – using the term "overt structure" instead of "surface structure":

[1] In fact, they considered one more constraint for aligning focused constituents.

(13)




Algorithms for performing productive parsing are presented in section 8, using techniques of dynamic programming. Given some radical restrictions, a computational complexity is established that is linear in the size of the input.

In the context of language comprehension, another mapping comes into play. It maps a given overt form to an optimal structural description *SD* whose overt portion matches the given form. The process of computing the optimal *SD* for an overt form is called *interpretive parsing.*

It is a common observation that competent speakers can often construct interpretation of utterances they simultaneously judge to be ungrammatical. Whereas it is notoriously difficult to account for this kind of "robustness" of natural language interpretation within rule- or principle-based models of language, the interpretation of ungrammatical sentences is much simpler when using an OT architecture.

Tesar & Smolensky's idea of *robust interpretive parsing* is the idea of parsing an overt structure with a grammar even when that structure is not grammatical according to that grammar. It is important to recognize that the presence of interpretable but ungrammatical sentences immediately corresponds to mismatches between productive and interpretive parsing. Consider an interpretive parse that starts with some overt structure *OS* and assigns an optimal structural description *SD*. Paired with *SD* is a certain underlying form *UF*. The grammaticality of *SD* (and its overt structure, *OS*) depends on whether the outcome of productive parsing leads us back to *SD*, when starting with *UF*. In case it does, then *SD* is grammatical; otherwise it is ungrammatical.

As a simple illustration we reconsider the earlier example from basic syllable theory – using the constraint hierarchy (7). Starting with VC being the overt part, the mechanism of *interpretive parsing* yields the optimal output .VC. (all other potential parses of the same overt string are suboptimal). On the other hand, the process of *productive parsing* – starting with the corresponding underlying form /VC/ – yields the parse .□V.+C, (with overt part C□V). The parse .VC. comes out as suboptimal from the production perspective. As a consequence, VC must be seen as ungrammatical in the given language.

Reconsidering the example from OT syntax, we take the constraint hierarchy (10) that accounts for Italian-like syntactic behavior. In Italian, sentences such as *he has sung* are unacceptable if the pronoun refers to a discourse topic. Using the hierarchy (10), this is demonstrated in tableau (11), where the sentence *he has sung* comes out as suboptimal. Despite its unacceptability, the sentence is parsed into a structural description, namely [$_{IP}$ he$_i$ has [$t_i$ sung]]. An important point in all these examples is that both in productive parsing and in interpretive parsing the same constraint hierarchies are used. The difference arises solely from the different candidate sets that are relevant for the different perspectives of optimization.

It will become clear in subsequent discussion that the idea of robust interpretive parsing is crucial for the mechanism of language learning in OT. Acknowledging the basic idea, I have a suspicion that the present form of interpretability and robustness is not yet the whole story, and the architecture indicated in (13) cannot be the last word. The stumbling block relates to the phenomenon of *unintelligibility*. This phenomenon corresponds to the observation that many grammatical forms are unintelligible, that is, they cannot be interpreted in a felicitous way. Using the basic architecture sketched in (13), each overt form gets an optimal interpretation supposing that *Gen* provides at least one. Obviously, the outcome of productive parsing decides the *grammaticality* of the form, and it has no influence on its *interpretability*. As a consequence, the only instrument that is available for the exclusion of interpretations is the manipulation of *Gen*. That is we have to stipulate hard constraints within *Gen* in order to account for the existence of unintelligibility. That this "solution" doesn't work, was argued by de Hoop (2000) by using examples such as the following:

(14) a. Few of the team members can drive, but every one of those will come to the match in his car.
b. #Few of the team members can drive, but every team member will come to the match in his car.

We have to account for both the intelligibility of (14a) and the unintelligibility of (14b). It is rather obvious that the use of hard constraints, corresponding to standard assumptions on discourse quantifiers, cannot solve the problem. The only way out of the problem, it appears, is to modify the architecture (13), a possibility that is pursued by de Hoop (2000).

The counterpart to the problem of *unintelligibility* is the problem of *ineffability*. Ineffability refers to the problem when a underlying form does not yield a well-formed structural description as its output. The problem raises parallel questions, for which answers

have been attempted within the architecture (13), for example by stipulating a *null parse,* a solution that was criticised by Pesetsky (1997).

4. The history of theoretical physics teaches us that audacious oversimplification paired with rigorous mathematical analysis sometimes yields incredible results. Tesar & Smolensky strictly follow this research strategy in the best passages of the book. An outstanding example is the development of the idea of *constraint demotion* (section 3 & 7) that builds the basis for a family of procedures for performing grammar learning in OT. The following simplifying conditions are necessary in order to make the idea feasible and applicable:

(A) UG = *Gen* + *Con.* The learning problem consists in inferring the ranking of the constraints in *Gen.* This excludes both the possibility that the constraints themselves are learned (in part at least) or that aspects of the generator are learnable. On the other side it excludes the possibility that the set of the possible rankings is constrained on a universal basis.

(B) The force of strict domination o: A relation of the form C o C' does not merely mean that the cost of violating C is higher than that of violating C'; rather, it means that no number of C' violations is worth a single C violation. The force of strict domination excludes cumulative effects where many violations of lower ranked constraints may overpower higher ranked constraints. In terms of a numerical representation of harmony, the idea corresponds to an exponential model of constraint combination.

(C) The OT grammar of the language that has to be learned is based on a *total* ranking of all the constraints: $C_1$ o $C_2$ o ... o $C_n$.
During learning the ranking of the constraints is not restricted to a total ranking. Instead, more general domination hierarchies are admitted which have the following general form:
$\{C_1,C_2, ...,C_3\}$ o $\{C_4,C_5, ...,C_6\}$ o ... o $\{C_7,C_8, ...,C_9\}$. ("stratified domination hierarchy")

For explaining the idea of constraint demotion we have to suppose a learner that is confronted with a grammatical structural description *SD* of a source language *L*, corresponding to a certain input *I*. Which information about the correct ranking of the constraints can the learner extract from this observation? Since OT is inherently comparative, the learner is not informed about the correct ranking by this positive information in isolation. Instead, the role of competing candidates must be considered. A set of these candidates is determined by *Gen*,

which is completely part of UG according to the condition (A). Given the learner has full access to *Gen*, it can be concluded that she has access to the competing candidates as well.

From the fact that *SD* is grammatical it can be concluded that *SD* must be more harmonic (less offending) than all the competing candidates that show a pattern of constraint violations different from that of *SD*. Taking the general philosophy of OT, it can be concluded that these competing candidates are ungrammatical. Each grammatical description, thus, brings with it a body of implicit negative evidence in the form of these competitors. This fact has to be emphasized as one of the main advantages of a comparative theory like OT.

With the help of assumption (B), it can be concluded that each winner-loser pair gives exactly the following information: the constraints that are lost by the grammatical structural description must be ranked lower than (at least) one constraint that is lost by the ungrammatical competitor. The basic learning mechanism of *constraint demotion* takes exactly this information and adjusts the ranking of the constraints accordingly. In short, the idea is that constraints that are lost by grammatical structural descriptions must be demoted in their ranking below constraints that are lost by competing structural descriptions. Constraints are only demoted as far as necessary.[2]

Simplifying a bit, the following example illustrates the case of learning OT syntax for an Italian-like source language. Given the sample input of tableau (15), it is assumed that the structural description (a) forms a grammatical candidate from the teacher's point of view (this is marked by the sign **U**). If it is further supposed that the constraints in the learner's mind are unranked with respect to one another (indicated by the dashed lines between the constraints), then the actual winner is the competing candidate (b), however.

(15)

| <sing(x), x=topic, x=he, T=pres perf> | SUBJ | DROP-TOP | FULL-INT | PARSE |
|---|---|---|---|---|
| **U** (a) [$_{IP}$ has [ sung]] | * ↓ | ↑ * | | * ↓ |
| **L** (b) [$_{IP}$ he$_i$ has [t$_i$ sung]] | | | | |

Notice that he role of the arrows is to point to the constraint-specific winner (and away from the loser.) Obviously, SUBJ and PARSE are lost by the grammatical structure (a) and DROP-

[2] In section 3.1.4 of the book there is a careful discussion why *constraint promotion* (promoting constraints toward the correct hierarchy) would not work in the general case. The point is that constraint promotion cannot solve the disjunction problem that arises if more than one constraint is lost by the ungrammatical competitor.

TOP is lost by the competing structure (b). This gives the following information concerning the "correct" ranking of the constraints: DROP-TOP o { SUBJ, PARSE }. The algorithm of constraint demotion then moves the constraints lost by the grammatical structure (SUBJ and PARSE) into a new "stratum" that is dominated by the rest of the constraints. The result is shown in tableau (16):

(16)

| <sing(x), x=topic, x=he, T=pres perf> | DROP-TOP | FULL-INT | SUBJ | PARSE |
|---|---|---|---|---|
| **UL** (a) [$_{IP}$ has [ sung]] | ↑ * | | ↓ * | ↓ * |
| (b) [$_{IP}$ he has [$t_i$ sung]] | | | | |

If more than one ungrammatical competitor is involved and/or we have to consider more than one grammatical target structure, then the procedure of constraint demotion has to be repeated in a well-defined way.

For the "correctness" of (iterative) constraint demotion the condition (C) is crucial. Tesar & Smolensky show that the iterative procedure of constraint demotion converges to a set of totally ranked constraint hierarchies, each of them accounting for the learning data. Interestingly, this result holds when starting with an arbitrary constraint hierarchy. Another important insight concerns the data complexity of constraint demotion. Consider a system with a fixed number of constraints, say N. Assuming the conditions (A)-(C) are satisfied, then the number of informative data pairs required for learning is no more than N(N-1)/2, independent on the initial hierarchy and the nature of the constraints.

This result is really surprising. Assuming that all rankings of the N constraints give different grammars, then the cardinality of the space of possible grammars equals N! – the number of possible total rankings. How can it be that at most N(N-1)/2 appropriate data pairs are sufficient to fix the ranking of the constraints and to select the correct grammar? The answer is that the (spectacular!) difference between the size of the OT grammar space and the number of data pairs that is needed to learn a grammar is due to the inherently comparative character of OT. Assuming N constraints, then for each pair 1# i, j # N it has to be decided whether $C_i$ o $C_j$ or $C_i$ o $C_j$. There are exactly N(N-1)/2 such decisions and each one can be brought about on the basis of one appropriate data pair triggering the corresponding set of winner-loser pairs. Consequently, no more than N(N-1)/2 appropriate data pairs should be necessary for learning the correct grammar.

It is instructive to compare the OT view of learning with the alternative view of parameter fixing. Although Tesar & Smolensky discuss the relevant literature on learning in the PP framework only in passing, the basic result is quite obvious. Let's assume a parameterized UG with n parameters. Then this system admits $2^n$ grammars when the parameters are binary. In the worst case, the average number of triggers before reaching the target grammar is $2^n$. This is due to the fact that the learner is informed about the correct value of the different parameters by positive data only, and that all parameters are interacting in the worst case. If we consider, for example, a 30-parameter space, then the number of possible grammars is $1{,}073 \times 10^9$, and this also is the number of triggers that are needed in the worst case to learn the grammar (assuming that the target language can only be described by one of the possible grammars.)[3] In order to compare this outcome with learning in OT grammar, let's consider a system with 20 constraints. Then the number of possible grammars is $20! = 2{,}43 \times 10^{18}$, much higher than in the PP example. However, in cases using constraint demotion, the data complexity is only 190. Tesar & Smolensky are modest enough to leave it to the readers drawing their own consequences from such straightforward calculations.

It has to be added that the PP model looks more appropriate when it is assumed that the parameters are independent and non-interacting, such that triggering data that indicate the appropriate value for a specific parameter are possible. Under such circumstances, the target grammar will be reached after $n^2$ triggers[4], which is comparable with the OT data complexity. The result is that models of learnability in the PP framework favor parameters that interact with each other as little as possible. Tesar & Smolensky conclude from this fact, quite correctly:

> Unfortunately, this results in a conflict between the goals of learnability, which favor independent parameters with restricted effects, and the goals of linguistic theory, which favor parameters with wide-ranging effects and greater explanatory power. (p. 85)

5. Before I come to an examination of the restrictions (A)-(C), I have to explain a really electrifying idea – the bootstrap mechanism for learning hidden structure. The most striking limitation of the algorithm of constraint demotion is that it must be provided with full structural descriptions in order to make learning possible. This situation is quite odd from the

[3] Cf. Boersma (1998), p. 318, for substantiating this fact.
[4] As discussed by Boersma (1998), this result holds at least for the triggering learning algorithm of Gibson & Wechsler 1994.

point of view of a natural learner, who is confronted with the overt parts of grammatical forms only. Now we are confronted with the puzzling setting that "the learner cannot deduce the hidden structure in learning data until she has learned the grammar, but she cannot learn the grammar until she has the hidden structure." (p.7). For example, "the learner cannot learn the metrical grammar until she knows where the feet lie, but she cannot know where the feet lie until she knows the grammar." (p.7).

In their introductory chapter, Tesar & Smolensky point out that this problem is very common in learning and extensively studied in the learning theory literature, mostly within connectionist / probabilistic frameworks (under labels such as *unsupervised learning, hidden Markov models, expectation maximization*). The proposed solution transfers the iterative algorithms found within some of these frameworks to the case of OT learning. In short, the projected algorithm proceeds as follows. An initial hierarchy of the given system of constraints is selected (typically one where the structural constraints dominate the faithfulness constraints). In a first step, the procedure of *robust interpretive parsing* is used and assigns – based on this hierarchy – a structural description SD to the overt learning datum. In the second step, the mechanism of *constraint demotion* is used on the basis of this structural description. The hierarchy is corrected if there are any competitors (with regard to the same underlying form that corresponds to SD) that have higher harmony than SD itself. With this improved grammar, return to the first step and repeat.

The idea of combining robust interpretive parsing and constraint demotion gives a plausible picture of children's language acquisitions. Becoming confronted with some overt datum, the child tries to understand this datum (on the basis of her current grammar). She performs interpretive parsing, resulting in a structural description that includes an underlying structure. Next, the child turns to the production perspective: she starts with the underlying form and performs productive parsing. If the results of productive and interpretive parsing are different, then this information is used to correct the grammar. The child applies constraint demotion presuming the interpretive parse as the winner (correct analysis) and the productive parse as the loser. The child has succeeded in learning the target grammar if interpretive and productive parsing always give the same structural descriptions. A noteworthy point is that an overt form will allow the learner to improve his grammar just in case the current grammar (incorrectly) declares it to be ungrammatical.

It is not certain that this algorithm always converges on a correct ranking for a language. Contrasting with the pure constraint demotion algorithm whose correctness proves to be independent of the initial constraint hierarchy, the convergence of the combined

algorithm is highly sensitive to the initial hierarchy. This is demonstrated by simulation experiments investigating the task of learning metrical stress grammar (based on a system of 12 constraints). Using a sample of 124 languages and a set of 62 overt forms for each of them, the results suggest that the full algorithm is practicable in case the initial hierarchy is appropriate. A general condition seems to be that faithfulness constraints are dominated by the structural constraints. This assumption corresponds to a general hypothesis in children's acquisition of phonology. It is used, for example, to explain that children‘s ability in production lags dramatically behind their ability in comprehension. Although there is much to do in order to find out the effect of specific starting hierarchies, the simulation experiments demonstrate convincingly that in the relevant cases the speed of convergence is surprisingly high and qualitatively different from that of general search procedures over parametric spaces (such as the triggering learning algorithm in the PP model).

It is an outstanding feature of the architecture of OT that it allows the integration of insights from certain statistical learning approaches and a symbolic learning theory making strong linguistic predictions. In this way, the problem of learning the grammar prior to assigning hidden structure, or vice versa, can be approached without falling into circularity. Overcoming the gap between the demands of strong linguistic theories and the requests of statistical approaches to language learning seems fruitful for a rather general reason: it helps to close the link between linguistic explanation and learnability in a rather realistic and practicable way. A related point is that the proposed iterative strategy – combining robust interpretive parsing and constraint demotion – can be applied to any linguistic domain that admits an OT analysis. This allows a triggered, linguistically informed learning without being so specific as to bind the learning algorithm to a particular linguistic domain.[5]

6. Let’s come to an examination of the conditions (A)-(C) explained above. One way to look at these conditions is to see them as oversimplifications that are made mainly for didactic and practical reasons. Oversimplifications may be needed to allow one to concentrate on a central problem and to sweep aside many problems that are less critical for understanding the central one (i.e., the problem of learning ‘hidden’ structure.) Moreover, oversimplifications may be necessary to achieve interesting mathematical results that simply aren’t possible without them. Another way to look at the conditions is to see them not as simplifications at all. Instead, they are perceived as conditions reflecting the true nature of the domain under discussion. As such the conditions are considered empirically justified.

I must admit that it is not quite clear to me which position Tesar & Smolensky really take with regard to the conditions (A)-(C): oversimplification or empirically sound restrictions. Concerning the condition (C), we find the following statement:

> From the learnability perspective, the formal results given for Constraint Demotion depend critically on the assumption that the target language is given by a totally ranked hierarchy. This is a consequence of a principle implicit in CD. This principle states that the learner should assume that the description is optimal for the corresponding input, and that it is the *only* optimal description. This principle resembles other proposed learning principles, such as Clark's Principle of Contrast and Wexler's Uniqueness Principle. (p. 47 ff)

> It appears likely to us that learning languages that do not derive from a totally ranked hierarchy is in general much more difficult than the totally ranked case. If this is indeed true, demands of learnability could ultimately explain a fundamental principle of OT: UG admits only (adult) grammars defined by totally ranked hierarchies. (p. 50)

Taking the condition (C) as a kind of principle that indicates when language learning is simple, however, is a different idea than taking it as a strict demand on theories of learning. In my opinion, the first idea is right and the second wrong. There are many examples where the target language produces synonymies (scrambling data in German an Korean may provide a case in point). I agree that this can delay learning in one case or the other.

In this vein, the suggestion is to take (C) as a kind of oversimplification, the acceptance of which is justified only for doing the first significant research steps.[6] In an advanced stage, the condition (C) should be given up and a more general theory should be developed, a theory that *explains* (C) as a principle about the complexity of language learning. In my opinion, the learning theory of Paul Boersma's (e.g. Boersma 1998) is on the right track for doing this job. With regard to the condition (B), Smolensky himself sees it as a "regimentation and pushing to extremes of the basic notion of Harmonic Grammar" (Prince & Smolensky 1993, p. 200.) And Gibson & Broihier (1998) argue that this restriction does not appropriately characterize the manner in which parsing preferences interact.

What about condition (A)? Many representatives of OT seem to consider it as a *conditio sine qua none*. Boersma's work on functional phonology (Boersma 1998), however,

---

[5] In Section 5 of the book a further example is discussed. It illustrates how the same iterative process that is adapting the grammar can be used for incrementally learning the lexicon of underlying phonological forms.

[6] For a similar view cf. Antilla and Yo Cho (1998).

puts forward convincing arguments exposing principle (A) likewise as a kind of oversimplification. Explaining the details of this challenging view falls outside the scope of this review.

7. More than ten years ago there was a hot debate in cognitive linguistics concerning the true architecture of cognition – the debate between connectionists and symbolists. The proponents of a symbolist architecture, among them Fodor and Polyshyn[7], had the clever idea to take the arguments for connectionism as proving that symbolic architecture is *implemented* in a certain kind of connectionist network. This idea corresponds to the strategy of maintaining classical architecture and reducing connectionism to an implementation issue. *Learnability in Optimality Theory* demonstrates that an opposite strategy is more exciting: augmenting and modifying symbolist architecture by integrating insights from connectionism.

8. In summary, the main aim of the book is to present a research program that attacks the crucial problem of learning 'hidden' structure, such as syllable structure or X-bar structure, which cannot be directly detected in the overt learning material. Tesar & Smolensky have managed to establish a landmark in grammar learning, convincingly connecting the learning problem with the core principles of OT. This book is obligatory reading for everyone who is interested in the application of optimality-theoretic learning theory to the problem of language acquisition. Moreover, it is to be highly recommended for everyone who sees learning as essential for language and cognition.

It's not superfluous to add that, to be sure, the content of this book has for the most part been published in diverse journals and anthologies. Owning an authorized compilation of this material, soundly revised, is undoubtedly worth the money.

REFERENCES

Antilla, Arto and Young-mee Cho: 1998, 'Variation and Change in Optimality Theory', *Lingua* **104**, 31-56.

Boersma, Paul: 1998, *Functional Phonology*, Holland Academic Graphics, The Hague.

Edward Gibson and Ken Wexler: 1994, 'Triggers', *Linguistic Inquiry* **25**, 407-454.

Edward Gibson and Kevin Broihier: 1998, 'Optimality Theory and Human Sentence Processing', in Pilar Barbossa, Danny Fox, Paul Hagstrom, Martha McGinnis, and

[7] The article of Fodor and Polyshyn and the dispute between Smolensky and them are reprinted in MacDonald & Macdonald (1995).

David Pesetsky (eds.), *Is the Best Good Enough. Optimality and Competition in Syntax*, The MIT Press, Cambridge, Mass., 157-191.

Grimshaw, Jane and Vieri Samek-Lodovici: 1995, 'Optimal Subjects', in *University of Massachusetts Occasional Papers in Linguistics 18: Papers in Optimality Theory*. GLSA, University of Massachusetts, Amherst, 589-605.

de Hoop, Helen: 2000, 'The Problem of Unintelligibility in OT Semantics', manuscript University of Potsdam.

Jacobson, Roman: 1962, *Selected Writings 1: Phonological Studies*, Mouton, The Hague.

Cynthia MacDonald and Graham MacDonald (eds.): *Connectionism: Debates on Psychological Explanation*, Blackwell, Oxford & Cambridge.

David Pesetsky: 1997, 'Optimality Theory and Syntax: Movement and Pronunciation', in Diana Archangeli and D. Terence Langendoen, Optimality Theory: An Overview. Blackwell, Oxford & Cambridge, 134-170.

Prince, Alan and Paul Smolensky: 1993, *Optimality Theory: Constraint Interaction in Generative Grammar*, Technical Report RuCCSTR-2, Rutgers Center for Cognitive Science (To appear, MIT Press).

Reinhard Blutner
Project Group "Interpretation of Dialogue"
Humboldt University Berlin
Prenzlauer Promenade 149-152
D-13189 Berlin
blutner@german.hu-berlin.de