THE ROLE OF THE LEXICON IN OPTIMALITY THEORETIC SYNTAX

Leonoor van der Beek Gerlof Bouma

Center for Language and Cognition (CLCG), University of Groningen

Proceedings of the LFG04 Conference

University of Canterbury

Miriam Butt and Tracy Holloway King (Editors)

2004

CSLI Publications

http://csli-publications.stanford.edu/

Abstract

Research in Optimality Theoretic syntax tends to focus on language universals and the prediction of systematic language-particular properties by means of constraint interaction, often appealing to the principle of *Richness of the Base*. However, this has left the role and formal status of the lexicon in such models largely uninvestigated.

In this paper we look at some existing architectures for OT syntax, notably the LFG based OT-LFG, and the consequences of these approaches for the syntax-phonology interface, lexical lookup, computational properties of the system and the ability to deal with non-systematic language particularities.

On the basis of this exposition we argue that the lexicon should be modeled as an extra argument of GEN, the universal function from inputs to candidate sets. This setup is able to deal with phenomena that were problematic for other architectures, while still respecting core aspects of Richness of the Base.

1 Introduction

Research in Optimality Theoretic (OT) syntax has largely focused on language universals and the prediction of systematic language-particular properties by means of constraint interaction. For example, the constraint ranking determines which contrasts in the paradigm of *to be* are expressed (Bresnan, 2002; Bresnan, 1999). The lexicon is assumed to have no influence on these phenomena. The morpho-syntactic features that we find in the lexical entries of *to be* are there *because* the constraint interaction determined that these contrasts are expressed.

Similarly, it is the language particular ranking of universal constraints that determines whether or not to leave the content of a pronominal unparsed – resulting in an expletive (Grimshaw and Samek-Lodovici, 1998) – not the presence or lack of an expletive lexical item in the lexicon.

The motivation for this line of research is found in the principle of *Richness of the Base* (RotB, Prince and Smolensky (1993). RotB is rooted in OT phonology and tells us that systematic differences between languages arise from different constraint rankings. This effectively bans the lexicon as a source of syntactic variation – which is in sharp contrast to lexicalist theories of language, such as GPSG (Gazdar et al., 1985), HPSG (Pollard and Sag, 1994) and LFG (Bresnan, 2001), that assume that lexical contrasts drive syntactic variation. As a result, the role and formal status of the lexicon, which maps sets of morpho-syntactic features to arbitrary phonological strings, have been largely left uninvestigated.

In this paper we explore the possibilities for encoding and accessing lexical information in current implementations of OT syntax that are in line with RotB, OT-LFG (Bresnan, 1999) in particular. We argue that these possibilities are not sufficient for accounting for syntax-phonology interface phenomena and unsystematic language particularities. Furthermore, we will look closer at the decidability of the OT-LFG system and some predictions made by the OT syntax model of Grimshaw and Samek-Lodovici (1998). Based on our findings we conclude that the lexicon is best modeled as an extra argument of GEN, the universal function from inputs to candidate sets. Such a setup

weakens RotB, but it facilitates accounts for phenomena that were problematic for the discussed models while respecting some of the core aspects of RotB.

The setup of this paper is as follows. We start with a brief discussion of RotB in its original form in OT phonology and its applications in OT syntax in section 2. We then discuss the consequences of marginalizing the lexicon (section 3). In section 4, we first discuss a possible solution for some of the problems that was suggested in the literature and then present our more general and principled solution of modeling the lexicon as an argument of GEN. Finally, we discuss the consequences of our proposal for RotB in section 5 and conclude in section 6.

2 Richness of the Base

Richness of the Base was originally formulated in the OT phonology literature to prevent analyses that appeal to systematic differences in the input. If a language only has .CV. syllables, then this is not because the input is restricted to that type of syllable but because the grammar thus restricts the output:

> [Under] *Richness of the Base*, which holds that *all* inputs are possible in all languages, distributional and inventory regularities follow from the way the universal input set is mapped onto an output set by the grammar, a language-particular ranking of the constraints. (Prince and Smolensky, 1993, p209)

Furthermore, it was assumed in the original OT phonology framework that the candidate generating function GEN and the constraint set CON are both universal. The universal input combined with a universal function from underlying phonetic representations to surface phonetic representations results in a universal set of possible candidates. This leaves only one source of linguistic variation: the ranking of the constraints. RotB thus became equivalent to 'all systematic differences between languages arise from differences in constraint ranking'. The RotB hypothesis has been widely accepted within OT phonology (but see van Oostendorp (2000) for some critical remarks). With the application of the OT framework to syntax, RotB had to be 'translated' to the field of syntax. As phonology and syntax are concerned with different types of objects, this translation is not straightforward.

2.1 RotB & OT syntax

In an OT syntax framework, the input consists of some semantic representation.[1] The output consists of a structured string and some link to the interpretation of that string. GEN in OT syntax thus differs crucially from GEN in OT phonology in that it is a function that changes the type of the object. This raises the question whether we can still conclude from the universal input and a universal function from inputs to candidate

[1]The formalization depends on the OT syntax implementation, but it is supposed to contain at least a predicate and argument structure as well as the information or discourse status of elements (Bresnan, 2002; Smolensky and Legendre, 2005; Grimshaw and Samek-Lodovici, 1998)

sets GEN that the set of all possible candidates is universal. Smolensky and Legendre (2005) answer this question affirmatively. They present a uniform treatment of phonology and syntax within the OT framework. The definition of RotB is the same for both modules:

> *Richness of the Base:* The space of possible interpretations – inputs to the production function f_{prod} – is universal. Thus all systematic language particular restrictions on what is grammatical must arise from the constraint ranking defining the grammar [...]. (Smolensky and Legendre, 2005, ch12)

This is a direct translation of RotB to OT syntax, which has been widely adopted by OT syntacticians and which has led to analyses of for example resumptive pronouns (Legendre, Smolensky, and Wilson, 2001), expletive pronouns, *pro* (Grimshaw and Samek-Lodovici, 1998), *do* (Grimshaw, 1997) and the paradigm of *to be* (Bresnan, 1999) in terms of constraint interaction. With respect to the role of the lexicon, Smolensky and Legendre (2005) say that "in any OT theory of syntax, [...] the lexicon is not an independent site of variation". In actual linguistic analyses, this has led to a conception of grammar in which lexical lookup takes place after optimization, thus excluding any influence of the lexicon on the selection of the optimal candidate.

Bresnan (2002) captures RotB by "viewing the morpho-syntactic input as arbitrary points in an abstract multidimensional space of dimensions". This input takes the form of the f-structures familiar from classical LFG. Both the candidates and the output consist of c-structure/f-structure pairs. In contrast to classical LFG, where the phonological string is read of the leaves of the c-structure, the candidates and the output do not include phonological material. Only optimal candidates are mapped onto a phonological string: "[...] it is the job of the lexicon to pair the inventory of abstractly characterized candidates selected by the constraint ranking with the unsystematic language-particular pronunciations by which they are used" (Bresnan, 1999). Instead of lexical or phonological material, the c-structure leaves consist of feature bundles, similar (but usually not identical) to the input feature bundles.

To illustrate, an input feature bundle may look like this: [BE PRES 1 SG] for the first person singular slot in the paradigm for the present tense of *to be*. These features may not be realized, violating FAITH constraints. In those cases, a more general form is realized, such as *are*: [BE PRES]. On the other hand, if the features *are* realized, they may violate certain markedness constraints, such as *SG or *1. The relative ranking of these markedness and faithfulness constraints determines which contrasts are expressed in a particular language. The tableaux in (1a) and (1b) illustrate how the constraint ranking for standard English correctly predicts that [BE PRES 1 SG] is realized with perfect faithfulness as *am*, while [BE PRES 2 SG] is realized as the more general *are*.[2]

[2]The constraint ranking in the tableaux is not fully fixed and effectively specifies a set of rankings. Strict domination is indicated by a vertical line between the constraints. A '!' indicates that a violation is fatal under at least one of the compatible rankings.

(1) a.

	[BE PRES 1 SG]	*PL	*2	FAITH$^{P\&N}_{be}$	*SG	*1	*3
☞	'am': [BE PRES 1 SG]				*	*	
	'is': [BE PRES 3 SG]			*!	*		*
	'are': [BE PRES]			*!			
	'art': [BE PRES 2 SG]		*!	*	*		

b.

	[BE PRES 2 SG]	*PL	*2	FAITH$^{P\&N}_{be}$	*SG	*1	*3
	'am': [BE PRES 1 SG]			*	*!	*!	
	'is': [BE PRES 3 SG]			*	*!		*!
☞	'are': [BE PRES]			*			
	'art': [BE PRES 2 SG]		*!		*		

Legendre, Smolensky, and Wilson (2001) are less explicit about the formal status of the lexicon. They assume that the input contains at least a target predicate-argument structure and note that arguments in an input structure are best viewed as bundles of features. Furthermore, they claim that the presence or absence of a lexical item in some particular language is a consequence of output of the grammar. Nevertheless, some lexical material seems to be present in the candidates: "In faithful parses [...] one position in a chain contains the *overt* lexical material of the corresponding element of the Index" (italics from the original). But if the candidates contain lexical material, the presence or absence of lexical items co-determines the candidate space and thus the output (which is supposed to determine the contents of the lexicon): RotB and pre-optimization lexical look-up do not go together.

Samek-Lodovici (1996) appears to assume that the input does not contain lexical material, but the candidates do. In other words, GEN introduces lexical items. He states: "[...] different lexicons give rise to distinct candidate sets language-wise" (Samek-Lodovici, 1996, p9). This is very much like our proposal in section 4.2. However, he follows Prince and Smolensky (1993) in that the lexicon should be derived from the grammar. In his analysis of expletives, he assumes that no language has a lexical entry for expletives: if constraint interaction determines that violating faithfulness constraints is better than not filling a certain position, than that will result in the expletive use of some pronoun (see section 3.4 for discussion on this topic).

The following section discusses various problems that follow from a direct translation of RotB from phonology to OT syntax. We will focus on the model described in Bresnan (1999;2002) because it is most explicit in its assumptions about the formal status of the lexicon.

3 Problematic consequences

In this section we discuss some consequences of a strict interpretation of Richness of the Base in OT syntax. We start with some syntax-phonology interface phenomena in section 3.1, which pose a problem for models with lexical look-up after morpho-syntactic optimization. We argue that these are counterexamples to the 'Principle of Phonology-free Syntax' (Zwicky, 1969; Zwicky and Pullum, 1986), and that these phenomena can only be modeled correctly by having phonological constraints influence syntactic optimization. In section 3.2, we discuss some linguistic phenomena that one may classify as unsystematic linguistic variation. As RotB in both its original phonological form and in the interpretations for syntax focuses on systematic variation, these issues are often set aside as uninteresting. We will argue that the topics under consideration are linguistically relevant and need explanation, even if the language model was designed to optimally account for systematic variation. A computational disadvantage of the absence of a finite lexicon that restricts the candidate set is discussed in section 3.3. Finally, we turn to a very specific piece of OT syntax research that builds on RotB, namely the work by Grimshaw and Samek-Lodovici (1998) on expletives, showing that some of their predictions are not borne out.

3.1 Syntax-phonology interface phenomena

Tromsø Norwegian V3 Standard Norwegian is, like most Germanic languages, a V2 language. This shows up for instance in sentences with a fronted non-subject. The subject then follows the verb. Consider the case of a wh-question with the wh-word in first position:

(2) a. Du sa noe.
you said something
'You said something.'
b. Hva sa du?
what said you
'What did you say?'
c. *Hva du sa?
what you said

However, several dialects of Norwegian allow for V3 word order in wh-questions. In such a dialect, not only the parallel to (2b), but also to (2c) is grammatical (Rice and Svenonius, 1998; Westergaard, 2003; Vangsnes, 2004).

One dialect in point is the Tromsø variety of the North Norwegian dialect. Interestingly, in Tromsø North Norwegian, the grammaticality of V3 co-varies with the prosodic features of the question word. Polysyllabic question words require V2. Monosyllabic question words, however, allow V3. The pattern for Tromsø North Norwegian thus is (from Rice and Svenonius (1998)):

(3) a. Koffor skrev han / *han skrev ikkje?
why wrote he he wrote not
'Why didn't he write?'

b. Ka du fikk?
what you got
'What did you get?'

Any way of making the wh-constituent prosodically heavier, whether it adds to *syntactic* weight or not, makes the V3 construction ungrammatical. So variations on (3b) with *kem eller ka* ('Who or what'), *ka slags* ('What kind') or *KA* ('WHAT', stressed), are ungrammatical.

Rice and Svenonius (1998) argue that this behavior is really due to the prosodic features of the wh-words in question,[3] and analyze it in terms of unstressed, monosyllabic wh-words not being able to support a foot on their own. They propose an architecture in which syntax remains indifferent to the V2/V3 constructions, and both are passed to the phonology component. So, as far as word order is concerned, syntax acts as a generator for phonology.

This setup requires that syntax is in fact indifferent to the two constructions. Using an universalist OT framework, in a somewhat more comprehensive grammar, this seems hard to accomplish, because it would require that the two candidates look the same to each and every constraint. Alternatively, one might consider an OT version that allows for variation in a more controlled way, that is by combining the outputs of different rankings (Anttila (1997) or Stochastic OT, Boersma and Hayes (2001)). But these versions of OT often relate the output-sets to occurrence frequencies. A phonological filter on these sets – the phonological optimization that comes after syntax – would disturb this correlation, and, in the case of Stochastic OT render the associated learning algorithm useless.

We conclude that a phenomenon like this is best modeled in OT by letting the interface constraints interact directly with the constraints of syntax. This means that the phonological string has to be present during optimization, and this, in turn, means that lexical lookup cannot occur only after optimization.

Dutch verb clusters In Dutch, all verbs of a subordinate clause and all non-finite verbs in a main clause reside in the verb cluster at the end of the clause. Non-verbal material related to a verb, in e.g. verb-PP or verb-adjective collocations, generally appears to the front of the verb cluster. However, to a limited extent, the non-verbal material is allowed to appear among the verbs in the cluster. The acceptability of this construction depends on the number of syllables, and therefore the *phonological heaviness* of the intervening material. For instance, from (4a)–(4c) acceptability is reduced.[4]

(4) a. Ik heb het niet **los** kunnen / kunnen **los** maken.
I have it not loose can.INF can.INF loose make.INF
'I didn't manage to loosen it'

b. Hij had haar **gerust** willen / ?willen **gerust** stellen.
he had her at ease want.INF want.INF at ease put.INF
'He meant to comfort her.'

[3] See (Vangsnes, 2004) for a different analysis.

[4] The non-verbal material concerned is highlighted.

c. . . . dat hij zich **ongerust** ging / *ging **ongerust** maken.
that he REFL worried went went worried make.INF
'. . . that he was getting worried.'

Diachronically, this effect can be seen even more clearly. Corpus study shows that there is a clear correlation between time and the average number of syllables of the non-verbal material in the verb cluster (Jack Hoeksema, p.c.). Over time, Dutch has become more intolerant of this material. To illustrate, the counterpart of (5) would be fully ungrammatical in modern Dutch.

(5) gewis hij zoude zijn heerlijk ontwerp hebben **ten uitvoer** gebragt.
certain he would his wonderful design have.INF to the execution brought
'He would have certainly implemented his wonderful design.' (*De werken van Jacob Haafner*, part 1, p336, ∼1810)

Phonology driven non-agreement in English Bresnan (1999) accounts for various neutralization effects in the paradigm of *to be* by means of universal constraints on the realization of morpho-syntactic features. These constraints do not explain why the synthetic negation *amn't* is ungrammatical in standard English (see example (6)). Something else is needed to account for the absence of this form. A possible explanation is offered by Dixon (1982). Dixon suggests that *am* is reduced to [ɑ:] before *n't* as to avoid the consonant sequence *-mn-* (subsequently leading to reanalysis and spelling of the sequence as an instance of *aren't*).

(6) a. I am silly. / I'm silly.
b. Aren't I silly?

(7) a. The lions are / 're / *is / *'s in the compound
b. Where are / 're / *is / 's the lions?

A similar phonological markedness constraint can explain the contrast between (7a) and (7b) (grammaticality judgments from Dixon): the infelicitous phonological sequence *where're* must be avoided and this may be done by using the copula *'s* with a plural subject (Dixon, 1982).

If we translate these phonological constraints into OT style constraints, we need them to interact with morpho-syntactic constraints. This means importing the lexical string into syntax, leading to a grammar that is no longer independent of the lexicon.

Alternatively, one could envisage an approach in the style of Rice and Svenonius (1998), where both alternatives are produced and a phonological filter prevents *amn't* and *where're* from surfacing. However, such an approach should also explain why the *are* is not allowed for first person singular elsewhere, and why *'s* for third plural can only occur after *where, there* and *here*.

The Dutch modifier *hoogst* A fourth and final example of phonology driven syntax is found in Dutch modification by intensifiers. The intensifiers *heel* and *erg* (Dutch, 'very') do not pose constraints on the modified adjective. In contrast, the intensifier *hoogst* (Dutch, 'highly') occurs with polysyllabic adjectives (Klein, 1998). Compare

heel+ADJ		*erg*+ADJ		*hoogst*+ADJ	
645	goed	98	goed	13	onwaarschijnlijk
330	erg	70	moeilijk	11	onzeker
214	lang	53	groot	10	ongebruikelijk
147	moeilijk	48	belangrijk	7	twijfelachtig
119	belangrijk	44	hoog	7	ongelukkig
106	snel	34	klein	6	noodzakelijke
94	klein	32	lang	6	irritant
89	hard	28	leuk	5	waarschijnlijk
80	mooi	26	populair	5	persoonlijke
75	sterk	24	sterk	5	merkwaardige

Table 1: Most frequent adjectives modified by *heel* (very), *erg* (very) *hoogst* (highly)

the co-occurrence data in the Volkskrant-newspaper 1998 volume ($\sim$ 17 mln words) for the three intensifiers in table 1. While many monosyllabic adjectives are extremely frequent, we did not find any occurrence of a monosyllabic adjective modified by *hoogst* in the Volkskrant corpus. Even on the web it is hard to find examples: of all monosyllabic adjectives in table 1, Google returned only one occurrence of one combination with *hoogst*: *hoogst leuk* (Dutch, 'very nice').

Like the previous examples, the distribution of the intensifiers in Dutch shows that the phonological form of a word – which is stored in the lexicon – influences more than just the phonological shape of the sentences: it influences at least the word order and the combinatory possibilities of the clause. That is, the lexicon influences grammar and therefore cannot be entirely derived from it.

3.2 Non-systematic language particularities

Smolensky and Legendre (2005), as well Bresnan (2002) and earlier formulations of RotB, make it very clear that the principle is concerned with systematic variation only. There are many (unsystematic) linguistic phenomena for which constraint re-ranking is an implausible explanation. Should we for instance conclude from the introduction of the word *hamburger* into the English language that its grammar changed as to the effect of suddenly disfavoring *ground beef sandwich*? What constraints prevented the single noun realization from becoming optimal before the noun was introduced? Similar questions can be asked for accounts of syntactically distinct realizations of a concept in different languages (Swedish *professorskan* vs *the professor's wife*), and syntactically distinct realizations of closely related concept within one language (*cause to die* vs *kill*). Why is it that the meaning of *cause to die (unvoluntarily)*, which is so close to *kill*, cannot be expressed by a single lexical item? Why does this pattern not extend to other lexical entries, i.e. why does *cause to sleep* not imply involuntariness and how could constraint ranking account for the correlation with the lack of a lexical entry for 'voluntarily cause to sleep'? The same line of reasoning can be applied to larger units such as idioms. Since these are semantically atomic but syntactically complex, the full construction has to be available at syntactic optimization.

These are examples of unsystematic linguistic particularities, just as arbitrary as phonological form. The lexicon is the ideal locus for such unsystematic information.[5] Only, this unsystematic information interacts with the grammar and thus has to be available before optimization takes place. This is problematic for OT syntax models that assume optimization takes place over sets of morpho-syntactic feature bundles. But *any* OT syntax model that claims that the lexicon is derived from the output of the grammar needs to say something more about these language particular, unsystematic properties that influence grammar. Smolensky and Legendre (2005, ch12) are willing to weaken the principle that all constraints are universal, in order to account for idiosyncratic language-specific phonological alternations. Other solutions that are more specifically aimed at OT syntax are discussed in section 5.1.

3.3 Decidability of OT-LFG generation

Kuhn (2003, and earlier work) develops a formalization of the OT-LFG framework and investigates its computational properties. Following Bresnan (2000), he models GEN as an over-generating LFG-grammar, taking an f-structure as input and specifying a set of c- and f-structure pairs with ϕ-mappings. This set is used as the candidate set. Kuhn also provides a syntax for specifying constraints in the formalization. One of Kuhn's important results is the *decidability of generation*. The generation task – for an underlying form, what is the optimal candidate according to an OT system? – is not trivial, because GEN may be unfaithful to the input, resulting in the infamous infinite candidate set.

We will gloss over the technical details of the decidability proof here. Suffice it to say that it involves factoring in the constraints into the LFG-grammar describing GEN, G_{base}. The result is also an LFG grammar, to which existing decidability of generation results can be applied.

The reason we can omit going into the algorithm here, is that the problem already arises in a preprocessing step. G_{base}, being a classic LFG grammar, is a set of annotated c-structure rules and a set of lexical entries. It is normalized to an equivalent G'_{base}, partly by moving the lexical entries into the c-structure descriptions; the lexical strings in G_{base} become terminals in G'_{base}. This crucially relies on having a finite lexicon. If the lexicon is infinite, we would end up with an infinite number of c-structure rules, and this step would never terminate. Note that the *base* lexicon needs to be finite. If aspects of morphology – say, compounding – can be captured in the c-structure rules, a *derived* lexicon may be infinite with decidability still holding.

The observation that the lexicon in such a system needs to be finite seems rather trivial. But some assumptions have to be made, or made explicit, in order to assure decidability and little attention has been paid to this fact.

Let us look at Bresnan's (2002) proposal, predicting the paradigma of *to be*. In order to preserve the universality of the candidate set, the trees describing the candidates have morpho-syntactic feature bundles as their leaves. This means that the terminals in

[5] Smolensky and Legendre (2005) state that learning *systematic* language particular grammatical information is 'utterly unlike' learning the phonological shape of a word, but ignore the role of the lexicon in the acquisition of *unsystematic* language particular grammatical information and do not offer alternative solutions for the problems described above.

the normalized grammar are also these morpho-syntactic feature bundles. The question is, therefore, whether one can make sure that the set of these bundles is finite. Because we are interested in the base lexicon, we will ignore features which have feature structures as their values.

So, is this set of *flat* feature bundles finite? Bresnan (2002) does not provide us with enough information to decide that for a general setting. However, we *can* specify the sufficient and necessary conditions. To have a finite set you need a finite number of features and a finite number of values. To have a finite number of values, they need to be discrete and to be drawn from a limited domain. In Bresnan's view, the features represent 'dimensions of possible grammatical or lexical contrast' (Bresnan and Deo, 2001, p6). It seems fairly uncontroversial to assume that there is only a certain number of these. And perhaps, for some of these dimensions it could be argued that only a finite number of contrasts has to be made. But for other features, such as PRED, this is less obvious. See Mohanan and Mohanan (2003) for some discussion of related issues.

3.4 Lexical expletives

As a final example we will look at the analysis of the distribution of expletive subjects in Samek-Lodovici (1996) and Grimshaw and Samek-Lodovici (1998). More specifically, we will argue that some of their predictions with respect to the lexicon are not borne out.

Grimshaw and Samek-Lodivici follow the approach to *do*-support in Grimshaw (1997). The distribution of other semantically empty or impoverished items has been analyzed in the same fashion (see e.g. Sells (2003)). Grimshaw and Samek-Lodovici assume that the lexicons of languages do not differ in ways relevant to expletives. Instead, an expletive is an unfaithfully used referential element. So, in English there is no difference between referential *it* as in "it howled" and the expletive *it* as in "it rained", the difference is that in the referential case the lexical meaning of the pronoun contributes to the meaning of the construction, whereas in the expletive case it does not. Whether a language allows such use of lexical resources is a matter of constraint ranking. As Grimshaw and Samek-Lodovici (1998, p205) write:

> [T]here cannot be a language which lexically lacks an expletive, any more than there can be a language which lexically lacks an epenthetic vowel. The occurrence of such items is not regulated by lexical stipulation; instead, the visible lexical items are the result of constraint interaction.

To illustrate, consider the case of English. The constraint SUBJECT, requiring sentences to have overt subjects, outranks the constraint FULL-INTERPRETATION, that puts a ban on using words that do not contribute their lexical meaning to the compositional meaning of the construction. The optimization of the input *rain′*, is summarized in the following tableau:

(8)

	rain′	SUB	F-INT
☞	it rains		*
	he rains		**!
	John rains		**!*
	rains	*!	

In the model, all NP's in the tableau are expletive NP's: their lexical content is ignored. That F-INT is a gradient constraint is highly relevant. Given some sufficient notion of information, the constraint is violated more when using an element of higher content. Presumably, in English, the third-person, singular, neuter pronoun carries the least content, resulting in its use in the winner *it rains*. This way, both the distribution and the choice of lexical elements is captured by the model.

'Expletiveness' not being part of the lexical specification also means that there cannot exist a language with *lexical expletives* – elements that are only used as an expletive. This prediction is borne out if we e.g. look at the main Germanic languages: the expletives are also used as referential third person pronouns or as locative adverbs.

In the South Norwegian dialect of Lyngdal, Vest-Agder, too, there is an expletive that is homophonous with the third-person, singular, neuter pronoun: *det*. However, a second expletive does not use the form for the locative adverb *der* ('there'), but has a similar but distinct form *dar* (Pål Kristian Eriksen, p.c.).[6] For instance:

(9) a. Dar snø. (Weather verbs)
dar snow.PRES
'It is snowing.'

b. Dar blei skutt ein rev (Impersonal passives)
dar became shot a fox
'Someone shot a fox.'

c. Der e dar ein katt (Existential sentences)
there be.PRES dar a cat
'Over there, there is a cat.'

Interestingly, *dar* is only used as an expletive, and cannot be explained as an allophonic variant of *der*. The conclusion therefore must be that Lyngdal South-Norwegian has a lexical expletive.

Helge Lødrup (p.c.) suggests that the vowel in *dar* is a remnant from the Old Norse form *þar*, used only as the referential, demonstrative locative adverb. Why the Lyngdal dialect has preserved this vowel in the expletive use alone, is unclear, but it is the fact that it could that poses a problem for Grimshaw and Samek-Lodovici's model, for want of a non-trivial lexicon.

[6]The phenomenon is as far as we are aware undocumented, and it is unclear how widespread the use of *dar* is. To give us a slight hint to the use and distribution of *dar* we have looked for instances of *dar* on the web. Crucially, instances were found were the same speaker also uses *der* for referential 'there'. The search is not meant as a representative study. We can report a couple of uses that appear to be from the south-west of Norway, up to and including the city of Stavanger. Apart from the spelling *dar*, *dår* and *dårr* – suggesting slightly different pronunciations – have also been found.

4 Solutions

Having reviewed some example cases that are problematic under the current conceptualization of the lexicon in OT syntax, in this section we will propose what we think is the proper place for the lexicon in OT syntax. Before doing so, we will look at an alternative that has been proposed in the literature and argue that it should be rejected.

4.1 A constraint called LEX

Several authors have modeled lexical information using constraints. Given the central role constraints play in linguistic explanation in OT, this approach almost suggests itself. Noyer (1993) proposes an inviolable constraint LEXICALITY, that disallows "signs" that are not composed of "morphemes". Kusters (2003, p69) refines and clarifies this constraint by adding "[LEX] rules out all strings of sound that do not consist of actual lexical material".[7] Finally, "[t]o model accidental lexical gaps" Bresnan (2002) assumes a highly ranked constraint LEX that says that "candidates [...] have pronunciations". Notice that, in these formulations, only Kusters commits himself to having the phonological string available during optimization. There, candidates are built up out of bits of associations of semantico-syntactic information and phonological form – i.e. morphemes. Associations that are not 'conventionalized', i.e. they are not in the language's lexicon, violate LEX. As we have seen before, Bresnan assumes that only the semantico-syntactic side plays a part during optimization. That said, with respect to the constraint, the three proposals are essentially the same, and we shall refer to them as LEX.

The amount of lexical information that LEX supplies is enough to solve some of the problems we mentioned in section 3. Like Bresnan's original application, one could use LEX to block forms that are missing from a paradigm for no apparent systematic (syntactic) reason.[8] If the form is missing for, say, phonological reasons, one may partly model the influence of phonology on syntax in that manner. Similarly one can model other cross-linguistic or diachronic idiosyncrasies, too. For instance – while radically changing the model – one could in principle assume that all languages *can* have lexical expletives, but most happen not to do so. Instead, these languages (mis)use regular lexical items for the job. Of course, in Lyngdal West-Norwegian, the candidate that uses the lexical expletive does not violate LEX.

However, LEX leaves some questions unanswered and introduces some conceptual problems of its own. For instance, real effects of phonology on syntax, where the actual phonological string plays a role, are not necessarily addressable by having LEX. Likewise, the potential decidability problem is not solved by LEX, since it is a solution in CON, and the decidability problems are associated with GEN. The lexical information in the system comes in too late to assure decidability.

[7] Without exploring it any further, Kusters does mention the possibility of having "LEX [as part of] the hardware of the grammar", instead of as a constraint (o.c., p70 fn33).

[8] The constraint LEX is used to achieve the same effect as the constraint **amn't* (Bresnan, 2002). It serves to block a certain candidate, to show that the model predicts the correct *replacements* for this form, without having to venture a guess as to why the form doesn't exist. As such, LEX is not a core part of the analysis.

More serious are the conceptual problems associated with LEX. Firstly, there will be a reduplication of information in the system. Lexical information is already needed for the phonological mapping, irrespective of where this mapping occurs. Now part of this information needs to be present in CON, too.

Secondly LEX makes for an atypical OT constraint. It is assumed never to be violated. It is also unclear what it would mean to violate LEX, because such a candidate contains either gibberish (Kusters model) or something by definition unpronounceable (Bresnan). As a result, LEX cannot be re-ranked. Also, although the abstract definition of the constraint is universal, actual uses have to be 'parameterized' for the lexicon of the language under scrutiny.

Furthermore, under Bresnan's conception, the constraint is used to model unsystematic properties of a language's lexicon, only. So, it should not block everything that is not in the lexicon of the particular language, but just the candidates whose non-optimality cannot be explained by the rest of the grammar. This, again, makes it an abnormal OT constraint. Whether a constraint is violated should depend only on its definition and on the candidate, but not on other candidates or other constraints, let alone on the outcome of an evaluation of the same candidate using the rest of the grammar. Such a constraint greatly increases the complexity of EVAL.

4.2 The lexicon as an argument of GEN

The problems associated with post-syntactic lexical lookup or with a constraint like LEX, suggest that the proper place for the lexicon is actually before CON. However, both the input and GEN are considered to be of a universal character. In order to retain this universality, we propose to model a language particular lexicon as an argument of GEN. That is, syntactic GEN is a universal function from a meaning-representing input *and* a lexicon to a candidate set.

We can adapt Kuhn's (2003) definition of GEN to include this argument. Remember that Kuhn based GEN on an LFG grammar G_{base}, describing universal properties of language. The candidate set is (roughly) the result of generating with G_{base} from an f-structure representing the input. To this conception of GEN, we easily add the lexicon by letting G_{base} take it as an argument.[9] Thus, the candidate set is defined as:

(10) Definition of GEN with lexicon

$$Gen(\Phi_{in}, \Lambda) \quad =_{def} \quad \{\langle T, \Phi' \rangle \in G_{base}(\Lambda) \mid \Phi_{in} \sqsubseteq \Phi'\}$$

(where Φ' and Φ_{in} are f-structures, T is a c-structure and Λ is a set of LFG-style lexical entries.)

Now GEN produces candidates that have lexical items on their c-structure terminals. Information about the phonological string is accessible for the constraints in CON. This allows for interaction of phonological and syntactic constraints, facilitating accounts for the interface phenomena we saw in section 3.1. For example, we can define linear

[9] Notice that we are only making a certain relation more explicit, by stating it as an argument. Kuhn seems to implicitly assume the lexical specification to be part of G_{base}. Our presentation only teases apart what in classical LFG are the annotated c-structure rules (i.e.: $\lambda x.G_{base}(x)$) and the lexical entries.

order constraints referring to syllable structure to account for the distribution of monosyllabic questions words in Tromsø Norwegian. An additional advantage of this setup is that it allows for simpler integration of morphology in the system. A phonology-free model of syntax makes impossible any form of interaction with morphology, which is highly string-sensitive. As a result, morphology has to be modeled as a separate module, connected to the rest of the language model in some way. Our model, on the other hand, allows for interaction of phonological, morphological and syntactic constraints.

Furthermore, we now have a locus for storing unsystematic language particular information, whether phonological or morpho-syntactical, and we do not have to apply a constraint ranking to account for the presence or absence of the word *hamburger* in a particular language at a particular time. As long as a language does not have a lexical entry for a particular concept, the one word realization is not in the candidate set, because GEN cannot generate it. Because English does not have a lexical entry for 'the professor's wife' (like Swedish does), it will use the genitive construction to express that meaning.

In our setup, the presence of a lexical item that is uniquely used as an expletive pronoun is not problematic. We simply treat it as an unsystematic language particularity and store it as such in the lexicon. This does not explain why so many languages unfaithfully use personal or demonstrative pronouns for expletives. In other words: we want to keep the explanatory power of the analysis in Samek-Lodovici (1996). It is not impossible to build in the analysis in our model: the Norwegian *dar* may have acquired its use as an expletive in exactly the way proposed, but the referential use may have disappeared or evolved while the lexicalized expletive use remained. Some additional assumptions have to be made, though, to explain the striking infrequency of these visibly lexicalized expletive pronouns.

Finally, the proposed model saves Kuhn's proof of decidability (Kuhn, 2003) without the need for any further assumptions.

5 RotB Revisited

We have shown that modeling the lexicon as an argument of GEN avoids many of the problems that a strict interpretation of RotB encounters. But at what cost? Do we throw out the most basic principle of the OT framework?

The model described in section 4.2 violates RotB in that it does not restrict all variation to differences in constraint ranking. Instead, we now have two places of analysis: the lexicon and the constraint ranking. But this does not mean that we throw out RotB altogether: the model is compatible with RotB in the sense that it assumes a) an unconstrained, universal set of possible inputs, b) a universal function GEN and c) a universal set of constraints.

5.1 Evaluating models

With two loci for linguistic analyses (the lexicon and the constraint ranking), the question arises which locus to choose. We have seen that analyses that crucially rely on information from the lexicon are necessary for some linguistic phenomena. However,

we agree with Smolensky and Legendre (2005) and Bresnan (2002) and most other work in OT syntax that explanations in terms of constraint ranking are to be preferred over lexicalist accounts. We therefore adopt a methodological principle as in Kuhn (2003):[10]

(11) *Methodological principle of OT*
Try to explain as much as possible as an effect of constraint interaction.

Recall that Smolensky and Legendre (2005) also realized that there are language particularities that cannot be explained by the ranking of universal constraints. For these phenomena, they weaken RotB by allowing language particular constraints. With language particular constraints, the ranking of the constraints is no longer the only source of linguistic variation. In order to save this idea as much as possible, they adopt a methodological principle very similar to the one above:

> One might say that the OT principle 'constraints are universal', is a violable meta-constraint on the explanatory value of substantive linguistic theories, the most explanatory theory of some domain being the one that best-satisfies the universal constraint. (Smolensky and Legendre, 2005)

We propose to view the OT principle 'all systematic variation is constraint ranking' as a violable meta-constraint in the same fashion Smolensky and Legendre (2005) propose for the universal constraint principle.

5.2 Kusters' diachronic perspective

A different approach can be found in the work of Kusters (2003), who also assumes that information about the lexicon of the language is available in syntax, albeit in the form of a constraint LEX. Interestingly, he exploits the resulting explanatory overlap to model language change in connection with social change.

Kusters posits that the content of the language particular lexicon is acquired by a new generation of speakers by inducing it from the output of the previous generation. Crucially, this output is not only the result of the previous generation's lexicon but also of their grammar. Consider the case in which a lexical item never surfaces because its use would involve some fatal violation of a markedness constraint. As a result, the item would never be incorporated into the lexicon of a language user from the next generation.

Another interesting case is when some lexical item is overloaded with meaning. To express the meaning $f'(g')$, a speaker may – again because of markedness constraints – be forced to just use the lexical entry for g': "gee", instead of uttering it together with the entry for f': "eff gee". This means that a language learner assigns the meaning $f'(g')$ to "gee". In the same fashion, an item can be stripped of content.

As a side effect, Kusters notes, the inter-generation change of the content of a morpheme mimics the Lexicon Optimization of OT phonology (Prince and Smolensky, 1993, p209). If "gee" alone lexically specifies $f'(g')$, no faithfulness constraints are

[10] Kuhn (2003) is more concerned with restricting the role of GEN than the role of the lexicon. In both cases the aim is to keep the candidate set as large as possible.

violated by using "gee" to express exactly $f'(g')$. This means that the Harmony of the optimal candidate in the the new generation is higher than the Harmony of the previous generation's optimal candidate. It shares this *Harmony maximization* with Lexicon Optimization. However it should not be forgotten that the lexicon does not relate to syntax as it does to phonology. In phonology, the lexicon supplies inputs. Lexicon Optimization in phonology is therefore *input optimization*. In syntax the input is meaning related, and if it has any relation to the lexicon at all, it is indirect.

As it stands, Kusters' model captures instances of lexical items changing content, or items being dropped from the lexicon from one generation to the other, and serves as a theory of grammaticalization in the sense that what used to be the result of optimization becomes entrenched in the lexicon. The addition of new words to a language, be they loan-words or inventions, does not readily follow. Nor is speaker internal language change catered for. Nevertheless the model looks like a good starting point for investigating these topics in OT syntax.

Finally, Kusters' approach directly carries over to the architecture we propose in the previous section. The lexicon is still learned from output forms, but the information enters the system in a different place.

6 Conclusion

Since the beginning of Optimality Theory, there has been a tendency to accentuate the universalist approach to grammar. This has led to the claim that all (systematic) linguistic variation should be explained by the ranking of universal constraints. We argued in this paper that this does not always give the right results in OT syntax. We focused on the role of the lexicon within this universalist approach to grammar and we showed how the lack of a language particular lexicon causes problems for different approaches to OT syntax.

In order to remedy these problems, we proposed to view a language particular lexicon as an argument in GEN, technically only a small formal adjustment to OT. While creating the possibility of solving the aforementioned problems, this keeps GEN universal. As a possible down-side, the new setup has as a consequence that the lexicon and constraint ranking as explanatory devices may have overlapping domains. We considered this situation in the light of Richness of the Base and of modeling language change and argued that this need not be a problem and may even be an asset of the theory.

This paper has only provided a formal sketch of what the framework should look like with a proper lexicon. Of course, many questions remain. For instance, the (extent of the) interaction between phonological and syntactic constraints offers a vast and mainly uncharted terrain of research. Also, although we assume that the lexicon provides information about the availability of terminals and maps from morpho-syntax to phonology, we have not considered what is specified about these items. An obvious question is whether argument structure should be coded in these lexicons.

Another open question is how the various proposals in the literature could be implemented in the proposed setup. Recasting Grimshaw and Samek-Lodovici's work in our framework offers some possible solutions, but also calls for extra assumptions

because of the less restricted nature of the setup. Furthermore, as Kusters' model is compatible with ours, it would be interesting to explore the issues brought forward in his work from a more formal perspective.

Acknowledgments

The authors would like to thank the audience of LFG04 in Christchurch, New Zealand for their comments. Also, the authors thank Helge Lødrup and especially Pål Kristian Eriksen for their data and input on the Lyngdal dialect. Leonoor van der Beek's research is carried out as part of the PIONIER project *Algorithms for Linguistic Processing*, under grant number 220-70-001 from the Netherlands Organisation for Scientific Research (NWO). Gerlof Bouma's research is carried out in the framework of the NWO Cognition Programme under grant number 051-02-071.

References

Anttila, Arto. 1997. Deriving variation from grammar. In F. Hinskens, R. van Hout, and L. Wetzels, editors, *Variation, change and phonological theory*. Benjamins, Amsterdam.

Boersma, Paul and Bruce Hayes. 2001. Empirical tests of the gradual learning algorithm. *Linguistic Inquiry*, 32(1):45–86.

Bresnan, J. 2001. *Lexical Functional Syntax*. Blackwell Publishers.

Bresnan, Joan. 1999. Explaining morphosyntactic competition. In Mark Baltin and Chris Collins, editors, *Handbook of Contemporary Syntactic Theory*. Oxford: Blackwell Publishers, pages 11–44. Also available on the Rutgers Optimality Archive: (ROA-299-0299).

Bresnan, Joan. 2000. Optimal syntax. In Joost Dekkers, Frank van der Leeuw, and Jeroen de Weijer, editors, *Optimality Theory: Phonology, Syntax and Acquisition*. Oxford University Press, Oxford, pages 334–385.

Bresnan, Joan. 2002. The lexicon in optimality theory. In Suzanne Stevenson and Paola Merlo, editors, *The Lexical Basis of Syntactic Processing: Formal,Computational and Experimental Issues*. John Benjamins, pages 39–58.

Bresnan, Joan and Ashwini Deo. 2001. Grammatical constraints on variation: 'be' in the survey of english dialects and (stochastic) optimality theory. MS. Available on `http://www-lfg.stanford.edu/bresnan/`.

Dixon, R.M., 1982. *Where Have All the Adjectives Gone? and other essays in Semantics and Syntax*, chapter Semantic neutralisation for phonological reasons, pages 235–238. Berlin: Mouton Publishers.

Gazdar, Gerald, Ewan Klein, Geoffrey Pullum, and Ivan Sag. 1985. *Generalized Phrase Structure Grammar*. Blackwell.

Grimshaw, Jane. 1997. Projection, heads, and optimality. *Linguistic Inquiry*, 28:373–422.

Grimshaw, Jane and Vieri Samek-Lodovici. 1998. Optimal subjects and subject universals. In Pilar Barbosa et al., editor, *Is the Best Good Enough?* MIT Press, Cambridge, MA, pages 193–219.

Klein, Henny. 1998. *Adverbs of Degree in Dutch and Related Languages*. John Benjamins, Amsterdam.

Kuhn, Jonas. 2003. *Optimality-Theoretic Syntax—A Declarative Approach*. CSLI Publications, Stanford, CA.

Kusters, Wouter. 2003. *Linguistic Complexity; The Influence of Social Change on Verbal Inflection*. Ph.D. thesis, Landelijke Onderzoeksschool Taalkunde, Utrecht. LOT dissertation series 77.

Legendre, Gérealdine, Paul Smolensky, and Colin Wilson. 2001. When is less more. In Pilar Barbosa, Danny Fox, Paul Hagstrom, Martha McGinnis, and David Pesetsky, editors, *Is the Best Good Enough? Optimality and Compitition in SYntax*. MIT Press, pages 249–289.

Mohanan, Tara and K.P. Mohanan. 2003. Universal and language-particular constraints in ot-lfg. In Miriam Butt and Tracy Holloway King, editors, *The Proceedings of the LFG'03 Conference*. CSLI Publications.

Noyer, Rolf. 1993. Optimal words: Towards a declarative theory of word-formation. MS. Rutgers Optimality Workshop-1.

van Oostendorp, Marc. 2000. Back to the base. on the richness of the base hypothesis. MS. Available on `http://www.vanoostendorp.nl/fonologie.php?lan=en`.

Pollard, Carl and Ivan Sag. 1994. *Head-driven Phrase Structure Grammar*. University of Chigago / CSLI.

Prince, Alan and Paul Smolensky. 1993. *Optimality theory: constraint interaction in generative grammar*. Rutgers University Center for Cognitive Science, Piscataway, NY.

Rice, Curt and Peter Svenonius. 1998. Prosodic V2 in Northern Norwegian. MS. University of Tromsø.

Samek-Lodovici, Vieri. 1996. *Constraints on subjects. An Optimality Theoretic analysis*. Ph.D. thesis, Rutgers University, New Brunswick, NJ.

Sells, Peter. 2003. Morphological and constructional expression and recoverability of verbal features. In C. Orhan Orgun and Peter Sells, editors, *Morphology and the Web of Grammar: Essays in Memory of Steven G. Lapointe*. CSLI Publications.

Smolensky, Paul and Géraldine Legendre. 2005. *The Harmonic Mind: From Neural Computation To Optimality-Theoretic Grammar Vol. 1: Cognitive Architecture; vol. 2: Linguistic and Philosophical Implications*. To appear at MIT Press.

Vangsnes, Øystein. 2004. On wh-questions and v2 across norwegian dialects: a survey and some speculations. In *Working Papers in Scandinavian Syntax 73*. Lund University, Institutionen för nordiska språk.

Westergaard, Marit Richardsen. 2003. Word order in wh-questions in a north norwegian dialect: some evidence from an acquisition study. *Nordic Journal of Linguistics*, 23(1):81–109.

Zwicky, Arnold. 1969. Phonological constraints in syntactic descriptions. *Papers in Linguistics*, pages 411–453.

Zwicky, Arnold M. and Geoffrey K. Pullum. 1986. The principle of phonology-free syntax: introductory remarks. In *Working Papers in Linguistics 32*. The Ohio State University, Columbus, OH, pages 63–91.

Leonoor van der Beek (vdbeek@let.rug.nl)
Gerlof Bouma (gerlof@let.rug.nl)

Center for Language and Cognition (CLCG)
Faculty of Arts, University of Groningen
P.O. Box 716
NL-9700 AS Groningen