

## Chapter 24

# LFG, Optimality Theory and learnability of languages

Jonas Kuhn

University of Stuttgart

Optimality-Theoretic accounts of grammar specification do not rely (exclusively) on rules and descriptions that every expression in the implied formal language has to satisfy strictly. Instead, a universal set of violable constraints is posited, and most grammatical expressions are not perfectly harmonic in all respects, but violate some of the constraints. A particular language is characterized by a certain priority ranking over the universal constraint set. This conceptualization of grammatical knowledge gives rise to the prediction of cross-linguistic variation: across languages, the same underlying meaning is realized by different expressions since different rankings of the constraints make different candidates optimal. The framework comes with a straightforward algorithmic formulation of the language learning problem – as error-driven constraint reranking.

Optimality Theory (OT) was combined with many linguistic description frameworks in the 1990s. LFG turned out to be a very appropriate base formalism for specifying the candidate representations in OT's competition-based definition of grammaticality. The novel way of characterizing formal languages prompted a range of debates regarding central assumptions in linguistic theorizing; various extensions of the competition-based setup were proposed; and the concept of violable constraints was taken over as an effective modeling device for managing ambiguities in broad-coverage computational grammar development. This chapter provides an introduction of the core concepts of OT as fleshed out on the basis of LFG, it illustrates the most influential extensions, and it reviews important conceptual debates triggered by the approach.

## 1 Introduction

The term Optimality Theory (OT) refers to a family of grammatical frameworks developed in various subfields of linguistics following the original proposal for phonology by Prince & Smolensky (1993). Bresnan (1996), in her keynote address at the first LFG conference,<sup>1</sup> showed that LFG constitutes a natural base formalism for an OT account of syntax, laying the foundation for substantial research activities in the late 1990s and early 2000s.

The key idea behind OT is that grammatical knowledge is not captured in distinct rule systems for each language, but theorists assume a universal set of constraints responsible for determining the harmony or markedness of potential realizations of some underlying input representation. For example, both the clause structure *you see who?* and *who do you see?* are candidate realizations for expressing a *wh*-question that asks for the object of a transitive verb. In some languages, the equivalent of the former is grammatical, whereas in English, the latter is. Neither of the two variants is perfectly unmarked; each violates certain constraints (in essence: an interrogative phrase is best realized at the beginning of the clause vs. the object of a verb is best realized in the canonical VP-internal complement position). What differs across languages is the relative prominence among the universal constraints, such that there are different winners in the competition for the most harmonic/least marked candidate, which is defined to be the language-particular output. Learning a language thus amounts to determining the correct prominence ranking over a known set of constraints to replicate the behavior observed in adult speakers. A more technical illustration of an OT system building on the LFG formalism will follow in Section 1.1.

The notion of violable grammar constraints inspired linguists to explore the explanatory potential of a competition-based definition of grammaticality for a whole range of linguistic subfields. Moreover as Section 1.2 will discuss, the notion turned out fruitful for computational work aiming at linguistically grounded broad-coverage grammars: expanding “classical” constraint-based grammar formalisms with a novel type of violable constraints enables a competition-based filtering of ambiguity sources and thus greatly facilitates the treatment of relatively rare lexical variants and grammatical constructions (Frank et al. 2001). An outline of the remaining sections of this chapter will be provided in Section 1.3.

---

<sup>1</sup>Details of the analysis presented in the keynote address are discussed in Bresnan (2001a) and Bresnan (2002).

## 1.1 Example of an OT-LFG analysis

The tableau in Figure 1 demonstrates more technically how an OT-LFG account predicts *who will she see?* to be the grammatical structure for realizing this type of question in English. It is adapted from Bresnan (2001a), a contribution that recasts Grimshaw’s (1997) OT analysis of English verb inversion in an LFG framework assuming Extended Head Theory (as detailed in Bresnan 2001b: ch. 7 and Bresnan et al. 2016: ch. 7).


Input: $\left[ \begin{array}{ll} \text{PRED} & \text{'see}(x, y)' \\ \text{GF}_1 & \text{["you"]} \\ \text{GF}_2 & \text{["who"]} \\ \text{Q-FOC} & \text{_____} \\ \text{TNS} & \text{FUT} \end{array} \right]$	OP-SPEC	OB-HD	STAY
a. $\langle [\text{IP you will } [\text{VP see who } ]], \dots \rangle$	*!		
b. $\langle [\text{CP who } e [\text{IP you will } [\text{VP see } e ]]], \dots \rangle$		*!	*
 c. $\langle [\text{CP who will } [\text{IP you } e [\text{VP see } e ]]], \dots \rangle$			**

Figure 1: Tableau for the competition that leads to *who will you see?* being the most harmonic candidate under the constraint ranking for English. The candidates’ c-structure trees are shown as bracketing expressions; their full f-structures is not shown here. Only a small subset of the candidate set is listed, omitting other realization variants. These are excluded for English by additional constraints, but could be grammatical in other languages and are thus included in the candidate set.

The tableau shows the comparison of three candidate realizations for the underlying content, the input – here the proposition  $see(x, y)$ , where  $x$  is a pronominal for the second person singular,  $y$  is the focus of a question, and the clausal realization expresses future tense. Particular syntactic choices, for instance the realization of the semantic arguments by grammatical relations such as SUBJECT and OBJECT are underspecified in the input. In the OT-LFG framework, the range of realization options across languages can be captured very well by assuming that the input is a partially specified f-structure, and the candidate set consists of all fully specified LFG analyses whose f-structure is subsumed by the input.

The excerpt from the full candidate set shown in Figure 1 focuses on c-structural realization alternatives for clause structure. In Extended Head Theory, functional categories (such as I and C) and their c-structure projections are mapped to the same f-structure as their corresponding lexical categories (V). The functional projections occur optionally, but they offer the possibility to realize additional head elements such as tense auxiliaries and they come with one c-structural specifier position each, which can for instance be used for *wh* operators. Also, it is a

possibility across languages for verbal head elements to be realized in a higher, c-commanding, position (e.g., V in I or I in C, as in candidate c.). By definition, elements in such a functional position are still *extended heads* of the lower projections. This nonderivational account of head mobility, which Bresnan (2001b: ch. 7) also formalizes in a non-OT setting of LFG, thus opens up a considerable spectrum of alternative candidate realizations for the basic clause structure.

Let us now consider what effect the three violable constraints have that Figure 1 shows in the order of the ranking for English (the constraint formalization will be discussed in more detail in Section 2.2). The harmony/markedness of a candidate is determined by its constraint violation profile. In the tableau notation, a violation is signalled by a star in the row for the corresponding constraint. Since OT assumes a *ranking* of the constraints (and not a more general weighing as for instance in Harmonic Grammar, Legendre et al. 1990), one can simply proceed from left to right through the constraints to determine the winning candidate(s): at each constraint, only the candidates that have the fewest violations for this constraint are kept. The other candidates have no more chance of winning the competition, even when they do not violate any of the lower-ranking constraints. It is common to mark the decisive constraint violations with an exclamation mark following the star; the cells for the lower-ranking constraints are often grayed out in tableaux. In our example competition, the OP-SPEC constraint demands that syntactic operators such as *wh* phrases be realized in c-structural specifier positions (highlighting the operator's prominence). Candidate a. violates this constraint, since the *who* is realized inside the VP. OP-SPEC is the highest-ranking constraint and the other candidates satisfy this constraint, hence candidate a. is excluded from further comparison. The OB-HD constraint (for obligatory head) says that every projected category ( $X'$ ,  $X''$ ) should have a lexically filled extended head. Candidate b., which includes a CP projection without a lexically filled head, incurs a violation of this constraint.<sup>2</sup> The third constraint, STAY,<sup>3</sup> states that categories should *dominate* their extended heads, thus punishing c-structural realizations in higher, c-commanding positions: In candidate c., the *wh* phrase in the CP-Spec position incurs one STAY violation (due to a lexically empty DP inside the VP), the *will* that is not realized within IP another one. Since STAY is low in the constraint ranking for English, candi-

<sup>2</sup>The notion of *extended heads* leads to the situation that the empty I head in candidate c. does not incur an OB-HD violation: The C head, which contains *will*, is in a c-commanding position and thus acts as an extended head for the I' and IP.

<sup>3</sup>The name STAY is carried over from Grimshaw's (1997) account, which assumed a derivational base formalism with upward movement in trees. The reconstruction is entirely nonderivational.

date c. is optimal despite the two violations: there are no surviving competitors with fewer violations.

Under different rankings of the same three constraints, different clause structures come out as optimal. When STAY outranks OP-SPEC, for instance, candidate a. is the most harmonic. This predicts non-echo matrix questions like in (1) to be grammatical, as is the case in a *wh*-in-situ language like Chinese.

- (1) Mandarin Chinese (Huang 1982: 253)  
 ni kanjian-le shei?  
 you see-ASP who  
 ‘Who did you see?’

According to OT assumptions, a language learner starts out with an uninformed initial constraint ranking. Whenever they hear an utterance by an adult speaker of the language, they check whether the observed output (for a contextually inferred input) is the same that their current ranking would have predicted. If not, the constraints responsible for the error are demoted in the ranking (Tesar & Smolensky 1998). Details of the OT approach to characterizing grammaticality and the OT-LFG formalization will be discussed Section 2.

## 1.2 Ranked soft constraints in computational grammar development

As mentioned, Joan Bresnan gave a keynote at the first LFG conference in 1996, presenting her proposal for integrating the competition-based grammatical framework of Optimality Theory with LFG’s declarative specification of parallel representations that are in imperfect correspondence (Bresnan 1996). The concept of violable constraints immediately prompted interest in the computational LFG community, which was working towards implementations of LFG grammars that robustly cover real-life text data, most notable in the Parallel Grammar Development project ParGram (see also Forst & King forthcoming [this volume]).

As Frank et al. (2001) discuss, violable constraints can alleviate a considerable practical problem with broad-coverage grammars that aim for linguistic precision: the *ambiguity management* problem. To reach an acceptable coverage of real corpus data with a linguistically precise grammar, the rules have to cater for essentially all conceivable realizations of variable phenomena – even for those that are infrequent and tend to underlie special contextual restrictions. In a grammar with such a comprehensive rule set however interactions among multiple variable phenomena can lead to an enormous degree of ambiguity in parsing. Although ambiguity packing techniques keep the computational parsing task itself

tractable (Maxwell & Kaplan 1989) (see also Cahill & Way forthcoming [this volume]), it can become hard for the grammar writers to keep track of whether the contextually justified readings are among the readings predicted by the grammar (King et al. 2004). The notion of violable constraints puts grammar writers in a position to overcome the ambiguity management problem via perspicuous filters.

The following examples may illustrate the proliferation of sources of ambiguity in a comprehensive grammar. In English, temporal adverbials can have the shape of a plain NP headed by a time expression such as *year*, *day*, *moment*, etc., as for instance in (2a,b). A carefully constrained grammar will rely on a lexical marking of the relevant nouns to avoid overgeneration. Without such a lexical restriction, any intransitive verb used with an NP in object position, such as (2c), would be incorrectly predicted to form a grammatical sentence.

- (2) a. The car had problems all weekend.<sup>4</sup>
- b. What are your future plans the next half year?<sup>5</sup>
- c. \*Lou yawned two big bones.

The rule in (3) uses a constraining equation to ensure that the adverbial NP comes with the marking [N<sub>TYPE</sub> TIME +].

$$(3) \quad \text{VP} \rightarrow \text{V} \quad (\text{NP}) \quad \{ \quad \text{PP} \quad | \quad \text{NP} \quad \}^*$$

$$\begin{array}{ccccccc} \uparrow = \downarrow & (\uparrow \text{OBJ}) = \downarrow & \downarrow \in (\uparrow \text{ADJ}) & & \downarrow \in (\uparrow \text{ADJ}) & & \\ & & & & (\downarrow \text{N}_{\text{TYPE}} \text{ TIME}) = \text{c} + & & \end{array}$$

Examples (4a,b) show however that adverbial uses of plain NPs can also occur with nouns that do not refer to time expressions.

- (4) a. Being prepared will save your soles from aching the entire trip.<sup>6</sup>
- b. And WoW had problems every expansion for years [...]<sup>7</sup>

By a semantic process of type coercion (see, e.g., Pustejovsky 1995), NPs referring to segments of a process (e.g., during spatial movement, as in (4a), or in repetitive processes, as in (4b)) can receive a temporal reading. A grammar that strictly ties adverbial NPs to a lexical marking of time expressions will fail on these examples.

<sup>4</sup><https://www.flickr.com/photos/speedcenter/19225841474>

<sup>5</sup>[https://www.reddit.com/r/UnderCards/comments/kdjbb1/ama\\_is\\_the\\_way\\_to\\_get\\_a\\_free\\_vieeeeews/](https://www.reddit.com/r/UnderCards/comments/kdjbb1/ama_is_the_way_to_get_a_free_vieeeeews/)

<sup>6</sup><https://vivaglammagazine.com/how-to-effortlessly-look-good-when-traveling/>

<sup>7</sup><https://forums.newworld.com/t/fix-your-product-this-isnt-an-indie-company/247609/59>

And since such coercion processes are relatively productive, grammar writers face a difficult tradeoff between maximizing coverage and risking a proliferation of ambiguity.

The issue is amplified when several variation phenomena interact, as the following considerations demonstrate. Completely independent of the type coercion process we just saw, many transitive verbs in English can occur with an understood object in appropriate contexts – i.e., without a syntactically realized direct object: *enjoy* and *announce* are both strictly transitive, but (5a,b) show that occasionally, an understood object can be filled in contextually.

- (5) a. [Text message with a video posting:]  
Here it is. Lean back and enjoy for the next 7 minutes.<sup>8</sup>
- b. [News headline:]  
Troy Williams will announce tomorrow<sup>9</sup>

To capture such uses, one could have transitive verbs optionally fill in a pronominal PRED value for their object (this can for instance be done by putting an optional equation in the lexical template for transitive verbs: { (↑ OBJ PRED) = ‘pro’ }). The pronominal PRED value indicates that its referent can be contextually inferred.

Both the temporal type coercion discussed previously and understood objects occur rather infrequently overall; yet, a broad-coverage grammar should cater for these constructions. Unfortunately, it is impossible to fully capture the contextual constraints on the use of such constructions within a knowledge engineering approach – individual uses may depend on inferences involving situational knowledge and general world knowledge.

By employing violable “soft” constraints inspired by Optimality Theory in a broad-coverage LFG grammar, the infrequent constructions can be captured in the grammar without causing a proliferation of ambiguity for the core constructions. Sentence (6a) is an example of a plain transitive use of the verb *announce*. A conspiracy of the two infrequent constructions we just discussed would however also lead to an understood object reading of *announced* (corresponding to the only reading available for (5b)) when *every expansion* would be analyzed as a coerced temporal adjunct (like in (2e)). With a comprehensive grammar allowing for these rare options, every single transitive VP (as in (6b)) would receive

<sup>8</sup><https://mobile.twitter.com/automobilirimac/status/765505874425180160>

<sup>9</sup><https://www.on3.com/teams/kentucky-wildcats/news/troy-williams-will-announce-tomorrow/amp/>

duplicate readings – although a temporal reading of the NP is extremely implausible in most cases. As (6c) shows, even an NP including a time expression will rather fill the position of the direct object than acting as an adverbial NP alongside an understood object (although in this particular case, the two readings are semantically hard to distinguish).

- (6) a. The company announced every expansion.  
 b. The dog chases every bird.  
 c. [From a comment page for a music band:]  
 Super talented, professional. We enjoyed every minute.<sup>10</sup>

The mechanism of ranked violable constraints discussed by Frank et al. (2001), which is implemented in the XLE system (Crouch et al. 2011), provides a highly convenient way of dealing with this situation. In (7), template definitions are seen that optionally<sup>11</sup> introduce the infrequent construction of a type-coerced common noun (such as *expansion*, *trip* and even *bird*) and an understood object with a transitive verb.

- (7) a. COMMON-NOUN(<sub>P</sub>) ≡ (↑ PRED) = ‘<sub>P</sub>’  
 { (↑ NTYPE TIME) = +  
 TYPECOERCIONTOTIMENOUN ∈ *o*(\*) }  
 b. TRANSITIVE(<sub>P</sub>) ≡ (↑ PRED) = ‘<sub>P</sub> < (↑ SUBJ) (↑ OBJ)>’  
 { (↑ OBJ PRED) = ‘PRO’  
 UNDERSTOODOBJ ∈ *o*(\*) }

Note that the optional functional descriptions go along with statements of the form ‘XYZ ∈ *o*(\*)’. These have the effect of introducing so-called optimality marks (XYZ etc.) to a data structure projected from c-structure to a novel *o*-projection (for Optimality projection).<sup>12</sup> Mathematically, the data structure is a multiset (i.e., a generalization of sets that can include two or more identical ‘copies’ as distinct elements). The purpose of this multiset is simply to implement a counting mechanism that keeps track of how many times a certain functional description was used in the construction of a particular LFG analysis. In c-structure rules

<sup>10</sup><https://www.esteemlivemusic.com/we-enjoyed-every-moment/>

<sup>11</sup>Enclosing functional descriptions in curly brackets has the effect of creating one option with the descriptions and one without.

<sup>12</sup>Like the  $\phi$ -projection, the *o*-projection starts out from c-structural entities (i.e., lexical or phrasal c-structure nodes, denoted by ‘\*’). This makes it possible to introduce distinct marks from several c-structure nodes that map to the same f-structure. This can for instance be relevant for economy constraints that favor structures with fewer nodes.



$A \rightarrow B_1 \dots B_n$ , the union of the multisets from all daughter categories  $B_1 \dots B_n$  forms the multiset projected from the mother category  $A$ . So the root node of a c-structure tree will always provide access to the aggregate counts of marks from the full tree.

The multiset of optimality marks provides the basis for filtering a set of candidate analyses, very much like in the OT tableaux illustrated in Section 1.1. The grammar writer can define the relative ranking of the violable constraints (the OPTIMALITYORDER) that guides the filtering:

- (8) OPTIMALITYORDER UNDERSTOODOBJ TYPECOERCIONTOTYPENOUN.

Starting with the highest-ranking type of optimality mark, the set of readings is successively reduced; for each type of OT mark only the candidates that have the fewest marks are kept. The effect of this is that the structures for infrequent constructions as illustrated in (4) and (5) will only be seen in the final output of the grammar when there is no alternative, more canonical way of analyzing the same string. In our example scenario, plain transitive sentences like (as (6b)) will hence receive only the canonical subject-object reading. At the same time however, the OT-style ranking system *does* cover uses of type coercion or understood objects. For instance when analyzing sentences that contain type-coerced adverbial NPs like (4a,b), *all* candidate analyses include at least one mark TYPECOERCIONTOTYPENOUN, hence they will survive this step of the filtering.

With such a filtering mechanism, the grammar writers can transparently control many of the interaction effects of the linguistic phenomena they are dealing with – without having to come up with explicit descriptions of contextual conditions for the rare special uses. The approach has hence been widely applied in the broad-coverage grammars of the ParGram project. As we will see in Section 2.6, some modifications of the plain ranking mechanism open up further functionality for grammar development.

To conclude this section we note however that there is a clear difference in the conceptual role played by the competition among candidate analyses (i) in OT-LFG, as discussed in Section 1.1) vs. (ii) in the “OT-style” filtering approach in grammar development, as discussed in this section: in the former case, the competition serves to define grammaticality in a given language, whereas in the latter case, all candidates are assumed to be grammatical in principle, and the competition serves to filter out implausible readings.

### 1.3 Chapter outline

The remainder of the chapter is organized into four sections and a conclusion: Section 2 discusses the OT framework and its combination with the LFG formalism in more detail. Section 3 provides a number of demonstrations of how the competition-based notion of grammaticality can systematically derive variation patterns within a language and predict typological patterns across languages. Section 4 goes into important extensions of the standard OT framework and reviews discussions of the status of key components of an OT system such as the violable constraints. It also provides a discussion of linguistic formalisms and learnability of languages. Section 5 reviews the developments after the mid-2000s, with an emphasis on a comparison of OT and other competition-based architectures. It also includes a discussion of the relationship between OT and the recent so-called neural modeling paradigm in Machine Learning and Natural Language Processing research. Section 6 concludes with a summary of important epistemological considerations regarding the OT approach.

## 2 OT: Formalism and computational considerations

Section 2.1 discusses the key components of an OT system in general and their motivation for the Theory of Grammar. On this basis, Section 2.2 addresses the application of OT systems in syntax and expands on the OT-LFG approach outlined in the introduction. Section 2.3 contrasts the OT-LFG approach with OT syntax approaches using a derivational base formalism.

Section 2.4 discusses the most important computational considerations regarding OT-LFG. An excursion in Section 2.5 addresses potential concerns regarding plausibility of OT-LFG that might arise if one assumes that the declarative specification of the candidate set has any direct psychological reality during human sentence processing. Section 2.6 provides some more details on the use of soft constraints in broad-coverage grammar development illustrated in Section 1.2.

### 2.1 Key components and assumptions of OT

The initial example shown in Figure 1 already provided an illustration of the key components of Optimality Theory in general and OT-LFG in particular. Within an OT system, the set of grammatical expressions for a particular natural language  $L$  is defined using (i) a language-independent function  $Gen$ , which maps a given input representation to a set of candidates, (ii) a universal set of violable constraints  $Con$ , and (iii) a harmony evaluation function  $Eval$ , applying a

language-specific prominence ranking  $\gg_L$  that is defined over the constraint set. We define an expression to be grammatical in  $L$  if it is (one of) the most harmonic candidate(s) in the set of competing realizations for some input according to  $\gg_L$ . A candidate  $C_1$  is more harmonic than a candidate  $C_2$  iff it contains fewer violations for the highest-ranked constraint in which the marking for  $C_1$  and  $C_2$  differs.

With the variable ranking relation over a universal set of constraints, an OT system does not just predict an individual language, but essentially an entire constellation of possible languages, the so-called factorial typology. The task for a theorist working on some family of phenomena is to find a set of constraints that plausibly predicts the spectrum of cross-linguistic variation as it is empirically attested in the languages of the world.

A key methodological assumption of linguistic work employing OT systems is that the input includes no language-specific information or restrictions. This is called the “richness of the base” (Smolensky 1996). A construction that is only available in some languages should not be explained by differences in the input, but must be derived as an effect of constraint ranking. In other words, the role of candidate generation, *Gen*, is limited to providing the space of possibilities; the substantive linguistic regularities should by assumption all be dealt with in the constraint set.

In phonology, the input can be thought of as the target sound sequence a speaker intends to realize. In the case of loan words, this target sequence may include sound clusters that cannot be expressed in the language. For instance, the Dutch loan word *plan* is realized as *pālāna* in Sinhalese:

- (9) Sinhalese (Indo-Aryan; Boersma et al. 2000: 5, citing Sannasgala 1976)  
 pālāna  
 ‘plan’

The OT explanation is that there is a constraint against complex syllable onsets and that in the constraint ranking for Sinhalese, this constraint is ranked higher than a constraint against the addition of epenthetical vowels. As a consequence, a surface realization that deviates from the target sound sequence arises as optimal. To capture the full spectrum of phonological systems, *Gen* must make far-reaching deviations from the input available.

Constraints that evaluate whether a property from the input is preserved in the output are called *faithfulness constraints*; there are three types, punishing (a) the insertion and (b) the deletion of a segment and (c) the alteration of some feature value (e.g., devoicing of a voiced consonant) (Prince & Smolensky 1993;

McCarthy & Prince 1995). Which part of the space of more or less faithful candidates a language uses for a certain phenomenon is determined by the relative ranking of faithfulness constraints and a second class of constraints, the so-called *markedness constraints*. These constraints assess the output shape of a candidate, irrespective of the input it is supposed to convey. By and large, OT predicts typological differences to arise from different strategies of reconciling conflicts among faithfulness and markedness constraints.

## 2.2 Optimality-Theoretic Syntax and its formalization

The predictive potential of OT systems regarding language typology as demonstrated in phonology soon attracted researchers working on different subfields of grammar, including syntax (Grimshaw 1997; Samek-Lodovici 1996; Pesetsky 1998; Legendre et al. 1998).<sup>13</sup> Technically, it is relatively straightforward to assume that some modular part of a conventional theory of grammar is replaced by an OT system: instead of the conventional output representation, a set of alternative candidates is assumed whose harmony is then assessed based on ranked violable constraints.<sup>14</sup> (It should be noted however that depending on which part of a conventional approach is being replaced by the output of an OT competition, there can be great differences in the predictive potential of the model component; we come back to this in Section 2.3.)

The fact that the explanatory power of an OT system comes largely from constraint interaction had an interesting effect in terms of sociology of science: Optimality-theoretic extensions of base formalisms from different schools of thought could be subjected to a meaningful comparison, even where diverging assumptions in the conception of the base formalisms themselves had before made comparisons very difficult. This led to an opening up of channels for exchange among theoretical frameworks (see for instance the collection of contributions in Legendre et al. 2000).<sup>15</sup>

Early work on OT syntax sometimes relied on an intuitive and informal conception of the syntactic structures that should compete with each other in the same candidate set. To preserve the idea from OT phonology that all conceivable

---

<sup>13</sup>The account of unaccusativity in French by Legendre et al. (1990) in the framework of Harmonic Grammar predates much of the work in OT phonology. Nevertheless, many syntacticians presumably only took notice of competition-based approaches in the aftermath of the success of OT phonology and reacting to Grimshaw's (1997) paper.

<sup>14</sup>The approach of "Harmonic Serialism" assumes an architecture that is broken down into a sequence of steps with local optimization. Müller (2020) for instance explores an extension to morphology.

<sup>15</sup>This unifying potential is discussed for instance by Newmeyer (2002: 44).

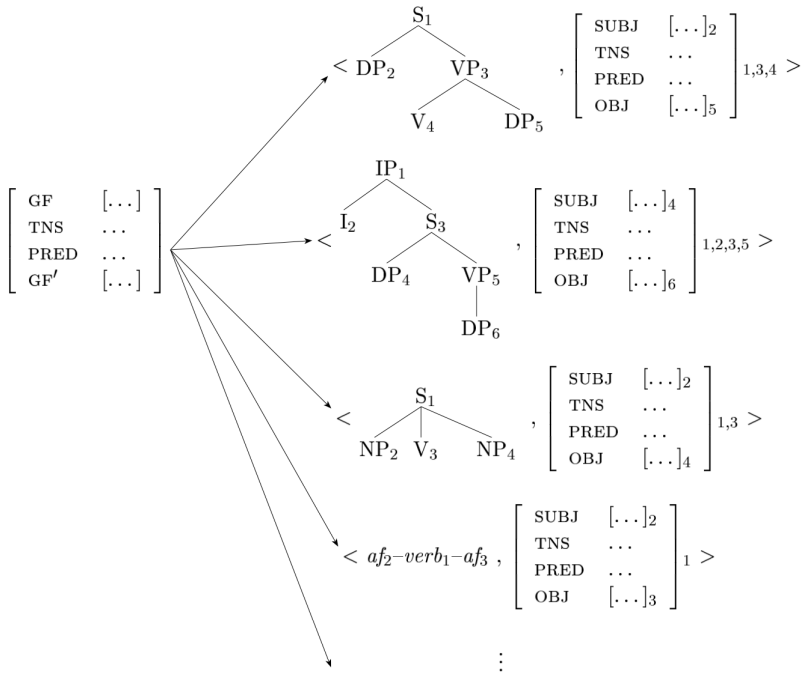


Figure 2: Illustration of candidate generation in OT-LFG (Bresnan 2001a): an underspecified functional structure (on the left) is the basis for generating full LFG analyses (c-structure/f-structure pairs) whose f-structure is subsumed by the input.

alternatives for saying the same thing are included in the candidate set, the natural assumption for the input in OT syntax is an abstract representation of the syntactically relevant *semantic content* of an utterance. As Bresnan (1996) points out, this intuition can be cashed out straightforwardly if candidate generation is based on a monotonic unification formalism like LFG: the input can be operationalized as a partial representation including the relevant semantic information, and the set of candidates as all the complete representations expanding this input representation – in the case of LFG pairs of c-structures and corresponding f-structures that are subsumed by the input presentation, as is illustrated in Figure 2 from Bresnan (2001a). Cross-linguistic variation at the level of constituent structure is captured by including all typologically different options for expressing some core semantic information in the candidate set.

A formal definition for the candidate generation function *Gen* in OT-LFG is shown in (10) (Kuhn 2003: 74).<sup>16</sup> *Gen* depends on an LFG-type base grammar  $G_{inviol}$ , which specifies the space of structurally valid candidate tuples from the full typological spectrum by means of the ‘inviolable principles’, and is applied to a partially specified f-structure  $\Phi_{in}$ .

(10) Restricted definition of *Gen*

$$Gen_{G_{inviol}}(\Phi_{in}) = \left\{ \langle T, \Phi' \rangle \in L(G_{inviol}) \mid \Phi_{in} \sqsubseteq \Phi', \text{ where } \Phi' \text{ contains} \right. \\ \left. \text{no more semantic information than } \Phi_{in} \right\}$$

The candidate set generated by *Gen* includes all c-structure/f-structure pairs from the base grammar whose f-structure  $\Phi'$  (i) is subsumed by the input f-structure  $\Phi_{in}$  and (ii) does not add any semantic information.<sup>17</sup>

The violable constraints in OT-LFG operate on the candidate representations. It is in principle conceivable to include the detection of constraint violations already in the specification of the base grammar  $G_{inviol}$ , using disjunctive f-annotations of the kind shown in Section 1.2 on soft constraints in computational grammar development and introducing marks to a special *o*-projection for counting. However, the specification would presumably become unmanageable fast, since the representational patterns addressed in the various constraints are interdependent; moreover, many constraints address not only patterns in f-structures, but make reference to portions of c-structure. The formalization of OT constraints in Kuhn (2003: ch. 4) therefore proposes to leave constraint marking out of  $G_{inviol}$  and assume a conceptually separate step for detecting the constraint violations for each c-structure/f-structure pair in the candidate set. This step can employ straightforward descriptive constraint *schemata* which use a special metavariable  $\star$  that is successively being instantiated to every structural element in a candidate analysis, i.e., to each c-structure node, and to each (sub) f-structure.

In (11), the schema-based formal capturing of Bresnan’s (2000) constraints OP-SPEC and OB-HD, which were discussed in Section 1.1, is seen (for details, see Kuhn 2003: 90ff):<sup>18</sup>

- (11) a. OP-SPEC (‘An operator must be the value of a DF in the f-structure.’)
- $$(f\text{-str}(\star) \wedge \exists v.[(\star \text{ OP}) = v]) \\ \rightarrow \exists f.[(f \text{ DF}) = \star]$$

<sup>16</sup>The language  $L(G)$  generated by an LFG grammar  $G$  is here defined as the set of tuples  $\langle T, \Phi \rangle$  such that  $T$  is a c-structure generated by the context-free skeleton in  $G$  and  $\Phi$  is a valid f-structure for  $T$  according to the functional descriptions in  $G$ .

<sup>17</sup>The restriction excluding surplus semantic information will be addressed in Section 2.4.

<sup>18</sup>The schemata assume appropriately defined auxiliary predicates *cat* (for category) etc.

“If  $\star$  is an f-structure bearing a feature  $OP$  (with some value  $v$ ), then there is some f-structure  $f$  such that  $\star$  is embedded in  $f$  under a DF function.”

- b. OB-HD (‘Every projected category has a lexically filled extended head.’)

$$(cat(\star) \wedge (bar-level(\star, 1) \vee bar-level(\star, 2)) \\ \rightarrow \exists n.[ext-hd(n, \star)])$$

“If  $\star$  is an X-bar or X-max category, then there is some node  $n$  which is the extended head of  $\star$ .”

The evaluation function *Eval* is operationalized as a trial of all schemata in the constraint set  $C$  on every structural element of a candidate analysis (c-structure node or partial f-structure). For the trial, the metavariable  $\star$  is instantiated to the element under consideration; the count for the relevant constraint is increased in case the proposition is satisfied. The language-specific ranking  $\gg_L$  over the constraints then controls the filtering of the most harmonic candidate(s), in the same way as illustrated in Figure 1 in Section 1.1.

All candidates that are optimal (under the constraint ranking for a given language  $L$ ) for some underlying partial f-structure  $\Phi_{in}$  then form the language (= set of c-structure/f-structure pairs) generated by an OT-LFG system for language  $L$  (Kuhn 2003: 117):

- (12) Definition of the language generated by an OT-LFG system

$O = \langle G_{inviol}, \langle C, \gg_L \rangle \rangle$  for language  $L$ :

$$L(O) = \{ \langle T_j, \Phi_j \rangle \in L(G_{inviol}) \mid \\ \exists \Phi_{in} : \langle T_j, \Phi_j \rangle \in Eval_{\langle C, \gg_L \rangle}(Gen_{G_{inviol}}(\Phi_{in})) \}$$

The OT-LFG formalization thus provides a declarative, fully operationalized framework for specifying a theory of the typological space of options for a range of grammatical phenomena, predicting specific realization patterns for languages associated with certain constraint rankings. Computational considerations, in particular regarding the complexity of *Gen*, will be addressed in Section 2.4.

Combining the OT concept of competition-based specification of grammatical knowledge with the LFG formalism for operationalizing the components of such an account was beneficial for both sides: For OT theorists interested in exploring the expressiveness of the Optimality-Theoretic approach when applied to phenomena from the broad field of morphology/syntax/semantics, the OT-LFG framework provided a basis whose formal and computational properties were well understood. Consequently, (i) concrete accounts for phenomena of interest

could be worked out and tested against attested linguistic data, and (ii) formal and computational implications of the novel conceptualization of modeling grammatical knowledge could be pinpointed. In the reverse direction, the LFG community benefited from the extension of the classical formalism, since OT-LFG provided a formalized framework for predicting patterns in the variation observed across languages (and within a language). By couching an LFG account for a phenomenon in the OT-LFG framework, its generality thus becomes testable against empirical data from language typology. And even when trying to capture variation patterns within a single language, an OT-LFG approach readily supports a reasoning that reduces the range of available realization options to a small set of independently justifiable directives.

### 2.3 OT-LFG's conceptual advantages over a derivational base formalism

A large proportion of work in OT syntax is not couched in the OT-LFG framework, but assumes a derivational base formalism such as Principle-and-Parameter Theory or Minimalism (see Grimshaw's 1997 original account and much subsequent work, e.g., Pesetsky 1998). With candidate analyses that are inherently derivational, an implementation of the original OT idea – that the input corresponds to the semantic content of the expressions under consideration – inevitably leads to a more complicated architecture. In the study of syntax, the expressions under consideration are full sentences; so, to implement the original OT idea, all alternative surface sentences expressing the same semantic content should be included in a candidate set. The complication for an OT architecture now arises from the fact that in a Chomskyan derivational approach, semantic interpretation is located at the level of logical form (LF), which is by definition one of the end points of a derivational process that starts out from a D-structure (as in the T-model underlying Principle-and-Parameter Theory, shown in Figure 3) or a set of lexical items, the “numeration”.

In a non-OT framework, the derivational processes that lead to a phonological form and a logical form are controlled by language-particular factors. When the derivational processes are taken to be the candidates of an OT system that adheres to the richness of the base, the derivational mechanisms have to be opened up in such a way that all language-particular restrictions on the derivations are lifted – generating all conceivable variants as alternative candidate derivations and leaving the calculation of language-particular effects to *Eval*, which selects the optimal derivation based on the constraint ranking. The input (for the definition of the OT candidate set) has to be (some relevant part of) an LF repre-



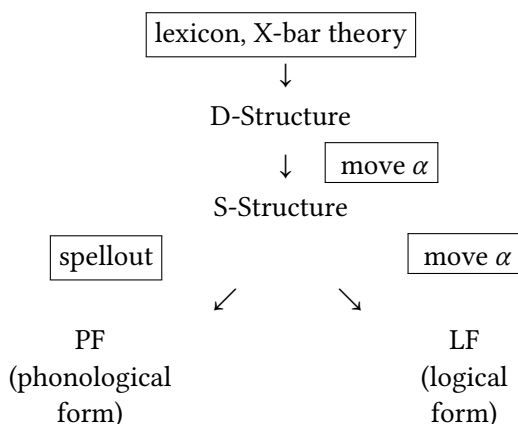


Figure 3: The T-model (or Y-model) of the derivation processes underlying a single analysis couched in Principle-and-Parameter Theory.

sentation, and the set of candidates that *Gen* assigns to such an “input” LF must be all derivations that *end in* this LF via unconstrained derivation – from any possible starting point (i.e., any numeration of lexical items that could arrive at this LF through unconstrained transformational derivations). Leaving aside concerns regarding the computational tractability of such a system, it is challenging to conceptualize the workings of the OT constraints if the candidate-internal derivation is taken literally as a (potentially destructive) structure-transforming process: if the evaluation of candidate derivations does not take place until a particular LF has been reached, how can the constraints be used to control the language-specific choices in the derivation steps happening early on in the process?

The constraint evaluation challenge can be resolved by viewing the derivational process inherent to the candidates as some abstract process that produces a *representation* including a record of all relevant steps (such as traces of a movement). With this representational strategy, a definition of the candidate set via a shared LF becomes possible. Let us call this an LF-as-input OT system, which preserves the original, meaning-related concept of the OT input (but enforces an abstract view of the derivations, with a representational record).

For a research approach starting out with a derivational framework, it may seem more natural however to resolve the constraint evaluation challenge in a different way: giving up the fully meaning-related notion of the OT input, one can turn to a different way of characterizing the set of competing candidates: since the derivations have their own technical starting point – D-structure or a

enumeration of lexical items – why not adopt a conceptualization in which this part of the candidate derivations constitutes the input for the OT system and have *Eval* compare the various possible derivations with identical D-structures/numerations? Important typological considerations regarding syntax and morphology *can* be addressed under this view. In fact, when a study focuses on a small set of specific phenomena – as is common practice both for typological and for language-particular studies in linguistics – the two approaches are often indistinguishable, since the (supposedly) *relevant* variation across candidates on which the study is centered originates from the same derivational subprocess across all candidates. This is presumably the explanation for the fact that the simpler system is often tacitly assumed. However, in situations where attention is not focused on a narrowly delineated range of phenomena, the two ways of conceptualizing the input *do* make an enormous difference. Only a semantically based input will leave the choice among realizations that differ in the lexical material in non-trivial ways to *Con* and *Eval*.<sup>19</sup>

A numeration-as-input approach generally imposes restrictions on the candidate space that are tied to particular languages (since semantically equivalent paraphrases of the same content which use different lexical material do not compete in the same candidate set). It is actually a consequence of the predominance of the numeration-as-input approach (and similar input conceptualizations) that *language-particular ineffability* was considered a central issue for OT accounts of syntax. The issue can be characterized as follows: since by definition there is always a most harmonic candidate in any given candidate set, the OT approach appears to systematically exclude the possibility that in some languages there is no grammatical realization at all for a conceivable linguistic construction. Fanselow & Féry (2019) provide the example in (13). While (13a) is grammatical in English, there is no grammatical way of saying (the equivalent of) (13b). They argue that a standard system of OT syntax as Grimshaw's (1997) will include a candidate set of alternative clause realizations for (13b), and one of them will inevitably be the most harmonic, incorrectly predicting that there is some grammatical realization used this set of lexical items.

- (13) a. Who did the president think that the foreign minister met in  
Afghanistan?

---

<sup>19</sup>Of course, not all accounts incorporating a competition-based subprocess are necessarily following the idea that all cross-lingual variation should be reduced to a global, fully meaning-based optimization. It is also conceivable to construe distinct derivational steps as separate, self-contained optimizations (compare e.g., Heck & Müller 2000; Müller 2003). Many approaches explicitly couched in a derivational setting indeed assume fairly restricted structural or derivational *domains* of local optimization (Müller 2012: sec. 4).

- b. \* Who did the president resign although the foreign minister met in Afghanistan?

Note however that with a sufficiently abstract semantic input, an LF-as-input OT system *could* be devised that makes reasonable predictions for all languages: Where there is no compact realization of a thought in a single clause, a more verbose, multi-clause paraphrase can be used (for (13b) maybe *Who was it that the foreign minister met in Afghanistan when the president resigned nevertheless?*). This appears like a plausible analogy to examples from phonology, where some languages enforce a very unfaithful realization of loan words or foreign names that fall outside the phonological patterns of a language (as exemplified by (9) above). Thus, the language-particular ineffability problem arising in numeration-as-input approaches is not a problem under a global meaning-based conceptualization of the input.

The considerations in this subsection have shown that while a competition-based construal of dedicated derivational subprocesses may be fruitful in order to systematically derive typological patterns for a certain submodule of a broader derivational theory of grammar, the original OT idea of deriving all cross-lingual differences – and thus learnability – as an effect of constraint reranking presupposes a comprehensive global competition among all alternative candidate expressions for a given underlying meaning. For capturing constraint evaluation in global competition with a derivational base formalism, there does not seem to be a good alternative to using representational traces of the (abstract) derivational process inherent to each candidate representation.

Bresnan's (1996) recasting of Grimshaw's (1997) account of extended projections – employing Extended Head Theory as sketched in Section 1 – can be viewed as a blueprint of a strategy that translates some relevant key aspects of a derivational approach into a representational approach. So, one can effectively view OT-LFG not only as an OT extension of the LFG framework; it also provides a feasible implementation for LF-as-input approaches, in particular a whole range of work that follows the general spirit of Grimshaw's (1997) proposal.

To sum up the previous and this section, OT-LFG as proposed by Bresnan (1996, 2000, 2001a, 2002) spells out a conceptualization of OT syntax that allows for a clean and comprehensive separation of language-independent candidate generation and violable constraints capturing the spectrum of typological variation. LFG's representational framework makes this separation conceptually simple. Global competition, for instance between morphological vs. syntactic means

of expressing for realizing an underlying feature is naturally accounted for since LFG assumes parallel correspondence among all representational structures.

## 2.4 OT-LFG: Computational considerations

Conceptually straightforward as it is, the idea of relying on feature structure subsumption to define the set of candidate analyses (i.e., tuples of c-structure, fully specified f-structure and potentially more LFG projections) in *Gen* as specified in Section 2.2 raised computational concerns. Johnson (1998) points to the issue that the parsing problem for an OT-LFG grammar is undecidable in the general case. To solve the parsing problem given a string *s*, all optimal analyses have to be found that have *s* as the yield of their c-structure tree.<sup>20</sup>

Figure 4 from Kuhn (2003: 173) illustrates the procedure with a semi-abstract example. In a first step (i) all potentially underlying input representations have to be found. This can be achieved by standard LFG parsing (which is decidable), using the base grammar that defines the set of all universally available candidate structures. The input information is by definition part of the f-structure in each candidate. This predicate-argument structure representation has to be filtered out (step (ii)). In a next step (iii), for each potential underlying input representation, the set of all candidates has to be generated (including the original as well as all alternative full c-structure/f-structure tuples); the constraints are applied to each candidate (iv) and the most harmonic candidate can be determined, given the relevant constraint ranking. This candidate is only included in the set of valid analyses for the original string *s* if *s* is indeed the yield of the optimal candidate (v).

We note that step (iii) involves generation with an LFG-type grammar from a partially specified f-structure. Wedekind (1999) shows that the general problem of generation from partial f-structures, given some LFG grammar, is undecidable. This implies that there could be cases in which the candidate set for a given input f-structure cannot be computed, so it would also be impossible to determine the effect of an OT system. However, Wedekind also points out that the decidability problem for generation with plain LFGs occurs only with certain technical feature representations that are not used to represent the semantics of natural-language sentences. How does this translate to the application of an

---

<sup>20</sup>Johnson (1998) assumes stronger conditions for the optimal candidate: the input representation determined in the first step needs to be included in the optimal analysis of the string. This corresponds to (strong) bidirectional optimality, as discussed in Section 4.4. For standard OT, only the production-based competition is relevant, as assumed in the remainder of this section.

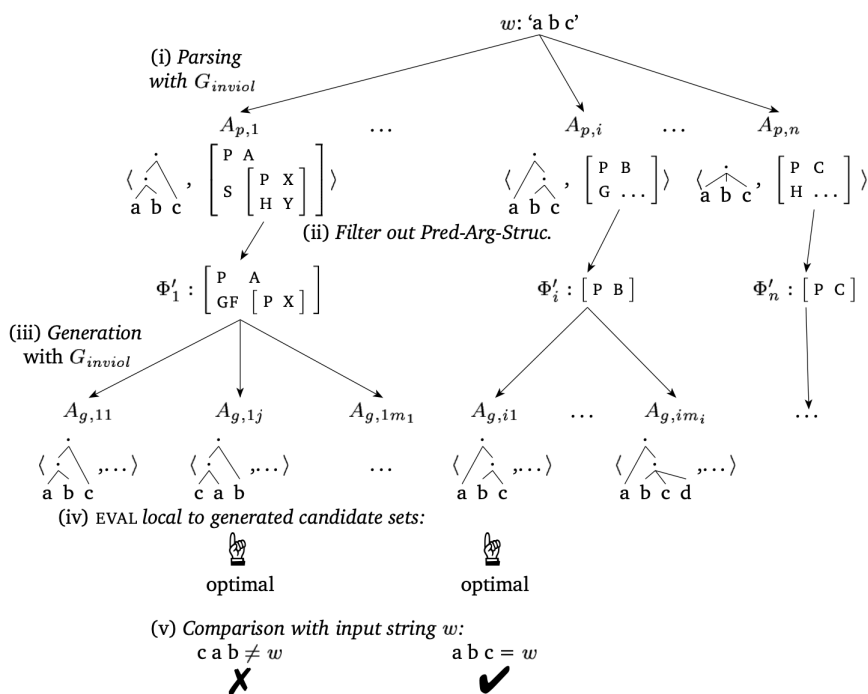


Figure 4: Illustration from Kuhn (2003: 173) for the parsing procedure for a standard OT system, with grammaticality based on production-based optimality.

LFG-style grammar in OT candidate generation? Is it possible to formulate restrictions on admissible partial f-structures and thus guarantee decidability of the parsing problem?

Potentially problematic cases are candidates that include violations of faithfulness constraints, as they are for instance assumed to derive *do*-insertion in English questions (Bresnan 2000: ex. 44). As we will see in the following, restrictions on the formalism can be devised that guarantee decidability but nevertheless permit the use of faithfulness constraints to derive syntactic variation of this kind. Taking advantage of the explanatory potential of OT, Bresnan's (2000) analysis does not stipulate a special, PRED-less lexicon entry in *do* insertion (like in standard LFG), but derives insertion of an additional verb as a consequence of the ranking of violable constraints. To achieve this effect, the *Gen* function underlying in this system has to be able to add "unfaithful material" quite freely. Does this mean that we are confronted with the decidability problem?

As Kuhn (2003: ch. 4) discusses, the option of adding unlimited amounts of material in a candidate is not only undesirable from a computational point of view. It also goes against the key idea of treating all candidates as potential verbalizations of some identifiable semantic content. The definition of *Gen* should therefore be restricted in such a way that the candidates' f-structures (i.e., the interpretable part of the representation) are not only subsumed by the input f-structure, they also may not contain any additional semantic information.<sup>21</sup> With this restriction, it can be guaranteed that the set of candidates stays computationally tractable (Kuhn 2002, Kuhn 2003: 199ff). The definition (10) above already incorporated the necessary restriction.

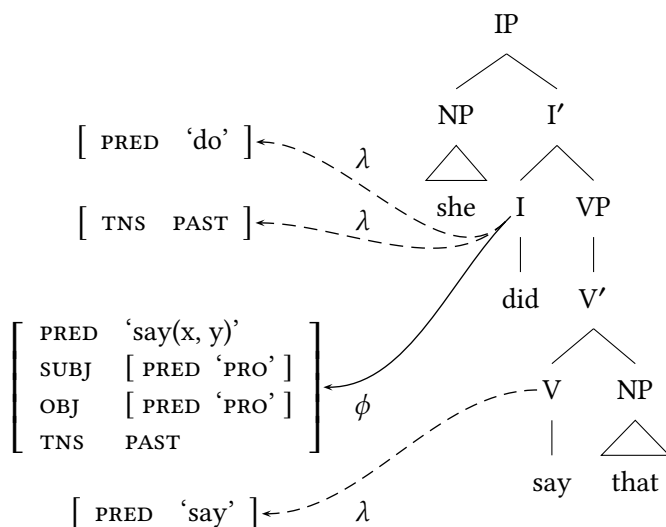


Figure 5: Illustration of a candidate incurring a faithfulness violation, following the technical formalization of Kuhn (2003: 112); c-structure nodes project not only to f-structure (via function  $\phi$ ), but also to a (very local) lexical structure (via function  $\lambda$ ).

Figure 5 illustrates a formalization of faithfulness violations that is compatible with this restriction and can be used to derive *do* insertion in English. The key idea from Bresnan (2000: ex. 44) is that the lexical contribution of elements that are inserted “unfaithfully” (as a way of satisfying some other constraint

<sup>21</sup>It is allowed for candidates to contain additional non-semantic f-structure information and material at the level of c-structure and other projections; this is unproblematic for decidability, since the amount of information that can be added is bounded by the size of the grammar.

that is higher-ranking than faithfulness) does not make it into the candidate's f-structure. The technical operationalization from Kuhn (2003: 112) uses a special  $\lambda$ -projection to keep track of all lexical contributions from the words, which may or may not re-appear in f-structure. In case they do *not* re-appear, a violation of a faithfulness constraint (DEP-IO) will be recorded by *Eval*.

The lexical entries underlying the analysis in Figure 5 are specified as shown in (14) (adapted from Kuhn 2003: ex. 140):

- (14) *did*            I      $g_1 = \lambda(\mathcal{M}^*)$   
                                        $(g_1 \text{ PRED}) = \text{'DO'}$   
                                        $\{g_1 = \uparrow\}$   
                                        $g_2 = \lambda(\mathcal{M}^*)$   
                                        $(g_2 \text{ TNS}) = \text{PAST}$   
                                        $\{g_2 = \uparrow\}$
- say*                V      $g_1 = \lambda(\mathcal{M}^*)$   
                                        $(g_1 \text{ PRED}) = \text{'SAY'}$   
                                        $\{g_1 = \uparrow\}$

In these entries, the functional annotations make use of local metavariables (such as  $g_1$  and  $g_2$ ) in the following way: The equation  $g_1 = \lambda(\mathcal{M}^*)$  defines  $g_1$  as the variable name for an attribute-value structure that is  $\lambda$ -projected from the current element's c-structural mother (= the node I for *did*). The equation  $(g_1 \text{ PRED}) = \text{'do'}$  introduces a PRED feature and value into the attribute-value structure. The third equation  $\{g_1 = \uparrow\}$  identifies the local l-structure with the c-structural elements normal f-structure (i.e., its  $\phi$ -projection; recall that  $\uparrow$  is defined as  $\mathcal{M}^*$ ). What is crucial is that the third equation is enclosed in curly brackets, which means that it is applied optionally. Hence, the lexical specification will either be included in the f-structure or not. (Since the rest of the LFG structure leads to an identification of the f-structure projected from *did* and from *say*, maximally one of the two entries can introduce their PRED value to f-structure.)

The enormous variation opened up by this optionality is controlled by the faithfulness constraint DEP-IO, which in this setup can be formalized as in (15):

- (15) DEP-IO (referred to as FILL in early OT work)  
 General OT formulation (Kager 1999: 68): 'Output segments must have input correspondents.'  
 (*atomic-f-str*( $\star$ )  $\rightarrow$   
 $\forall n, P. [(cat(n) \wedge feature-path(P) \wedge (\lambda(n)P) = \star) \rightarrow (\phi(n)P) = \star]$ )

“For all categories  $n$  and feature paths  $P$ , if  $\star$  is an atomic value under  $P$  in the  $\lambda$ -projection from  $n$ , then  $\star$  is also the value under  $P$  in the  $\phi$ -projection from  $n$ .”

Recall that in the application of the *Eval* function, every structural element in a candidate representation – i.e., every c-structure node, every local f-structure, and also every local l-structure – will be tested for all constraints. For the top-most l-structure seen in Figure 5,  $\star$  will be instantiated to the PRED-value ‘do’, which is an atomic f-structure and which is also in the  $\lambda$ -projection of a node  $n$  (namely I) under a path  $p$  (namely PRED). So to satisfy the DEP-IO faithfulness constraint, PRED ‘do’ should also be in the f-structure for I, which it is not. This leads to the desired effect of capturing an insertion in c-structure which has no correspondence in f-structure as a DEP-IO violation.

To sum up, the insertions that are required to implement a generalized OT account are compatible with the restricted definition of *Gen*. By definition, the semantic information in a candidate’s f-structure is always the same as the semantic information in the underspecified input f-structure  $\Phi_{in}$ . Additional material can only occur in locally projected l-structures. While it is not forbidden to have infinitely many distinct candidate structures for an input f-structure, the restriction keeps the candidate set computationally tractable (Kuhn 2003: ch. 6).<sup>22</sup>

## 2.5 Excursion: The cognitive status of “directional” candidate set specification

As established in the previous subsection, it is possible to provide a declarative formal characterization of the language generated by an OT system. Thanks to the non-derivational character of (tuples of) LFG structures, it is possible to use the sharing of the semantic part of the structures as the defining element for candidate sets (independent of the question of how a computational system might be implemented that takes a string of words as an input and produces all fully specified LFG structure tuples with that surface string that are optimal for some input according to the OT system for a given language). Moreover, it can be shown that the task of determining whether a string of words is in the language generated by an OT system is not computationally undecidable. This subsection is an excursion that discusses a reaction that LFG practitioners might have when confronted with the multi-step breakdown of the abstract parsing task for an (LFG-based) OT system sketched in Figure 4: the assumption of a back and forth

---

<sup>22</sup>Exploiting results from Kaplan & Wedekind (2000), recursive loops that lead to infinite candidate sets can be captured in a tractable way.



of generation and parsing processes seems to go against the goal of developing a cognitively plausible model architecture to capture grammatical knowledge.

The intuitive algorithmic breakdown of parsing with plain LFG does not carry over to parsing in the OT-LFG framework, which necessarily has to incorporate a production-based directionality, while parsing models comprehension. Parsing with a plain LFG grammar follows the logic of the structure specification step by step: given a string, matching lexicon entries and c-structure rewrite rules are used to construct a set of c-structure trees spanning the string. To narrow the set down to the trees which correspond to a valid f-structure, functional descriptions from the lexical annotations and the rule annotations are then taken into account in a process of model construction (in a feature logic), and wherever we find a valid f-structure, we have one possible analysis for the input string. The intuitive simplicity of the algorithmic breakdown has presumably played a role in the attraction of the LFG formalism as a psychologically realistic framework – in particular as it contrasted with the model of Chomskyan derivational approaches, which work with a notion of an underlying deep structure (D-structure or numeration). The theoretically motivated transformation of an underlying deep structure into a surface structure does capture intuitions regarding the highly systematic relationship between expressions like *the cat drinks milk* and *what does the cat drink?* or between *she saw the cat* and *the cat was seen*. But in comprehension, a listener is confronted with the linear string from a surface structure (say, *the cat was seen*), and there is no cognitively intuitive algorithmic process that leads back to conceivable underlying deep structures. When spelling out a parsing procedure with OT-LFG, we now find ourselves in a similar situation: if we translate the bidirectional characterization of the set of valid structures for a given string into an algorithmic procedure, we do not arrive at a plausible rendition of what could be going on in comprehension.

One might suspect that these considerations challenge the cognitive plausibility of Optimality Theory. It should be noted however that there is an oversimplification in the reasoning that assumes a direct conceptual mapping of a declarative specification of some function (such as the function from a string of words to a set of c-structures associated with valid f-structures) to the seemingly straightforward algorithmic breakdown of this function. It is misleading in the general case to assume that an intuitively appealing translation of a composite function into some procedure is the only option for realizing the theoretically motivated function in a cognitive system. As a matter of fact, there would be no computationally tractable parsers for standard LFG if one relied on the simplest translation of the conceptual steps underlying grammatical specification

in LFG into an algorithm. As the discussion in Maxwell & Kaplan (1996) illustrates, LFG parsing performs a highly sophisticated interleaving of the various sources of grammatical and lexical knowledge. Vice versa, Edward Stabler's work on the formalization of Chomskyan derivational grammar models (e.g., Stabler 2011, 2013) shows that algorithmic solutions do not have to stick to the seemingly counterintuitive procedures in the underlying theoretical characterization.

The complicated procedural breakdown of the parsing task for OT systems follows in fact from the decision to incorporate an important consideration into the conceptualization of grammatical knowledge that has nothing to do with the procedural knowledge needed for parsing a particular given input string, but rather reflects a language learner's and adult speaker's conscious or tacit knowledge of the expressive *potential* that lies in the language system: for a speaker of language X to know the grammatical way of expressing something in X amounts to knowing which other potential ways of expressing the same thing are not available in X.<sup>23</sup> During the acquisition of X, the speaker will have learned from exposure to adult speakers' language behavior which realizational variants can be completely excluded. So, the adult speaker's grammatical knowledge may well be thought of computationally as a "hard-wired" input-output mapping that freezes the patterns which have stabilized in the competition system (superseding a dynamic acquisition phase, during which the relevant constraint rerankings were triggered for language learner).

As a side note, the "pre-compilation" of a cascade of optimality constraints with their step-by-step filtering effect into a single input-output function was subject to research in the context of OT systems that can be fully formalized with finite-state systems. Karttunen (1998) proposed a special finite-state operator (so-called lenient composition) that has the effect of turning a sequence of individual constraints formulated as transducers into a single transducer with the same effect as an OT competition.<sup>24</sup> To achieve a similar computational effect for a syntactic OT-LFG system, the internal data structures built up during parsing (following a chart-based or dynamic programming approach) would have to be re-designed to simultaneously incorporate a production-based and a comprehension-based directionality – which could then be instantiated in a single bottom-up algorithmic pass, avoiding direct bidirectionality processing (compare the "interleaved" bidirectional processing approach proposed in Kuhn

---

<sup>23</sup>When working with a classical grammar formalism, the formal model of a speaker's grammatical knowledge provides no way of making this differential knowledge explicit, yet to arrive at a particular formal grammar in language acquisition, the learner must have pruned away certain realization options that the language under consideration does not exploit.

<sup>24</sup>A similar approach is discussed by Frank & Satta (1998).

2000b). Since work in computational linguistics focusing on learning without any language-specific prior knowledge had already reached a relatively advanced state using different approaches, there was no substantial practical interest in putting the full OT-LFG account to use on a larger scale.

## 2.6 OT-style constraint ranking in broad-coverage grammars

Section 1.2 provided an illustration of the constraint ranking mechanism implemented in the XLE system, which is widely used in grammar development. The examples showed how violable constraints allow grammar writers to include rule variants for infrequently occurring constructions without causing a proliferation of implausible readings for canonical constructions.<sup>25</sup>

The XLE implementation of violable constraints via a special *o*-projection provides some further mechanisms that are of high practical value in broad-coverage grammar development. The specification of the ranking of the optimality marks in (16) illustrates some of these mechanisms. We will shortly go through a number of details, but first of all we note that the ranking is specified in the configuration section of the grammar code. This means that the relative ranking of the marks (and hence of the soft constraints) can be adjusted for different application scenarios of the grammar without changing the grammar code itself. Thus the ranking specification can be used to flexibly adjust the grammar to peculiarities of certain language registers or text domains.

- (16) OPTIMALITYORDER +PPASOBL (UNDERSTOODOBJ  
TYPECOERCIONTO TIME NOUN) NOGOOD MISSING3SGMARKING  
STOPPOINT FRAGMENT.

In (16), the highest-ranking mark PPASOBL is preceded by a plus sign. This indicates that in the filtering of analyses, the mark is not considered to be negative, but positive. When the available readings differ in the count of PPASOBL, the ones with the *maximal* number of marks survive. This provides grammar writers with a way of giving preference to a certain variant rather than having to “punish” a different variant, which, depending on the feature representation adopted, may be impractical. Rule (17) exemplifies the introduction of PPASOBL as a preference mark for the oblique object analysis of PPs (like in *wait for someone*). The oblique reading will be preferred over the alternative of analyzing the

---

<sup>25</sup>Computationally, we can note in the light of Section 2.4 that XLE’s OT-style constraint ranking is not normally used to modify the notion of grammaticality from the base grammar; hence, the additional complexity of a two-way application of the grammar in parsing and generation mode does not arise.

PP as an adjunct whenever both are available. This has the effect of reducing the number of parses for a considerable number of input sentences. In real-world uses of grammars, the available information makes it often hard to make an informed decision. So, in many application scenarios, it may be preferable to make a (more or less arbitrary) decision in favor of one of the variants.<sup>26</sup>

$$(17) \quad VP \longrightarrow V \left( \begin{array}{l} \text{NP} \\ (\uparrow \text{OBJ})=\downarrow \end{array} \right) \left\{ \begin{array}{l} \text{PP}^* \\ \downarrow \in (\uparrow \text{ADJ}) \\ (\uparrow \text{OBL}_\theta)=\downarrow \\ \text{PPasOBL} \in o(*) \end{array} \right\}$$

The ranking in (16) also illustrates the workings of parentheses and of some specially defined keywords. The marks UNDERSTOODOBJ and TYPECOERCIONTO-TIMENOUN, which were introduced in Section 1.2, are now jointly enclosed in parentheses to treat them as equally ranked. In grammar writing practice, such constraint ties are frequently used for phenomena that are independent from each other, since there is no grammar-internal justification for giving preference to one of them.

To the right of the two we see the mark NOGOOD. This is a predefined keyword that has the effect that all marks that follow receive a special interpretation that is best explained with a concrete example. Consider MISSING3SGMARKING: this mark is introduced in the definition of the template SUBJNON3SGAGR in (18). This template is used in present tense verb forms of English like *I laugh* or *they laugh*. There is a similar template for third-person singular forms like *she laughs*. The third disjunct in (18) covers the use of the form *laugh* with a third-person singular NP like in *\*she laugh*, which is ungrammatical in standard English. By providing this option, the grammar will robustly cover agreement mistakes or it can be used for varieties of English that include this variant.

$$(18) \quad \text{SUBJNON3SGAGR} \equiv \left\{ \begin{array}{l} (\uparrow \text{SUBJ NUM}) = \text{PL} \\ | (\uparrow \text{SUBJ NUM}) = \text{SG} \quad (\uparrow \text{SUBJ PERS}) \sim = 3 \\ | (\uparrow \text{SUBJ NUM}) = \text{SG} \quad (\uparrow \text{SUBJ PERS}) = 3 \\ \text{MISSING3SGMARKING} \in o(*) \end{array} \right\}$$

<sup>26</sup>Contrary to the situation with very infrequent constructions that were discussed in Section 1.2, the filtering may here have the effect that a contextually inappropriate analysis is chosen over the more appropriate analysis (for instance in *Sue was waiting for hours*). But in a range of applications this may not be too problematic, while a reduction of the sources of ambiguity can be extremely helpful during the process of extending the grammar or fixing a certain problem. When the grammar is used in application scenarios in which it can be harmful to occasionally choose the contextually inappropriate variant, the preference mark can simply be taken out of the ranking specification, so the parser outputs both variants.

With the ranking specified as in (16), *she laugh* will receive an analysis; however due to the use of the NOGOOD mark, it will be labeled as ungrammatical by the parsing system.

A last point illustrated by (16) is the predefined mark STOPPOINT. This mark offers a way for grammar writers to control the computational behavior of the parser. Putting an optimality mark to the right of STOPPOINT (like FRAGMENT in the example) has the effect that the functional descriptions that it marks will not be used at all in the first pass of running the parser. However, if parsing leads to an empty set of valid analyses, the parser will be reset for a second pass, this time including marks like FRAGMENT. This mechanism can be used to make the grammar more robust without compromising the runtimes of the parser for “well-behaved” input sentences. A typical example for using STOPPOINT in the ParGram grammars is as part of a fall-back option for covering strings that do not receive an analysis with the standard root symbol of the grammar. An artificial category like FRS (for fragments) is provided as an alternative root symbol. (19) is a schematic depiction of a recursive rule for such a category. Its purpose is to collect c-structure fragments such as NPs, PPs and certain incomplete verb projections. With such a rule, problematic input sentences (e.g., with misspellings or rare constructions) will still receive an analysis for the parts that are covered correctly. The partial f-structure contributions are collected in a first/rest data structure. Note that each fragment introduces one FRAGMENT mark, so the filtering mechanism will output the option with the fewest (i.e., on average the largest) fragments. It is even possible to use several instances of STOPPOINT in a ranking to potentially trigger several resets.

$$(19) \text{ FRS} \longrightarrow \left\{ \begin{array}{l} \text{NP} \quad | \quad \text{PP} \quad | \quad \text{VP} \\ (\uparrow \text{FIRST})=\downarrow \quad (\uparrow \text{FIRST})=\downarrow \quad (\uparrow \text{FIRST})=\downarrow \\ \text{FRAGMENT} \quad \text{FRAGMENT} \quad (\uparrow \text{SUBJ PRED})=\text{'pro'} \\ \in o(*) \quad \in o(*) \quad \text{FRAGMENT} \in o(*) \end{array} \right\} \left( \begin{array}{l} \text{FRS} \\ (\uparrow \text{REST})=\downarrow \end{array} \right)$$

As the various mechanisms we briefly discussed show, a flexible ranking for specially marked parts of the grammar conveniently puts grammar writers in a position of exerting control over the set of valid structures that an LFG grammar will assign to a given input string in parsing (and similarly for a given input f-structure in generation).<sup>27</sup>

<sup>27</sup>One way of looking at the addition of rankable constraints to an LFG grammar writer’s means of expression is to include ideas from grammatical frameworks that never factorized the task of disambiguation out, most notably Constraint Grammar (Karlsson 1990).

An alternative approach to controlling the space of ambiguities is the use of probabilistic techniques, which have been widely applied in the context of structural analysis of realistically occurring language (Collins 1997; Riezler et al. 2000, 2002). Here, the approach to the ambiguity problem is to rely on supervised training of probabilistic models that predict the distribution of alternative linguistic representation structures, dependent on a variety of contextual factors. If a sufficiently large training corpus is available that was manually disambiguated by competent speakers of the language, a so-called treebank, the complex interaction of knowledge sources on the contextually appropriate choice of readings can be captured quite reliably. When the sentences parsed in the application scenario are similar enough to the training corpus, the disambiguation quality that can be reached is typically higher than in a knowledge engineering approach of classical grammar writing, since statistically relevant patterns of all kinds (e.g., word order preferences, lexical-semantic argument selection preferences, but even statistical effects unrelated to grammatical knowledge) are learned “in passing”.

The XLE system offers both the optimality ranking approach discussed in this section and a probabilistic filtering approach that relies on supervised treebank training. In the practice of broad-coverage grammar development with a highly expressive formalism such as LFG, both mechanisms have their place and a combination is arguably the most effective way to go: a probabilistic approach exploits the empirical distribution of interacting factors, such that a sufficiently expressive probabilistic formalism (or machine learning model) will induce implicit statistical knowledge even about patterns that could not (yet) be captured in symbolic terms (see, e.g., Cahill et al. 2007; Forst 2007). On the downside however, a plain probabilistic approach leaves little leeway for grammar writers to inject specific symbolic knowledge about certain constructions. By using symbolic optimality ranking as a pre-filter for the set of candidate analyses going into treebank training, the grammar writers can easily experiment with alternative strategies (King et al. 2000).<sup>28</sup>

---

<sup>28</sup>It has to be noted that in the combined setup, the grammar writer is not in a very informed position to determine the relative ranking among the optimality constraints for multiple different linguistic constructions – for instance UNDERSTOODOBJ and TYPECOERCIONTO TIME NOUN from Section 1.2. This is something that an empirically informed training procedure can do better. The utility of symbolic soft constraints for linguistically informed ambiguity management lies more in the flexibility of experimenting with preference, dispreference and delayed execution (via one or more STOPPOINT marks) of constraints. By leaving the ranking of OT marks within a section (before NOGOOD, between NOGOOD and STOPPOINT, etc.) very flat through the use of parentheses, the ranking decision is postponed to the subsequent probabilistic filtering module.

There are a number of publications that report on the use of ranked constraints in various contexts of grammar specification, for instance Zaenen & Crouch (2009); Bögel et al. (2009); Dione (2014).

### 3 Linguistic applications of the competition-based concept of grammaticality

Section 2.6 provided details on the OT-style ranking mechanism that offers very effective functionality for ambiguity management in broad-coverage grammar development. We now go back to the original, theoretically motivated OT syntax model that employs competition among candidates to determine the grammatical way of expressing an underlying meaning. The model has been broadly employed to (a) systematically derive variation patterns within a language and (b) predict typological patterns across languages.

A broad range of syntactic phenomena have been addressed with OT syntax approaches. We will take a closer look at a few accounts in this and the subsequent sections – mostly to illustrate some specific properties of OT systems, in particular in the guise of OT-LFG and extensions that have been proposed. A full overview of all important phenomena addressed in the literature is beyond the scope of this chapter.<sup>29</sup> This section focuses on two important predictive schemes that the OT approach offers at the interface between syntax and morphology. Section 3.1 shows how morphological blocking phenomena can be derived in a very general way. Section 3.2 reviews the harmonic alignment account of the typological spectrum of differences in argument linking.

#### 3.1 Generalizing blocking accounts to incorporate morphology-syntax competition

LFG's system of corresponding parallel representations can be straightforwardly integrated in the competition-based grammaticality account of OT. This opens up a path for formulating a generalized theory of morphological blocking that was described before a generic mechanism of comparison was included in the overall formal framework. Prominent examples are the accounts by Andrews (1982, 1990), building on top of the Elsewhere Principle from phonology (Anderson 1969). The idea of morphological blocking offers an explanation of how

---

<sup>29</sup>Important areas excluded for space reasons are for instance positional alignment accounts of phrase structure such as Sells (1999, 2001). Sten Vikner and collaborators have worked out a detailed account of object shift in OT (Vikner 2001; Engels & Vikner 2014).

within a morphological paradigm such as the inflection of English verbs in present tense, an unmarked form (like *laugh*) can fill all cells in which no explicitly marked form (like the third person singular form *laughs*) is available. When alternative forms are available that express different degrees of specificity with respect to certain morphosyntactic features, the existence of a marked form for some specific feature combination (or position in a morphological paradigm) blocks the use of an unmarked form for this particular combination: *laughs*, which is marked for PERSON and NUMBER (to the values 3 and SINGULAR, respectively), blocks the unmarked *laugh*, which is underspecified for PERSON and NUMBER. Technically, Andrews (1982) proposes a blocking condition which states that a less specifically marked form A cannot be used in a position X if there is a form B that comes with a more specific marking, subsumed by A's specification.

With a competition-based definition of grammaticality, blocking effects can be construed as a consequence of general constraint interaction (Bresnan 2002, 2001a): the unmarked form is assumed to incur faithfulness violations, since it does not explicitly realize the underlying feature information in the input. For each inflectional category, a faithfulness constraint (e.g., FAITH<sup>NUM</sup> for number) checks whether a surface form accurately marks the underlying feature. On the other hand, a markedness constraint is assumed for each specific feature value, punishing the explicit marking of this value (e.g., \*PL "avoid marking the plural explicitly", \*SG "avoid marking the singular"). The markedness constraints implement the tendency in natural language to keep expressions as concise as possible. On the basis of these two antagonist constraints, learning their relative ranking for a given language<sup>30</sup> has the effect of learning in which paradigm cells to use a marked form vs. the unmarked form:<sup>31</sup> if \*PL outranks FAITH<sup>NUM</sup> in a language, the plural (of the word class under consideration) is realized by an unmarked form in this language; under the reverse ranking, a marked form is used for plural. For fusional morphology, in which a single morpheme (like for instance the -s in English present tense verb forms) can realize person and number simultaneously, conjunctive faithfulness constraints to sets of inflectional categories have to be assumed besides faithfulness to an individual inflectional category, for instance FAITH<sup>PERS&NUM</sup> (Bresnan 2001a: ex. 22). The verb inflection paradigm for present tense in modern English can be predicted by the following ranking: \*PL, \*1, \*2 ≫ FAITH<sup>PERS&NUM</sup> ≫ \*SG, \*3 ≫ FAITH<sup>PERS</sup>, FAITH<sup>NUM</sup>. Plural as well as first

<sup>30</sup>The learner acquires the constraint ranking through exposure to output produced by adult speakers; the speakers' underlying input has to be inferred from the situational context.

<sup>31</sup>To capture the finegrained differentiations inherent to inflectional paradigms, the faithfulness constraints have to be parametrized, for instance to specific verbs/verb classes for which learners have to learn distinct patterns.



and second person is never marked, since the markedness constraints for these feature values outrank all faithfulness constraints. For the combination of PERSON SINGULAR and NUMBER 3, a fusional form is used, since faithfulness to the combination of PERSON and NUMBER outranks the markedness constraints \*SG and \*3. For the other PERSON/NUMBER combinations, the fully unmarked form (*laugh*) is used, since FAITH<sup>PERS</sup> and FAITH<sup>NUM</sup> rank lower than the markedness constraints \*SG and \*3.

Given this characterization of the task of learning inflectional paradigms, the following typological spectrum is opened up by the interaction of faithfulness and markedness constraints: (i) when the markedness constraints outrank all faithfulness constraints, a paradigm with no inflectional distinctions follows; (ii) when faithfulness outranks all markedness constraints, all paradigm cells go along with an explicit forms; (iii) blocking effects occur when faithfulness is ranked in between certain markedness constraints. The account then predicts features (or feature combinations) whose markedness constraints outrank faithfulness to be realized by an unmarked form.

The OT-LFG framework makes it even possible to generalize the OT account of blocking to situations where it is not just alternative synthetic word forms that could be used to express an underlying feature bundle, but syntactically complex expressions are an additional alternative. Speakers of English have learned for instance when to use the analytical realization of a comparative adjective or adverb (such as *more quickly*) instead of a synthetic realizations (such as \**quicklier*). In LFG's system of imperfect correspondence among parallel representations (Bresnan 2001b), such alternatives are just different surface realizations of the same f-structure. Now, the set-up in OT-LFG is to have such alternatives compete for the status of the most harmonic candidate. It is clearly possible for a theorist to find constraint sets that will lead to analytical realization of a phenomenon in one language an synthetic realization in another. This alone may not be considered a strong argument in favor of a competition-based framework using parallel representation structure like LFG. However, when it can be shown that having analytical and synthetic alternatives side-by-side in the candidate set for realizing an input (i.e., expanding the same partial f-structure) leads to a systematic explanation of variability in inflection paradigms that mix analytical and synthetic realizations – via a generalization of the morphological blocking effect – this constitutes persuasive evidence that the architecture of the theoretical account does capture aspects of the human cognitive system quite well. This is exactly what Bresnan (2001a) achieves by the account of negation in varieties of English she proposes.

The tableaux in Figure 6 illustrate competitions among different analytic c-structural realizations of negation; in Bresnan (2001a: ex. 43), these tableaux serve to motivate the constraint set for an account in which an analytical form blocks a synthetic form. The analysis assumes two alternative realizations for the negation of verbs or auxiliaries in English: *not* can adjoin to the auxiliary itself, which is realized in I<sup>0</sup>, or it can adjoin to the VP. (For the modal verb *can*, the orthographic rules of English happen to make the distinction visible in the written form.) By hypothesis, both alternatives can have the meaning of a wide-scope negation, but only the latter can mean negation of the VP. Bresnan assumes one markedness constraint for each of the two possible sites for adjoining negation, \*NEG-VP and \*NEG-I; in English \*NEG-VP is ranked higher than \*NEG-I. Faithfulness to the negation scope (i.e., the constraint FAITH<sup>NEG</sup>) however is ranked higher than both markedness constraints. As an effect, the NEG-I option (*cannot*) arises as the optimal realization for a wide-scope reading of negation, whereas the more marked analytic form (*can not*) is required to express the VP scope of negation.

$\neg(\text{POSS}(\text{work}(\text{he})))$	FAITH <sup>NEG</sup>	*NEG-VP	*NEG-I
☞ a. <i>he cannot have been working</i>			*
b. <i>he can not have been working</i>		*!	
POSS( $\neg(\text{work}(\text{he})))$	FAITH <sup>NEG</sup>	*NEG-VP	*NEG-I
a. <i>he cannot have been working</i>	*!		*
☞ b. <i>he can not have been working</i>		*	

Figure 6: Two tableaux from Bresnan (2001a: ex. 43)

For the realization of negated forms of the auxiliary *be* in various varieties of English, analytical forms compete with synthetic forms: the negated third person singular can be realized as *is not* or as *isn't*. Moreover, a synthetic form that is unmarked for person and number is available: *aren't*. Interestingly, although Standard English has a marked form for declarative first person singular (*am*), there is a lexical gap for the negated first person singular.<sup>32</sup> In negated interrogative clauses, this gap is – for many speakers – filled by the unmarked *aren't* (examples from Bresnan 2001a: ex. 14-15):

- (20) a. \* Am I not going?  
 b. I am not going.

<sup>32</sup>For synchronic learnability of such an idiosyncratic gap, it is not relevant how the gap came about. Bresnan (2001a: fn. 26) mentions stigmatization of an older synthetic form *ain't* as a potential explanation.

- (21) a. Aren't I going?  
 b. \*I aren't going.

This effect can be derived in an OT-LFG analysis that assumes high-ranking markedness constraints punishing analytic negation adjoining either to  $C^0$  (which would yield \**Am not I going?*) or to VP (yielding (20a) (Bresnan 2001a: ex. 61-63). These markedness constraints outrank the constraint  $\text{FAITH}_{be}^{\text{PERS\&NUM}}$ , which regulates faithfulness to the person and number feature for the auxiliary *be*. A hypothetical synthetic form \**amn't* marked for person and number is unavailable due to the idiosyncratic lexical gap in English that was just discussed.

How can the grammatical framework model that a person acquiring English learns about such an idiosyncratic gap? In the OT framework, it has been proposed to assume a constraint *LEX* parametrized for specific lexical material and incurring a violation whenever it is used. Learners of a language with an idiosyncratic gap will rank the respective *LEX* constraint above all other constraints (because adult speakers never use this material when one would expect them to, based on the context).<sup>33</sup> For the English speakers using (21a) rather than (20a) to fill the lexical gap in the interrogative case, the third analytical option, adjoining negation to  $I^0$ , is ranked lower than  $\text{FAITH}_{be}^{\text{PERS\&NUM}}$ . This has the effect that the pattern in (20)/(21) is predicted: when *be* can be realized in  $I^0$  as is the case in a declarative clause, the fully marked analytic form is the most harmonic; in a question however, where *be* is in  $C^0$ , the unmarked form *aren't* wins out.

This analysis demonstrates the explanatory potential coming from a competition-based account of grammaticality that makes distinct grammatical means available in candidate sets based on an input corresponding to the underlying content.

### 3.2 Harmonic alignment

In many languages, certain properties that argument phrases like subjects and objects can bear (e.g., first person vs. third person, full NP vs. pronoun, overt case marking, but also the choice of grammatical relation itself) are correlated with the availability of grammatical syntactic realization options, for instance in

<sup>33</sup>Note that assuming constraints sensitive to specific lexical material in a language does not go against the principle of richness of the base, which excludes language-particular restrictions on the candidate set. However, it is not fully compatible with the assumption of a (finite) universal set of constraints. From the point of view of learning algorithms, it seems quite plausible however that instances of some constraint schema can be parametrized by lexical items that the learner has added to their inventory. See also van der Beek & Bouma (2004) for a discussion of language-particular lexicon properties within OT-LFG.

a clause with a transitive verbs. In the Australian language Dyirbal for example, the case marking patterns for transitive verbs are sensitive to such properties: “1st/2nd person pronouns are marked when they are objects, but not when they are subjects” (Aissen 1999: 674). When looking at the distribution of the relevant properties across languages, typologists have made the following observation: it is possible to organize these properties along markedness scales in such a way that the different scales tend to align with each other (Silverstein 1976). In a relational scale, subjects are used for the most salient/central arguments, followed by objects for less salient arguments, followed by obliques. In terms of thematic roles, agents are more prominent than patients. Animacy hierarchies have first and second person pronominals at the top of the scale, followed by third person pronouns, common nouns referring to humans, to animate referents and finally inanimate referents. Aissen (1999, 2003) develops an influential OT syntax account<sup>34</sup> demonstrating that many fine-grained observations from typological studies can be explained when the following assumption is made: the OT constraints that make reference to the various markedness or prominence scales cannot be arbitrarily (re-)ranked, but there are universal subhierarchies that are imposed over families of related constraints. These subhierarchies, technically implemented by the mechanism of harmonic alignment, have the effect that the various different markedness or prominence scales are systematically aligned.<sup>35</sup> For certain pairs of constraints, the relative prominence is fixed *a priori*,<sup>36</sup> while their interaction with other factors can still be freely learned from the observations.

For instance, Aissen’s (1999) account explains the split ergativity patterns in Dyirbal, where under specific conditions argument phrases are realized without case marking: the subject is generally unmarked when it is first or second person; the object is unmarked when it is third person. When the subject is third person, case has to be marked; likewise when the object is first or second person. (22) shows the OT subhierarchies that ensure the alignment of the relational scale and the person scale (combining first and second person as “local”). In essence, it is more marked to align a high element from one scale with a low element

---

<sup>34</sup>Aissen does not explicitly couch her account in an OT-LFG setting, but it is fully compatible and has greatly influenced subsequent OT-LFG work.

<sup>35</sup>The technique of harmonic alignment across prominence scales was already introduced by Prince & Smolensky (1993) for phonological features (sonority and syllable structure).

<sup>36</sup>Zeevat & Jäger (2002) demonstrate that the effect of the subhierarchies may also follow empirically from a systematic skewedness in patterns of usage. To the extent that this skewedness follows from invariant aspects of human social interaction, etc., it is presumably hard to tell empirically whether *a priori* rankings should be assumed within in the language faculty.

from another scale than to align either two high elements (Su/local) or two low elements (Ob/3rd).

- (22) a. \*Su/3rd >> \*Su/local  
 b. \*Ob/local >> \*Ob/3rd

To capture the case marking patterns, each of the alignment constraints is locally conjoined with the constraint  $*\emptyset_{\text{CASE}}$ , which punishes expressions that do not use overt marking for the respective combination – similar to faithfulness constraints. Local conjunction ( $C_1$  &  $C_2$ ) of two distinct OT constraints  $C_1$  and  $C_2$  within a given local domain  $D$  is a mechanism that captures the fact that in certain cases, it can be more marked when the two constraints are violated within the same local domain, for instance the same argument phrase, than when there are two independent violations of  $C_1$  and  $C_2$  (Smolensky 1995). Since the local conjunction  $C_1$  &  $C_2$  can be ranked independent of the individual constraints, special markedness patterns that are sensitive to the conjunction can be learned.

In Aissen's (1999) account, the universal subhierarchies assumed for alignment constraints like \*Su/3rd carry over to the family of their local conjunctions:

- (23)  $*\emptyset_{\text{CASE}} \& *Su/3rd \gg * \emptyset_{\text{CASE}} \& *Su/local$

Learning the case marking patterns for a particular language amounts to learning where within the universal subhierarchy a structural markedness constraint for the relevant grammatical feature is placed in that language – here the constraint  $*\text{STRUCT}_{\text{CASE}}$ . In Dyrbal,  $*\text{STRUCT}_{\text{CASE}}$  splits up both of the two hierarchies from (22):

- (24) a.  $*\emptyset_{\text{CASE}} \& *Su/3rd \gg * \text{STRUCT}_{\text{CASE}} \gg * \emptyset_{\text{CASE}} \& *Su/local$   
 b.  $*\emptyset_{\text{CASE}} \& *Ob/local \gg * \text{STRUCT}_{\text{CASE}} \gg * \emptyset_{\text{CASE}} \& *Ob/3rd$

Hence, it is more harmonic to avoid using a structural case marking on an argument when this is the subject and first or second person, while for third person subjects the high ranking of the conjunction  $*\emptyset_{\text{CASE}} \& *Su/3rd$ , excludes an unmarked subject in favor of a case-marked one.<sup>37</sup>

---

<sup>37</sup>We will come back to Aissen's harmonic alignment account in Section 4.2, as it forms the central target of Newmeyer's (2002) critique of functionally motivated constraint sets.

## 4 Extensions and debates

The competition-based definition of grammaticality in OT with its typological predictions attracted considerable attention in linguistic research communities, at the same time triggering debates about consequences of the new formal framework. Extensions were proposed to capture aspects of language(s) that the standard OT setup does not bring out. One such phenomenon is free variation, which one would expect not to exist under a plain OT approach. Section 4.1 discusses stochastic OT, an extension of the OT approach that does capture free variation.

Section 4.2 addresses a debate regarding the motivation of OT constraints. A substantial part of the OT community has been following the practice of providing a motivation for each constraint they assume which is grounded in functional considerations (for instance physiological considerations in OT phonology). This led to controversies, which are illustrative for the conceptual status different researchers assign to the constraints formulated in a theory of grammar. Section 4.3 steps back and discusses some of the cognitive considerations that led to the proposal of a competition-based account of knowledge of grammar in the first place. The section links this specifically to the question of learnability.

A second important extension of the basic architecture, bidirectional OT, is discussed in Section 4.4. It is motivated, *inter alia*, by the so-called phenomenon of word order freezing in languages with a (relatively) free order. In word order freezing, a clausal pattern that one would actually expect to be ambiguous *de facto* receives only one interpretation.

### 4.1 Extension I: Stochastic OT

One counterintuitive prediction of the competition-based definition of grammaticality is that languages should display very little (if any) free variation among two equally grammatical ways of expressing the same thing – for example *we need a more catchy title* vs. *we need a catchier title*. Even when it is just some minor and low-ranking constraint in which the two realization options differ, the OT system will by definition predict one variant to be ungrammatical for the relevant input.<sup>38</sup> Most natural languages *do* however offer free variation for certain lexical or grammatical means – there is for instance a certain amount of free word order variation in many languages.

---

<sup>38</sup>Of course, two equally harmonic candidates *can* arise when the relevant constraints are tied. In more comprehensive grammatical accounts however, this modeling option will only be available in exceptional circumstances, since most constraints play a role in multiple constraint interactions. So other phenomena will enforce a resolution of the ties.

The circumstance that a standard OT system covering a non-trivial range of phenomena is extremely unlikely to ever predict free variation is concealed by the fact that linguistic studies commonly focus their attention on a particular family of phenomena. When isolating a specific set of constraints that is relevant for deriving the observations regarding these phenomena, the possibility of ranking two or more constraints at the same level for a given language (= a constraint tie) *does* create a basis for explaining the systematic occurrence of free variation. For instance, Bresnan's (2001a) account of English auxiliaries discussed in Section 3.1 predicts variation among *cannot* and *can't*. However, for a more comprehensive account of grammatical knowledge, we have to assume that all the constraint sets posited for certain phenomena are combined in one larger constraint set. As a consequence, the effect of most constraint ties will go away – since the alternatives will differ in properties that are of relevance for some independent account (e.g., *cannot* has an extra syllable).

The problematic implications that standard OT has for free variation triggered several independent proposals for an extension of OT systems that will naturally predict free variation (see, e.g., Müller 2014; Asudeh 2001). One of the most influential proposals is known as stochastic OT (Boersma 1998; Boersma & Hayes 2001).<sup>39</sup> It preserves most of the original OT architecture; the key modification is that the ranking of the OT constraints is no longer viewed as fixed and discrete, but (1) the rank of a given constraint is a value on a continuous scale, and (2) the constraints are assumed to oscillate stochastically around their (mean) rank. Hence, at the time of harmony evaluation for a particular OT competition, it is possible with a certain probability for two constraints that are close in rank to effectively swap on the prominence scale. As an effect, a stochastic OT system (SOT) can predict variation patterns. The learning algorithm proposed for SOT operates with incremental error-driven adjustment of the constraints' mean rank, which even puts the learner in a position to replicate the quantitative distribution pattern in the observed variation data. Asudeh (2001) shows that in combination with a harmonic alignment analysis following Aissen (1999), SOT can derive optionality patterns in Marathi (Indo-Aryan, India): for non-volitional transitive verbs like *saapaḍṇe* ('to find'), either of the two arguments can be realized as the subject (while the other is realized as the object).

Bresnan et al. (2001) demonstrate with an analysis of the corpus distribution of passives that the SOT approach, again combined with Aissen's (1999) harmonic alignment analysis, can explain strong quantitative effects in a language like English (which does not categorically enforce passive for certain person constellations among the arguments of transitive verbs) in parallel to strictly categorical

<sup>39</sup>A similar account was proposed by Anttila (1997).

patterns in Lummi (Straits Salish, British Columbia). In Lummi, the realization of a transitive verb with a third person agent and a first or second person patient in active voice is ungrammatical. The underlying input has to be realized in passive voice. Now, although in English a clause like *she invited me* is not ungrammatical, Bresnan et al.'s (2001) study showed that in corpora of spontaneous spoken English, there is a significant statistical effect of an elevated passive use in this constellation (i.e., *I was invited by her*) – a circumstance that is exactly what one would expect when the standard OT treatment of Lummi is extended to English in a stochastic OT framework.<sup>40</sup>

## 4.2 Functional motivation of the constraint set

In any linguistic account that makes use of abstract descriptive categories such as ‘syllable’, ‘subject’, ‘passive’, ‘quantifier’, etc., the symbolic expressions assumed to describe formal and content-related properties of linguistic utterances are theoretical constructs. They are not directly observable. However, most linguistic theories attempt to choose their central descriptive representations in a way that permits a mapping to empirically observable properties based on as few assumptions as possible: phoneme representations are chosen based on semantically distinguishable minimal pairs, for the definition of syntactic notions like subject, operationalized tests are advanced, etc. In the same vein, the candidate representations in most OT work (and definitely in OT-LFG) are chosen under an operationalized regime assuming that all candidate distinctions can be derived from the surface distribution and contextual clues reflecting semantic distinctions.

What about the choice of constraints used to drive the typological predictions? It is important to notice that even with perfectly uncontroversial, empir-

---

<sup>40</sup>In their criticism of Boersma & Hayes (2001), Keller & Asudeh (2002) argue that viewing a stochastic OT system not just as a model of variation/optionality, but of some notion of graded grammaticality tied to corpus frequencies, is conceptually problematic since it blurs the standard distinction between competence and performance. It should be noted however that certain systematic observations regarding quantitative distributions in corpus data seem to make it inevitable to revise some standard assumptions. Compare Bresnan's (2011) autobiographical notes: “Strikingly, the rare, marginal, and ‘incorrect’ construction types in large collections of English language usage parallel the rare grammatical phenomena that can be found across languages of the world. Moreover, judgments of ungrammaticality are often unstable and can be manipulated simply by raising or lowering the probability of the context. Most remarkably, language users have powerful predictive capacities, which can be measured using statistical models of spontaneous language use. From all these discoveries I have come to believe that our implicit knowledge of language has been vastly underestimated by theoretical linguistics of the kind I had practiced.” (Bresnan 2011)



ically grounded candidate representations, there are generally many extensionally equivalent choices of the constraint set for deriving the same distribution of optimal candidates (= the predicted language typology). For a given constraint set, an alternative set of *ad hoc* constraints with the same empirical prediction can be constructed. OT systems are empirically underdetermined in this respect (and this is no design fault; many formal systems employed as scientific models are underdetermined along certain dimensions). Therefore, to convince oneself that an OT analysis does indeed reflect a linguistically valid pattern, it is important to exclude that one or more of the constraints are ill-justified and merely play the role of getting the predictions right. Hence, starting out in phonological OT work, it has become good practice to provide a plausible motivation for each constraint, independent from this constraint's role *within* the constraint set – a “functional” motivation. In phonology, many constraints can be given a motivation based on considerations of articulation, aerodynamics or perception, for instance “a constraint against voiced obstruents, e.g., the Voiced Obstruent Prohibition (VOP, \*[+voice], \*Laryngeal) can be said to be functionally grounded, since vocal fold vibration is difficult to sustain if the outgoing airstream is blocked.” (Krämer 2017: sec. 3.2.3). In syntax, the argumentation often needs to be more indirect, for instance by putting forward general considerations of economy to motivate a constraint against movement in a derivational framework, such as STAY in Grimshaw's (1997) and similar frameworks. But throughout the application fields of OT, a wide-spread argumentation practice sees the need for independent justification of constraints beyond the fact that it reaches a particular effect within the factorial typology. Many researchers welcomed that OT syntax triggered a confluence of the formalist and the functionalist perspective on language and grammars.<sup>41</sup> Still, Newmeyer (2002) presents a vigorous argument against the conception of “functionally-based optimality theory” (FOT) accounts, specifically targeting Aissen's (1999; 2003) harmonic alignment account. As the argumentation shows however, and as Bresnan & Aissen (2002) argue in detail in their rebuttal of Newmeyer (2002) in the same volume of *Natural Language & Linguistic Theory*, the notion of constraints that Newmeyer's argumentation is based on is not the one inherent to the standard OT conception (which completely shifts the definition of grammaticality away from a rule-based system to the interaction of violable constraints). Newmeyer criticizes that by requiring constraints to be paired with an external functional motivation, FOT “incorrectly locates the form-function interplay in the mental grammar itself, rather than see-

---

<sup>41</sup>Haspelmath (1999) argues for the additional need to take diachronic evolutionary processes into account.

ing the response of form to function as emerging from language use and acquisition” (Newmeyer 2002: 43). Newmeyer draws into question whether “the claim that every constraint has a functional motivation” (Newmeyer 2002: 56) is empirically contentful for OT syntax. But if we take into account that constraints are abstract constructs which are not directly observable in any particular language, this formulation somehow reverses the need for justification that a theoretician should provide for their assumptions. Putting forward some plausible motivation for theoretical constructs, beyond the technically desired effect, responds to principles of scientific practice precluding arbitrariness in an underdetermined formal system. Providing a functional motivation for an OT constraints does not amount to an empirical *claim*, but it is part of the argumentation that the abstract choices made adhere to second-order principles of good scientific practice. It is only a full OT system that is empirically falsifiable; inadequacy may come to the surface when there is no way of extending a system which plausibly covers a core set of phenomena to clearly observable additional evidence.

### 4.3 Learnability in the context of the broader cognitive architecture

As has become clear from the discussions up to this point, a theorist’s decision to move from a conventional generative grammar formalism to the competition-based formal model of grammaticality underlying OT does not merely mean that instead of working with hard constraints they now work with violable constraints. The status of familiar descriptive devices becomes fundamentally different with the different characterization of the set of well-formed analyses. Some of the debates that this development triggered have been already addressed in the previous sections – for instance the question how ineffability might be modeled and how optionality/free variation can be accounted for.<sup>42</sup>

But it is worthwhile to pause and consider the status of the components of an OT system as a model of the language faculty within our broader cognitive system. The competition-based frameworks of Harmonic Grammar (Legendre et al. 1990) and Optimality Theory (Prince & Smolensky 1993, 2004) were originally developed to reconcile (i) the potential of connectionist networks for learning complex input-output functions from exposure to data following up on work in the Parallel Distributed Processing framework (Rumelhart et al. 1986)<sup>43</sup> with (ii) the insights from linguistic theory in the generative tradition, which models

---

<sup>42</sup>Wunderlich (2006), in his encyclopedia article on OT in morphology and syntax, provides a list of fundamental questions that arise when adopting an OT approach.

<sup>43</sup>The collection edited by Smolensky & Legendre (2006) is devoted to this perspective, but it does not seem to be very prominent in linguistic debates.

systematic generalizations in a formal system. Optimality Theory is an attempt to gain ground towards resolving some of the most central challenges for the cognitive sciences: understanding how abstract systematic knowledge (which is best described using recursive symbolic systems, relying on logical inference) is implemented in connectionist architectures and how it blends with associative knowledge (which is best captured in subsymbolic terms). We know that the neurophysiological basis for all our cognitive systems, the human brain, is a large and complex connectionist network. For artificial connectionist networks, the potential to pick up complex patterns from empirical learning data has been convincingly demonstrated for many scenarios. However, the abstract symbolic concepts that are at the core of many linguistic accounts of grammar – allowing for a very compact characterization of very far-reaching generalizations – turn out to be hard learning targets for a bottom-up empirical learning procedure with the comparatively simple artificial networks available (Marcus 2001).<sup>44</sup>

Given the complexity of the brain and our very preliminary understanding of the interaction of cognitive subsystems available to humans, it is no surprise at all that there is still a gap between our current understanding of systematic knowledge, captured via abstract concepts, and what we know about the level of neurophysiological implementation. Yet, when looking more specifically at knowledge about language, there seems to be a certain degree of impatience in the research communities taking a connectionist vs. an abstract symbolic approach – possibly because on both sides of the gap, mature theories and modeling frameworks have evolved substantially over the past decades, yet the key question of how the views go together remains rather open.

From the point of view of linguistics, it may seem that the concerns about the missing path across scientific levels are just a problem for cognitive scientists who believe that a connectionist implementation of linguistic knowledge needs to be spelled out in concrete terms. Under the working assumptions of many linguistic frameworks, details of a technical implementation are of subordinate importance as long as one can convince oneself that a symbolic approach could be implemented in principle. This thinking ignores however that many of the established frameworks, including LFG, had (and have) a major weakness when it comes to capturing the learnability of languages. This weakness is no embarrassment *per se*, since it was never in the focus of research interests; but it implies that it is not even clear in principle how an abstract description of grammatical knowledge could be implemented at the neurophysiological level – where it can-

---

<sup>44</sup>Compare e.g., the systematicity debate started by Fodor & McLaughlin (1990) (Buckner & Garson 2019).

not be the case that the relevant language-specific knowledge is symbolically encoded, but some representation needs to be learned from observable indications. Formally, the Principles-and-Parameters Theory (Chomsky 1981) has an answer to the question of learnability: the framework assumes an articulate structural system to be innate, such that learning amounts to setting a small number of switches, the parameters. The Minimalist Program (Chomsky 1995) attempts to derive the complex structural displacement patterns (which are required to reduce the complex sound/meaning relationship available in natural languages to a uniform construction plan and which by assumption must be predetermined by universal grammar) as a consequence of a small set of assumptions that are justified on the grounds of simplicity of the theory. Through innate universal grammar, the learner of a language has access to the principles that govern the displacement patterns and thus the learner only needs to learn the language-particular choice of linearizing the displacement configurations (Chomsky et al. 2019). Language learning is here conceptualized as a process of setting a number of discrete parameters. What is not very clear however is how the language acquisition system interacts with other parts of the cognitive learning system, many of which are quite clearly responding to the statistical distribution of cues. But linguistic knowledge could not be acquired and put to use if it had no effective interfaces with the statistically sensitive parts of cognition. It is hard to imagine that a learner can find out about the space of possible constructions in a language when they cannot rely on expectations regarding *typical* (= highly frequent) ways of saying something; when the learner notices that certain expectations were incorrect, this will trigger a highly informative learning step regarding the relationship between the language system and contextual factors. For instance, a learner can only learn about a rare phenomenon like heavy NP shift as in (25) if they have a notion of a canonical order, so they can reconstruct the conditions for deviations.

(25) I gave [<sub>PP</sub> to Sue ] [<sub>NP</sub> the books that I found on my aunt's attic ]

The aspect of statistical learning is where connectionist approaches turn out to capture the empirical behavior quite realistically – but a theory that construes language acquisition as an entirely different process provides no grounds for capturing such triggers and any frequency-related patterns.

The classical constraint-based theories of grammar, such as LFG and Head-driven Phrase Structure Grammar (see Przepiórkowski forthcoming [this volume]), *do* emphasize the objective of finding psychologically realistic models of the cognitive processes of production and comprehension (i.e., parsing and generation algorithms that start from an information state that corresponds to what

is available to a human listener or speaker). However, the goal of coming up with realistic algorithmic accounts capturing the information processing of real cognitive agents does not traditionally extend to the cognitive process of *learning* a language, i.e., acquiring the lexical and grammatical knowledge necessary to perform the tasks of production and comprehension. The grammar formalisms are not designed in such a way that there is a formal learning procedure that always starts from the same initial state, and then takes in observations from adult speakers of Bulgarian, English or Mandarin. But this would be required to have a falsifiable theory of the way grammatical knowledge is instantiated in cognitive agents. In the classical paradigm, the precise grammatical knowledge representation for a particular language is (still) specified by a scientific observer, the linguist, who is able to “look behind the scenes” and make far-reaching decisions about the use of certain descriptive means from a meta perspective – which is clearly not a realistic rendering of the information available to human learners (who however nevertheless reach the knowledge state robustly and fast). Of course, there are good research-strategic reasons why the classical paradigm stops short of also trying to model the cognitive process of acquisition: the grammatical knowledge that is available to adult speakers of the languages of the world is complex and the systematic workings of most language-particular systems are far from being understood. So one might say that the research community is still at the stage of clarifying what the exact targets for the acquisition process are, thereby avoiding a situation where it could not be truly judged whether a learning algorithm is on the right track from a theoretical point of view. An opponent could however argue that it is not clear whether the formalism that was designed so meta observers can specify a theory of adult linguistic knowledge provides the right concepts and interfaces to ever support a realistically learnable knowledge representation of grammar and the lexicon (for instance because it is unclear how associative knowledge merges in, as mentioned above). If there is any truth in this objection, the best strategy is probably to adopt a parallel strategy: advance systematic accounts of the adult linguistic knowledge and at the same time try to explore architectures that are better suited for modeling learning and for interfacing with knowledge that is more readily captured in a connectionist framework.

The design of Optimality Theory provides a link between the abstract symbolic level, tying in with established concepts from linguistic theory, and the level of connectionist implementation. From the point of view of LFG, the greatest conceptual gain from adopting an OT perspective might lie in the fact that this provides a fleshed-out learning algorithm (Tesar & Smolensky 1998), which is compatible with insights about the low-level generalizing behavior of connectionist approaches.

#### 4.4 Extension II: Bidirectional optimization

The standard definition of grammaticality in an OT system is based on what is often called a speaker-oriented, or production-based competition. Here, the most harmonic candidate analysis realizing some underlying content representation is determined. The characterization of the competing candidates reflects the fundamental knowledge that a competent speaker of the language has: they know which is the grammatical way of expressing some thought in their language. When they learned their language, they had to exclude other ways that would be possible in principle. The candidate set and rerankable constraints are thus a straightforward rendering of the mathematical search space for learning an input-output function capturing the speaker's competence.

A competent language user however has an additional ability: the listener is (mostly) able to reconstruct what the speakers wanted to express in the given context. Formally, the disambiguation problem for a given surface form can be construed as the mirror image of the task of realizing some underlying representation with the grammatical means of some particular language. Hence, it is tempting to explore to what extent the same competition-based architecture can be applied to model our ability to disambiguate. The representational setup of OT-LFG makes it particularly easy to implement the reverse competition: for a listener-oriented, or comprehension-based optimization, all that needs to be altered from the standard scenario is the basis for defining the candidate set. Here, all candidate analyses sharing the same surface string are compared.

A realistic model of our cognitive ability to make disambiguation decisions of course has to incorporate a lot more than just grammatical and lexical knowledge. On reading an ambiguous request/instruction like (26) (compare (27a) vs. (27b)), a reader may exploit frequency knowledge about the phrasal verb *read in* vs. the simple verb *read*, which one might argue is part of their extended lexical knowledge. However the utterance context of the request will probably play a much more important role: Does the building have a physical library? Or is the reader receiving programming hints? Ultimately, reasoning about what the speaker/author meant will draw upon any available world knowledge.

- (26) Read in the library.
- (27) a. Read in the comma-separated data file.  
b. Read in the living room.

So, taking all relevant knowledge sources into account, there is an asymmetry between production-based and comprehension-based optimization. Nevertheless, it is a valid question what disambiguation decisions are made when there

are no grammar-external clues. If we can isolate such cases, we might learn a lot about the way our grammatical knowledge is organized.<sup>45</sup>

A case in point are so-called word-order freezing phenomena, which have received considerable attention in OT frameworks. For languages with relatively free word order, the following type of effect is often reported: when there are no other disambiguating clues, listeners interpret examples as unambiguous that due to freedom of word order should actually have two possible interpretations. Bouma & Hendriks (2012), who provide a detailed discussion of OT treatments of word order freezing, give the Dutch example in (28):

(28) Dutch (West Germanic; Bouma & Hendriks 2012: ex. 4)

Fitz zag Ella.

Fitz saw Ella

Only ‘Fitz saw Ella’ (SVO), although structurally compatible with ‘Ella saw Fitz’ (OVS)

The interpretive preference for cases like (28) is directly predicted if one assumes that not only the determination of grammatical (surface) outputs follows a (production-based) optimization, but also the distinction among potential underlying interpretations in comprehension – assuming the very same constraint set.<sup>46</sup> Accounts that make use of this idea are called bidirectional optimization accounts. Early discussions of such an account in an OT-LFG setting are found in Lee (2001) and in Kuhn (2000a, 2003).

Assuming that comprehension-based competition plays a role in a model of grammatical knowledge immediately raises questions about the relationship between the two directions: in a standard OT setting, one would not want to assume in general that listeners *only* apply constraint evaluation on the possible analysis candidates of a surface string to retrieve the semantic interpretation; they should also double check that the surface string is also optimal in the reverse direction<sup>47</sup> – otherwise, a listener may end up with a candidate structure that is not even grammatical in the language that the speaker used. For illustration, let us

<sup>45</sup>Recall from Sections 1.2 and 2.6 that comprehension-based optimization is also predominantly used with the OT-style constraint ranking scheme implemented in the XLE system (Frank et al. 2001).

<sup>46</sup>Zeevat (2006) argues that bidirectional optimization is not an adequate account for word order freezing; but see Bouma & Hendriks (2012).

<sup>47</sup>The simple comprehension-based optimization *does* play a role in work in OT phonology. The mechanism of lexicon optimization, assumed by Prince & Smolensky (1993), is such a competition. Also, Tesar & Smolensky (1998) propose a procedure of robust interpretive parsing during the learning process.

consider what optimizations listeners have to take into account when processing the English sentence (29a) and the German sentence (29b), which is superficially a close correspondence of the English sequence.

- (29) a. Lou hopes not to arrive before Sunday.  
b. German  
Lou hofft nicht vor Sonntag anzukommen.  
Lou hopes not before Sunday to-arrive  
'Lou doesn't hope to arrive before Sunday.' or 'Lou hopes not to arrive before Sunday.'

If it was enough in comprehension to determine the optimal most harmonic structure for the observed surface string, we would expect that there is no great difference between the processing of the English and the German example. However, since in German simple negation of full verbs is grammatical, whereas in English it is not, there is a difference in the number of readings available. For (29a), a listener of English would never come up with the matrix negation reading – even in a context that would strongly favor it. This observation can be captured if we assume that a listener verifies that the most harmonic candidate in the comprehension-based optimization (with meaning  $\hat{m}$ ) is also the most harmonic in a production-based optimization (taking  $\hat{m}$  as the input). Demanding that a form-meaning pair has to be optimal both among all production-based and among all comprehension-based candidates is called *strong bidirectional OT*.

An additional variant of combining the outcome of the two optimizations was pioneered by Blutner (1998, 2000). With a so-called *weak bidirectional optimization*, not only the optimal form-meaning pairs from a competition play a role in defining meaningful expressions. The “runner-up” form-meaning pair after removing the overall winner from the competition is defined to express a more specialized form-meaning relation, and this can continue down a scale of expressions. The idea is best illustrated with the well-known example (30):<sup>48</sup>

- (30) a. John killed the sheriff.  
b. John caused the sheriff to die.

---

<sup>48</sup>Although the competitions assumed in bidirectional OT accounts in pragmatics like the one sketched here are inspired by the OT systems used to characterize the grammar of a language, it should be noted that a weak bidirectional approach would presumably not work if the full set of candidates was assumed, most of which are ungrammatical. OT pragmatics work focuses on relating grammatical expressions of different structural shape to the spectrum of potential meanings that can express.



(30a) is the unmarked way of expressing that John killed the sheriff. Logically, (30b) is equivalent; however, on hearing this sentence (and not (30a)), most listeners will take the speaker to implicate that the sheriff died in an indirect way from John's actions, assuming Gricean maxims. Weak bidirectional optimality predicts this by assuming that the overall competition will make the form/meaning pair of sentence (30a) and a proposition with a plain instance of a killing event the most harmonic candidate in both directions. The more marked form (30b), paired up with a "specialization" of the meaning is weakly optimal.<sup>49</sup>

The fields of OT semantics and OT pragmatics which in essence build on top of the idea of bidirectional OT are among the most prolific areas in terms of publications (see for instance the *Stanford Encyclopedia of Philosophy* entry van Rooij & Franke 2020). It has to be noted that the role that the constraints play in OT systems modeling systematic patterns in pragmatics is slightly different from the grammaticality-defining role of the constraint set in OT phonology and OT syntax; nevertheless, there are many systematic connection points between OT syntax and OT semantics/pragmatics, discussed for instance by Beaver & Lee (2004).

## 5 Developments after the mid-2000s

This section starts out with a discussion of the developments in the OT framework after the phase of the highest research activity in the late 1990s and early 2000s (Section 5.1) and furthermore asks how aspects of language learning are captured in OT vs. in mainstream work in Computational Linguistics and Natural Language Processing since the 2000s (Section 5.2). Section 5.3 addresses the relatively recent developments of neural modeling, i.e., the application of "deep" artificial neural networks.

### 5.1 Developments after the peak research activity

Müller (2012), in the conclusion to his survey paper on OT syntax, notes that after a phase of very high research activity, there have been comparatively few OT contributions in the area of syntax – contrary to the situation in morphology, semantics/pragmatics, and most notably phonology. To a certain degree, this development also holds for OT-LFG work on syntax. According to Müller's analysis of the situation, many ideas from derivationally based work in OT syntax live on

---

<sup>49</sup>The formal properties of bidirectional OT are discussed in Jäger (2002).

in the research strand of the Minimalist Program, in which for instance the decision to apply one of a number of competing elementary operations like Agree, Merge, Move, Delete etc. is resolved in different ways across languages, which can be captured in a ranking account (Müller 2012: sec. 5).

In this section, we address the question: what was the further development in the part of the OT syntax community that assumes nonderivational candidate analyses – in particular OT-LFG? Section 4.3 above ended with the observation that the couching of LFG in an OT framework adds a plausible algorithmic account of learnability of grammars, which the original formalism was missing. Against this background, one might have expected an increase of activity rather than a reduction, for instance in the computationally oriented LFG subcommunity. One can only speculate about the exact reasons for trends in research communities, but there seem to have been framework-internal factors, which are discussed in this section, and external factors in the broader computational community, which will be discussed in Section 5.2.

Looking at the framework itself, the introduction of a competition-based definition of grammaticality in syntax research, or in a subcommunity of syntax research, did not go along with a fixed catalogue of new principles that readily leads to the formation of a homogeneous and focused research paradigm. Rather, the introduction of OT can be seen as the starting point for addressing a broad bundle of phenomena and conceptual aspects that conventional formal grammar models were not able to capture. These aspects include quantitative aspects of linguistic knowledge that draw the strictly binary notion of grammaticality and the established split of grammatical competence and performance into question.<sup>50</sup> The fanning out of community activities into a whole range of development lines should probably not be seen as a failure of the new framework, but as a sign of the true complexity of empirical interrelations that transcended the scope of the conventional approach due to its strategic idealizing assumptions. Under such circumstances, new developments committed to the elimination of overly far-reaching idealizations cannot be couched in a single formal framework. Regarding the quantitative aspects for instance, extensions like stochastic OT (Section 4.1) provide certain alternatives, but in particular work by Bresnan (2007); Bresnan & Ford (2010); Bresnan & Nikitina (2010) went on to explore a broader range of modeling options. For other aspects, bidirectional optimization and other extensions of the architecture promise the most explanatory account.

---

<sup>50</sup>In her acceptance speech of the Lifetime Achievement Award by the Association for Computational Linguistics, Joan Bresnan shares her view of the development of the various descriptive frameworks that points in this direction (Bresnan 2016).

In parallel to these architectural considerations, linguistic work continues to study empirically robust grammatical patterns for which symbolic abstractions have been established – exploring a phenomenon in new languages and/or interactions among phenomena. It is hence quite natural that depending on the focus of particular studies, authors would choose to use an OT-LFG framework or a classical LFG setting.

## 5.2 The shift of computational syntax research in the 2000s

As just discussed, the OT approach unlocked a diverse range of new conceptual and empirical questions, which explains in part why we did not witness the development of a single, comprehensive research paradigm that couched all competition-based work in theoretical linguistics. But in addition, there were also framework-external factors that presumably precluded a streamlining of activities into a single coherent research paradigm. One of them is that the linguistically motivated exploration of the formal OT architecture was not accompanied by major synchronized efforts in a computationally oriented community<sup>51</sup> – contrary to the situation in the 1980s and early 1990s where formal and computational results for linguistic grammar formalisms like LFG received significant attention at the computational linguistics conferences. At first glance, this is surprising. OT as a move to a competition-based framework in theoretical work in linguistics *did* lead to an architecture that is structurally very similar to the architectures underlying the dominant data-driven approaches in computational linguistics and natural language processing (NLP) (Eisner 2000: 287). A high-level objective in both contexts lies in developing a model framework capable of using empirical (“training”) data from a natural language (in principle an arbitrary one) to induce a language-particular instantiation of the framework that replicates the competence/language behavior of a speaker of that language. This is reflected by the fact that the same families of learning algorithms could be employed (which led to interesting discussions between theoretical and computational linguistics, see Keller & Asudeh 2002; Goldwater & Johnson 2003; Jäger 2007). What differs across the research contexts despite the great similarities are three interleaved points.

First, linguistic OT work by design uses a strict constraint ranking to limit the degrees of freedom in theory development and to be able to connect to the

---

<sup>51</sup>The use of the idea of violable constraint for postfiltering the readings predicted by a classical grammar discussed in Sections 1.2 and 2.6 is so different conceptually from OT as a grammaticality-defining device that many of the implications of the latter do not arise for the former.

long tradition of work on linguistic generalizations at the level of symbolic descriptions of linguistic structures.<sup>52</sup> In practically motivated machine learning approaches in NLP however, there is no reason for positing such restrictions on the learnable input-output functions. Hence, the actual learning processes for some given language input are quite different in the two research areas. Key insights from decades of research on the theory of grammar have nevertheless been incorporated in most NLP modeling efforts of the time, since supervised learning on annotated corpora (“treebanks”) is the most effective approach, and the annotation categories in available corpora reflect many of the important distinctions from theoretical linguistics. Again however, the processes of generating new insights in theoretical vs. computational work is somewhat disconnected: improvements in NLP are based on enrichments of the competition and evaluation models – which is excluded by assumption in linguistic OT work.

Second, while both in linguistics and in NLP, competitions in both processing directions play a role, it is the production-based view that plays the defining role (for grammaticality) in linguistics, whereas it is the comprehension-based view (=disambiguation in parsing) that is the most fundamental in NLP.<sup>53</sup>

Finally, third – and related to the first and second point – there is a difference in the scope of empirical phenomena related to language and text targeted by linguistic vs. contemporary NLP work. A linguistic OT approach targets grammatical phenomena and therefore controls for context factors which would discriminate among candidates but are considered extralinguistic (or orthogonal to the phenomenon under consideration). This is important to be able to isolate the effect of the abstract grammatical categories responsible for systematic generalizations. NLP work on the other hand does not aim to isolate the linguistic knowledge involved in solving processing tasks – it pursues the objective of maximizing the predicted system scores for the tasks as they realistically occur (e.g., replicating the structure assignment decision human annotators made on real-life corpus sentences, where the annotators recurred to their linguistic knowledge as much as they exploited any explicit or implicit contextual clue). From the NLP perspective, it is therefore not only legitimate, but good practice to try to exploit correlations of the actual target task with any other trends reflected in the empirical training data. For instance, a syntactic parser, say, of English, trained with NLP techniques on a treebank from a certain domain will presumably make its decisions to a large part because it has picked up grammatical knowledge regarding the positioning of grammatical subjects and objects from the treebank. To a

---

<sup>52</sup>Stochastic OT assumes oscillating rank values for a constraint, but each evaluation is still based on the strict ranking from standard OT.

<sup>53</sup>Compare Sections 1.2 and 2.6 and the discussions in Frank et al. (2001) and Kuhn (2001, 2003).

certain degree, it will however also exploit non-linguistic domain knowledge reflected in the distribution of factual statements (e.g., a certain brand name being mentioned a lot more than others in a context like *buy a ... watch*). The difference in the scope of empirical phenomena targeted across the subfields implies very different foci in the modeling work.

The three differences in the configuration details and application of competition-based learning architectures imply different key research challenges in theoretical vs. computational work and explain how separate agendas have evolved. Nevertheless, there are numerous connecting points that become particularly relevant when the (often tacit) simplifying assumptions underlying the respective standard approaches are relaxed. For example, the idea of bidirectional optimization from theoretical work can be translated straightforwardly to machine learning models applied to disambiguation in parsing and choice in generation (Cahill & Riester 2009; Zarri   et al. 2011; Yu et al. 2019). An example of developments in which linguistic insights and considerations re-gained attention in the past years is the Universal Dependencies project (Nivre et al. 2016), which assembles treebanks for a growing number of languages in a cross-linguistically uniform dependency format (compare Haug forthcoming [this volume]).

### 5.3 OT and recent neural models in natural language processing

Section 4.3 addressed the connectionist motivation behind the original proposal of OT. Interestingly, over the course of the 2010s successes in machine learning with artificial neural networks (Henderson 2020) and the broad availability of high-capacity computing resources brought about a major shift in NLP research (as well as most other areas of applied machine learning), often associated with the buzz word of “Deep Learning”: neural network models replaced the conventional machine learning (ML) architectures discussed here in Section 5, which required the design and optimization of the set of ML features to reach the best generalizations from training data. (The machine learning features are the equivalent to the rankable constraint in OT – the only difference is that they are used in more general mathematical functions than in the strict relative ranking of OT.)

The training of “deep” neural networks does not rely on pre-designed ML features. Instead, it employs hidden layers of neurons that are densely connected with a neuron-based input or output representation and with other hidden layers. Weight parameters on all the connections are iteratively adjusted by supervised training based on input/output data, where weight adjustments for links involving hidden layers are percolated from the ends using backpropagation (which can be thought of as spreading out the activation of some neuron to all connected

neurons in the neighboring layers). The effect is that during empirical learning, the hidden layers receive the role of couching emerging internal representations of systematic patterns observed in the training data. For instance, when certain input neurons tend to be activated when a specific output occurs, some neuron in the next hidden layer can take over the role of an internal “feature” recording the constellation. With appropriately chosen network architectures that capture task-specific properties of the input and output representation (e.g., cross-talk among neighboring elements in a sequential input such as character or word token sequences), the neural approach led to substantial improvements over conventional ML approaches for a great variety of learning tasks.

The conceptual relationship between recent computational neural networks and the OT approach as popularized in the 1990s has not been discussed prominently in the research literature, but it is worthwhile comparing the major components in these architectures – in particular since the inherent black-box characteristics of neural models have prompted wide-spread efforts into making the emerging model representations scientifically interpretable (see, e.g., Belinkov & Glass 2019). At first glance, the move away from human-designed ML feature sets, which parallel OT constraint sets, seems to have widened the gap between NLP work and a linguistic notion of OT competition. Moreover, the effective ability of neural models to induce task-relevant generalizing representations which bridge between some input and some output representation has made it possible to train neural architectures for complex input/output mapping tasks for which a conventional approach would have crucially involved a linguistically informed intermediate representation. An illustrative example is machine translation. Conventionally, it was considered unquestionable that the best possible machine translation approach would for instance exploit a parser on the source side and a generator on the target side, which take advantage of all accumulated insights regarding the grammatical systems in these languages (potentially mediated through ML-based parsers/generators trained on treebanks that capture the linguistic knowledge). Now, neural models for translation can be trained on very large numbers of pairs of input/output sentences without providing any abstract characterization of linguistic properties or composing a pipeline of sub-steps such as parsing, transfer and generation (for an overview, see Zhang & Zong 2015). The model parametrization resulting for the end-to-end translation process captures many of the relevant grammatical regularities in the hidden layers. The great advantage of the free induction of internal representations as systematic patterns occur in the observed input/output relation is that the model will capture not only regularities along the major dimensions underlying established modular descriptions of knowledge of language, but any other trends that

are manifest in the data. And since all aspects are incorporated in a single, densely connected model architecture, cross-talk among the various knowledge sources and contextual clues is captured very effectively.<sup>54</sup>

But how could then a comparison with the classical OT architecture be of any use or lead to new insights? OT was designed to have an abstract symbolic representation level (OT constraints) in a quasi-connectionist model architecture that implements an input/output function with supervised training. However, with the assumed fixed set of grammatically relevant constraints, one of the key strengths of current connectionist models – the free induction of internal representations at hidden levels, without preconceived features or violable constraints – is excluded by design. Accommodating for the tension between OT’s fixed set of symbolic constraints and the current neural model’s completely unrestricted representational space could be the key to advances on both the linguistic and the computational side. The unrestricted neural models lack scientific interpretability, but they can use representation learning to capture cross-talk among any factors influencing the observed input/output behavior, while the OT approach enforces the exclusion of non-linguistic factors as an influence on the competitions. It is conceivable to reconcile the shortcomings on either side by building up symbolically informed diagnostic tools operating on trained neural models: such tools would in a first step allow researchers to inspect in a controlled way how a neural model trained on corpus data captures known symbolic generalizations. In a second step, it could be explored empirically how factors from additional, non-linguistic knowledge sources affect the model’s “own account” of the symbolic generalizations beyond the controlled setup. Recent work in NLP has started looking at various techniques (probing tasks, causal intervention) that can serve as such diagnostic tools for inspecting neural models (Ettinger et al. 2016; Gulordava et al. 2018).

To give an idea of the potential that lies in symbolically guided analysis of the internal representations of trained neural models, Figure 7 shows an analysis of the behavior of a neural syntactic parser from Falenska & Kuhn (2019). The parser, a neural graph-based dependency parser, is trained in a supervised way to predict dependency arcs (essentially representing the grammatical relation between the words in a sentence). The model can build up and exploit its own internal representation to capture emerging systematic patterns that guide

---

<sup>54</sup>Recent advances in the neural modeling of text, in particular transformer-based contextual word embeddings like BERT (Devlin et al. 2019) have led to an often surprising level of emerging generalizations. Even before the advent of this generation of models, Loula et al. (2018) reported “impressive generalization capabilities” of neural sequence-to-sequence models in phrasal composition underlying the meaning of commands like *turn left twice*.

the decision for assigning a particular dependency arc. It may for example notice the influence of intervening words (as in the contrast *some prefer water* vs. *some cold water*, where the intervening word influences what is the most likely grammatical relation between *some* and *water*).

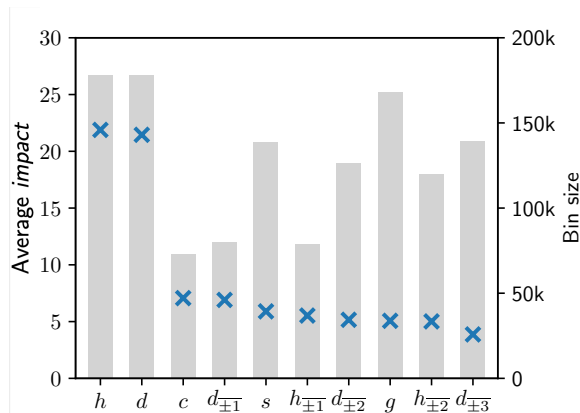


Figure 7: Analysis of trained neural dependency parsers from Falenska & Kuhn (2019), aggregating results from nine treebank training experiments (Ancient Greek, Arabic, Chinese, English, Finnish, Hebrew, Korean, Russian and Swedish): impact (blue crosses) of the information from words in various functional positions (occurring with a frequency plotted by the gray bars) on the prediction score of a multi-layer perceptron for a dependency arc in a graph-based dependency parser. The position types compared are the following: heads ( $h$ ), dependents ( $d$ ), children of  $d$  ( $c$ ), siblings ( $s$ ), grandparents ( $g$ ),  $h, d_{\pm i}$  tokens at distance  $\pm i$  from  $h$  or  $d$  which are none of  $h, d, c, s$ , or  $g$ .

Figure 7 shows a diagnostic analysis of the neural model that emerged from training on treebanks for nine languages. The model was specifically trained to decide whether two given tokens in a sentence should stand in a head-dependent relation, taking into account all tokens in the sentence. For the trained model, it is possible to quantify what impact on the decision the model attributes to individual tokens in the data.<sup>55</sup> For the analysis shown in Figure 7, the token impact was systematically aggregated based on the position of the token in the syntactic gold-standard configuration, as labeled in the treebank. (This information is not available to the model itself.) The blue crosses in the diagram indicate what impact the model itself has decided to assign to tokens in the various configurational positions. The two tokens with the greatest impact are – as one would

<sup>55</sup>The paper introduces a special normalized measure for the impact of specific word representations on the prediction.



expect – the head (*h*) and the dependent (*d*) themselves. But other tokens in the sentence have an impact too. The gray bars represent the frequency of tokens in the various configurational positions in the training corpora. In a structurally uninformed model, our expectation would be that the impact of certain types of tokens correlates with their frequency, so the blue crosses should be high on the tall gray bars and low on the short bars. The diagram reveals that the parser has for instance learned from exposure to data to be substantially more sensitive to the children of dependents (*c*) than to siblings (*s*), although the latter occur much more frequently in the training data.<sup>56</sup> In fact, the neural model with its self-induced internal representation achieves a better prediction quality than the best classical machine learning models, for which carefully designed configurational features are provided. Since direct information about the representational configuration is not in the input representation that the parser receives in training or application, we have an indication that the emerging internal, connectionist representation does encode implicit knowledge about important aspects of functional structure.

The linguistic questions that can be addressed with the dependency parsing example are relatively limited, but it is conceivable to generalize the approach to a more complex interplay of factors affecting linguistic expressions. Diagnostic tools of this kind may thus provide an informed view on the ability of a connectionist model to pick up patterns for which a theory of syntax has posited symbolic abstractions – capturing their systematic significance. A central challenge for theoretical interpretability of the models' predictive capacity is to further develop the diagnostic machinery for disentangling the overlaid effects of very different knowledge sources.<sup>57</sup>

---

<sup>56</sup>It should be noted that there is no delimitable locus in the model's parameter space where this information is represented – but based on the theoretically informed diagnostic tests, we can observe that whatever the model has learned correlates with the distinction.

<sup>57</sup>To make progress in this process of disentangling, it is necessary to cross long-established disciplinary boundaries that provided clear-cut subspaces in the study of language, text and corpora of discourses from specific contexts: theory of grammar with its subdisciplines, psycholinguistics, sociolinguistics, but also literary studies, media studies, etc.; many relevant patterns fall into the realm of scholarly disciplines that do not study language and text *per se*, such as cultural studies, history and social science. The delimited subspaces have so far justified convenient idealizations in the working assumptions – in the theory of grammar for example the idealizing assumption of a shared body of grammatical and lexical knowledge that makes up the linguistic competence of all native speakers of language X. An extra challenge comes from the divergent methodologies that have developed in the subfields as a response to the very distinct idealizations; this becomes clear in work in subareas of digital humanities and computational social science which has recently explored corpus-based modeling techniques for addressing research questions from literary studies (Kuhn 2019) or political science (Padó et al.

## 6 Conclusion: Epistemological considerations

Optimality Theory was introduced to the study of language in the 1990s as a symbolic description framework for linguistic representations which can make stronger empirically testable predictions about the human language faculty than a classical formal grammar. It can be seen as an attempt to overcome an epistemological limitation of the established formalisms underlying most work in linguistics. The limitation affects plain rewrite systems from the Chomsky hierarchy as much as extensions such as for instance transformational systems, tree-adjointing grammars and unification-based grammars.

With a classical framework, the predictions that can be tested “directly”<sup>58</sup> against empirical observations are always tied to the formal grammar for a particular language, i.e., a specific instantiation of the class of formal systems, for example, a specific context-free grammar  $G_{137} = \langle N_{137}, \Sigma_{137}, P_{137}, S_{137} \rangle$  with a concrete set of non-terminal symbols, terminal symbols, rewrite rule productions and start symbol (and similarly for more expressive grammar formalisms).

One such instantiation is typically viewed as a scientific model for (aspects of) the grammatical competence of speakers of, say, Japanese, Swahili or English. The formal system predicts a set of terminal strings, which can be experimentally compared against the linguistic behavior of competent speakers. A grammatical theory about a range of phenomena, e.g., in Japanese, is then falsifiable in the sense of Popper (1959)<sup>59</sup> because it is conceivable that relevant types of terminal strings predicted to be excluded from the formal language according to the theory *do* in fact occur in an experiment (potentially using corpus studies that compare against very similar string types predicted to be included).<sup>60</sup> The theoretical scope is however rather limited: if  $G_{137}$  fails to predict an empirically observable

---

2019). Considerable effort is needed to appropriately incorporate findings from computational models in the respective question contexts and theoretical frameworks (Reiter et al. 2020).

<sup>58</sup>“Directly” is set in scare quotes, since what goes through as direct empirical evidence in linguistic work always depends on methodological preassumptions that a community agrees on. The nature of language data in communication is such that very few event types are truly observable in a direct way. However, appeal to certain theoretical constructs and certain contextually triggered inferences is typically considered uncontroversial since they are orthogonal to research questions under debate.

<sup>59</sup>Falsifiability is a prerequisite for a theory with any predictive power.

<sup>60</sup>Most theories of grammar work with the stronger assumption that (some part of) the internal symbolic structures used by the formalism also represent relevant aspects of meaning, i.e., they can be viewed as logical forms. This provides an additional basis for testable predictions; the experiment then has to access speakers’ (and listeners’/readers’) interpretation of given strings. What is directly testable are however still the system’s predictions for one fully parametrized (language-particular) instance of the grammar formalism.

opposition of acceptable vs. unacceptable data in Japanese, all that the theorist can conclude is that some aspect in the specification of this formal grammar has been inappropriate. When there are multiple ways of fixing the problem (e.g., by different strategies of introducing additional non-terminal symbols in a variant  $G'_{137}$  and by modifying some existing productions and adding some new ones in  $G''_{137}$ ), there is no theoretically forceful way of distinguishing between them. Assume for the sake of the argument that only one of the modified grammars makes structural similarities with Korean, which is modeled in some grammar  $G_{214}$ , explicit. As long as the other modified formal grammars predict the same formal language, a theoretically well-founded statement differentiating between the options cannot be made (for lack of falsifiability).

In contrast to classical formal systems, an OT system can make testable predictions for quite a different type of experiment: the theory implemented in a particular OT system, with a spelled-out candidate generation function *Gen* and a set of rankable constraints *Con*, does not predict a single formal language, but – via the factorial typology implied by all possible rankings of *Con* – a whole class of formal language approximations of natural languages. Fully formalized OT systems come with a spelled-out empirical learning algorithm (Tesar & Smolensky 1998 and subsequent work in the community), which provides the basis for falsifiability of a theory about the human language faculty as such: a formal OT system predicts how a learner of any specific language in their learning behavior responds to exposure to language behavior by adult speakers of the language in question.<sup>61</sup> A concrete OT system can thus be falsified by evidence from *any* of the languages of the world: a particular observed adult language behavior could in principle trigger a sequence of constraint rerankings that make it impossible for the learner to converge on the constraint ranking needed for this language. If this is the case for some conjectured theory (with a constraint set, etc.) and observed language data, then the theory of some part of the human language faculty counts as falsified.

Of course, most research communities employing classical formalisms, including the LFG community, have established agreed-upon meta principles and methodological research practices which effectively ensure that empirical evidence from a particular natural language is accepted by the community as evidence affecting *all* formal systems adhering to the shared conventions (to continue with our context-free grammar illustration, the theory could be characterized as the set of formal grammars  $\{G_i \mid \text{the components of } G_i \text{ satisfy all meta$

---

<sup>61</sup>Adult behavior is assumed to be observed in contexts that provide enough extra-linguistic clues to support inferences about the intended meaning where there is ambiguity.

principles }). There are countless examples of such meta principles: X-bar theory, extended projections, the concept of lexical redundancy rules, constraints on legitimate transformations in a derivational framework, a universal layering of functional projections, etc. With such principles it becomes possible that the grammatical theory constituted by the meta principles is falsified by language-particular evidence. Even the language learning/acquisition problem has been formulated on the basis of such framework, presumably most prominently in the Principles-and-Parameters theory (Chomsky 1981): the learner (guided by some language acquisition device that is assumed to be part of the cognitive equipment) has the task of adjusting a number of free parameters in an otherwise highly constrained innate grammatical system – in response to observed linguistic behavior by adult speakers. (A rigorous formalization as for OT learning algorithms is typically not provided.)

However, although sophisticated research practices have been established that ensure a far-reaching consensus about plausible meta principles assumed in a community, the epistemological relationship between a particular instantiation of the meta framework (say, formal grammar  $G_{137}$  for Japanese) and the framework itself, with its meta principles, remains contestable. Typically, the representational constructs developed to capture generalizations across natural languages have been established in a long process of cautious plausible reasoning – yet it is almost always conceivable that there are alternative, empirically indistinguishable ways of predicting surface-level divergences across languages. To explain unexpected patterns in some language X, a modification in the formulation of one or another meta principle could be made, or idiosyncratic lexical knowledge could be posited. The falsifiability issue of general theoretical statements is not completely resolved by the assumption of meta principles. This circumstance explains in part why there have been many controversial debates about the representational locus for capturing cross-linguistic variation in a phenomenon – take for instance Binding Principles, which one might construe along a configurational tree structure or along a functional hierarchy (Asudeh & Dalrymple 2006). In the same vein, a re-occurring type of argument against specific linguistic accounts is the accusation for overly strong theory-internal assumptions. In other words, it happens quite frequently that members of a research community develop reservations with respect to the falsifiability of parts of the established consensus framework.<sup>62</sup>

---

<sup>62</sup>Given the underdetermination of theories of grammar by direct empirical evidence, aesthetic arguments regarding the simplicity of a theory are often advanced, most notably in Minimalism (Chomsky 1995). But even this strategy cannot escape controversies, since there are different possible starting points for seeding a theoretical accounts in fundamental propositions.

As just noted, OT can be technically seen as the move towards an approach that meets higher standards of falsifiability (when fully formalized). One may ask oneself then why it has not replaced classical formalisms in mainstream linguistic research? Section 5.1 discussed this question under the perspective of the concrete course of research activities in the 1990s and 2000s. But there are also relevant meta-theoretical considerations: since most aspects of representational choice in linguistic modeling are not directly observable (only the linear sequence of surface units of expression and semantic entailments of the content of utterances are directly observable), the space of possible OT theories remains vastly underdetermined. This means that (unless a community decides to change their research paradigm completely), plausible argumentation for abstract intermediate representations remains an important part of linguistic theorizing. And since substantial groundwork in linguistic research has always lain in the systematic capturing of regularities in variation patterns for particular languages, the justification for the use of classical formalisms has not disappeared. By choosing a strict ranking approach over violable constraints captured in terms of established symbolic representations, the OT endeavor was from the outset designed to stay connected with work using the classical formalisms; the effect of relevant constraints on a phenomenon under consideration can be calculated in manually constructed tableaux.<sup>63</sup> Insights from a specific OT account may thus feed back into the more general debate of what are appropriate theoretical constructs for systematically capturing a particular aspect of linguistic knowledge.

Against this background, the most important contribution of OT to generative linguistics might have been to increase the awareness in (part of) the community that a comprehensive, falsifiable account of the human language faculty has to include a formalized account of learnability of language from exposure to data.

## References

- Aissen, Judith L. 1999. Markedness and subject choice in optimality theory. *Natural Language & Linguistic Theory* 17(4). 673–711. DOI: 10.1023/A:1006335629372.
- Aissen, Judith L. 2003. Differential object marking: Iconicity vs. economy. *Natural Language & Linguistic Theory* 21. 435–483. DOI: 10.1023/A:1024109008573.

---

<sup>63</sup>As was discussed in Section 5, recent machine learning approaches open up alternative avenues for modeling human language behavior. Computational models incorporating a very large parameter space can be trained on large corpora to achieve higher prediction accuracy on most tasks; however, the emerging representations (in the hidden layers of “deep” neural network models) provide no direct basis for a falsifiable theory of aspects of linguistic competence. Some connection to hypotheses that can be expressed symbolically seems to be indispensable.

- Anderson, Stephen R. 1969. *West Scandinavian vowel systems and the ordering of phonological rules*. Cambridge, MA: Massachusetts Institute of Technology. (Doctoral dissertation).
- Andrews, Avery D. 1982. The representation of case in Modern Icelandic. In Joan Bresnan (ed.), *The mental representation of grammatical relations*, 427–503. Cambridge, MA: The MIT Press.
- Andrews, Avery D. 1990. Unification and morphological blocking. *Natural Language & Linguistic Theory* 8(4). 507–557. DOI: 10.1007/bf00133692.
- Anttila, Arto. 1997. *Variation in Finnish phonology and morphology*. Stanford: Stanford University. (Doctoral dissertation).
- Asudeh, Ash. 2001. Linking, optionality, and ambiguity in Marathi. In Peter Sells (ed.), *Formal and empirical issues in Optimality Theoretic Syntax*, 257–312. Stanford: CSLI Publications.
- Asudeh, Ash & Mary Dalrymple. 2006. Binding theory. In Keith Brown (ed.), *Encyclopedia of language and linguistics*, 2nd edn., 23–31. Amsterdam: Elsevier. DOI: 10.1016/b0-08-044854-2/01955-6.
- Beaver, David I. & Hanjung Lee. 2004. Input-output mismatches in OT. In Reinhard Blutner & Henk Zeevat (eds.), *Optimality Theory and pragmatics*, 112–153. London: Palgrave Macmillan.
- Belinkov, Yonatan & James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics* 7. 49–72. DOI: 10.1162/tacl\_a\_00254.
- Blutner, Reinhard. 1998. Lexical pragmatics. *Journal of Semantics* 15(2). 115–162. DOI: 10.1093/jos/15.2.115.
- Blutner, Reinhard. 2000. Some aspects of optimality in natural language interpretation. *Journal of Semantics* 17(3). 189–216. DOI: 10.1093/jos/17.3.189.
- Boersma, Paul. 1998. *Functional phonology: Formalizing the interactions between articulatory and perceptual drives*. Amsterdam: University of Amsterdam. (Doctoral dissertation).
- Boersma, Paul, Joost Dekkers & Jeroen van de Weijer. 2000. Introduction — Optimality Theory: Phonology, syntax, and acquisition. In Joost Dekkers, Frank van der Leeuw & Jeroen van de Weijer (eds.), *Optimality Theory: Phonology, syntax, and acquisition*. Oxford: Clarendon Press.
- Boersma, Paul & Bruce Hayes. 2001. Empirical tests of the gradual learning algorithm. *Linguistic Inquiry* 32(1). 45–86. DOI: 10.1162/002438901554586.
- Bögel, Tina, Miriam Butt, Ronald M. Kaplan, Tracy Holloway King & John T. Maxwell III. 2009. Prosodic phonology in LFG: A new proposal. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '09 conference*, 146–166. Stanford: CSLI Publications.

- Bouma, Gerlof & Petra Hendriks. 2012. Partial word order freezing in Dutch. *Journal of Logic, Language, and Information* 21(1). 53–73.
- Bresnan, Joan. 1996. LFG in an OT setting: Modelling competition and economy. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '96 conference*. Stanford: CSLI Publications. Handout from keynote address.
- Bresnan, Joan. 2000. Optimal syntax. In Joost Dekkers, Frank van der Leeuw & Jeroen van de Weijer (eds.), *Optimality Theory: Phonology, syntax, and acquisition*, 334–385. Oxford: Oxford University Press.
- Bresnan, Joan. 2001a. Explaining morphosyntactic competition. In Mark Baltin & Chris Collins (eds.), *Handbook of contemporary syntactic theory*, 11–44. Oxford: Blackwell.
- Bresnan, Joan. 2001b. *Lexical-Functional Syntax*. Oxford: Blackwell.
- Bresnan, Joan. 2002. The lexicon in Optimality Theory. In Paola Merlo & Suzanne Stevenson (eds.), *The lexical basis of sentence processing: Formal, computational and experimental issues*, 39–58. Amsterdam: John Benjamins. DOI: 10.1075/nlp.4.03bre.
- Bresnan, Joan. 2007. Is syntactic knowledge probabilistic? Experiments with the English dative alternation. In Sam Featherston & Wolfgang Sternefeld (eds.), *Roots: Linguistics in search of its evidential base* (Studies in Generative Grammar), 77–96. Berlin: Mouton de Gruyter.
- Bresnan, Joan. 2011. Linguistic uncertainty and the knowledge of knowledge. In Roger Porter & Robert Reynolds (eds.), *Thinking Reed: Centennial essays by graduates of Reed College*, 69–75. Portland, OR: Reed College.
- Bresnan, Joan. 2016. Linguistics: The garden and the bush. *Computational Linguistics* 42(4). 599–617. DOI: 10.1162/COLI\_a\_00260.
- Bresnan, Joan & Judith L. Aissen. 2002. Optimality and functionality: Objections and refutations. *Natural Language & Linguistic Theory* 20(1). 81–95. DOI: 10.1023/A:1014222605182.
- Bresnan, Joan, Ash Asudeh, Ida Toivonen & Stephen Wechsler. 2016. *Lexical-Functional Syntax*. 2nd edn. (Blackwell Textbooks in Linguistics 16). Malden, MA: Wiley-Blackwell.
- Bresnan, Joan, Shipra Dingare & Christopher Manning. 2001. Soft constraints mirror hard constraints: Voice and person in English and Lummi. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '01 conference*, 13–32. Stanford: CSLI Publications.
- Bresnan, Joan & Marilyn Ford. 2010. Predicting syntax: Processing dative constructions in American and Australian varieties of English. *Language* 86. 168–213. DOI: 10.1353/lan.0.0189.

- Bresnan, Joan & Tatiana Nikitina. 2010. The gradience of the dative alternation. In Linda Ann Uyechi & Lian-Hee Wee (eds.), *Reality exploration and discovery: Pattern interaction in language and life* (CSLI Lecture Notes 197), 161–184. Stanford: CSLI Publications.
- Buckner, Cameron & James Garson. 2019. Connectionism. In Edward N. Zalta (ed.), *The Stanford encyclopedia of philosophy*, Autumn 2019. Stanford: Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/fall2019/entries/connectionism/>.
- Cahill, Aoife, Tracy Holloway King & John T. Maxwell III. 2007. Pruning the search space of a hand-crafted parsing system with a probabilistic parser. In *ACL 2007 workshop on deep linguistic processing*, 65–72. Prague: Association for Computational Linguistics. DOI: 10.3115/1608912.1608924.
- Cahill, Aoife & Arndt Rieger. 2009. Incorporating information status into generation ranking. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, 817–825. Suntec, Singapore: Association for Computational Linguistics. DOI: 10.3115/1690219.1690261.
- Cahill, Aoife & Andy Way. Forthcoming. Treebank-driven Parsing, Translation and Grammar Induction using LFG. In Mary Dalrymple (ed.), *Handbook of Lexical Functional Grammar*. Berlin: Language Science Press.
- Chomsky, Noam. 1981. *Lectures on government and binding*. Dordrecht: Foris Publications. DOI: 10.1515/9783110884166.
- Chomsky, Noam. 1995. *The Minimalist Program*. Cambridge, MA: The MIT Press. DOI: 10.7551/mitpress/9780262527347.001.0001.
- Chomsky, Noam, Ángel J. Gallego & Dennis Ott. 2019. Generative grammar and the faculty of language: Insights, questions, and challenges. *Catalan Journal of Linguistics Special Issue: Generative Syntax: Questions, Crossroads, and Challenges*. 229–261. DOI: 10.5565/rev/catjl.288.
- Collins, Michael. 1997. Three generative, lexicalised models for statistical parsing. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European chapter of the Association for Computational Linguistics*, 16–23. Madrid: Association for Computational Linguistics. DOI: 10.3115/976909.979620.
- Crouch, Richard, Mary Dalrymple, Ronald M. Kaplan, Tracy Holloway King, John T. Maxwell III & Paula Newman. 2011. *XLE Documentation*. Xerox Palo Alto Research Center. Palo Alto, CA. [https://ling.sprachwiss.uni-konstanz.de/pages/xle/doc/xle\\_toc.html](https://ling.sprachwiss.uni-konstanz.de/pages/xle/doc/xle_toc.html).



- Dalrymple, Mary, Ronald M. Kaplan, John T. Maxwell III & Annie Zaenen (eds.). 1995. *Formal issues in Lexical-Functional Grammar*. Stanford: CSLI Publications.
- Devlin, Jacob, Ming-Wei Chang, Kenton Lee & Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (naacl-hlt 2019)*, 4171–4186. Association for Computational Linguistics. DOI: 10.18653/v1/N19-1423.
- Dione, Cheikh M. Bamba. 2014. LFG parse disambiguation for Wolof. *Journal of Language Modelling* 2(1). 105–165. DOI: 10.15398/jlm.v2i1.81.
- Eisner, Jason. 2000. Book reviews: Optimality Theory. *Computational Linguistics* 26(2). <https://aclanthology.org/J00-2014>.
- Engels, Eva & Sten Vikner. 2014. *Scandinavian object shift and Optimality Theory*. Basingstoke: Palgrave Macmillan. DOI: 10.1057/9781137431646.
- Ettinger, Allyson, Ahmed Elgohary & Philip Resnik. 2016. Probing for semantic evidence of composition by means of simple classification tasks. In *Proceedings of the 1st workshop on Evaluating Vector-Space Representations for NLP*, 134–139. Berlin: Association for Computational Linguistics. DOI: 10.18653/v1/W16-2524.
- Falenska, Agnieszka & Jonas Kuhn. 2019. The (non-)utility of structural features in BiLSTM-based dependency parsers. In *Proceedings of the 57th annual meeting of the Association for Computational Linguistics*, 117–128. Florence: Association for Computational Linguistics. DOI: 10.18653/v1/P19-1012.
- Fanselow, Gisbert & Caroline Féry. 2019. Ineffability in grammar. In *Resolving conflicts in grammars: Optimality theory in syntax, morphology and phonology. Linguistische Berichte / Sonderhefte 11*. Helmut Buske Verlag.
- Fodor, Jerry A. & Brian P. McLaughlin. 1990. Connectionism and the problem of systematicity: Why Smolensky’s solution doesn’t work. *Cognition* 35. 183–204. DOI: 10.1016/0010-0277(90)90014-b.
- Forst, Martin. 2007. Filling statistics with linguistics: Property design for the disambiguation of German LFG parses. In *ACL 2007 workshop on deep linguistic processing*, 17–24. Prague: Association for Computational Linguistics. DOI: 10.3115/1608912.1608916.
- Forst, Martin & Tracy Holloway King. Forthcoming. Computational implementations and applications. In Mary Dalrymple (ed.), *Handbook of Lexical Functional Grammar*. Berlin: Language Science Press.
- Frank, Anette, Tracy Holloway King, Jonas Kuhn & John T. Maxwell III. 2001. Optimality Theory style constraint ranking in large-scale LFG grammars. In

- Peter Sells (ed.), *Formal and empirical issues in Optimality Theoretic syntax*, 367–397. Stanford: CSLI Publications.
- Frank, Robert & Giorgio Satta. 1998. Optimality theory and the generative complexity of constraint violation. *Computational Linguistics* 24(2). 307–316. <https://aclanthology.org/J98-2006>.
- Goldwater, Sharon & Mark Johnson. 2003. Learning OT constraint rankings using a maximum entropy model. In Jennifer Spenader, Anders Eriksson & Östen Dahl (eds.), *Proceedings of the Stockholm workshop on variation within Optimality Theory, Stockholm University*, 111–120.
- Grimshaw, Jane. 1997. Projection, heads, and optimality. *Linguistic Inquiry* 28. 373–422.
- Gulordava, Kristina, Piotr Bojanowski, Edouard Grave, Tal Linzen & Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, volume 1 (long papers)*, 1195–1205. New Orleans: Association for Computational Linguistics. DOI: 10.18653/v1/N18-1108.
- Haspelmath, Martin. 1999. Optimality and diachronic adaptation. *Zeitschrift für Sprachwissenschaft* 18(2). 180–205. DOI: 10.1515/zfsw.1999.18.2.180.
- Haug, Dag. Forthcoming. LFG and Dependency Grammar. In Mary Dalrymple (ed.), *Handbook of Lexical Functional Grammar*. Berlin: Language Science Press.
- Heck, Fabian & Gereon Müller. 2000. Successive cyclicity, long distance superiority, and local optimization. In Roger Billerey & Brook D. Lillehaugen (eds.), *Proceedings of the 19th West Coast Conference on Formal Linguistics*, 218–231. Somerville, MA: Cascadilla Press.
- Henderson, James. 2020. The unstoppable rise of computational linguistics in deep learning. In *Proceedings of the 58th annual meeting of the Association for Computational Linguistics*, 6294–6306. Association for Computational Linguistics. DOI: 10.18653/v1/2020.acl-main.561.
- Huang, C.-T. James. 1982. *Logical relations in Chinese and the theory of grammar*. Cambridge, MA: Massachusetts Institute of Technology. (Doctoral dissertation).
- Jäger, Gerhard. 2002. Some notes on the formal properties of bidirectional Optimality Theory. *Journal of Logic, Language and Information* 11(4). 427–451.
- Jäger, Gerhard. 2007. Maximum entropy models and stochastic Optimality Theory. In Annie Zaenen, Jane Simpson, Tracy Holloway King, Jane Grimshaw, Joan Maling & Chris Manning (eds.), *Architectures, rules, and preferences: Variations on themes by Joan W. Bresnan*, 467–479. Stanford: CSLI Publications.

- Johnson, Mark. 1998. Optimality-theoretic Lexical Functional Grammar. Unpublished manuscript, Brown University; Paper presented at the 11th Annual CUNY Conference on Human Sentence Processing, Rutgers University.
- Kager, René. 1999. *Optimality Theory*. Cambridge, UK: Cambridge University Press. DOI: 10.1017/cbo9780511812408.
- Kaplan, Ronald M. & Jürgen Wedekind. 2000. LFG generation produces context-free languages. In *COLING 2000 volume 1: The 18th International Conference on Computational Linguistics*, 425–431. DOI: 10.3115/990820.990882.
- Karlssoon, Fred. 1990. Constraint grammar as a framework for parsing running text. In *COLING '90 volume 3: Proceedings of the 13th International Conference on Computational Linguistics*, 168–173. DOI: 10.3115/991146.991176.
- Karttunen, Lauri. 1998. The proper treatment of optimality in computational phonology. In *Finite state methods in natural language processing*, 1–12. <https://aclanthology.org/W98-1301>.
- Keller, Frank & Ash Asudeh. 2002. Probabilistic learning algorithms and Optimality Theory. *Linguistic Inquiry* 33(2). 225–244. DOI: 10.1162/002438902317406704.
- King, Tracy Holloway, Stefanie Dipper, Anette Frank, Jonas Kuhn & John T. Maxwell III. 2000. Ambiguity management in grammar writing. In *Proceedings of the Linguistic Theory and Grammar Implementation workshop at European Summer School in Logic, Language, and Information (ESSLI-2000)*.
- King, Tracy Holloway, Stefanie Dipper, Anette Frank, Jonas Kuhn & John T. Maxwell III. 2004. Ambiguity management in grammar writing. *Research on Language and Computation* 2. 259–280. DOI: 10.1023/b:rolc.0000016784.26446.98.
- Krämer, Martin. 2017. Current issues and directions in optimality theory. In *The routledge handbook of phonological theory*. Routledge. DOI: 10.4324/9781315675428-3.
- Kuhn, Jonas. 2000a. Faithfulness violations and bidirectional optimization. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '00 conference*, 161–181. Stanford: CSLI Publications.
- Kuhn, Jonas. 2000b. Processing optimality-theoretic syntax by interleaved chart parsing and generation. In *Proceedings of the 38th annual meeting of the association for computational linguistics*, 368–375. Hong Kong: Association for Computational Linguistics. DOI: 10.3115/1075218.1075264.
- Kuhn, Jonas. 2001. Generation and parsing in Optimality Theoretic syntax – Issues in the formalization of OT-LFG. In Peter Sells (ed.), *Formal and empirical issues in Optimality Theoretic Syntax*, 313–366. Stanford: CSLI Publications.

- Kuhn, Jonas. 2002. OT syntax: Decidability of generation-based optimization. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 48–55. Philadelphia: Association for Computational Linguistics. DOI: 10.3115/1073083.1073094.
- Kuhn, Jonas. 2003. *Optimality-Theoretic Syntax – A declarative approach*. Stanford: CSLI Publications.
- Kuhn, Jonas. 2019. Computational text analysis within the Humanities: How to combine working practices from the contributing fields? *Language Resources and Evaluation* 53(4). 565–602. DOI: 10.1007/s10579-019-09459-3.
- Lee, Hanjung. 2001. Markedness and word order freezing. In Peter Sells (ed.), *Formal and empirical issues in Optimality Theoretic Syntax*, 63–128. Stanford: CSLI Publications.
- Legendre, Géraldine, Jane Grimshaw & Sten Vikner (eds.). 2000. *Optimality-Theoretic Syntax*. Cambridge, MA: The MIT Press. DOI: 10.7551/mitpress/5161.001.0001.
- Legendre, Géraldine, Yoshiro Miyata & Paul Smolensky. 1990. *Can connectionism contribute to syntax? Harmonic grammar, with an application*. Tech. rep. 467. Institute of Cognitive Science, University of Colorado at Boulder. CU-CS-485-9, Computer Science Technical Reports.
- Legendre, Géraldine, Paul Smolensky & Colin Wilson. 1998. When is less more? Faithfulness and minimal links in wh-chains. In Pilar Barbosa, Danny Fox, Paul Hagstrom, Martha McGinnis & David Pesetsky (eds.), *Is the best good enough? Optimality and competition in syntax*, 249–289. Cambridge, MA: The MIT Press & MITWPL.
- Loula, João, Marco Baroni & Brenden M. Lake. 2018. Rearranging the familiar: Testing compositional generalization in recurrent networks. *CoRR* abs/1807.07545. DOI: 10.18653/v1/w18-5413.
- Marcus, Gary. 2001. *The algebraic mind: Integrating connectionism and cognitive science*. Cambridge, MA: The MIT Press. DOI: 10.7551/mitpress/1187.001.0001.
- Maxwell, John T., III & Ronald M. Kaplan. 1989. An overview of disjunctive constraint satisfaction. In *Proceedings of the 4th International Workshop on Parsing Technologies (IWPT 1995)*, 18–27. Also published in Tomita (1991) as ‘A Method for Disjunctive Constraint Satisfaction’, and reprinted in Dalrymple, Kaplan, Maxwell & Zaenen (1995: 381–402).
- Maxwell, John T., III & Ronald M. Kaplan. 1996. Unification-based parsers that automatically take advantage of context freeness. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG ’96 conference*, 1–31. Stanford: CSLI Publications.

- McCarthy, John J. & Alan Prince. 1995. Faithfulness and reduplicative identity. In Jill N. Beckmann, Laura Walsh Dickey & Suzanne Urbanczyk (eds.), *Papers in Optimality Theory*, 249–384. Amherst: GLSA.
- Müller, Gereon. 2003. Local vs. global optimization in syntax: A case study. In Jennifer Spenader, Anders Eriksson & Östen Dahl (eds.), *Proceedings of the Stockholm workshop on variation within Optimality Theory*, Stockholm University, 82–91.
- Müller, Gereon. 2012. Optimality-theoretic syntax. Unpublished manuscript, Universität Leipzig.
- Müller, Gereon. 2014. Optionality in optimality-theoretic syntax. In *The second GLOT international state-of-the-article book*, 289–322. Berlin: De Gruyter Mouton.
- Müller, Gereon. 2020. *Inflectional morphology in harmonic serialism*. Sheffield, UK: Equinox Publishing.
- Newmeyer, Frederick J. 2002. Optimality and functionality: A critique of functionally-based optimality-theoretic syntax. *Natural Language & Linguistic Theory* 20(1). 43–80. DOI: 10.1023/A:1014290005775.
- Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty & Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'16)*, 1659–1666. Portorož: European Language Resources Association (ELRA). <https://aclanthology.org/L16-1262>.
- Padó, Sebastian, Andre Blessing, Nico Blokker, Erenay Dayanik, Sebastian Haunss & Jonas Kuhn. 2019. Who sides with whom? Towards computational construction of discourse networks for political debates. In *Proceedings of the 57th annual meeting of the Association for Computational Linguistics*, 2841–2847. Florence: Association for Computational Linguistics. DOI: 10.18653/v1/P19-1273.
- Pesetsky, David. 1998. Some optimality principles of sentence pronunciation. In Pilar Barbosa, Danny Fox, Paul Hagstrom, Martha McGinnis & David Pesetsky (eds.), *Is the best good enough? Optimality and competition in syntax*, 337–383. Cambridge, MA: The MIT Press & MITWPL.
- Popper, Karl. 1959. *The logic of scientific discovery*. London: Hutchinson. Translation of *Logik der Forschung* (1935), by the author.
- Prince, Alan & Paul Smolensky. 1993. *Optimality Theory: Constraint interaction in Generative Grammar*. Tech. rep. 2. New Brunswick, NJ: Rutgers University Center for Cognitive Science. Published as Prince & Smolensky (2004).

- Prince, Alan & Paul Smolensky. 2004. *Optimality theory: Constraint interaction in Generative Grammar*. Oxford/Malden: Blackwell. DOI: 10.1002/9780470759400.
- Przepiórkowski, Adam. Forthcoming. LFG and HPSG. In Mary Dalrymple (ed.), *Handbook of Lexical Functional Grammar*. Berlin: Language Science Press.
- Pustejovsky, James. 1995. *The generative lexicon*. Cambridge, MA: The MIT Press. DOI: 10.7551/mitpress/3225.001.0001.
- Reiter, Nils, Axel Pichler & Jonas Kuhn (eds.). 2020. *Reflektierte Algorithmische Textanalyse. Interdisziplinäre(s) Arbeiten in der CRETA-Werkstatt [Reflected Algorithmic Text Analytics: Interdisciplinary Works from the CRETA Workshop]*. Berlin: De Gruyter.
- Riezler, Stefan, Tracy Holloway King, Ronald M. Kaplan, Richard Crouch, John T. Maxwell III & Mark Johnson. 2002. Parsing the Wall Street Journal using a Lexical-Functional Grammar and discriminative estimation techniques. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. Philadelphia: Association for Computational Linguistics.
- Riezler, Stefan, Detlef Prescher, Jonas Kuhn & Mark Johnson. 2000. Lexicalized stochastic modeling of constraint-based grammars using log-linear measures and EM training. In *Proceedings of the 38th annual meeting of the Association for Computational Linguistics*, 480–487. Hong Kong: Association for Computational Linguistics. DOI: 10.3115/1075218.1075279.
- Rumelhart, David E., James L. McClelland & the PDP Research Group (eds.). 1986. *Parallel distributed processing, volume 1: Explorations in the microstructure of cognition: Foundations*. Cambridge, MA: The MIT Press.
- Samek-Lodovici, Vieri. 1996. *Constraints on subjects: An Optimality Theoretic analysis*. New Brunswick, NJ: Rutgers University. (Doctoral dissertation).
- Sannasgala, P. B. 1976. *Sinhalese vocables of Dutch origin*. Colombo: Kularatne.
- Sells, Peter. 1999. Constituent ordering as alignment. In Susumu Kuno et al. (eds.), *Harvard studies in Korean linguistics* 8. Harvard University.
- Sells, Peter. 2001. *Structure, alignment and optimality in Swedish*. Stanford: CSLI Publications.
- Silverstein, Michael. 1976. Hierarchy of features and ergativity. In R. M. W. Dixon (ed.), *Grammatical categories in Australian languages*. Australian Institute of Aboriginal Studies.
- Smolensky, Paul. 1995. On the internal structure of the constraint component Con of UG. Handout of talk, UCLA, 7 April, Rutgers Optimality Archive (<http://roa.rutgers.edu>), ROA 87.
- Smolensky, Paul. 1996. *The initial state and ‘richness of the base’ in Optimality Theory*. Tech. rep. 4. Department of Cognitive Science, Johns Hopkins University.

- Smolensky, Paul & Géraldine Legendre (eds.). 2006. *The harmonic mind: From neural computation to optimality-theoretic grammar*. Cambridge, MA: The MIT Press.
- Stabler, Edward P. 2011. Computational perspectives on Minimalism. In Cedric Boeckx (ed.), *Oxford handbook of Linguistic Minimalism*, 617–642. Oxford: Oxford University Press. DOI: 10.1093/oxfordhb/9780199549368.013.0027.
- Stabler, Edward P. 2013. Two models of Minimalist, incremental syntactic analysis. *Topics in Cognitive Science* 5(3). 611–633. DOI: 10.1111/tops.12031.
- Tesar, Bruce & Paul Smolensky. 1998. Learnability in Optimality Theory. *Linguistic Inquiry* 29(2). 229–268. DOI: 10.1162/002438998553734.
- Tomita, Masaru (ed.). 1991. *Current issues in parsing technology*. Dordrecht: Kluwer Academic Publishers. DOI: 10.1007/978-1-4615-3986-5.
- van der Beek, Leonoor & Gerlof Bouma. 2004. The role of the lexicon in Optimality Theoretic syntax. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '04 conference*, 60–80. Stanford: CSLI Publications.
- van Rooij, Robert & Michael Franke. 2020. Optimality-theoretic and game-theoretic approaches to implicature. In Edward N. Zalta (ed.), *The Stanford encyclopedia of philosophy*, Summer 2020. Stanford: Center for the Study of Language & Information (CSLI). <https://plato.stanford.edu/archives/sum2020/entries/implicature-optimality-games>.
- Vikner, Sten. 2001. The interpretation of object shift and Optimality Theory. In Gereon Müller & Wolfgang Sternefeld (eds.), *Competition in syntax*, 321–340. Berlin: De Gruyter Mouton. DOI: 10.1515/9783110829068.
- Wedekind, Jürgen. 1999. Semantic-driven generation with LFG- and PATR-style grammar. *Computational Linguistics* 25(2). 277–281. <https://www.aclweb.org/anthology/J99-2006>.
- Wunderlich, Dieter. 2006. Syntax: Optimality Theory. In Keith Brown (ed.), *Encyclopedia of language and linguistics*, 2nd edn., 408–418. Amsterdam: Elsevier. DOI: 10.1016/b0-08-044854-2/04213-9.
- Yu, Xiang, Agnieszka Falenska, Ngoc Thang Vu & Jonas Kuhn. 2019. Head-first linearization with tree-structured representation. In *Proceedings of the 12th International Conference on Natural Language Generation*, 279–289. Tokyo: Association for Computational Linguistics. DOI: 10.18653/v1/W19-8636.
- Zaenen, Annie & Richard Crouch. 2009. OBLs hobble computations. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG '09 conference*, 644–654. Stanford: CSLI Publications.
- Zarriß, Sina, Aoife Cahill & Jonas Kuhn. 2011. Underspecifying and predicting voice for surface realisation ranking. In *Proceedings of the 49th annual meeting of the Association for Computational Linguistics: Human Language Technologies*,

- 1007–1017. Portland, OR: Association for Computational Linguistics. <https://aclanthology.org/P11-1101>.
- Zeevat, Henk. 2006. Freezing and marking. *Linguistics* 44(5). 1095–1111. DOI: 10.1515/LING.2006.035.
- Zeevat, Henk & Gerhard Jäger. 2002. A reinterpretation of syntactic alignment. In *Proceedings of the fourth international Tbilisi Symposium on Language, Logic and Computation*.
- Zhang, Jiajun & Chengqing Zong. 2015. Deep neural networks in machine translation: An overview. *IEEE Intelligent Systems* 30(5). 16–25. DOI: 10.1109/MIS.2015.69.